Statistics for Biology and Health

Nicole Lazar

The Statistical Analysis of Functional MRI Data

Springer

Statistics for Biology and Health

Statistics for Biology and Health

Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Duchateau/Janssen: The Frailty Model
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival AnalysisL A Self-Learning Text, 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Lazar: The Statistical Analysis of Functional MRI Data
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielsen: Statistical Methods in Molecular Evolution
O'Quigley: Proportional Hazards Regression
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/LanWittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analysing Ecological Data

Nicole A. Lazar

The Statistical Analysis of Functional MRI Data

Springer

Nicole A. Lazar
Department of Statistics
University of Georgia
Athens, GA 30602-1952
USA
nlazar@stat.uga.edu

ISBN: 978-0-387-78190-7 e-ISBN: 978-0-387-78191-4
DOI: 10.1007/978-0-387-78191-4

Library of Congress Control Number: 2008930095

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To David and Benjamin

Preface

The study of brain function is one of the most fascinating pursuits of modern science. Functional neuroimaging is an important component of much of the current research in cognitive, clinical, and social psychology. The excitement of studying the brain is recognized in both the popular press and the scientific community. In the pages of mainstream publications, including *The New York Times* and *Wired*, readers can learn about cutting-edge research into topics such as understanding how customers react to products and advertisements ("If your brain has a 'buy button,' what pushes it?", *The New York Times*, October 19, 2004), how viewers respond to campaign ads ("Using M.R.I.'s to see politics on the brain," *The New York Times*, April 20, 2004; "This is your brain on Hillary: Political neuroscience hits new low," *Wired*, November 12, 2007), how men and women react to sexual stimulation ("Brain scans arouse researchers," *Wired*, April 19, 2004), distinguishing lies from the truth ("Duped," *The New Yorker*, July 2, 2007; "Woman convicted of child abuse hopes fMRI can prove her innocence," *Wired*, November 5, 2007), and even what separates "cool" people from "nerds" ("If you secretly like Michael Bolton, we'll know," *Wired*, October 2004). Reports on pathologies such as autism, in which neuroimaging plays a large role, are also common (for instance, a *Time* magazine cover story from May 6, 2002, entitled "Inside the world of autism"). The 1990s were designated "The Decade of the Brain" by the National Institute of Mental Health and the Library of Congress; the 2003 Nobel Prize in Medicine was awarded for research that lies at the foundation of functional magnetic resonance imaging (fMRI), one of the most prevalent and popular tools used for studying brain function.

Statisticians have a key role to play 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 might be hundreds of thousands of time series for a single subject, one for each "voxel," or volume element, of the brain). The number of subjects on the other hand is generally small, a situation that creates challenges for statistical inference. Statisticians have already made many important contributions

to the field, and as more universities set up imaging centers of their own, the presence of on-site statistical experts becomes more important. Obtaining the necessary background in neuroimaging and neuroscience, however, can take many years of intense study. My goal in writing this book was to provide an introduction to functional magnetic resonance imaging, aimed at statisticians, that would highlight the important scientific issues and survey the common (and some not so common) analysis pathways.

The primary intended audience is statisticians who are interested in this growing field and who wish to gain an understanding of the major problems and current solutions. A secondary audience is cognitive psychologists and other neuroscientists who use fMRI as a research tool. This book can also serve them as a summary of the major statistical questions in the analysis of functional neuroimaging data and of the commonly used methods. Readers need only be familiar with basic graduate level statistics – linear models, general and generalized linear models, nonparametric statistics, Bayesian theory, and the like.

The first three chapters of this book give the scientific background: a brief introduction of how fMRI data are acquired appears in Chapter 1, followed by chapters on experimental design and data preprocessing. Chapter 4 is a "bridge" chapter, summarizing the major statistical issues and setting the stage for the core of the book, chapters 5 through 10. These chapters describe the various statistical approaches that have been taken for analyzing fMRI data, from the popular general linear model (Chapter 5), through spatiotemporal models (Chapter 6), multivariate approaches (Chapter 7), analyses using basis functions (Chapter 8), and Bayesian analysis (Chapter 9). Chapter 10 covers the important problem of multiple testing in fMRI. Chapter 11 is the other end of the "bridge" connecting to Chapter 4 – a look back at additional statistical questions in light of the knowledge acquired in the previous chapters. Finally, Chapter 12 presents analysis of a real data set as a simple case study.

It is worth emphasizing that no book of this nature can ever be completely comprehensive, nor can it be totally current. The pace of statistical research and innovation is such that, almost by definition, such a book would be out of date before it could be published. I have aimed instead to give readers an overview, with some detail, of the most commonly used methods, sprinkled with an accounting of some of the more idiosyncratic approaches. In this way I hope to show the richness and creativity of existing statistical analyses and make new researchers aware of what has already been attempted.

I have been fortunate in my more than ten years of working in this field to have learned from and interacted with many talented statisticians and psychologists. My thanks go to Jeongyoun Ahn, Jim Becker, Yoav Benjamini, Dulal Bhaumik, DuBois Bowman, Pat Carpenter, Bill Eddy, Chris Genovese, Robert Gibbons, Marcel Just, Ming-Hung (Jason) Kao, Tim Keller, Christine Krisky, Yehua Li, Beatriz Luna, Jennifer McDowell, Rebecca McNamee, Abhyuday Mandal, Stephen Miller, Ana Moura, Tom Nichols, Todd Ogden,

Cheolwoo Park, Sumitra Purkayastha, Lynne Seymour, Taniya Sikdar, Andrew Sornborger, Lin Sun, John Sweeney, Keith Thulborn, Joel Welling, Nathan Yanasak, Jun Ye, and Qun Zhao for helpful conversations over the years and for reading parts of this manuscript as it was in progress. Several anonymous reviewers provided useful comments and suggestions. Special thanks to Heidi Sestrich for help with Latex. My appreciation also to John Kimmel at Springer for his patience and technical advice.

As always, none of this would have been possible without the help, encouragement and love of my parents, Morty and Rita Lazar, my brother Michael, and my husband David Sidore. My heartfelt thanks to you all for seeing me through this latest endeavor.

Nicole Lazar
Athens, Georgia
January 2008

Contents

1

The Science of fMRI

1.1 A Quick Tour of the Brain

A challenging aspect of starting to work as a statistician in the area of brain research is learning about neurophysiology. This is necessary if one is to hold meaningful conversations with psychologists and neuroscientists.

We start with orientation. This refers to common terms in use by scientists to describe the different perspectives for looking at the brain. The brain is divided into two hemispheres, the left and the right, separated by the *corpus callosum*. The corpus callosum can be thought of as the *midline* of the brain. The direction away from the midline is *lateral*, whereas the direction toward the midline is *medial*. *Proximal* and *distal* mean closer and farther away, respectively.

The *rostral* or *anterior* position refers to the front end of the brain (behind the forehead); the other end of this axis, the hind end, is *caudal* or *posterior*. The final axis is *dorsal/superior* (the top of the brain) versus *ventral/inferior* (the bottom side of the brain). These are depicted in Figure 1.1.

It is helpful to be familiar with these terms, since many specific brain areas are described by their positions along the various axes of orientation.

The brain comprises three main parts – *forebrain*, *midbrain*, and *hindbrain*. The forebrain is made up of the cerebrum, the thalamus, and the hypothalamus. The cerebrum, also known as the cortex, is the largest part of the human brain and it is responsible for higher brain function, the sort that is of interest in fMRI. The cortex is often compared to a crumpled-up handkerchief - just as a folded-up cloth can fit into a smaller physical space than one laid out flat, so too does the "crumpling" of the cortical "cloth" allow the cortex to fit inside the human skull (the unflattened cortex has an area of around 2500 cm^2, on average; see for example Huettel et al. 2004). Furthermore, the crumpling of a cloth, and likewise of the cortex, creates folds – *sulci* in neuroanatomical terms, surrounded by bulges of cloth, or *gyri*. A consequence of having many folds and bulges is that a crumpled-up cloth has a much larger surface area than a smoothed one. The complex structure of the cortex has

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_1,

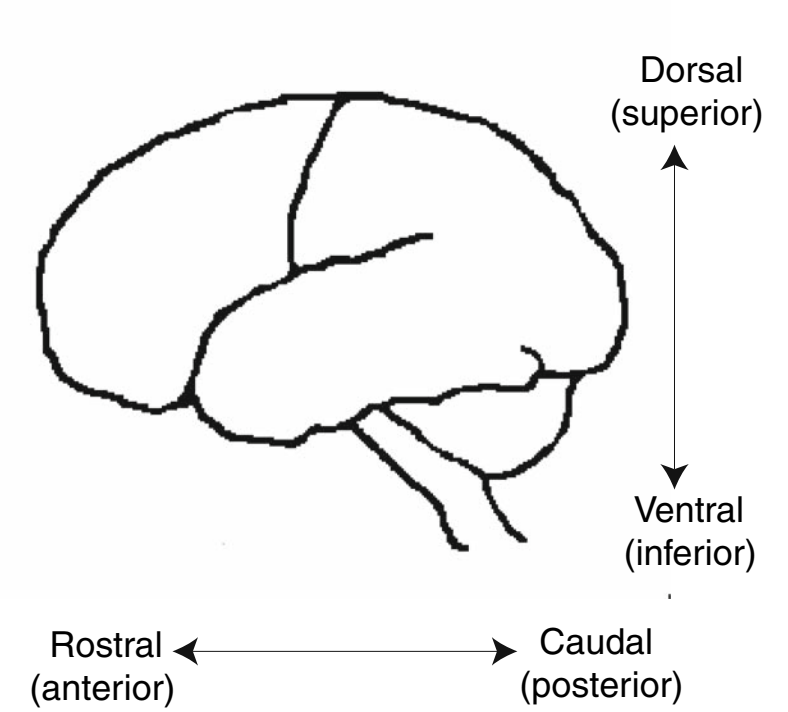


Fig. 1.1. Sketch of a brain with some of the orientation axes drawn in.

been related to higher brain function: higher animals tend to have more cortex (Carlson, 1981). The thalamus is involved in sensory and motor function; it acts as a "relay" for neural input to the cerebral cortex. The hypothalamus is associated with regulating homeostasis in the body, thirst, hunger, circadian rhythms. It is also involved in the modulation of reflex reactions and behaviors (such as fighting and fleeing) that are related to survival. The thalamus and hypothalamus, together with the amygdala and hippocampus, make up what is known as the *limbic system* (Carlson, 1981). The limbic system plays a central role in emotional behavior.

The hindbrain contains the cerebellum, pons, and medulla. These latter two, together with the midbrain, make up the *brain stem*. The cerebellum ("little brain") handles movement, posture, and balance, as well as fine motor coordination. The brain stem is responsible for basic life functions, such as breathing and regulation of the cardiovascular system (Carlson, 1981).

The cerebral cortex consists of four *lobes*: *frontal*, *parietal*, *occipital*, and *temporal*, as seen in Figure 1.2. Each lobe is broadly responsible for different functions of the brain. The frontal lobe, which is at the front of the brain, behind the forehead, is involved with higher function such as reasoning and planning. It also has a role in problem solving, emotion, and motor control.

The parietal lobe, which is caudal to the frontal lobe, and separated from it by the *central sulcus*, is associated with recognition, perception, and orientation. At the back of the cortex is the occiptal lobe, which is responsible for visual processing. The temporal lobe, located on the ventral part of the cerebrum, is involved with memory, speech, and the recognition of auditory stimuli (Huettel et al., 2004). It is separated from the parietal and frontal lobes by the *Sylvan fissure*.

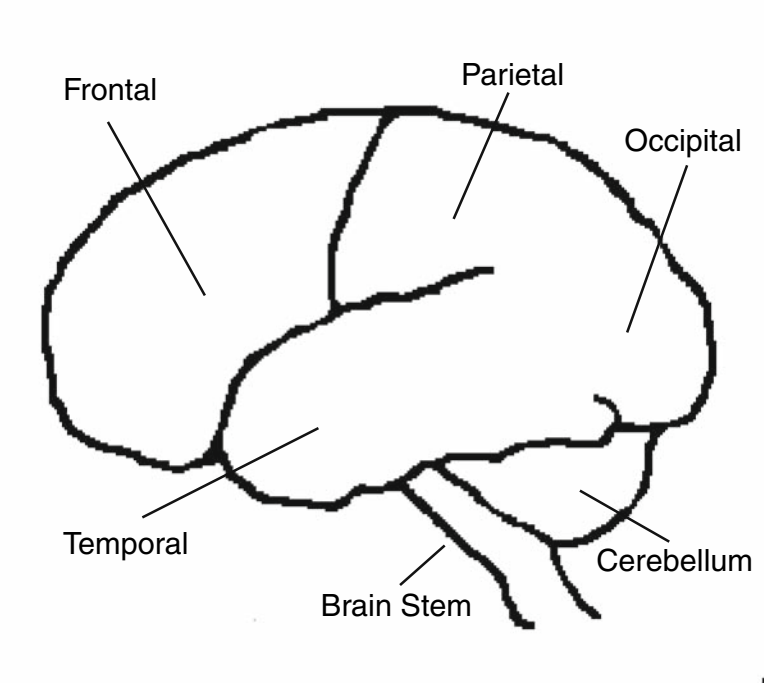


Fig. 1.2. Sketch of the brain showing the four lobes, the cerebellum, and the brain stem.

1.2 The Science of fMRI

How does functional MRI work? To answer this question we first need to understand the concept of magnetic resonance (MR). From it we can advance to an understanding of magnetic resonance imaging as a general technique for studying anatomical structure, and finally explore applications to functional imaging. More detailed explanations can be found in Brown and Semelka (1995) and Hashemi et al. (2004).

1.2.1 Introduction to Magnetic Resonance

Our bodies are made up of atoms, which in turn consist of protons (possessing a positive charge), electrons (possessing a negative charge), and neutrons (having no charge). The protons and neutrons are in the nucleus of the atom, and the electrons are outside of the nucleus. The number of each type of particle (proton, electron, neutron) determines the specific characteristics of the atom. Two characteristics that are of particular interest for MR are the *atomic number* and the *atomic weight*. The atomic number refers to the number of protons, and is the same for all atoms of a particular element. For instance, all carbon atoms have an atomic number of 6 since carbon atoms have six protons. The atomic weight is the sum of the number of protons and neutrons. This may differ for different atoms of an element, or *isotopes* of the element. Continuing the carbon example, one isotope of carbon has six neutrons as well as six protons, giving it an atomic weight of 12. Another isotope has seven neutrons, for an atomic weight of 13. These are denoted ^{12}C and ^{13}C, respectively. Both of these isotopes have an atomic number of 6.

The importance of the atomic weight and number for MR stems from the fact that these two determine a third property of an atom, namely its *spin*. The possible values of spin depend on the atomic weight and the atomic number as follows: Spin is 0 if both the atomic weight and the atomic number are even. The ^{12}C isotope of carbon has 0 spin. Nuclei with 0 spin cannot be studied using MR. On the other hand, spin is of half-integral value (1/2, 3/2, 5/2, etc.) if the nucleus has an odd atomic weight. And spin is of integral value (1, 2, 3, etc.) if the nucleus has an odd atomic number and even atomic weight. Since most elements have at least one isotope that possesses spin, in principle almost any element can be studied using MR. In practice, however, it is common to study the ^{1}H isotope of hydrogen, as it: possesses spin of 1/2; is the most abundant isotope of hydrogen, and hydrogen is found in abundance in the tissues that are the target of magnetic resonance imaging; and is very sensitive to the magnetic field into which the body is placed for the purpose of imaging. Unless stated otherwise, we will assume henceforth that images are based on the ^{1}H isotope. For any other isotope the basics of how the imaging would proceed are the same.

An MR scanner is a large magnet that generates a magnetic field that is many times more powerful than the natural magnetic field of the earth. The strength of a magnetic field is measured in *Tesla* (T). One Tesla is 10,000 *Gauss*. The fields generated by MR scanners range from 1.5T and up. By contrast, the magnetic field of the earth is around 0.5 Gauss, or 0.00005T (Hashemi et al., 2004). Atoms that are placed in a magnetic field of given strength, usually denoted $\mathbf{B}_0$, absorb photons of frequency ω if the atoms have nonzero spin. The frequency of absorption depends on the *gyromagnetic ratio* γ of the nucleus, via the *Larmor equation*

$$\omega = \gamma \mathbf{B}_0.$$

γ differs for different elements, and also for the different isotopes of the same element; for ^{1}H, $\gamma = 42.58$ megahertz per Tesla. This high value of γ (for comparison, the value of γ for ^{13}C is only 10.71), together with its natural abundance in the body, is what makes hydrogen particularly easy to image using magnetic resonance.

Particles with spin naturally spin, or *precess*, around a central axis, much as a gyroscope or spinning top. In the natural state, the nuclei in the body precess in random directions, and the net magnetization of the body is zero, since the spins in the different directions cancel each other out. When the body is exposed to a magnetic field, the hydrogen atoms in particular have a tendency to line up in the direction of that field, with about half going parallel and half going anti-parallel. Significantly, the alignment isn't exactly half and half, and it is the slight preference of the particles to go in the direction of slightly lower energy (the parallel orientation) that allows magnetic resonance imaging to work at all. As more nuclei line up in the direction that is parallel to the magnetic field, the object inside the scanner becomes slightly magnetized. The magnetization of the object inside the field is denoted $\mathbf{M}$, which can be thought of as a vector with direction and magnitude (length). The components of the vector are *longitudinal*, aligned in the direction of the magnet (this is usually depicted on the z-axis), and *transverse*, aligned orthogonal to the main magnetic field (in the (x, y)-plane). The nuclei continue to precess, but now the directions are no longer random – precession is along the axis that is parallel to the magnetic field. Whereas the atoms are lined up in the direction of the field and rotate along the axis defined by the direction of the field, each atom is precessing at a different phase. Returning to the spinning top image, we can think now of looking at many tops, all gyrating around central axes that are the same along the z dimension. The different tops, however, aren't spinning all in tandem, so they have varying values in the x and y dimensions. The result of these processes is that the spins in the (x, y)-plane cancel each other out, so that the net magnetization in the transverse direction is zero (or close to it), and the total magnetization is derived from the difference in alignment (parallel to the field or anti-parallel) of the atoms. As the proportion of spins in the parallel alignment increases, so too does $\mathbf{M}$.

There are two ways of increasing the number of protons in the low energy state (Huettel et al., 2004). One is to decrease the ambient temperature, since the number of spins in the parallel alignment increases as the temperature goes down. However, as noted by Huettel et al., in order to observe a meaningful change in the net magnetization, a large drop in temperature is required, so that this approach for increasing $\mathbf{M}$ is not feasible, in general. The second way is to increase the strength of the external magnetic field – as the field strength goes up, more protons, proportionally, align in the parallel, lower energy direction, a phenomenon known as the *Zeeman effect*.

1.2.2 Acquiring MR Images

The next step is to inject additional energy into the system, in the form of radiofrequency (RF) pulses. When the RF pulse is at the right frequency (the so-called *resonance frequency*), the protons absorb the energy, and gradually release it, to return to their initial state. Effectively, some of the spins in the low energy state are excited by the RF pulse, jumping to the higher energy state. When the pulse is turned off, excited protons emit energy as they return to the parallel orientation. The emitted energy is detected in turn by radiofrequency coils in the MR scanner. Magnetic resonance imaging takes advantage of these dual processes of absorption and re-emission (or relaxation): by applying the RF pulses in an appropriate fashion, it is possible to uniquely identify each location in the space that is being imaged. This is accomplished through the use of *gradient pulses*, or small perturbations to the main magnetic field, which are applied in each of the three directions x, y, and z. The differential application of magnetic gradients is fundamental to the formation of MR images. Hence, as we will see in the next chapter, there are many ways in which the gradient pulses can be applied, and finding optimal strategies is a major area of research among MR physicists. Here, we survey the basic principles that underlie all different pulse sequences.

Atoms that are precessing near the frequency of the RF pulse (as determined by the Larmor equation) are the only ones that will be affected. For hydrogen atoms in a 1.5T scanner, applying an RF pulse of 63.87 MHz (the Larmor frequency of (42.58 MHz/Tesla) times 1.5T) will move some nuclei from the low energy state to the high energy state. When energy is injected into the system at this frequency, the affected protons are "tipped," or aligned in a uniform angle (the *flip angle*). As a result, $\mathbf{M}$ is flipped away from its orientation at equilibrium (i.e., when there is no extra energy in the system) and toward the transverse plane that is orthogonal to the axis of the original field. The tipped protons all precess in phase, by contrast with the situation at equilibrium, where they precessed at random in the direction of the magnetic field. The resultant dynamic field generates current in the receiver coils proportional to the number of hydrogen atoms in the tissue, as the atoms emit energy (once the RF pulse is turned off) as they return to equilibrium.

Specific application of the gradients proceeds as follows. Recall that the body in the scanner is parallel to the field of the magnet. By convention, this direction is denoted as the z-axis. If we don't apply any extra energy beyond the basic RF pulse, there is no spatial discriminatory ability – an echo signal of the object in the scanner will be obtained, but there will be no way of distinguishing locations in the object. If that RF pulse is not at the Larmor frequency, there will be no excitation of protons. By using a gradient coil we can vary the strength of the magnetic field, so that each location in the brain has its own resonance frequency. The first step is to apply a gradient in the z direction; this will cause the strength of the magnetic field to vary, usually in a linear fashion, from the top of the brain to the bottom, for instance. To

exemplify this numerically, suppose the subject is inside a 3T scanner. The effect of the gradient coil is to make the field slightly stronger than 3T at, say, the top of the head, and slightly lower than 3T at the bottom of the head. If we now apply a radiofrequency pulse at the frequency corresponding to 2.9T, say, only spins in those parts of the brain that are exposed to the 2.9T field will be precessing at the appropriate rate as determined by the Larmor equation. The other protons won't be affected at all.

Therefore, after application of a gradient G_z in magnitude to the field in the z direction, a slice of the brain is selected. Call this slice z_1. All protons in the particular slice affected by the gradient will precess at the same rate, and this rate in turn matches the frequency of oscillation of the RF pulse. Nuclei in other slices will be precessing either too quickly or too slowly (depending on their location), and hence will not be able to absorb the RF energy – they won't *resonate*. Changing the gradient shifts the focus to different slices of the brain. Furthermore, as the gradient is applied, energy is added to that already in the field, so that protons in two different slices, say z_1 and z_2, will have different resonant frequencies, ω_1 and ω_2. This process is demonstrated schematically in Figure 1.3.

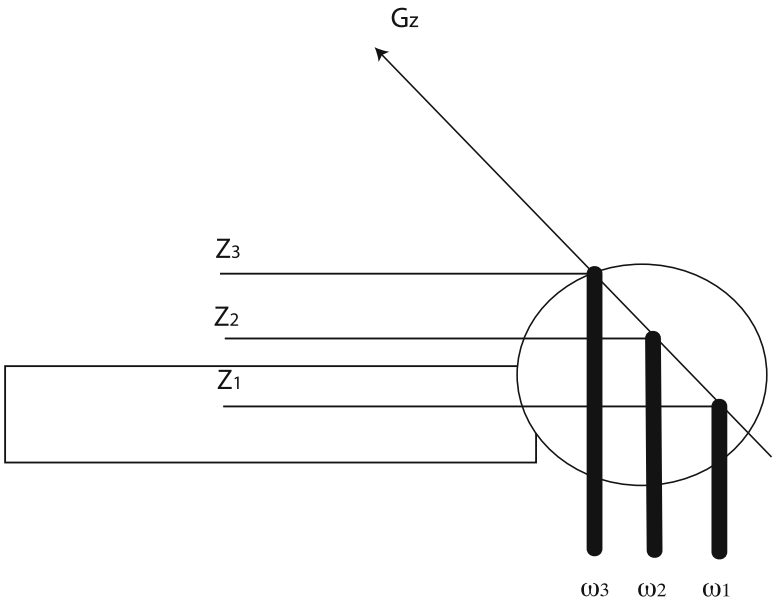


Fig. 1.3. The slice selection process. When the gradient is applied, the total magnetic field to which a proton is exposed will depend on its location, according to the Larmor equation. At location z_i tissue absorbs energy with frequency centered at ω_i.

A similar principle holds for localizing in the transverse plane. After the slice is chosen, we have to encode in the other two directions, that is, along the (x, y)-plane perpendicular to the field. This occurs in much the same way: a gradient is applied in the x direction, and another in the y direction. To understand more clearly how this works, we can think of the tissue in the chosen slice as being divided into an array with rows and columns. The protons in the slice are precessing in phase. Now, apply a *phase encoding* gradient to the field. This G_y gradient will be increased (linearly, in general) as we move from the bottom row of the array to the top row. Now, nuclei that sit in different rows precess at slightly different phases, however, within a row, the protons are still precessing in phase. Finally, a *frequency encoding* (or readout) gradient G_x is applied, which increases as we move from the leftmost column of the slice to the rightmost. As a result, the nuclei in different columns are precessing at different frequencies, although within a column the frequency of precession is constant. Within the array defined by the slice, each element can be distinguished from all the others, since each location has a different phase (the y-axis coordinate) and frequency (the x-axis coordinate), as shown in Figure 1.4.

The localization within the field can be made more explicit by considering an expanded version of the basic Larmor equation:

$$\omega_i = \gamma(\mathbf{B}_0 + \mathbf{G} \times \mathbf{r}_i).$$

In this expanded equation, ω_i is the frequency of the proton at position $\mathbf{r}_i$ and $\mathbf{G}$ is a vector summarizing the amplitude and direction information of the gradient (Hinshaw and Lent, 1983). As can be seen from the equation, in the presence of the gradients, each proton will resonate at its own frequency. Hence, the MR image is, in effect, a map of these frequencies. The intensity of the image at a given pixel is proportional to the number of protons in the corresponding voxel (volume element), weighted by the relaxation times for the tissues that are in that voxel.

1.2.3 Relaxation

Recall that when RF energy is injected into the system at the correct frequency, it is absorbed by the protons. Once the RF pulse is turned off, the protons start to *relax*, or emit the energy in an attempt to return to their equilibrium state. There are three relaxation times that are relevant in magnetic resonance imaging; these are denoted T_1, T_2 and T_2^*.

T_1 is the *spin-lattice* or *longitudinal* relaxation time. This is the time required by the z component of the excited magnetic field to return to 63% of its original value, after an excitation pulse. At equilibrium the magnetized field $\mathbf{M}$ is parallel to the static magnetic field $\mathbf{B}_0$. Absorption of energy causes $\mathbf{M}$ to rotate relative to the static field. The T_1 relaxation process is therefore the mechanism by which the protons release energy and return to their equilibrium (low energy) orientation. The T_1 relaxation curve can be described by

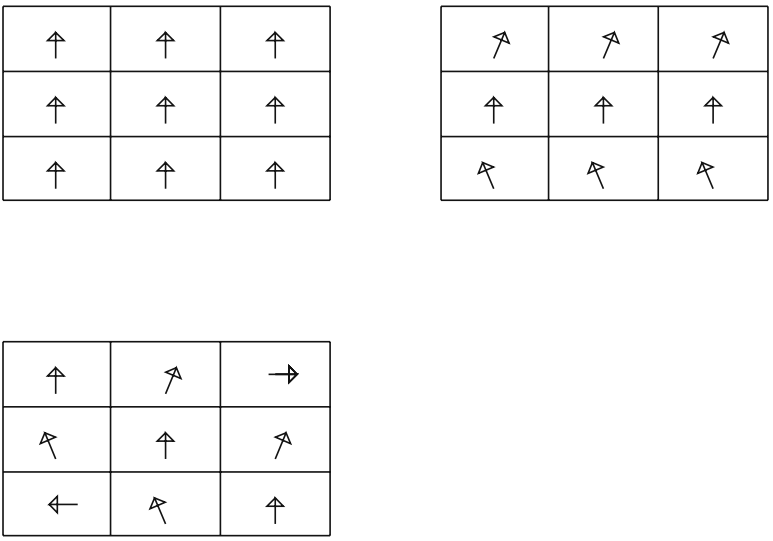


Fig. 1.4. Schematic of the effect of the phase and frequency gradients. Prior to the application of the gradients, all of the elements in the slice are precessing at the same phase and frequency, as indicated by the arrows in the top left matrix all pointing in the same direction. Application of the gradients in sequence causes each location in the (x, y)-plane to precess at a distinctive phase and frequency – first the rows are shifted relative to each other, as seen in the top right matrix, then the columns are shifted within each row, as seen in the matrix on the bottom left. This allows localization of the areas selected for imaging.

an exponential function, $1 - e^{-t/T_1}$, where t is the elapsed time; if $\mathbf{M}_0$ is the original magnetization, then $\mathbf{M}_z$, the amount of longitudinal magnetization at time t following an excitation pulse, is given by

$$\mathbf{M}_z = \mathbf{M}_0(1 - e^{-t/T_1}).$$

Hence, for instance, after two T_1 time periods, $1 - e^{-2*T_1/T_1} = 0.86$; the magnetized field will be at 86%, relative to the level prior to the excitation pulse (see Figure 1.5).

The second relaxation time T_2 is the *transverse*, or *spin-spin*, relaxation time. T_2 represents the time needed for the transverse component of the magnetized field to return to 37% of its initial value (see Figure 1.6). This type of relaxation is the result of the gradual loss of phase coherence, and hence is a result of inhomogeneities in the tissue (so-called "internal inhomogeneities"). It can also be described by an exponential function, e^{-t/T_2}, or

$$\mathbf{M}_{xy} = \mathbf{M}_0 e^{-t/T_2},$$

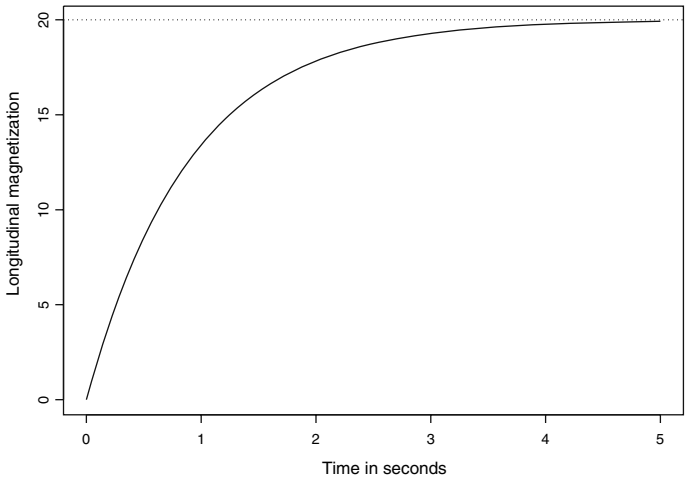


Fig. 1.5. Curve of longitudinal relaxation. T_1 is the time required by the z component of the field to return to 63% of its original value following an excitation pulse.

where $\mathbf{M}_0$ is as before and $\mathbf{M}_{xy}$ is the signal loss at time t in the transverse plane. Since loss of phase coherence must occur before equilibrium can be reached, T_2 is usually much shorter than T_1.

T_2^* is a relaxation time related to a phenomenon called *free induction decay*, or FID. As soon as the RF pulse is turned off, the imposed structure also begins to fade: the spins return to precessing at random (or "freely"), and there is a decay of the signal over time. The typical pattern of the FID is in Figure 1.7.

Note that the general form is sinusoidal, but the amplitude gets dampened down over time. The relaxation in the spins induces a current in the receiver coil. And the rate of decay due to FID is denoted T_2^*. In contrast to T_2, T_2^* depends on the external field, as well as the spin-spin interactions. That is, it is a function of both external (magnet-related) and internal (tissue-related) inhomogeneities. Magnets with less homogeneous fields have higher values of T_2^*, regardless of the value of T_2. Hence T_2^*, which again represents a rate of decay, is always smaller than T_2, unless perfect homogeneity of the main magnetic field is achieved. As systems improve the amount of field inhomogeneity is reduced; however it is probably not possible to attain total homogeneity of the external magnetic field. Hence there will always be some T_2^* effect.

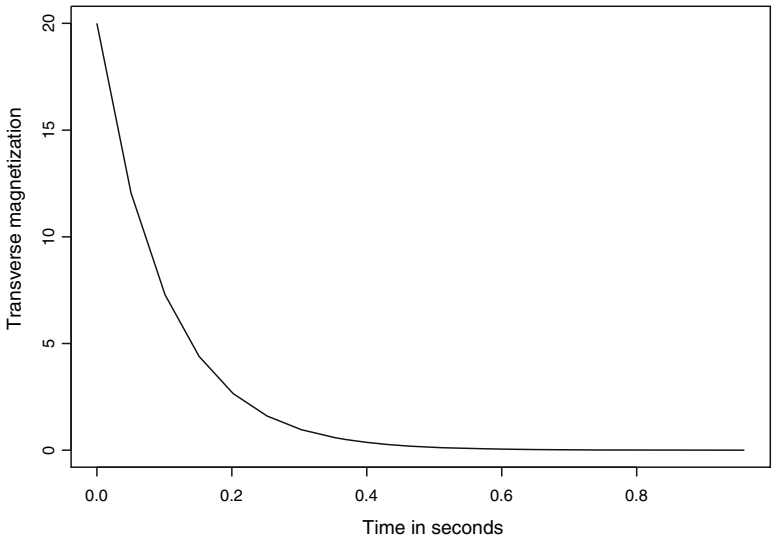


Fig. 1.6. Curve of transverse relaxation. T_2 is the time required for the transverse component of the field to return to 37% of its initial value.

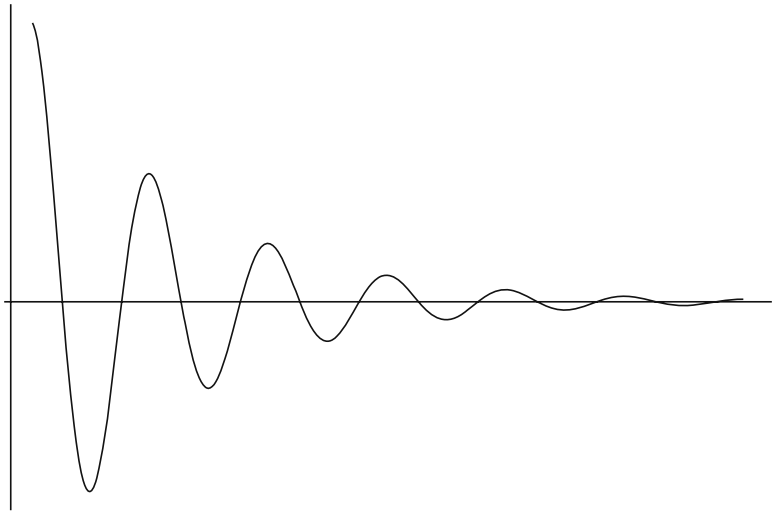


Fig. 1.7. The free induction decay curve.

The values of T_1 and T_2 differ for gray matter, white matter, and cerebrospinal fluid, and for magnets of differing strengths (although in principle, as we have seen, T_2 should depend only on the type of tissue, in practice there are fluctuations in the observed value). For instance, at 1.5T the values of T_1 for gray matter, white matter, and cerebrospinal fluid are roughly 900 ms, 600 ms, and 4000 ms, respectively, whereas the T_2 values are 100 ms, 80 ms, and 2000 ms, respectively. In all cases, as we would expect, the T_2 times are much shorter than the T_1 times. It is possible to take advantage of the differences among tissue types to weight images in such a way that greater contrast is achieved.

1.2.4 From MRI to fMRI

It remains to be seen how the technique described in the previous sections can be used to image the working brain. What is the connection between brain function, as exhibited in neuronal activity, and magnetic resonance, and how is this connection exploited to create functional magnetic resonance images?

Although the story is still not completely known, neuroscientists have a good idea of the mechanisms at work. When the brain becomes active in response to a particular task or stimulus, the rate of blood flow to the regions involved in the task or affected by the stimulus increases. The increase in blood flow occurs because glucose needs to be delivered to the relevant areas. The metabolism of the neurons in the affected areas also changes, as their rate of firing increases. As a result of the increase in metabolism, more oxygenated blood arrives in the relevant regions. However, active neurons do not require much more oxygen than do inactive neurons, and hence there is an increase in the oxygen levels, not of the neurons themselves, but rather of the proximate blood vessels. The increase in metabolic demand of the active neurons, and not the activity of the neurons per se, is what is measured by fMRI.

The changes in the ratio of oxygenated to deoxygenated blood are measured via the *hemodynamic response*, estimation and characterization of which is the focus of much of the statistical research in fMRI. Figure 1.8 shows a schematic of a typical hemodynamic response function (HRF) for a voxel in an active part of the brain. Upon presentation of the stimulus, there is a delay of approximately 2 seconds before any change is observed, as blood is delivered to the relevant area. A gradual increase in the response peaks at about 6 seconds following the stimulus. If there is no further stimulation, the HRF starts to slowly decay, returning to baseline levels. Often a dip below baseline is observed before complete recovery. It takes approximately 15 to 20 seconds to return to baseline levels, depending on the experimental task.

Magnetic resonance enters into this description because blood contains iron, which is paramagnetic. A paramagnetic material has the property that when it is placed in a strong magnetic field, the atoms in that material try to align themselves with the field, thereby increasing the field strength. In other words, the paramagnetic material becomes a magnet, as long as the

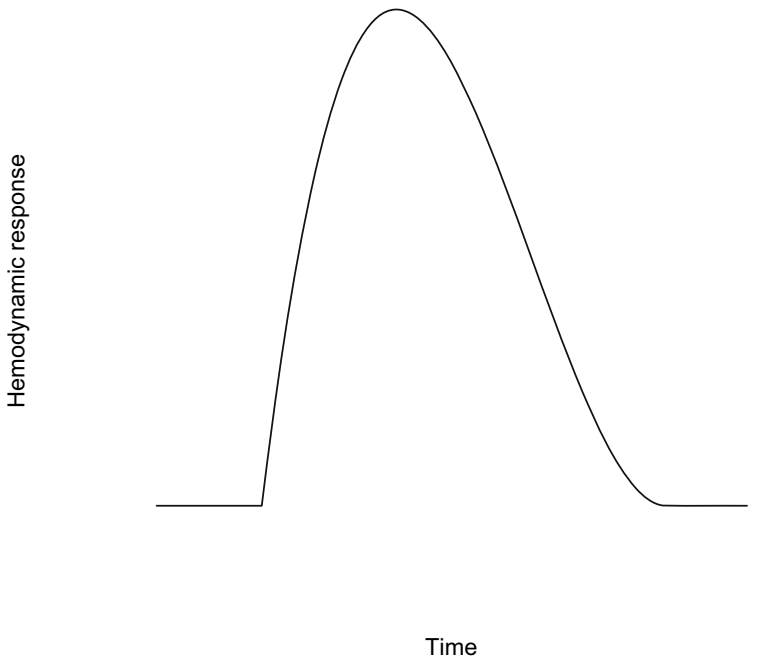


Fig. 1.8. The hemodynamic response. Following presentation of a stimulus there is a delay of approximately 2 seconds before a change in the signal is observed. The response increases gradually, peaking at around 6 seconds after the stimulus. If no further stimulus is presented, the response will decline to baseline. The whole process takes 15 to 20 seconds. Dips below baseline are sometimes observed before the initial increase in signal, as well as prior to the final return to baseline.

field is present (in our context, as long as the subject is inside the scanner). As shown by Pauling (Pauling and Coryell, 1936), the magnetic properties of oxygenated and deoxygenated blood differ; deoxygenated blood is more paramagnetic than oxygenated blood. This affects the measured MR signal through the Blood Oxygenation Level Dependent (BOLD) contrast effect, since changes in oxygenation of the hemoglobin will cause changes in the local magnetic field applied to the body.

More specifically, deoxygenated blood has a *magnetic susceptibility* that is 20% greater than that of oxygenated blood. Magnetic susceptibility is a measure of the intensity of magnetization of a substance when it is placed in an external magnetic field, or, thought of another way, the distortion induced in the field with the introduction of the substance. The effect of introducing

a substance with magnetic susceptibility into a magnetic field is a decay of transverse magnetization, which is related to the times T_2 and T_2^*.

In a seminal work, Thulborn et al. (1982) showed that the rate of T_2 relaxation of blood is related, via an exponential relation, to the proportion of deoxygenated hemoglobin. As the strength of the applied magnetic field increases, furthermore, so too does this BOLD effect. At low fields (less than 0.5T), there is not much difference in the transverse relaxation times for oxygenated and deoxygenated blood. Hence, high fields (that is, 1.5T and above) are necessary for BOLD fMRI. Also, and of critical importance for the development of fMRI as an imaging technique, it is possible to determine the level of blood oxygenation directly from the MR signal of the blood.

Ogawa et al. (1990) took this further, in the first "true blood oxygenation level dependent (BOLD) contrast experiment" (Matthews, 2001, p. 11). They showed that, following deoxygenation of the blood, magnetic susceptibility of blood vessels (relative to the surrounding tissue in the brain) increased. The effect of this is twofold: first, local field gradients are generated, and second, T_2^* in tissue water around the blood vessels decreases. This change is exploited in BOLD-fMRI, the most common type of fMRI, and the focus of this book.

The implication of these early experiments is that, as the concentration of oxygenated blood in the vicinity of a neuron changes, the measured MR signal should be affected. And, MR imaging is indeed sensitive enough to detect these changes, which are induced by the function of the brain (Ogawa et al., 1992; Kwong et al., 1992), although there is still uncertainty surrounding the precise mechanisms and the connections among the various processes. Crucially, fMRI does not measure brain activity directly, but rather correlates of brain activity. See Huettel et al. (2004) for a more in-depth review of some of the outstanding issues.

1.2.5 From Data to Image

After all of this, we still don't have an image of a brain that would be recognizable as such. That is because the data are collected in Fourier space, known as *k-space* in the fMRI literature. The next step in our journey is to gain some understanding of the data in k-space, how they relate to the data in "image space," and how to move from the former to the latter. Generally, statistical analysis takes place in image space, not k-space, although some preprocessing may be done in k-space.

It is helpful to recall the Larmor equation for a particular (x, y)-coordinate in a slice. See Jezzard and Clare (2001) for more detail.

$$\omega(x, y) = \gamma \mathbf{B}_0 + \gamma G_x x + \gamma G_y y$$

Recall that the y-axis corresponds to phase encoding, and the x-axis to frequency encoding. As a matter of convention, the signal obtained with no phase encoding gradient is placed in the center of k-space. As we move along

the y-axis, the phase encoding gradient increases. When the G_x gradients are applied, we *sample* a row in k-space for a given phase. As we do this for each phase in turn, we fill in a "data matrix" of k-space values corresponding to changes in both frequency and phase at a chosen slice.

Let $\rho(x, y)$ denote the density of hydrogen nuclei at location (x, y); this is the ultimate quantity of interest. The signal induced in the receiver of the system is a vector with magnitude $\rho(x, y)$ times the size of the elemental unit, $dxdy$, and phase, denoted by $\phi(x, y, t) = 2\pi\omega(x, y)t$. Thus the contribution to the signal from a particular position is:

$$\rho(x,y)\{\cos[2\pi(\gamma\mathbf{B}_0 + \gamma G_x x + \gamma G_y y)t] + i\sin[2\pi(\gamma\mathbf{B}_0 + \gamma G_x x + \gamma G_y y)t]\}dxdy$$

(Jezzard and Clare, 2001). Here, the cosine is the "real" part, or the contribution along the x-axis, while the sine is the "imaginary" part, the contribution along the y-axis. When the signal is received, there is an additional demodulation, which allows for the static field $\mathbf{B}_0$ to be ignored, hence the signal that is actually stored by the scanner at time t (integrating now over x and y) is

$$I(t) = \int\int \rho(x,y)\{\cos[2\pi(\gamma G_x x + \gamma G_y y)t] + i\sin[2\pi(\gamma G_x x + \gamma G_y y)t]\}dxdy.$$

The significance of this expression is made clearer by introducing the following standard notation for locations in k-space: $k_x(t) = 2\pi\gamma G_x t$ and $k_y(t) = 2\pi\gamma G_y t$. With this notation, we have

$$I(t) = \int\int \rho(x,y)[\cos(k_x x + k_y y) + i\sin(k_x x + k_y y)]dxdy.$$

In other words, the measured signal is the Fourier transform of the ρ values of interest. It is therefore possible to recover the $\rho(x, y)$ densities by applying the inverse Fourier transform to the signal that is measured and stored by the scanner.

The center of k-space contains the strongest signal (recall that in the phase encoding direction, the location with no phase encoding is in the center). As we move out to the periphery in either direction, the signal becomes weaker. In fact, the signal in the precise center of k-space overwhelms the rest of the recorded signal and its effect has to be removed in order for the Fourier transform to result in an image that looks like a brain. Even though the signal in the periphery of k-space is weaker than that in the center, it cannot be ignored. Information on fine details of the image is contained on the edges of k-space, and ignoring it results in blurry reconstructed brain images.

2

Design of fMRI Experiments

In addition to the usual questions of experimental design that are relevant for any psychological experiment, fMRI studies present a unique set of design parameters that need to be determined. These parameters refer to the details of the data acquisition from the scanner. In this chapter we explore issues relating to the design of fMRI experiments from imaging and statistical perspectives.

2.1 Imaging Design Issues

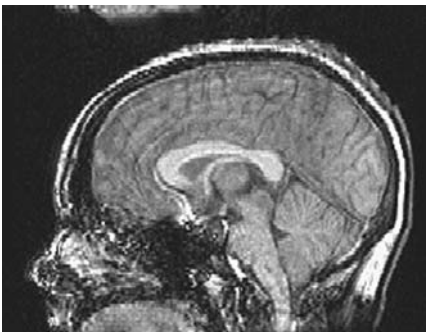
The imaging parameters are fixed at the start of any study. By changing the values of the imaging parameters, the technologist who operates the scanner controls the spatial and temporal resolution of the data, and hence the overall quality of the resultant images. It is standard to report the values of these parameters in the discussion of fMRI studies.

2.1.1 Description of Parameters

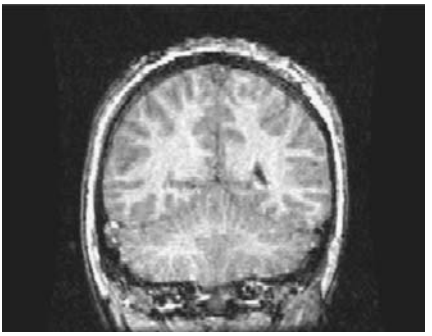
Images can be acquired in one of four *imaging planes*: axial, coronal, sagittal, and oblique. The first three are more common in practice. Axial slices are perpendicular to the longitudinal axis of the body – a series of axial slices will go from the top of the brain, down. Coronal, or frontal, slices are parallel to the front of the body – a series of coronal slices will proceed from the front to the back of the brain. Sagittal slices are obtained in parallel to the midline of the body – a series of sagittal slices will proceed from one side of the brain (say, the left side) to the other. Finally, an oblique slice is taken by tilting any of the three standard views. Examples of the three primary perspectives are shown in Figure 2.1.

The *field of view* (FOV) is the physical size of the image, measured in mm^2, and specifies the region from which the image was sampled. For instance, if the FOV is 20 cm in each direction, this means that the slices encompassing

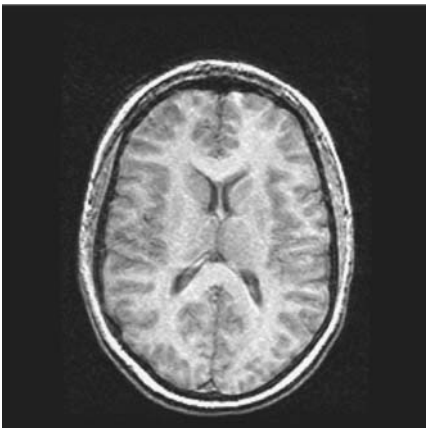
N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_2,



(a) Sagittal slice



(b) Coronal slice



(c) Axial slice

Fig. 2.1. Examples of the three primary slice orientations. *Figures courtesy of Rebecca McNamee, University of Pittsburgh.*

the regions of interest are contained within a 20 cm x 20 cm region in the plane.

The *acquisition matrix size* is the size of the grid into which the plane of the FOV is divided for each slice. The acquisition matrix is usually square, often 64 x 64 or 128 x 128. The FOV and the size of the acquisition matrix determine the two dimensions of a voxel in the plane of a slice. Thus, if the field of view is 20 cm in a particular direction, and the matrix size is 64 in that same direction, voxels will be of length $200/64 = 3.125$ mm on a side. These are typical values.

Slice thickness is the thickness, measured in mm, of an individual slice. If the x and y dimensions of the three-dimensional object are given by the FOV

and the acquisition matrix, the z dimension is given by the slice thickness. The *slice gap*, also in mm, is the space between consecutive slices. The reason for leaving a gap between slices, historically, is that the radiofrequency pulses were imperfect, and there could be "crosstalk" between slices if there was no gap. This results in contamination of the signal, since the slices aren't perfectly separated from each other (Hashemi et al., 2004). A typical study might use slices that are 5 mm thick with a gap of 1 mm. Together with the FOV and matrix size values described in the previous two paragraphs, this results in voxels that are 3.125 mm x 3.125 mm x 5 mm, which are often seen in practice. With improvements in imaging techniques, many studies now do away with the slice gap altogether.

The order in which slices are acquired is fixed by the *excitation sequence*. While it might be intuitively appealing to image slices in a physically sequential fashion, this leads to crosstalk. Hence slices are usually acquired in an interleaved manner, for instance, all even slices, then all odd slices. Arbitrary sequencing is also possible with some software.

All of the parameters described above relate to the details of the data collection, such as the size of a voxel. Brown and Semelka (1995) call these *extrinsic parameters*. There are also *intrinsic parameters*, which affect the voxel's signal.

Repetition time (TR) is the time, in milliseconds, between successive applications of the radiofrequency pulses to a particular volume of tissue. Suppose then that we have applied a single 90^o pulse; the time until we apply the next one is the TR. What happens between the two pulses?

Immediately after the first pulse is applied, the magnetization flips from being aligned in the z direction to being in the (x, y)-plane, as was described in Chapter 1. Call the net magnetization in the (x, y)-plane $\mathbf{M}_0$. As soon as the pulse is turned off the magnetization in the transverse plane starts to decay, with a concomitant recovery of magnetization in the z direction, according to the formula

$$\mathbf{M}_z(t) = \mathbf{M}_0(1 - e^{-t/T_1}),$$

as we saw previously. At time $t =$TR, i.e., when we apply the *next* pulse (and hence flip the magnetization back into the (x, y)-plane),

$$\mathbf{M}_z(\mathrm{TR}) = \mathbf{M}_0(1 - e^{-\mathrm{TR}/T_1}),$$

which is less than $\mathbf{M}_0$, the initial magnetization in the z direction before the first pulse was applied. Once this second pulse is turned off, again there is recovery toward the z-axis, but as a new pulse will be applied at time $t = 2\mathrm{TR}$, again the recovery will not be complete. That is, for each successive pulse, the system starts with less than total magnetization $\mathbf{M}_0$.

As the TR increases, there is clearly more time for the radiofrequency energy to dissipate, through the relaxation process described in the previous chapter, and we get closer to the initial magnetization $\mathbf{M}_0$ at the start of each successive pulse.

In terms of the received signal, at time $t = 0$ the signal is a strong FID, also as described in Chapter 1. At successive TRs the signal is still an FID, but with dampened amplitude. Ideally, one would want to measure the signal immediately after the RF pulse, with no delay. While this isn't feasible in practice, if we could do so, the FID signal would be proportional to $1 - e^{-\mathrm{TR}/T_1}$. Similar statements hold for each of the successive FIDs, that is, they are maximal if they can be measured right after the application of the RF pulse.

However, limitations on the hardware make it impossible to measure the signal immediately upon the application of the RF pulse. Instead, there is a brief waiting time between the originating pulse of the image, and the peak of the echo, that is, the acquisition of data from the center of k-space (the maximum of the signal). This short time is called *echo time* or TE, and is also measured in milliseconds. Now, the FID in the transverse plane decays at the rapid rate T_2^* according to the function e^{-t/T_2^*}. If we could take the measurement immediately, before any signal decay, the measured signal would be the initial magnetization, $\mathbf{M}_0$, flipped into the (x, y)-plane. After time TE the measured signal is slightly smaller, namely $\mathbf{M}_0 e^{-\mathrm{TE}/T_2^*}$.

Note that both processes of T_1 and T_2 relaxation are occurring simultaneously, so in fact the measured signal is proportional to

$$\mathbf{M}_0(1 - e^{-\mathrm{TR}/T_1})e^{-\mathrm{TE}/T_2^*}.$$

Thus, it is possible to manipulate the MR signal by changing TR and TE, which are under the control of the experimenter. Roughly speaking, TR is related to T_1 effects in the image, and TE is related to T_2^* (or T_2) effects. We will see this in more detail in the next section.

The *excitation* or *flip angle* sets the amount of rotation away from the equilibrium axis following the radiofrequency excitation pulse. The default value for the flip angle is 90° in most scanners, since this gives the maximum magnetization in the transverse plane. Excitation angle is the other parameter, along with TR, that determines the amount of T_1 weighting in the image.

2.1.2 How Are Resolution and Image Quality Affected by Changes in the Parameters?

Many of the acquisition parameters, such as the flip angle, the size of the acquisition matrix, and the field of view (and hence the voxel size), are fixed by convention, although not by necessity. Others, such as TR, are given more to the control of the experimenter. Yet, in any case it is important to be aware of the tradeoffs in terms of resolution and image quality (for instance, as measured by signal to noise ratio), as well as in terms of total scan time, that result from changes in the values of these parameters. A useful summary of the effects of manipulating the different values on resolution, signal to noise ratio, and scan time can be found in Brown and Semelka (1995).

We consider first the direct effect of manipulating TR and TE. This is seen via the contrast between two different tissue types, call them A and B, with different values of T_1 and T_2^*, which we will denote T_{1A}, T_{1B}, T_{2A}^*, and T_{2B}^*, respectively. The contrast between the two tissue types is given by

$$c_{AB} = \mathbf{M}_{0A}(1 - e^{-\mathrm{TR}/T_{1A}})e^{-\mathrm{TE}/T_{2A}^*} - \mathbf{M}_{0B}(1 - e^{-\mathrm{TR}/T_{1B}})e^{-\mathrm{TE}/T_{2B}^*}.$$

Now, suppose that $T_{1A} > T_{1B}$, in other words, tissue type A has a longer T_1 recovery time than does tissue type B. Put another way, it takes tissue A longer to reach equilibrium than it takes tissue B, so that at any given point in time t the recovery curve for A will be below that for B. Furthermore, at different points in time, the distance between the two recovery curves will be different, as demonstrated in Figure 2.2 – near $t = 0$, the curves will be close together (no recovery in either case), there is no signal from either tissue, and $c_{AB} = 0$. As t goes to infinity, the curves will be close together again (full recovery in both cases), the effect of T_1 is reduced (in the limit vanishing altogether). So, at very short or very long TRs, the contrast between the tissues is not large. Between those two extremes there is a difference between the tissues. Since we have assumed that T_{1B} is the smaller of the two T_1 times, tissue B recovers more quickly and hence has the stronger signal at intermediate TRs. Indeed, for any two tissues that differ in their value of T_1, we have therefore shown that there is a time point that maximizes the distance between the two recovery curves, giving optimal contrast, and in general, shorter TRs will give greater T_1 contrast.

A similar analysis shows that shorter TE values diminish the effect of T_2^* (or T_2), whereas longer TEs enhance the contrast from T_2^*. Often, in the case of fMRI, the TE is set at the gray matter T_2^* time, since this enhances the BOLD effect, although recent studies have employed variable TEs (see, for example, Chen et al. 2003a).

The above results can be used to weight images differentially, according to T_1, T_2, or T_2^*. T_1 weighted images will result from an experiment with an intermediate TR and a small TE. Tissues that have long T_1, such as cerebrospinal fluid, will be downweighted and hence will not be as visible in the image, as tissues with shorter T_1 times, such as white matter. T_2 weighted images are often used as structural reference points in functional studies. These are generated from a long TR and a medium TE; the resultant images highlight regions of cerebrospinal fluid and downweight the white matter. Similar to T_2 contrast, T_2^* contrast is obtained by long TR and intermediate TE. Images weighted in this way are sensitive to the amount of deoxygenated blood in the tissue, which is itself a function of the changing metabolism of the active neurons. In general, it is possible to take advantage of the T_1 and T_2 times of different types of tissues, to tailor images that highlight the areas of interest (see Hashemi et al. 2004, for more detail).

Another important aspect of an fMRI experiment is to control the *signal to noise ratio* (SNR). In fMRI, SNR is proportional to the product of volume of a voxel and $\sqrt{N_y \mathrm{NEX}/\mathrm{BW}}$, where N_y is the number of phase encoding steps

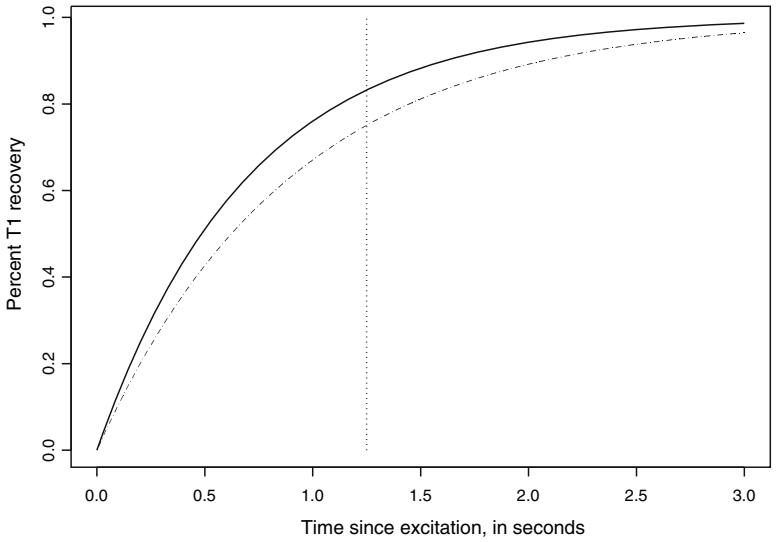


Fig. 2.2. The differences in relaxation times for different tissues can be exploited to weight images, in such a way that different aspects of the brain are emphasized.

(that is, in the y direction), *NEX* is the number of excitations (the number of times the scan is repeated) and *BW* is the bandwidth (the range of frequencies in the slice selection). From this formula a number of things are clear. First, as we increase the voxel volume, the SNR increases linearly. This is reasonable, since increasing the size of a voxel also increases the number of protons in a voxel, and the measured signal is related to the number of protons. Also, as the number of excitations increases, so too does the SNR, however, the increase is not linear. Rather, it follows a square root law: increasing NEX by a factor of 2, for instance, increases SNR by a factor of $\sqrt{2}$, since the signal doubles but the noise goes up only by a factor of $\sqrt{2}$ (typical variance behavior). In a similar fashion, doubling the number of phase encoding steps, N_y, increases the SNR by $\sqrt{2}$. The last component in this formula for SNR, the bandwidth BW, enters in an inverse relationship. Moving to a larger bandwidth increases the noise, and diminishes SNR. Again, the change follows a square root law.

Different ways of looking at SNR provide insight into the various tradeoffs among the parameters. SNR depends on two elements: the volume of an individual voxel and the total time to sample all signals. We can rewrite the expression for SNR analyzed in the previous paragraph, as follows. The size of a voxel along the y-axis, call it Δy, is equal to the field of view (FOV) in the y direction divided by the number of phase encoding steps. Similarly, Δx is the FOV in the x direction divided by the number of frequency encoding steps,

N_x. Hence the volume of a voxel can be rewritten as $\text{FOV}_x\text{FOV}_y\Delta z/N_xN_y$. Combining this with the previous expression for SNR, we have that SNR is proportional to

$$(\text{FOV}_x/N_x)(\text{FOV}_y)\Delta z\sqrt{\text{NEX}/N_y\text{BW}}.$$

With this formulation, we can explore the tradeoffs between resolution of the image and signal to noise ratio. Namely, for a fixed FOV, as we increase the number of phase encoding steps, SNR decreases. Conversely, if the FOV increases along with N_y (thus keeping the voxel size, and the spatial resolution, constant), SNR increases. However, there is a tradeoff in that the acquisition time will also increase. Increasing the slice thickness increases signal to noise ratio, but decreases resolution.

The scan, or acquisition, time is given by $\text{TR} \times N_y \times \text{NEX}$. Looking now at the effect of modifying the TR, it is clear that increasing TR increases the scan time. It also increases coverage – with a longer TR, it is possible to acquire more slices of the brain in a single scan. And, increasing TR leads to better SNR. On the other hand, increasing TE has no effect on the scan time, but does cause a worsening of SNR.

In summary, there are tradeoffs between the various desiderata from an imaging perspective – good signal to noise ratio, short scan times (to prevent discomfort of subjects in the scanner, for instance), and adequate spatial resolution. It is not always possible to achieve all of these goals simultaneously, and it is the work of MR technologists to figure out what parameter settings will balance the differing needs of a given experiment. Configuring the scanner parameters appropriately is a key component of fMRI experimental design.

2.1.3 Filling in k-Space

Another aspect of fMRI design relates to how the k-space data are themselves acquired – the planning of *pulse sequences*. Here, too, there are various tradeoffs, each trajectory through k-space having advantages and disadvantages. The planning of optimal sequences is still a topic of research among MR physicists. This section describes two of the more common types of data acquisition for fMRI: echo-planar imaging (EPI) and spiral imaging.

With *echo-planar imaging*, the gradients move through k-space in a *boustrophedonic* pattern, i.e., using alternate left to right and right to left lines (literally, as the oxen turns while plowing). EPI, originally proposed by Mansfield (1977), is the fastest MRI imaging technique currently available (Hashemi et al., 2004). However, it requires specific hardware to perform, namely, special gradients that can be turned on and off rapidly. *Single-shot EPI* allows for all of k-space to be filled in following a single RF pulse, whereas other fast imaging techniques require multiple pulses. In the most widely used version, "blip EPI," a large phase encoding gradient places the first echo at the edge of k-space. The readout, or frequency encoding, gradient forms a trail

of echoes, that is, we acquire a line of k-space data in the x-coordinate. The direction of the trail alternates back and forth, as subsequent phase encoding "blips" move the gradient along to the opposite side, thus covering all of k-space. In this version of EPI, the phase encoding gradient is turned on only when the readout gradient is 0 (i.e., at one or the other end of the k_x-axis in k-space). An advantage of the blipped EPI, compared to earlier versions of EPI in which the phase encode gradient was on continuously, is that the resultant trajectory through k-space is truly rectilinear. This makes it easier to Fourier transform the data to obtain an image. Figure 2.3 shows the k-space trajectory for blipped EPI.

Fig. 2.3. The k-space trajectory for standard, single-shot blipped echo-planar imaging (EPI).

It is worth emphasizing a technical point here, namely that since all of k-space is filled in following a single RF pulse, the data must be acquired quickly, before there is significant T_2 or T_2^* decay. On the other hand, the experimenter also wants to sample a large part of k-space in order to have adequate spatial resolution of the images, and this takes time. That is, we see again the types of tradeoffs that appear so frequently in fMRI experimental design. The way the question manifests itself here is that there is a need to compromise on the number of lines of k-space that are acquired over the space in question, a sacrifice of spatial resolution to the demands of time. Typical EPI images will be 64 x 64, or at most 128 x 128 voxels in a slice, whereas other pulse sequences can produce images of as much as 512 x 512 pixel resolution. The latter take longer to acquire, since they use more than a single pulse to

acquire all of k-space. A single image using EPI can be acquired in 30 to 50 milliseconds, with a whole volume scan (multiple slices) in only 2 to 4 seconds. By contrast, a particular sequence called FLASH, with resolution of 256 x 256 pixels, acquires a single image in 2.5 to 10 seconds, and a multislice volume can take up to 4 minutes (Jezzard and Clare, 2001). The limiting factor on EPI is the ability to rapidly alternate the gradients back and forth, which seriously taxes the hardware.

For single-shot EPI, TR refers to the time between successive images. The speed of single-shot EPI is offset by a number of drawbacks, aside from the relatively low spatial resolution and the stress the method places on the gradient hardware. First, the data that are collected during the transitions from one line of k-space to the next are not used in the creation of the images, so the technique entails an intrinsic loss of data. Second, EPI data are prone to various artifacts, in particular, *susceptibility artifacts*, which manifest themselves as distortions in the interfaces between air sinuses and brain tissue. Third, the long readout time of single-shot EPI may result in *geometric distortions*, which are seen as stretching or shearing when there is a distortion in the (x, y)-plane, or a dampening of signal when the distortion is in the z direction.

Multi-shot EPI is similar to single-shot EPI, with the main difference that the readout is divided into multiple segments. The segments are acquired in an interleaved manner, with one shot applied to each, as demonstrated in Figure 2.4. This variant of EPI puts much less stress on the gradient hardware, but images take longer to acquire.

Spiral imaging, like EPI, is a fast image acquisition technique. It differs from EPI in how the gradients are used to traverse k-space. Whereas EPI takes a rectilinear trajectory, based on rapidly switching gradients, spiral imaging uses sinusoidal gradients to induce a spiraling path through k-space. The spirals may be from the edge of k-space inward, or from the center of k-space out (see Figure 2.5). It is also possible to combine the two trajectories in a "spiral in/out" pattern, which increases SNR and decreases susceptibility artifacts (Glover and Law, 2001); the advantages of the spiral in/out method over conventional spiral methods appear to be greater at higher field strengths (1.5T versus 3T) for a variety of tasks that activate a range of brain regions that are particularly susceptible to artifacts (Preston et al., 2004). Spiral imaging induces less stress on the gradient hardware, since it samples k-space continuously, rather than switching back and forth as in EPI. The continuous sampling also means that all the k-space data are used in creating the image. Finally, spiral imaging is faster than EPI. A disadvantage of spiral acquisition is that the data no longer fall on a Cartesian grid, hence, in order to apply the Fourier transform, the data must be *resampled* prior to image reconstruction. This involves interpolating the data in order to map from a spiral to a rectangle. Like EPI, spiral imaging is prone to certain artifacts arising from inhomogeneities in the magnetic field, and spatial distortions, although these can be mitigated by using the spiral in/out trajectory. The patterns of the

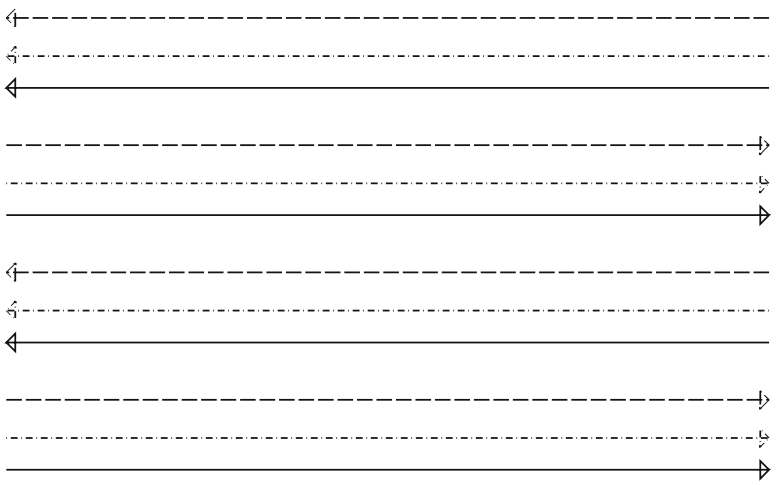


Fig. 2.4. The k-space trajectory of multi-shot EPI.

artifacts are somewhat different for the two types of trajectories, however (Jones et al., 2001).

2.2 Statistical Design Issues

The other aspect of experimental design for fMRI is statistical, namely, how should the study itself be carried out? While this question is common to all scientific studies, fMRI presents some interesting challenges.

2.2.1 Common Experimental Designs

There are two main approaches to the design of fMRI experiments, from the perspective of stimulus presentation. The first, *block design*, will be familiar to statisticians as a traditional way of designing an experiment. The second, usually called *event-related design*, arises from functional neurology studies, in particular, *event-related potentials*, or ERP. More recently, hybrid, or mixed, designs, which combine aspects of block and event-related, have been used.

Traditionally, due to limitations in resolution, fMRI experiments utilized simple block designs, in which periods of rest (or fixation) alternated with periods of task, or periods of different tasks were alternated. This was necessary in order to accumulate enough data to make a statistical analysis feasible.

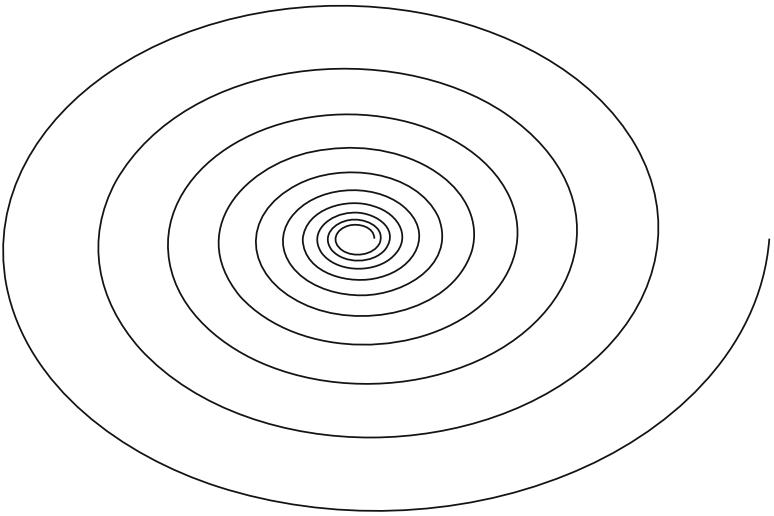


Fig. 2.5. The k-space trajectory for spiral imaging.

These "boxcar designs" (so-called because of their "on-off" nature, which can be depicted graphically as "up-down"; see Figure 2.6) lend themselves to a variety of statistical approaches. Block design experiments are easy to carry out, with presentation of the stimulus taking place in blocks of a fixed length, say 30 seconds. In each block one stimulus type is presented. A simple example of this type of experiment is the presentation of a flashing checkerboard during the task blocks, alternating with blocks of fixation, in which no visual stimulus is presented. More complex designs to accommodate more than one type of task are, of course, possible.

Block designs are powerful for locating voxels in which the level of activity is significantly different in the task versus the control conditions. To understand this recall the hemodynamic response described in Chapter 1. Following presentation of a stimulus there is a gradual rise in the signal until a peak is reached, after which the system returns to baseline levels in the absence of further stimulation. In a block design there is constant stimulation for the duration of the task blocks, meaning that the hemodynamic response does not return to baseline during this time. Instead, as the stimulus is repeatedly presented, the hemodynamic response in the active voxels accumulates, rising to a plateau instead of a short-lived peak. Decay back to baseline occurs only when the stimulus presentation is turned off, that is, during the control blocks. Voxels that are not active do not exhibit the characteristic response, and so

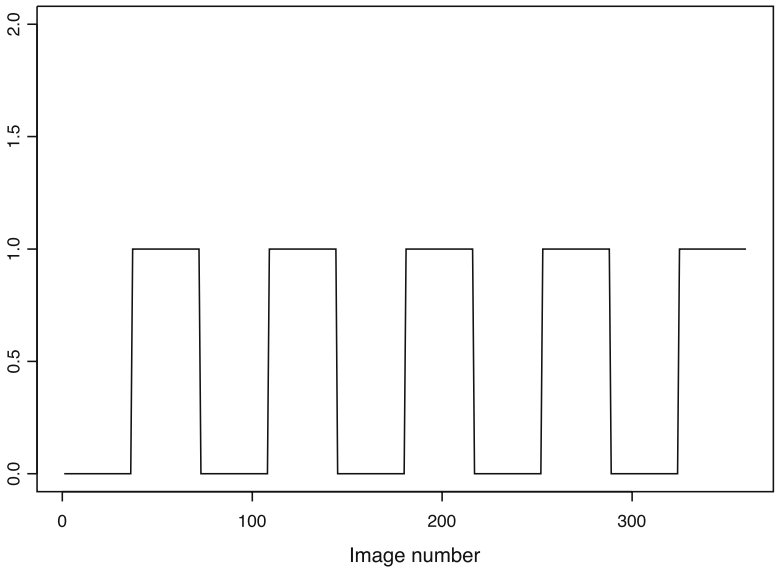


Fig. 2.6. Stimulus path of a simple block design. Blocks of the experimental task alternate with blocks of rest, or control. Of interest is the comparison between levels of activation during the task and during rest.

will not show increased levels of signal during the task blocks compared to the control blocks.

As should be evident even from this brief discussion, block design experiments are amenable to a range of statistical analyses, most of which involve comparison of (time-averaged) activity across experimental conditions. Since analysis is driven in large part by comparison, much of the focus in devising a block design experiment is on finding suitable controls. Ideally, the scientist wants the conditions to differ only in the variable of interest, that is, the control task should be identical, insofar as possible, to the experimental task. With complicated, and even not so complicated, cognitive processes such as are measured in a typical fMRI experiment, it isn't always clear what an appropriate control should be.

A study by Newman et al. (2001) demonstrates the potential impact of the baseline, or control, condition on the results of an fMRI experiment. They compared three baselines – rest, passive, and task-related – against the same task. The *rest* baseline, in which the subject lies in the scanner with no stimulus being presented, is the simplest, and is often used as the default control condition. In the *passive* baseline, stimuli are presented just as in the task condition, but subjects are instructed not to process them. As pointed out by Newman et al., it is impossible to know whether or not subjects truly

follow the instruction not to process the stimuli, but the rationale behind this control is clear. Finally, the *task-related* baseline presents the subject with a well-defined task. The experimental task in the study at hand was one of "phoneme discrimination" – subjects were presented with pairs of three letter nonsense words (consonant-vowel-consonant) and had to decide whether or not the two nonwords in each pair ended with the same sound. The three baseline (control) conditions were rest (relax and don't think of anything), passive listening (listen to the nonsense words but don't do anything in response), and monitoring tones, in which subjects were presented with triplets of high and low pitched tones, and had to judge if the final tone was high pitch. Note that the third baseline condition is similar to the experimental task, but with tones replacing the nonsense words. The authors found that there were clear differences in the activation patterns, depending on the control that was used. In general, more activation was detected in the resting baseline condition and less in the passive listening baseline. In related work, Marx et al. (2004) found that the choice of rest baseline – eyes open versus eyes shut – was critical for a simple visual task.

The event-related design, known also as *single-trial fMRI*, moves away from blocking the experimental conditions. In these studies, trials, or stimuli, are presented individually, separated by an *interstimulus interval* (ISI), which can be fixed within an experiment, or may vary from trial to trial. For example, instead of showing a flashing checkerboard continuously for 30 seconds, as we would do in a blocked design, in an event-related design, the checkerboard is flashed only once, for a short period of time. Another short checkerboard burst may follow some time later, or perhaps a completely different stimulus may be presented (Figure 2.7). This paradigm greatly expands the flexibility of the fMRI experiment, since researchers are no longer bound by the constraints of a formal block design. For instance, it is possible to let the stimulus presentation depend on the response of the subject: A correct response to a question could lead to a more difficult question on the next trial. Studies can be "self-paced" in the sense that the subject himself controls when stimuli are presented (e.g., Maccotta et al. 2001). Stimuli can also be presented randomly, again in contrast to the block design, wherein the stimulus within a block is fixed, and stimuli across blocks alternate. Furthermore, if the ISI is long enough, the neural activation following a stimulus will return to baseline, hence the event-related paradigm allows researchers to learn about the hemodynamic or BOLD response (the time course of activity) at a single voxel. This is not feasible with a block design, as it averages over hemodynamic responses, thereby blurring the individual features. Therefore, whereas block designs are good for *detection* of activated voxels, event-related designs are more effective at *estimation* of the hemodynamic response function. Another feature of event-related analysis is that it allows for separation depending on the response to the task (for instance, trials in which the subject responded correctly versus trials in which the subject responded incorrectly). As a result

of the greater flexibility afforded by event-related studies, their statistical analysis is often more challenging.

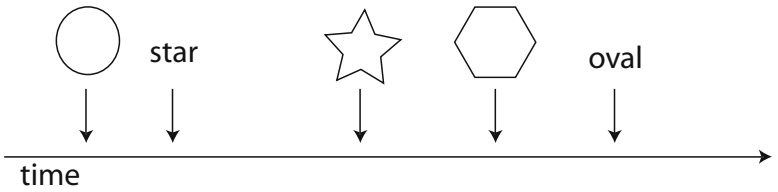


Fig. 2.7. Schematic of an event-related fMRI design. Instead of a "boxcar" describing the stimulus series, as in the block design, the event-related design can be described as a sequence of (possibly) irregularly spaced spikes. Stimuli of different types are presented in random order and at different time lags. In this example, a researcher might be interested in the differences between visual and verbal presentations of shapes.

A major design issue in event-related studies is the interstimulus interval, specifically the optimal length of the ISI, and whether it should be fixed or random. Early studies (for instance, Buckner et al. 1996) used constant ISI and allowed the hemodynamic response to fully evolve; thus, for a stimulus of duration 2 seconds, the optimal span of time between the end of a stimulus and the beginning of the next was 12 seconds (Bandettini and Cox, 2000). The resultant design – 2 seconds of stimulus followed by 12 seconds of ISI in which the subject would fixate, repeated multiple times – has the flavor of a "quasi-block" design and doesn't fully exploit the flexibility of the event-related paradigm. In addition, such studies are time consuming since the stimulus is presented very infrequently, they lack statistical power, and subjects become distracted during the long ISI. On the other hand, as noted by Bandettini and Cox (2000),with shortened constant ISI, there is an associated decrease in information (functional contrast) compared to the traditional block design.

Design with variable ISI is a more complex question. Early studies with variable ISI revealed somewhat contradictory results regarding how often the stimulus or task should be presented and how rapidly the design should alternate between task and control conditions. Dale (1999) found that the optimal design for hemodynamic estimation is one that alternates rapidly between conditions (short average ISI) and varies the time that passes between them (varying ISI). Friston et al. (1999c) explored a range of designs, from the

purely stochastic to the purely deterministic (block design), and found the latter to be the most efficient. The difference in findings, as pointed out by Birn et al. (2002), could be postulated to lie in the different goals that were being assessed in the two studies: estimation of the hemodynamic response function in the first case, and detection of active voxels in the second. Birn et al. (2002) examine the situation in more detail via simulation. They place the block design and the event-related design with short ISIs as two extremes – the former has stimuli presented continuously over a (relatively) long period of time, alternating with long periods of control; the latter switches frequently between task and control. In between these two, with variable ISI, it is possible to have a wide range of designs, and in all cases keep the proportion of time spent in the task condition constant. As expected from the discussion above, and indeed not surprisingly, their simulation studies reveal that there is no one optimal design when using variable ISI; rather, the "best" design depends on the goals of the study (detection versus estimation).

This issue is explored more fully by Liu et al. (2001), who develop a theoretical framework of the relationship between detection power and estimation efficiency. The tradeoff between the two is fundamental, in the sense that increasing one perforce degrades the other. Hence it is not possible to design a study, no matter how the ISI and stimulus presentation are manipulated, that will be able to achieve both goals simultaneously and efficiently. An additional factor included in their model is the "predictability" or perceived randomness of the design. Block designs have a high degree of predictability, since they alternate between two conditions, opening up the possibility (or even certainty) that subjects will anticipate stimuli, with concomitant confounding on the effects of interest. By deriving bounds on the estimation efficiency, Liu et al. demonstrate theoretically that this quantity is maximized for the case of two conditions (stimulus and control) when each condition is equally likely and the stimulus time course (i.e., the string of "stimulus" and "control" labels that describe the state of the experiment at each time point) is obtained as a sequence of Bernoulli trials, with probability $p = 0.5$ of success. Likewise, by deriving theoretical bounds on the detection power, they show that this quantity is maximized in block designs, again with equal probability of being in the task and control conditions. Similar results hold when there is more than one trial type (Liu and Frank, 2004).

In between the two extremes of maximizing detection power at the expense of estimation efficiency and maximizing estimation efficiency at the expense of detection power lie a range of what Liu and colleagues term *semirandom designs*, which have the potential to be useful when both goals are desired in the same experiment. The cost of trying to achieve a balance between the two extremes via a semirandom design is time; the semirandom designs take longer because the researcher is now trying to both estimate and detect. Essentially, in the semirandom design the probability of being in the task condition is allowed to vary over time, as opposed to being fixed, as in the purely random design.

Finally, the relationship between predictability and the other two quantities is examined. Liu et al. (2001) demonstrate that as the ability to estimate the hemodynamic response goes up, average predictability goes down. This makes sense if we recall that random designs are optimal for estimation, and random designs are less predictable, by definition. Another finding is that small increases in predictability may be worthwhile, since they can lead to gains in detection power without seriously impairing the estimation efficiency.

The design of event-related studies is an active field in the fMRI community and the best way to learn about the most recent state of the art ideas is to read the current literature. To wit, we describe one final twist on the topic, studied by Visscher et al. (2003), the *mixed block/event-related design*. In this design, blocks of task are alternated with blocks of control, as in a standard block design. However, within the task blocks, trials are assigned at random (that is, with varying ISIs), as in an event-related study. The advantage of this design, as the authors showed in an extensive simulation study, is that it allows for the separation of transient activity, related to the stimulus as it is presented, from sustained activity, which carries across tasks and stimuli. Many other design variants are possible (Liu, 2004).

Which of these various designs should be used in a given study? It is evident from the discussion above that there is no one "correct" design, and even optimality of a chosen design is contingent on the specific questions of scientific interest. My experience has generally been that the design of the experiment, and particularly the choice between block design and event-related design, depends more on the constraints of the study than on purely statistical considerations, which in any case can accommodate any of the differing paradigms described above. To some extent scientific trends play a part as well – as event-related, mixed, and semirandom designs become more popular and more accepted in the literature, researchers are going to want to exploit their strengths and flexibilities in order to get the most out of the data. As statisticians, it is important that we be aware of the latest designs, their advantages and disadvantages, and steer our neuroimaging colleagues away from designing experiments that will not allow them to answer the questions in which that truly interest them.

2.2.2 Additional Issues

In this section, a number of miscellaneous issues relating to the statistical design of fMRI experiments are explored.

We start with two specific questions on the use of block designs: first, the timing of data acquisition within the block, and second, identifying activation that is the result not of the task, but of the transition from task to control or vice versa. These questions are of potential impact on both the statistical analysis and the interpretation of block design studies.

Veltman et al. (2002) look at the first question in the context of language processing. Conventionally, the stimulus is presented at the onset of a TR,

but Veltman and colleagues argue that this might produce bias. To check this, they varied the length of the fixation condition in a block design study, thereby affecting the relationship between stimulus presentation and the TR; instead of the two always coinciding, the presentation of the stimulus was shifted relative to the onset of the TR. They found that effect sizes were larger in the "no shift" condition than in the conditions where the stimulus presentation was decoupled from the TR. Furthermore, activation was detected in some areas only at certain timing conditions, and not others. The authors conclude that even for block designs, it is important to distribute the sampling of the hemodynamic response, as is done in event-related studies.

The question of activation during transitions between conditions in the block design is taken up by Konishi et al. (2001). They found a set of regions that were transiently activated at the transitions between blocks, consistently for a variety of conditions involving different visual stimuli (verbal and facial) as well as different trial presentation rates. Interestingly, these regions did not always coincide with regions in which task-related activation was detected by the usual statistical analysis. Moreover, the size of the transition effect was similar in all four of the conditions considered by the authors. Perhaps most significant, the transient activation was sometimes detected even when there was no detected activity in the relevant task block, indicating that it is not the activation in the task block alone that is pertinent.

We next turn to efficient and optimal experiment design, in particular for event-related studies, a focus of two recent papers. Block designs can be incorporated into the framework of finding a good design, since they can be seen as one extreme of a continuum of designs. The search for an optimal design involves an exploration of the space of possible experimental setups, which can be accomplished in many different ways. When we discuss "sequences" below, the intention is to reference the string of conditions representing the presentation of stimuli at each time point. For conceptual simplicity, think of each condition in the experiment being assigned an integer value. Then a design is a string, or sequence, of integers.

Buračas and Boynton (2002) consider efficient estimation of the hemodynamic response curve, noting that, by use of a more efficient design it is possible to reduce scan time without loss of signal, compared to a less efficient design. They point out that researchers often pick an event-related sequence of trials in a rather arbitrary fashion, namely, generating many sequences at random, and picking the one that yields the best estimation efficiency. Clearly this is not a satisfactory approach in general, as researchers are left without any guidelines for choosing the design sequence in their next study, and of course there is no guarantee that the best sequence from among a set of randomly generated sequences is optimal in any other sense. Buračas and Boynton propose instead to use *maximum length shift register sequences*, or *m-sequences*, to find good stimulus presentation sequences at little computational cost (unlike the random sequence approach). The emphasis here is on identifying the specific sequence(s) that should be used in an event-related

study. m-sequences are sequences of integers (each integer representing here a condition type, for instance, task or control) which are generated using modulo arithmetic, where the modulus depends on the number of conditions. Sequences are created using the formula

$$s_k = \sum_{i=1}^{r} c_i s_{k-r+i-1},$$

where c_i are set coefficients, s_k is the next value in the sequence (appended to the previous $s_1, \ldots, s_{k-1}$), and r is the *order* of the shift, the number of previous terms included in generating a new element of the sequence. For the simplest fMRI experiments, with only two conditions, both c_i and s_k take on the values 0 and 1, and the generation of new elements uses arithmetic modulo 2.

Buraccas and Boynton show, based on simulations, that generation of event-related designs via m-sequences results in higher estimation efficiency than randomly generated designs, especially for shorter sequence lengths. Even using as many as one million random designs and picking the best among these, estimation efficiency was less than for the m-sequence. Having more than one event type also improves the advantage of m-sequences over random sequences. Indeed, Liu (2004) proves that m-sequences come close to attaining his theoretical upper bound on estimation efficiency and have low predictability, although they have low detection power. The drawback of m-sequences is that there is less flexibility in design, since length and type of allowable sequence are restricted. In practice, the authors claim that this is not any real barrier to using their approach, since the class of acceptable m-sequences is large.

Wager and Nichols (2003) propose the use of a genetic algorithm to accomplish the search over design space. Genetic algorithms start with an initial set of designs and evolve new designs via three steps that mimic the ways changes occur in DNA: *selection*, *crossover*, and *mutation*. Selection of designs is akin to natural selection – the genetic algorithm tests the initial set of designs (which are usually randomly generated) according to some prespecified goodness criterion, selects the best ones, and creates "offspring" from them (new designs). In this way, the best features of the existing designs are passed on. Crossover is similar to the biological exchange of DNA across chromosomes – two designs will exchange "material," or sequence patterns, from a randomly chosen point onwards. And in mutation an element of the design sequence is randomly switched to take another value. These processes are iterated until an appropriate model is identified. Wager and Nichols consider three goodness criteria: effectiveness at detection, estimation efficiency, and counterbalancing of the design. The genetic algorithm is flexible enough to optimize multiple criteria simultaneously (this is accomplished by creating a new, weighted, score, which combines individual criteria), and can also take account of both statistical and psychological requirements (for instance, it might be undesirable

to have the same stimulus repeated too many times in a row, for reasons of anticipation on the part of the subject; the genetic algorithm can incorporate such constraints by rejecting designs that have unwanted properties).

Via a series of simulation experiments, the authors demonstrate the efficacy of the genetic algorithm over random search, although they note that it is hard to achieve all three goodness criteria with one design, in agreement with findings by others. They are also able to give specific recommendations on the length of the ISI in an event-related design, and the length of a block for a block design (subject to the particular model assumptions made in their simulations), suggesting that the use of a genetic algorithm coupled with simulation will be a useful pilot tool in the early stages of planning an experiment. In this scenario, the researcher would run versions of the genetic algorithm with different assumptions and different weighting of the criteria of interest, specific to the research questions at hand, to arrive at an optimal experimental design.

As a final point we mention briefly the problem of using fMRI in clinical populations (that is, groups of subjects with neuropsychological disorders, such as autism, schizophrenia, or Alzheimer's disease) and children, a theme that we will revisit in Chapter 3. These groups present special challenges, among them the need for the researcher to devise experiments that the subjects will be able to perform. For example, it might not be reasonable to expect children or patients with attentional difficulties to carry out tasks that put a heavy load on memory. Choice of task, then, should be suited to the experimental group of interest, as discussed in Jessen et al. (2002)for clinical groups and by Gaillard et al. (2001)(see also references therein) for children.

3

Noise and Data Preprocessing

One of the notable, but surprising, aspects of fMRI for new statistical researchers is that the data are very noisy. By this we mean both that the data are prone to noise from a wide variety of sources, which we shall survey in this chapter, and also that the BOLD signal is but a small portion of the overall measured MR signal. Hence it is crucial, prior to any formal statistical analysis, to clean up as much of the extraneous sources of variation as possible, and to isolate the signal. Fortunately, the noise in fMRI is well understood, and there are standard methods and tools for preprocessing the data, so that new researchers coming into this field do not have to start from the very beginning. This chapter presents the major sources of noise that are of concern in fMRI, and outlines the approaches for preprocessing.

3.1 Sources of Noise

Noise in fMRI data may be roughly characterized into three groups: thermal noise, system noise, and subject- and task-related noise. The first two types of noise are related to the properties of the scanner and are intrinsic to the imaging process. The third type is derived from the inescapable fact that the experimental subjects are human, and as such will breathe and move around while in the scanner. Both of these activities, among others, have the effect of introducing noise into the image.

Thermal noise is an intrinsic part of MR imaging. It reflects changes in the strength of the MR signal over the course of an imaging session, caused by *thermal motion* of the electrons in the sampled tissue and in the electronic components of the scanner. Thermal motion occurs when electrons collide with atoms, for example, in the scanner hardware. As the temperature of the system increases, the rate of collisions goes up, resulting in greater distortion of the signal. In theory, then, it would be possible to eliminate thermal noise by reducing the temperature in the MR system as a whole. As a matter of practice, this approach is not feasible, however, because thermal noise will

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_3,

always be present, unless the temperature is reduced to absolute zero. Thermal noise also increases with the strength of the main magnetic field (Edelstein et al., 1986). In general, thermal noise is not of much concern. Since it is random, and not related specifically to the experimental task, its effects can usually be mitigated by simply averaging over data points.

As the name implies, *system noise* is introduced by fluctuations in the functioning of the MR hardware, that is, noise that comes from the system itself. Two common causes of system noise are inhomogeneities in the static magnetic field and instabilities in the gradient fields. Problems with the radiofrequency coils are also a source of system noise. An important form of system noise is drift in the signal, whereby over the course of an experiment, the signal intensity at any given voxel gradually and systematically changes, as shown in Figure 3.1.

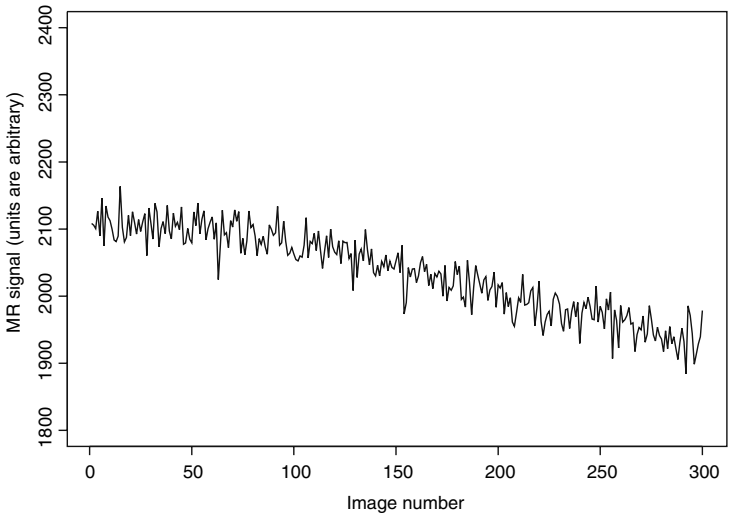


Fig. 3.1. Over time, the underlying signal intensity gradually changes. This phenomenon is known as signal drift and is one of the sources of noise in fMRI data.

Thermal and system noise are intrinsic in that even if the object scanned is a so-called inert "phantom," such as a ball or a cylinder filled with liquid, these types of noise will still be apparent. By far the more interesting forms of noise from a statistical perspective are those arising from the fact that the object in the scanner is not a phantom, but rather a living, breathing, thinking human being, performing an experimental task – *subject- and task-related noise.*

While in the scanner, the subject receives instructions to remain perfectly still. The reason for this is related to the way in which the data are acquired, as was described in the previous chapters. Recall that a series of radiofrequency (RF) pulses are sent through the tissue, with the goal of localizing activity. If the subject moves, activation from one location will be contaminated with activation from nearby locations, resulting in a blurring of signal. Because the spatial resolution of fMRI data are at the level of individual voxels, and voxels, as we have seen, are typically 1-3 mm per side, even very small amounts of movement can have a serious effect on the signal. In addition, a typical fMRI scanning session can last as long as an hour, and it is very difficult to remain stationary for such a long time. As a result, *head motion* is a major source of the noise in fMRI data. Data that are corrupted by too much head motion are often discarded; while this may involve only eliminating scans that are too noisy, in some extreme cases all the data from a single subject are rejected by the experimenter. Given that many fMRI studies include a small number of subjects, throwing away even one is very wasteful.

In contrast to thermal noise, noise caused by head motion is far from random. Since the entire head moves as one, motion induces extra spatial correlation. More importantly, perhaps, motion is often related to the experimental task (Hajnal et al., 1994; Hajnal et al., 1995). As an example, consider a simple eye movement task, called a visually guided saccade. The subject looks at a screen inside the scanner, fixating perhaps on a crosshair in the center. A dot appears in the subject's peripheral vision, and the experimental task is to direct the eyes to the location of the dot on the screen. The head is to stay stationary as the eyes move, but there is a very natural tendency to move the head as well, even if ever so slightly. Aside from task-related movement, head motion may result from swallowing, blinking, and head bobs as subjects fall asleep in the scanner. Clinical populations of patients, as well as children, may have particular difficulty following the instruction to remain perfectly still over the course of an experiment, but it should be emphasized that even among healthy, willing adults it is difficult to eliminate the problem altogether. Shutting the eyes (Stephan et al., 2002) and holding the breath (Abbott et al., 2005) are hypothesized to lead to effects similar to those of head motion. Breath holding is evidently more common among adults at the start of a challenging task, and even holds of short duration (on the order of several seconds at most) have been found to affect the measured MR signal, so this is also a type of task-related noise, in some circumstances.

Seto et al. (2001) report an interesting simulation study that examined task-related head motion in three different groups: patients recovering from stroke, age-matched controls, and young controls. In this study, subjects were put into an environment that very closely matched a typical MR scanner, but actual images were not acquired. Instead, they were required to do a variety of motor tasks and the amount of head motion was recorded, using an accurate position tracking system that would not be feasible inside an actual scanner. The researchers found that the stroke patients had more head motion than

the other two groups, and were also more variable. The age-matched controls had more head motion than the young controls. They also found that head motion was greater during the performance of the motor tasks than during the rest periods, although the differences in head motion across tasks were not consistent. The direction of the head motion, in terms of translations and rotations, differed for some tasks, but not others. These results highlight that head motion is a very complex issue; especially with clinical populations, care should be taken to minimize the opportunities for movement.

A second major source of subject noise is *physiological*. *Physiological noise* usually refers to noise in the signal induced by respiration and heartbeat (Hu et al., 1995). When the subject breathes, this of course causes motion, as does the beating of the heart. One effect of heartbeat and respiration is the actual movement of the brain inside of the skull, quite apart from any movement in the body of the subject. These are small scale motions, faster and more regular than head motion. It is possible to exploit these characteristics of physiological noise in the *data preprocessing* step. Respiration leads to increased blood volume inside the skull, which causes the brain to expand (Hu et al., 1995). The resultant motion and its effects on fMRI are rather complex. Beyond motion, respiration also gives rise to systematic changes in the magnetic field, thus directly affecting the measured MR signal.

We also include under the rubric of subject and task-related noise, noise that is literally due to the scanning environment, which is very loud. While this might seem to be an intrinsic source of noise, in fact it is the interaction here as well with the subject inside the scanner, and how the literal noise of the scanning environment affects the response in some experiments, that is of interest. Numerous studies have shown that the effect of the noise made by the scanner can be severe, owing to distraction of the subject, evocation of the BOLD response in task-irrelevant areas, and the like (see Moelker and Pattynama 2003, and the references listed therein, for an overview of this problem).

Moelker and Pattynama (2003) list four main sources of acoustic noise in the MR imaging environment that are of potential concern. These are: the gradient currents (readout and phase encoding) that are used for localizing activation within a slice; eddy currents induced by fluctuations in the system; the radiofrequency and slice selection pulses (which generally occur simultaneously); and background noise from air conditioning and ventilation. The major noise source are the gradient currents, where for EPI at 1.5T, peak levels can be as high as 130 decibels. Gradient current acoustic noise increases with the strength of the main magnetic field as well as with the strength of the gradient currents themselves.

The effect of acoustic noise can be either direct or indirect. Direct effects are those that confound the experimental task, hence these are mainly a concern for studies that use auditory stimuli. Noise from the scanner induces a BOLD response in auditory cortex that is similar to that of any other auditory stimulus. The baseline level of activation in those areas, in the presence

of scanner noise, is higher, making it harder to detect activation that is task-related by any statistical method (making this a true source of noise in the statistical sense). Indirect confounding is mostly due to attention (Moelker and Pattynama, 2003); effects include distraction, increases in activation in areas of the brain related to attention, temporary hearing loss (which again will affect auditory tasks), and masking (difficulty in hearing an auditory stimulus over the noise of the scanner itself). Research indicates that acoustic noise is most likely not related to motion, but the studies are not conclusive on this point.

A simple psychophysical experiment conducted by Scarff et al. (2004) demonstrates some of these effects. Seven healthy adults with normal hearing were presented five tones of equal loudness, ranging in frequency from 250 to 4000 Hz. Acquisition parameters were manipulated so that the peak scanner noise was close to the frequency of one of the tones. For the frequency that was close to the frequency of the peak scanner noise, the perception of loudness, as well as the detectable fMRI activity, were both decreased. The experimenters considered two scanning protocols, producing two different frequencies of peak scanner noise, with consistent results. Although there was no control condition in this study (it not being possible to remove scanner noise completely), and the sample was quite small (typical of fMRI experiments, due among other factors to the cost of the scan), this outcome offers some evidence for acoustic masking and the confounding effect of background noise on the outcome of an auditory experiment.

As a final example of subject related noise, consider the fact that while an individual is in the scanner, that person is actively thinking about many things, and inactively thinking about others, in spite of any instructions that may have been given by the experimenter. The human brain is always active, always attending (consciously or not) to the surroundings. No matter how a researcher tries to control all of these external (to the experimental conditions of interest) factors, it is simply quite impossible to do so. These mostly transient elements should be related to as noise, because they will manifest themselves in task-irrelevant areas and will also affect baselines levels throughout the brain, making it potentially harder to detect true effects of interest.

3.2 Dealing with Noise by Manipulating the Scanning Environment

There are two main approaches to handling the noise inherent in fMRI data, and these are usually used in tandem. The first approach, which is the focus of this section, is to prevent noise, insofar as this is possible, by suitable manipulation of the environment. Clearly, not all noise can be approached in this way, and even the types of noise that can will not necessarily be eliminated.

The second approach, which we take up in the next section, thus involves estimating and removing the various sources of noise from the recorded signal, a procedure that is known as *preprocessing*.

Noise related to the subject and task are the primary targets when the scanning environment is manipulated. Researchers have concentrated in particular on head motion, since this is one of the major sources of fMRI noise in general. It is routine, for instance, to pad the inside of the magnet around the head with pillows or foam padding to make it harder for the subject to move suddenly and to any great extent. Subjects can go through training prior to the experiment, in either a simulated or real scanner, to become adjusted to the environment and learn how to minimize head motion. Masks and plaster head casts which are molded to the subject's head and face, and then bolted into position, are an effective means of reducing motion, but some subjects find them uncomfortable and anxiety-inducing (Savoy, 2001). Also in wide use are bite bars, in which the subject bites down on a dental mold throughout the course of the scanning session. Subjects may also find bite bars hard to tolerate, but they are an efficient way of preventing head motion. See Edward et al. (2000) for a brief review of methods currently used to fix the head in one location throughout the course of a study, as well as the advantages and disadvantages of each technique.

As is apparent from Edward et al. (2000), the main problem with methods for restraining motion such as bite bars and masks, is that they can provoke anxiety in the subject, even in healthy young adults. This in turn can affect performance on the task of interest. For this reason, Thulborn (1999) proposes a more "user-friendly" procedure for reducing head motion, namely, a visual feedback system. A visor is mounted to the RF coil in the line of sight of the subject. A screen inside the visor provides the subject with visual feedback on head position, so that if motion does occur, the subject is able to see this and realign. A small-scale study on six healthy volunteers showed that the visual feedback is helpful in reducing head motion. Furthermore, although the feedback system is visual, it apparently does not interfere significantly with activation resulting from visual processing tasks.

Environmental manipulation is also used to reduce ambient scanner noise, as described in the review article by Moelker and Pattynama (2003). Acoustic noise can be reduced by adjusting the timing of image acquisition, by using pulse sequences that generate less noise, and by improving the scanner hardware. Most common, however, is to reduce noise through the use of earplugs, helmets, vacuum cushions or other devices to muffle sound; Moelker and Pattynama call this "passive noise reduction." This is by contrast with "active noise reduction," in which noise is eliminated via the application of a sound that is the exact inverse of the original noise. The two approaches can be used together, although for technical reasons active reduction does not seem to be as effective for fMRI. Passive reduction is cheap, effective, and probably the easiest method currently available for decreasing acoustic fMRI noise.

Finally, for dealing with problems of noise arising from distraction, attentional drift, and the like, various methods are used. Eye trackers can help verify that subjects remain on-task for visual processing experiments and identify problematic trials, although this is most frequently done a posteriori rather than in real time. Simply having the MR technologist speak to the subject in between tasks or scans will refocus the attention, and this too can be a highly effective mechanism.

3.3 Preprocessing fMRI Data

Not all noise can be removed by careful control of the imaging environment. Instead, it is always necessary to perform some preprocessing, or cleaning, of the data, prior to statistical analysis. Different software packages accomplish this in different ways; even different laboratories using the same software preprocess their data differently. Within this, however, there are certain commonalities. Motion correction, for example, is almost always carried out, although the details of how the amount of motion is estimated, and how its effects are eliminated, may differ. Some preprocessing steps are carried out in k-space, others are applied after image reconstruction. In this section we survey some of the big issues in data preprocessing.

First, recall from the discussion in Chapter 1 on how data are acquired that there is a high frequency component in the center of k-space. A critical (if trivial) preprocessing step is, therefore, to perform a correction to the baseline of the data, that is, subtract the mean. If this step is not performed, the first Fourier component will overwhelm everything else, and the image will consist of a single bright spot in its middle.

Recall too that in EPI the trajectory through k-space reverses direction on alternate lines. This is corrected via a simple flipping of alternate lines so that all proceed in the same direction. When this is done, it is important to also align the data properly, to correct for possible mistimings in the gradients. If the reversed lines are not realigned, "ghosts" will appear in the reconstructed image (Lazar et al., 2001). Ghosts are faint, shifted images of the brain that appear in the air outside the head in the reconstructed image. While ghosts are often corrected in k-space at the time of line reversal, it is also possible to remove them in image space (Buonocore and Zhu, 2001).

Slice timing correction accounts for the fact that slices of data are acquired one at a time, rather than all at once, yet most data analysis schemes assume that every voxel was sampled at exactly the same time. When slices are acquired in an interleaved fashion, for example, all even slices then all odd slices, this assumption is clearly not tenable. But even when the slices are acquired sequentially, there is a delay in moving from one to the next. Thus, different voxels are sampled at (slightly) different times. It is generally desirable to correct for this time shift by reshifting the voxel time series, and this is easily accomplished on the image space data (Smith, 2001).

Scanner drift is corrected by *detrending.* There are many different ways of detrending the data, ranging from simple mean correction (normalizing to the mean signal intensity of each run), to linear and polynomial models, wavelets and splines. These methods, as well as an automated detrending procedure, are compared by Tanabe et al. (2002). For the polynomial techniques each voxel time course is fit by the appropriate model and the detrended signal acquired as the difference between the fit and the original time series. The spline model is a cubic spline with five knots, three of them internal to the time course at the voxel. The wavelet model uses a scale of 1 to track low frequency noise. The automatic detrending procedure chooses the detrending method that fits best on a voxel-by-voxel basis (that is, instead of applying the same detrending method to the entire data set, the procedure that fits each voxel best is used to detrend that voxel).

Tanabe and colleagues find that simple mean correction greatly enhances the number of active voxels detected in the subsequent statistical analysis, compared to no correction for drift. Hence mean correction is always a valuable part of the preprocessing to adjust for scanner drift. Indeed, in the rest of their comparison, Tanabe et al. first apply mean correction, followed by the detrending algorithm. With this approach on two subjects, they find that the spline method performs best overall in terms of increasing the number of active voxels that are detected. Further gains can be achieved using the automated algorithm to pick the optimal procedure for each voxel. When they look specifically at the automated method within regions of interest relevant to the visual task in this experiment, they find that almost half of the voxels do not benefit from additional detrending beyond the mean correction; for the rest of the voxels the spline model is chosen most often.

Just as head motion is a major focus of environmental manipulation within the scanner, so too, much effort has concentrated on estimating head motion and correcting the data for its effect, prior to formal statistical analysis. The first step, estimation, involves aligning each scan to a "target" scan, which may be the first image, an image taken from the middle of the series, an average of images, and so on. A common simplifying assumption in the estimation process is that head motion is "rigid body," that is, the head and the brain do not change their shape, only their position and orientation. Translations and rotations in each direction are then estimated by minimizing the distance between the scan and the target. This can be done in two dimensions (on a slice by slice basis) or in three dimensions (over the whole brain simultaneously), using a variety of optimization procedures and different measures of distance. Typical distance measures are least squares and least absolute difference between the image and the target. In both cases the goal is to find values of the translation and rotation parameters that minimize the distance between the two. See Brammer (2001) for a nontechnical overview of motion estimation.

Once the motion parameters have been estimated, the second step of the procedure, namely correction, can be carried out. Again, there are various ways to perform the correction, or registration, step. The estimated translation

and rotation parameters are used to shift the image to the target, using various interpolation schemes such as sinc interpolation, polynomials of different orders (cubic or quintic, for example), and linear interpolation. These methods are performed on the reconstructed images, rather than on the raw k-space data. Eddy et al. (1996) propose, instead, performing motion correction on the k-space data and using Fourier interpolation. They note that translations can be achieved in k-space without interpolation, but that rotations in either k-space or image space generally require interpolation. Rotations in k-space are accomplished through a series of *shearing matrices*; since a rotation matrix can be written as a product of three shearing matrices, rotations are easily and efficiently performed. Furthermore, performing motion correction in k-space with shearing matrices for rotations results in no loss of the statistical information in the images, which is not the case for interpolations and corrections on image space data (Eddy and Young, 2000).

Several groups have also considered motion correction in real time, as the scan is being performed. This is much more computationally burdensome than motion estimation and correction after the fact as part of a general preprocessing of the data, but is made possible by advances in computer memory and speed. Cox and Jesmanowicz (1999), for example, generalizing the shearing matrix ideas of Eddy et al. (1996) from two dimensions directly to three, develop a quick algorithm for image rotation and shift. Using least squares distance and a gradient descent method, motion can be estimated and corrected for as fast as the images themselves are acquired.

An alternative real-time approach is given by Ward et al. (2000), who use "navigator echoes" before each image acquisition to determine if there have been any translations or rotations of the brain in the three-dimensional space, and if so, to what extent. These estimated translations and rotations are then used to alter the image acquisition parameters in order to compensate for any motion that may have occurred. Obtaining the information from the navigator echoes and correcting the image acquisition parameters can be accomplished very quickly, making this another feasible method for real-time motion correction.

While not carried out in real time, the method proposed by Caparelli et al. (2003) is similar in spirit to the real-time techniques, namely, to develop a procedure by which motion parameters can be quickly estimated so that problematic scans can be detected immediately. When a scan with excessive motion is identified, it is then possible to give feedback to the subject and repeat the scan. Unlike other procedures, in which the motion parameters are carefully estimated and the images subsequently registered to the target, Caparelli et al. define and monitor instead several "quality measures" of the image; furthermore, these measures are calculated on the k-space data, not out of any sense of optimality, as in Eddy et al. (1996), but rather because the system with which these authors work "...does not provide real time image reconstruction..." (p. 1412). The quality measures are essentially

squared differences between the image at time t and the target image, taken to be the first image.

It should come as no surprise that motion correction may cause artifacts when in fact there has been no subject motion, as pointed out by Freire and Mangin (2001). According to these authors, the use of least squares for estimating the motion parameters introduces bias because areas of activation will be treated as outliers. They therefore advocate the use of more robust distance measures (they suggest mutual information, but others are possible) to avoid this problem. Presumably this would not be an issue for motion correction techniques that preserve statistical information. In a series of simulation experiments designed to test some of the popular motion correction algorithms, but notably not the one introduced by Eddy et al. (1996), Freire and Mangin demonstrate that, in the absence of motion, least squares-based methods introduce spurious activation.

Finally, we mention briefly two comparisons of motion correction tools. The first, by Morgan et al. (2001), uses a computer-generated phantom to evaluate and compare motion correction algorithms in three popular fMRI software packages. A computer-generated phantom is much more flexible than a standard, physical one, making it possible that different types of motion be mimicked, together with different patterns of activation and noise. With the exception of one algorithm implemented in one version of the popular software SPM (see Appendix A for more discussion of SPM), there are few differences among the procedures. Oakes et al. (2005) perform a more extensive comparison, using five different software packages, different configurations of the algorithms within each package (default options exist, but these can be manipulated and customized by users), and real as well as simulated data. While the different packages, and different implementations within each package, have advantages, they find that no one algorithm dominates the others, and in practical terms of detecting activation, all are quite similar. Hence the choice of software and procedure for motion correction does not seem to be critical; what matters is that motion correction be performed.

Perhaps because its effects are more subtle, and not as pervasive, physiological noise correction has not been the subject of as much research as correction of head motion. An early paper by Hu et al. (1995) provides a general method for estimation and correction of physiological effects. Two main ideas are involved in their approach. First, estimation of the physiological noise is carried out on the k-space data. Second, physiological activity of the subject is monitored simultaneously as functional data are being acquired. This allows for the two types of data to be synchronized and is key to this preprocessing step. An explicit assumption of the method is that the noise induced by heartbeat and respiration is pseudo-periodic in nature, which makes it more natural for one to estimate those effects in k-space. However, there is no reliance on the periodicity of the respiration and heartbeat themselves.

The first step of the algorithm is to synchronize the k-space data with the physiological motion. This is achieved by determining for each sampled

point the respiratory or cardiac cycle into which it falls. The next step is estimation; Hu et al. use as a model a truncated Fourier series of relatively low order, to reflect the smoothness in the variations within a physiological cycle. Following estimation, it is possible to remove the effects from the phase and magnitude of the measured data, as we have seen in other preprocessing steps. The authors note that the effects of respiration are more prominent than those of heartbeat, so the latter are first removed, then the cardiac effect is estimated and removed from the respiration-corrected data.

Correcting for physiological noise has a number of consequences, according to results presented in Hu et al. (1995). First, the standard deviations at each voxel are significantly reduced. Second, voxel time courses that showed periodicity due to cardiac or respiratory influences lose their cyclic behavior, allowing the time course of activation to manifest itself. Third, physiological correction enhances the detection of low level activations, as expected by the reduction in voxel variability. Fourth, physiological effects do not appear to be correlated with activation.

More recent work by Glover et al. (2000) introduces a method similar to that of Hu and colleagues, but based in image space rather than k-space. The essential idea is the same, with the main change being that the estimation of physiological effects is carried out in image space, that is, after reconstruction of the data. Hu et al. estimate the physiological effects prior to reconstruction. The functional form of the estimation model in both algorithms is the same. The authors compare the two methods on only three subjects, but in all three cases the image-based procedure is much more effective in reducing noise. Glover et al. speculate that this might be due to the fact that for k-space correction, estimation can only proceed in parts of k-space where the signal to noise ratio is high enough to provide a good fit to the Fourier series model, that is, near the origin. This limitation introduces correlations in the reconstructed image data, so that some regions are undercorrected and others are overcorrected. This is not a problem in image space, since each voxel is treated separately.

Chuang and Chen (2001) propose another image-based method for physiological noise correction. By contrast with the previous two procedures, Chuang and Chen's algorithm does not require simultaneous monitoring of heartbeat and respiration during the functional scan. Rather, they suggest estimating the physiological effects directly from magnitude changes in the voxels that exhibit strong physiological fluctuation. Their method uses information on the typical frequencies of heartbeat and respiration, that is, how many times the heart beats or breath is inhaled, in a unit of time. Finding the strong frequency components within this typical range provides the estimate. Correction proceeds as in the other algorithms. Results are comparable to the k-space method.

McNamee and Eddy (2005) consider the problem of physiological correction from a very different perspective than previous works. Their two underlying considerations are that it is feasible to estimate directly the relationship

between the physiological data and the fMRI data, so that no functional form for the former needs to be assumed; and that the physiological effect on the BOLD signal may not be instantaneous, but rather will take place after some time lag. First, they calculate simple correlations for the time series of frequencies of the data in k-space (phase and magnitude separately) and the physiological measures (heart beat and respiration separately) on a voxel by voxel basis. Working in this time series framework, a simple generalization is then to calculate correlations at lag j, that is, to shift one of the series relative to the other. In this way, it is possible to determine the lag at which the physiological data are most highly correlated with the fMRI data. The time lag showing the strongest correlation between the cardiac and the fMRI data is slice dependent in a predictable way, which has implications for the way in which the correction should be performed. Specifically, correlations are cyclic with a period equal to the TR. The same is not true for the respiration data. The strongest correlation between respiration and signal is often at no lag, reflecting that breathing has an immediate effect, whereas the effect for heartbeat is delayed. Both of these findings are then taken into account in the modeling process to remove the physiological noise. Cardiac and respiratory effects are entered as independent variables in a simple regression model, with the magnitude or phase of the k-space data the dependent variable. A separate model is fit at each point in k-space. Time lags can also be easily accommodated in this linear regression to reflect the different effects of heart beat and respiration. Residuals from these models then give the fMRI data corrected for the various physiological effects.

Although not a means of removing any specific type of noise, many groups include *smoothing* as part of their preprocessing of fMRI data, particularly spatial smoothing of the images. The most common approach is to smooth the data with a Gaussian filter; the usual way of characterizing the filter is through a measure called *full width half maximum* (FWHM). Formally, for a Gaussian distribution with variance σ^2, the FWHM is defined as $2\sqrt{2\log 2}\sigma \approx 2.36\sigma$. The width of the spatial filter is expressed in millimeters at half of the maximum value, with typical values being between 3 and 10 mm FWHM (one to three voxels). Less formally, if the maximum value of the Gaussian kernel is ν, then to obtain the FWHM, go down to 0.5ν and find the length of the kernel at that height, from one end to the other. See Figure 3.2 for an example. Clearly, as the FWHM increases, the kernel becomes wider, resulting in more smoothing.

There are two main reasons generally given in the fMRI literature for smoothing (see Smith 2001). First, is that small amounts of smoothing improve the signal to noise ratio, since the effect of smoothing is to blur the measured signal in neighboring voxels. If the size of the filter is small, the noise will get averaged away, but the signal of interest should not be adversely affected. The second reason given for smoothing is to improve the quality of the data for statistical analysis by making the data "more normal."

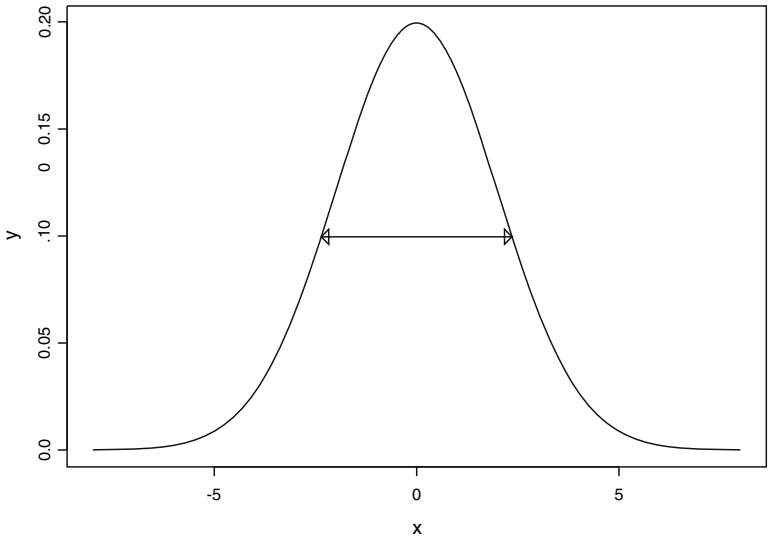


Fig. 3.2. Full width at half maximum, the common way of quantifying the amount of smoothing in fMRI. Here, the kernel has a standard deviation of 2. The maximum height at $f(0)$ is approximately 0.1995; half of that is approximately 0.0997. The x value corresponding to that height is approximately -2.355, hence application of this kernel results in a smooth of roughly 4.7 voxels.

The disadvantage of spatial smoothing is that if the size of the filter is not appropriate, that is, too large or too small, there can be a deleterious effect on the statistical analysis. For instance, if the filter is too large, say larger than the region in which activity takes place (this can happen if the regions are very small), then the activation will be smoothed out and hence will not be detected. If the filter is too small, the signal to noise ratio won't be improved, and the spatial resolution will degrade. Smoothing can also cause regions that are functionally different to merge together (Fransson et al., 2002).
One also needs to be aware that filtering will change the nature of the spatial correlation among voxels. For these reasons not all groups filter as part of the preprocessing. A reasonable compromise might be to analyze the data both with and without smoothing, and in the latter case with kernels of varying FWHM, in order to gain a clearer understanding of how critical the smoothing is to the conclusions that are drawn from the data.

3.4 Assessing the Effects of Preprocessing

Once the preprocessing steps have been decided upon and implemented, we might also consider evaluating how useful they have been. This issue has been addressed by only a few groups.

One approach to assessing the effects of preprocessing has been proposed by Strother and colleagues (Strother et al., 2002; LaConte et al., 2003; Shaw et al., 2003), using a paradigm that they call the *nonparametric prediction, activation, influence, and reproducibility* (NPAIRS) framework. The paradigm is called nonparametric because it uses cross-validation ideas, by splitting an fMRI data set into two equal parts; one half is set aside for estimating the parameters of a model, which are then used for prediction on the second half. The process is repeated, switching the roles of training and testing samples. These two steps give a measure of prediction accuracy. Reproducibility (the ability to repeat an fMRI experiment and get the same result) is measured by comparing the end result of the statistical analysis on the two training sets.

As input to the system, the authors set up different preprocessing streams, which may include no preprocessing, as well as adjusting the preprocessing parameters (for instance, varying the FWHM in the smoothing step). Running each of the preprocessing streams plus a final statistical analysis, and calculating the prediction accuracy and reproducibility of each, provides the information necessary to assess the value added by each processing step or each level of preprocessing. The authors suggest a graphical approach, for instance, plotting prediction accuracy and reproducibility for each combination of preprocessing steps and parameters. The NPAIRS framework is also used to determine optimal preprocessing streams on an individual basis, that is, which preprocessing steps are carried out and to what extent can be decided for each subject separately. For example, different amounts of filtering might be optimal for different subjects, some subjects have more motion than others, and so forth. Thus it is not surprising that individualized processing streams will generally be more advantageous than application of the exact same preprocessing to all subjects in a study. The advantage of NPAIRS is that it helps to identify, for each subject, which preprocessing steps should be performed.

McNamee and Eddy (2001) evaluate the efficacy of preprocessing steps by extending analysis of variance (ANOVA) ideas to what they term *visual ANOVA*, or VANOVA. Just as ANOVA partitions variability into different sources so that the effect of each can be isolated and understood, so too does VANOVA provide a visual "partitioning" of the variability in an fMRI data set after each preprocessing step so that the effect of each one can be isolated and understood. The method tracks changes in the mean and variance of the intensity of each voxel over time, as each preprocessing step is carried out. In this way, it is easy to evaluate which preprocessing steps have an effect on the mean of the image, which have an effect on the variance, and which affect both. It is also possible to quantify those effects numerically. Furthermore, the visual technique allows the user to identify where in the

brain specific preprocessing steps have the greatest impact. For instance, in the study reported by McNamee and Eddy, motion correction and physiological correction had the largest effect on the mean and the variance at the edges of the brain. Other uses for VANOVA, as pointed out by the authors, are to optimize the order in which the preprocessing steps are applied and to assess whether a given preprocessing step is needed. Of course, the analogy with ANOVA is only conceptual, borrowing the idea of partitioning sources of variability. The formal mathematics behind ANOVA, such as projections and orthogonality, do not hold here.

4

Statistical Issues in fMRI Data Analysis

Having surveyed the science of fMRI, specifically how the data are collected, experimental design, and noise, and before moving to an in-depth discussion of the many statistical techniques that have been developed (and continue to be developed) for the analysis of fMRI data, it might be helpful to take a step back and consider what the major statistical analysis questions are. This will help frame the discussion in the coming chapters, as well as provide context for that discussion.

4.1 Characteristics of the Data

fMRI data acquired on a single subject are characterized by the following features: they are abundant, they are noisy, and they are highly correlated both spatially and temporally.

We can think of the data with which we have to work as a time series, or more generally a movie, of the human brain in action. At each voxel of the brain, the measured data are the MR signal as it evolves over the time course of the experiment. In a typical experiment this time course may be hundreds of time points long, with an image being acquired at each time point. The number of spatial points for which data are available will usually be in the tens or hundreds of thousands, if not higher. With standard acquisition parameters, a single slice contains 4096 voxels (it is a 64 x 64 grid); the number of slices varies greatly from study to study and from MR center to MR center. A lower limit on the number of slices might be 7-10; an upper limit might be in the 30s or 40s. The number of slices, of course, will depend on how much of the brain needs to be imaged, the TR, the voxel size, and other factors, as discussed in Chapter 2. Hence for each subject in a study, we will have potentially hundreds of thousands of time series, each several hundred time points in length.

Chapter 3 described the many sources of noise that are prevalent in fMRI data. Indeed, the signal of interest represents only a small part of what is

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_4,

measured by the MR scanner, and the changes we are looking for are similarly small in scale, on the order of 3%. Even after preprocessing, there remains considerable variability (statistical noise) in the data, presenting a challenge to statistical methods and limiting, to some extent, the types of analysis that will be effective.

Another complication is the spatial and temporal structure in the data, which even now is not fully understood. What is evident is that the signal is inherently correlated both in time and space. The temporal correlation arises from the fact that stimuli are presented continuously or periodically over time, and the reaction to a stimulus at time t will clearly be affected by the stimulus at time $t-1$, and the reaction to it, and also by stimuli and reactions farther back in time; indeed one could plausibly argue that all of a person's past history affects his response to each and every stimulus presented during an fMRI experiment. Spatial correlation comes about because all of the voxels are in the brain of a single person. However, the correlation is most likely more complicated than if it were based on simple physical distances. For instance, while it is reasonable that voxels that are close to each other in the brain might be correlated in their activation patterns, it is also likely that voxels distant from each other can co-activate. As one example, voxels in the different language processing areas might be highly correlated, even if they aren't in physical proximity.

In an extensive examination of the statistical characteristics of the BOLD response, Chen et al. (2003b) studied seven subjects during a rest ("null hypothesis") condition, in order to gain a deeper understanding of fMRI noise. The subjects were scanned in two different centers, one on a 1.5T magnet and six on two separate 3T magnets. Finally, in one of the 3T scanners, functional data were acquired using a spiral sequence, while in the other, they were acquired using EPI. In all cases, subjects were simply asked to recline in the scanner with their eyes shut.

For the analysis, the authors only considered voxels in white or gray matter. The data were not motion corrected so that they could also explore the effect of motion on the statistical properties of the BOLD response. A number of interesting observations can be made. First of all, for most of the subjects in the study, the distribution of the mean level of the voxels had a long left tail; this was more apparent in the gray matter than in the white matter, with the latter being less skewed in general. The means were more spread out in the gray matter than in the white, especially in the data acquired via spiral imaging. By contrast, the distributions of the voxel standard deviations exhibited longer right tails; for all of the subjects in the study, the average standard deviation was smaller in the white matter than in the gray. As Chen et al. point out, the implication of these two sets of findings is that the white matter has smaller between-voxel variation over space than does the gray matter, and furthermore, it also has smaller within-voxel variation over time, on average.

Using simple q-q plots and correlations, the authors next examined the normality of the temporal variation within each voxel, for each subject. For

most of the subjects the percentage of the voxels that were significantly different from normal according to the correlation analysis was small, on the order of 5% or less in both white and gray matter. However, for one subject, as much as 35% of the voxels exhibited significant non-normality. Based on these findings, the authors tentatively concluded that temporal variation in the BOLD response noise is approximately normal, whereas spatial variation appears to be well fit by a gamma distribution instead.

Regarding head movement, Chen et al. observed that the subjects with a higher motion index also had more non-normally distributed voxels in the gray matter. They concluded that head motion is a major contributor to the observed non-normality of the temporal variation of the BOLD signal. No such relationship between head motion and spatial variability was found.

As we shall see in later chapters, many of the standard statistical analyses make assumptions about the normality of the noise; while this may be justified in the time dimension (i.e., within a voxel), the work of Chen et al. seems to cast doubt on the viability of the assumption along the spatial dimension (i.e., across voxels). Clearly, these issues warrant further exploration.

These features of the data make them especially challenging, and interesting, for statisticians to work with. In Chapter 5 we will examine, among other topics, the linear model approach to analyzing fMRI data, and the implications of the normality assumption. Chapter 6 will take up the more realistic, and complex, spatiotemporal modeling that is becoming more prevalent in the literature.

4.2 Detection or Estimation?

As we saw in Chapter 2, the two main experimental designs used in fMRI are capable of effectively addressing two different questions of interest. Block designs are especially useful for *detection*, that is, locating which voxels are "activated" in response to a given task, compared to a control condition. Event-related designs, by contrast, provide a means of *estimating* the hemodynamic response function. This, in turn, can also lead to detection, as in (for just one example) Gibbons et al. (2004).

Since fMRI studies have traditionally employed block designs, detection of activity has naturally been the focus of much research. Most of the commonly employed statistical techniques have as their end product a map, usually referred to as a *statistical parametric map* (or *nonparametric* if the map was obtained as the result of a nonparametric analysis) 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. The brain maps that the public often sees on the covers of magazines and in the science pages of the newspaper are one step removed from these statistical parametric maps; those "pretty pictures" show only those voxels which have

been declared active. As we shall see, reaching this point involves a combination of statistical models, statistical tests, and corrections for multiple testing; the latter is known in the neuroimaging community as the problem of "thresholding" (see Section 4.3).

With the growing popularity of event-related designs due to their greater flexibility, statistical methodology is being extended to permit efficient estimation of the hemodynamic response function, and also to use this to arrive at conclusions regarding which voxels should be considered active. Ideally, then, our statistical procedures should permit us to answer both sets of questions. Issues of detection and estimation, including the various modeling decisions that need to be made for both, are addressed in Chapter 5.

4.3 Thresholding

The question of thresholding, or determining which voxels are active, is obviously intricately linked to the question of detection described in the previous section. Most analysis and testing is still done on a voxel by voxel basis (although more sophisticated approaches are becoming common), thereby necessitating at each voxel a decision as to whether it should be considered active or not. With the many tens of thousands of voxels in a typical scan, the result is a large multiple testing problem. Various solutions exist already in the fMRI literature, ranging from completely ad hoc arbitrary thresholds to thresholds based on high-level mathematical theory. These are explored in more detail in Chapter 10.

4.3.1 Is "Voxel Activation" the Right Criterion?

One might claim that applying thresholds and designating some voxels to be "active," with the rest "inactive" is not, in fact, a realistic way of thinking about the data. Indeed, the application of hard thresholds is misleading in the sense that voxels that are close to, but below, the threshold, will not be shown in an activation map, whereas voxels that are close to, but just above, the threshold, will be shown. Yet the difference between those two voxels is minimal. Some authors argue, plausibly, that instead of hard binary thresholding, we should create gradated maps that display the change in the extent of activated regions as thresholds are modified (Jernigan et al., 2003).

We should also note that a thresholded map is a static object, summarizing the entirety of an experiment that evolved over time in a single "yes-no" decision (active-inactive). The dynamic nature of brain activity, and in particular any information about the times at which different regions become active, is lost. Questions about regional *connectivity* cannot be answered from this approach, yet these are often more interesting and critical than the detection question that is answered. One of the fascinating and difficult questions about

brain science is understanding the circuits in the train of function – which regions activate first, which regions do they then affect, and so on. More flexible and sophisticated statistical methods are needed to incorporate and examine the dynamic, temporally changing activation patterns that more accurately describe what is truly going on in the brain during the course of an imaging experiment. Friston (1998) phrases this dichotomy as a question of looking for "principles" (that is, organizing principles of brain behavior) versus drawing "maps" (the common practice). This distinction is an important one.

4.3.2 Reliability and Reproducibility of Activation

Another element of importance when thinking about issues of thresholding relates to the consistency of activation for individual voxels. If, in response to a given stimulus, the same voxels in the brain of a particular individual always activate, then the search for "active voxels" has a concrete meaning: We are looking for those particular voxels in the brain of an individual that become active when that person taps his fingers, or solves a math problem, or sees a face. However, research has shown that in fact the "test-retest reliability" of activation is quite low; a voxel that is active in one experimental run has only about a 50% chance of being active in a later repetition of the same experiment, carried out on the same individual (see, for example, Genovese et al. 1997; Noll et al. 1997; Maitra et al. 2002). Likewise, activation volume, that is, the overall number of voxels detected as active, has been found to vary considerably from scan to scan, apparently in a fashion that is uncorrelated with changes in levels of signal, levels of noise, and subject performance on the task (Saad et al., 2003b). Liu et al. (2004) also found significant differences in the numbers of active voxels across scanning trials within a session, but noted that averaging the trials led to a lack of significant differences *across* sessions separated in time. This lack of reliability of activation indicates that the target is much more elusive, and hence the focus on voxel level analysis is misguided. Researchers should instead concentrate on the behavior of clusters, or regions, of voxels – can activation be reliably detected at this aggregate level? Indeed, this is usually the ultimate goal, but it is only achieved indirectly in many commonly used statistical procedures.

When regions of activation, as opposed to individual voxels, are considered, one no longer needs to be concerned with the status of each and every voxel (active or not). Instead, the focus will be on the overall patterns of activation that are observed in the test and retest maps. It is easy to assess the test-retest reliability visually by simply looking at the two maps, a qualitative rather than a quantitative comparison, as concluded by Liu et al. (2004). More formally, as part of a more organized statistical analysis, differences between test and retest could be evaluated through *contrasts* in a linear model. Taking this view, Kiehl and Liddle (2003) demonstrated strong reproducibility of the hemodynamic response in an event-related study of an auditory "oddball" task, where the two sessions, carried out on 10 subjects, were separated by

approximately six weeks. In the oddball task, a "standard stimulus" is presented with relatively high probability (in this experiment, 75% of trials were the standard stimulus); with lower probability, a "target" and a "novel" stimulus are also presented (frequency 12.5% each in the Kiehl and Liddle study). The subject is only expected to respond to the target stimulus, ignoring both the standard and the novel stimuli when they are presented. Kiehl and Liddle (2003) give both detailed functional maps as well as a contrast-based statistical analysis to back up their claim that the hemodynamic response, at least for this task, is quite consistent across testing sessions separated in time.

Other researchers have not found such consistent patterns across testing sessions, even when the focus has been on regions of activity. McGonigle et al. (2000) scanned a single subject 33 times over the course of two months on a series of three simple motor (finger tap), visual (flashing checkerboard), and cognitive (generate random digits from 1 to 9) tasks. In what the authors term a "Groundhog Day" scenario, each scanning session was treated as though it were the first time the subject was performing the experiment, to mimic the usual paradigm in which each subject is seen only for one session, and the results of that session are taken as representative of the subject's brain activation patterns as a whole. In each session of the "Groundhog Day" scenario, the subject performed all three tasks, in counterbalanced order, block design experiments. By contrast with the results of Kiehl and Liddle, McGonigle et al. report statistically significant differences in activation across the 33 scanning sessions. This can also be seen in the thresholded maps they present for the three tasks over the sessions, which vary greatly, and not in any systematic fashion. The authors caution against relying too heavily on the outcome of a single scanning session, as discovered effects may be idiosyncratic to the particular day or time at which the images were acquired. It is worth noting that in a follow-up analysis of the same data set, this research group softened their conclusions somewhat (Smith et al., 2005).

At the present time, the test-retest reliability and reproducibility of fMRI results seem to still be somewhat of an open question. Evidently, some experimental paradigms are more robust than others, and will reveal consistent patterns of activity across subjects and across scanning sessions within a subject. Others are more elusive. A complicating factor is the different ways in which one can think about reliability of activation. Among the types of reproducibility that have been of interest are: consistency for a single subject across sessions, or across runs within a single session; consistency across subjects, within or between sessions; and consistency across imaging centers, on different scanners, using the same or different subjects. Aside from the activation status of a voxel, characteristics of the BOLD response itself have also been examined for their reliability on repeated trials. See Casey et al. (1998) (reproducibility across imaging centers); Salli et al. (2001) (effect of classification rule for determining the activation status of a voxel on reproducibility); Neumann et al. (2003) (reliability of BOLD response patterns across imaging sessions) for more detailed explorations of specific aspects of this topic.

4.4 Multiple Subjects

Most imaging experiments include a small number of subjects, but in current studies that number is still greater than one, and the trend is toward larger groups (Thirion et al., 2007). Considering each subject on an individual basis is relatively easy, but not powerful in any statistical sense. Furthermore, it does not permit researchers to draw meaningful conclusions about the behavior of the group of subjects as a group, to extrapolate to the population from which the subjects were sampled, or to compare groups of subjects (for instance, young children, older children, adolescents and adults; or schizophrenic patients and healthy controls; or men and women). It is therefore crucial to have statistically valid and powerful methods of combining the information obtained from multiple subjects in order that these higher-level scientific questions can be investigated.

The issue of combining information across subjects has two aspects, only one of which is statistical. The other dimension is physiological and has to do with the fact that each person's brain is different. Brains differ in size, shape, and in the relative positions of identifiable regions and landmarks. It only makes sense to combine brain images from different individuals if they are put on an equal basis, in the form of a common space. Thus a preliminary step prior to any statistical combination work is to warp the brain images onto such a common atlas, of which several possibilities are available, most notably the Talairach coordinate system (Talairach and Tournoux, 1988). Once this is done, we can consider the various methods of combining information across independent subjects, and of comparing independent groups of subjects, depending on the ultimate inferential goals. In Section 5.5 we examine and compare methods for combining subjects.

4.4.1 Consistency Across Subjects

Just as reliability of activation is an important consideration for model fitting in general and thresholding of statistical maps in particular, so too is consistency across subjects, at least in some broad sense, when the focus is on combining data to create *group maps*. One would naturally expect between-subject variability to be greater than within-subject variability; subjects will differ more from each other than an individual will differ from himself scanned at different points in time (barring any brain trauma or similiar phenomenon, of course). Still, even in this context some amount of reproducibility across subjects of the effect of interest is clearly desirable. To take an extreme (and unrealistic) example, if every subject has a completely idiosyncratic brain response to a simple motor task such as tapping the fingers on the right hand, then any meaningful statistical analysis will be ultimately unattainable. We cannot gather strength from disparate sources of information (subjects) if there is no commonality among them, hence a fair amount of consistency as we move from subject to subject is needed. The issues, between subject and

within subject variability, are of course related, and it is important to assess the relative strength of these two effects. At the very least, statisticians approaching the analysis of group level data need to be aware of the special complications that arise in the neuroimaging setting.

How consistent are patterns of activation across subjects? Typically, the combination of subjects to create a group map is not performed on the thresholded maps of the individuals (see Section 5.5.2 for a detailed discussion), so it makes little sense to worry about the reproducibility of the binary active/inactive images. Furthermore, since the hemodynamic response underlies what is observed in the data, whatever the experimental paradigm from which they were generated, we can focus for the purpose of this discussion on variability of that response across subjects.

In an early study on the stability of the hemodynamic response, Aguirre et al. (1998) had 41 subjects perform a simple motor task (bilateral button press). Thirty-two of these subjects were scanned only once; four of the subjects were scanned five times, with each scan on a different day, spread out over several months; the final five subjects were also scanned five times, but all five scans were on the same day, in the same scanning session. For the first group of subjects, therefore, only one estimate of the hemodynamic reponse could be obtained; the scans on the other two groups yielded five separate estimates. This allows for comparison both within and between subjects. Analysis was not performed on the whole brain, rather the authors concentrated on a specific area, composed of approximately 200 voxels. The hemodynamic response was estimated by averaging together the time courses of the active voxels in the identified search region of interest.

Looking first at the subjects who were scanned multiple times during one session, the authors found significant differences in the hemodynamic response on only one (one subject was also dropped from this part of the analysis due to insufficient overall activation in the prespecified region of interest). This indicates that scan-to-scan variability within a single session is probably low (although the small number of subjects makes any such conclusion tentative at best). For those subjects who were scanned multiple times over a period of months, by contrast, significant differences were found for three. Finally, the subjects who were scanned only once exhibited highly significant differences in their hemodynamic responses. This is noteworthy especially since the search region was highly localized and did not include many voxels, a stark contrast to the common whole brain analyses. Putting these results together, it appears that the hemodynamic response within subjects and scanning sessions is much more stable than the response across subjects, as one would expect. However, it is also evident that the response within a subject is not necessarily stable over time, as we also saw in the discussion in Section 4.3.2.

Based on their findings, Aguirre et al. suggested that analysis might be improved if the fitted hemodynamic response function is unique to each subject. This notion was further explored in Handwerker et al. (2004), who also examined the effects of shifting the parameter values of the models used for

the HRF, on the outcome of a statistical analysis. In addition, Handwerker and colleagues considered multiple brain regions of interest, which were all assumed to be activated by a single task. Across 20 subjects who had enough activation in the regions of interest to warrant further analysis, it was found that the form of the HRF indeed differed from subject to subject, as well as from region to region within a subject. Some subjects exhibited more variability across regions than did others. The empirically derived HRFs of many subjects also differed significantly from the "canonical model" implemented in the SPM software (we will see more on this model in the next chapter); the peak in the BOLD response tended to occur earlier for these subjects than allowed for in the standard model.

In general, there was more variability across subjects within each region than within each subject across regions, again indicating that behavior within an individual is relatively stable. Note that this stability is only relative; as Handwerker et al. point out "Although intersubject variability is larger, there is a large intrasubject variability" (p. 1649). Compared to using the canonical HRF for all subjects and regions, the authors found improved detection ability (in simulated data) by using an empirically derived function for each subject; in their study, the empirical HRF was derived in one region and applied to others, and this still resulted in better performance. Presumably, further improvement could be achieved if the time and effort were taken to estimate region-specific response functions, although this tactic might not be practical in all situations.

A recent exploration of this question on a large cohort of 81 subjects (Thirion et al., 2007) gives some interesting further insight on a scale that is not possible with smaller groups of subjects. For example, by splitting their data set into smaller subsets of 10 to 16 subjects (the typical size for multi-subject studies) Thirion and colleagues show that the resultant group maps obtained by some of the common models exhibit a great deal of variability. This variability is strong enough that in some cases different scientific conclusions would be reached. This is a pretty strong indication that across-subject variability should not be ignored. Group maps are more reproducible as the number of subjects on which they are based increases, providing an impetus to move toward larger studies in general. As has been found by other researchers, Thirion et al. (2007) also note that reliability or reproducibility of activation is highly task-dependent.

4.5 Regional Versus Whole Brain Analysis

Should fMRI data be analyzed on a region by region basis, with, for instance, different models or different forms for the HRF assumed in different parts of the brain? Or should we take the stance that one unified analysis should suffice for the brain as a whole? Both approaches have advantages as well as drawbacks. A region-based analysis is likely to be more realistic. As we shall

see in Chapter 5, researchers have repeatedly found variations in, for example, the details of the BOLD response (time to peak, rate of decay back to baseline, etc.) when different tasks and regions of the brain are considered. Hence it is unlikely that one model will be appropriate for all voxels. Certainly at the very least one ought to consider that the behaviors observed in the gray matter, where most brain activity occurs, will differ from those in white matter or cerebrospinal fluid, and perhaps the latter should be modeled differently than the former. Even within the gray matter itself, however, it might be reasonable to posit region-specific analyses. The drawbacks of such an approach are equally obvious. First, it necessitates having some conception of what these models should look like, which might not be feasible especially when new experimental tasks are being explored. Second, the process will perforce be more involved, requiring for example a partitioning of the brain into gray/white/cerebrospinal fluid, and perhaps further within the gray matter into specific regions of interest (ROIs). This is a very time-consuming and arduous task, demanding expert knowledge, and it introduces an element of subjectivity via the definitions of the regions for each subject in the study; furthermore, since the size of the voxel is based on the imaging parameters and has no intrinsic physiological meaning, a single voxel may contain more than one type of tissue (for example, both gray and white matter can appear in the same voxel). Fitting region-specific models is more computationally intensive than fitting one model for the whole brain, especially if model selection is also part of the procedure. Finally, if a different model is fit to each region, or potentially even to each voxel in the analysis, issues of thresholding become more complicated, since a single cutoff point can no longer be applied, and interpretation of the resultant map is more difficult.

The advantages of a whole brain analysis mirror the disadvantages of the regional analysis. First, it is computationally easier to fit a single model to the entire brain, even if model selection is involved. Not having to segment the brain into the different tissue types or partition it into ROIs saves effort and time (defining and outlining the regions of interest can take weeks of work), permitting a more detailed focus on the statistical analysis proper. One threshold can be applied to the entire brain and the binary map of active/not active voxels is easily interpeted. On the other hand, as noted above, the assumption that one model or one type of analysis will be suitable for every location is questionable and probably not believed by most practitioners. Whole brain analyses are often, although not always, done on a voxel-by-voxel basis; this introduces the additional unrealistic assumption that voxels are independent.

It is in fact quite common to have both approaches represented in a single analysis: An analysis will be performed of the entire brain using a single model, in conjunction with a more specific analysis of several ROIs, focusing on particular aspects of behavior (for instance, percentage of activated voxels in a region). The questions of level of analysis and the assumptions implicit in the various choices are starting to receive more attention in the fMRI literature, but the area is still rather underdeveloped.

4.6 Summary of Statistical Challenges

The opportunities inherent in the analysis of fMRI data are many. I have touched here on some of the most obvious and immediate issues: developing useful models for detection and estimation of the HRF, thresholding, combining information from multiple subjects, and comparing results from groups of subjects. As we shall see in the upcoming chapters, each of these basic questions has provided the impetus for methodological/statistical research, yielding many and varied (partial) solutions.

As would also be expected from a data source as rich as fMRI, the interesting statistical questions deepen the more we learn. For instance, moving away from the simple and simplistic voxel-by-voxel analysis, we can consider incorporating spatial and temporal information garnered from previous studies to build more realistic models. This presents challenges of its own, both conceptual and computational, which have yet to be fully explored. Methods from computer science, engineering, and even data mining can be applied to fMRI data, representing different views of the scientific problems, and these can (and should) be compared to more statistical approaches. Model selection and model choice, visualization of large high-dimensional data sets, efficient processing when the number of subjects is also large (a situation that is becoming increasingly feasible), Bayesian versus frequentist analyses: all of the questions at the center of modern statistical practice can, and do, find expression in one form or another in fMRI. Many of these topics are taken up in the following chapters.

5

Basic Statistical Analysis

In this chapter, we consider some of the basics of the statistical analysis of fMRI data. The goal is to study the predominant fundamental approaches to the data analysis problem so that we can build up to an understanding of the more statistically and conceptually advanced methodologies being currently developed. We start with a brief discussion of exploratory data analysis (EDA). The discussion then moves on to the basic statistical analysis of block design studies and event-related studies. This will set the stage for a more general survey of the general linear model as it is used in fMRI data analysis. The chapter ends with an examination of some simple methods for combining multiple subjects.

5.1 Exploratory Data Analysis

Exploratory data analysis (EDA) as we commonly use it in statistics is not very prevalent in fMRI. I believe that there are several reasons for this. First, the nature of the data plays an important role. Since the raw data obtained from the scanner have to be preprocessed significantly before they begin to look anything like a brain, there is a sense in which the lowest level information to which we have easy and interpretable access has already undergone a fair amount of analysis. A statistical map showing levels of activation at each voxel, for instance, has already been subjected to statistical analysis, with an assumed underlying model, and in this situation classical EDA is not relevant. Yet this is the form in which practitioners are used to seeing their data, often for the first time.

This would argue for taking the data down a level, bringing us to the second point, namely, the massively complex spatial and temporal characteristics of fMRI data render many of the standard EDA visualization tools difficult, if not impossible, to apply. For example, what would a boxplot of the 4096 time courses in a typical slice look like? How should the data from different slices be presented – as each slice individually or all slices collectively? How would

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_5,

the data from subjects within a group, or for different groups of interest, be compared? Some preliminary statistical analysis seems to be necessary prior to performing EDA, or alternatively new EDA methods for large, complicated, spatiotemporal data sets need to be developed and applied in this setting.

Finally, and related to the second point, is that at the "most raw processed level" (that is, after preprocessing to clean up the data but prior to any substantial statistical analysis), it is not even entirely clear what we would hope to discover via traditional EDA. Regardless of the experimental design, be it block or event-related, one would be dealing with time courses of the hemodynamic response function (HRF), either in aggregate (block design) or distinctly for each trial (event-related design). "Interesting patterns" in these time courses would be indicative perhaps of localized activation, the subject of more formal statistical analysis. Again, we come back to the question of how exactly should large, spatiotemporal data be visualized? Histograms and other plots to assess normality are clearly not relevant or directly applicable. Likewise, many of the tools we are used to employing seem to have no direct utility or applicability to fMRI data.

In summary, the development of appropriate and relevant tools for the exploration and visualization of "most raw processed" fMRI data, together with the formulation of suitable target questions for EDA in this context, is an area that has not yet, in my view, been adequately addressed. As with much of the work in fMRI, it is possible, indeed likely, that here, too, importing ideas from other fields with similar data problems (such as geostatistics) could be fruitful.

An interesting attempt at performing some EDA steps for fMRI appears in Luo and Nichols (2003). These authors propose a set of diagnostics and interactive graphical tools for simple assumptions on the error structure (such as normality or autoregressive of various orders), independence, constant variance, and model fit. The process involves a series of steps, some relating to scan summaries and others to model summaries. The aim of the former is to identify particular scans (images) that are potentially problematic, for instance, very noisy or showing a lot of motion; the aim of the latter is to check whether model assumptions are violated, and if so when and where. That is, there is both a temporal and a spatial element to the EDA, as well as a modeling component, as there must be. Finally, after remediation, the diagnostics provide a measure of the validity of the analysis at each voxel. Note that this EDA is different from what a statistician might typically deem "exploratory" in that it includes modeling as an integral part.

5.2 Block Designs: Basic Analysis

Section 2.2.1 described how in a block design experiment the constant stimulation during the task blocks results in a characteristic time course for activated voxels, whereby elevated levels of activation are observed roughly during the

task blocks and lowered levels of activation are observed during the control blocks. By contrast, the observed activation for voxels that do not respond specifically to the stimulus is not related so directly to the structure of the design.

An example of several voxels from a well-understood visual task appears in Figure 5.1. In this block design, periods of fixation (a black screen with a red crosshair in the center) alternated with a black and white flashing checkerboard at 8 Hz (also with a red center crosshair). The experiment started and ended with a period of fixation. Each block lasted 30 seconds and the experiment lasted 4.5 minutes. The plot shows the time courses for eight voxels: three from occipital cortex that are very robust and task-related; three from the edges of occipital cortex that are weakly related to the task; and two from unrelated areas in the cortex. The first panel shows representative voxel time courses from each of the three regions, plotted on the same scale. As can be seen in this panel, the general levels of activation are the same in all voxels, making this alone a useless criterion for discriminating between active and inactive voxels. On the other hand, the voxels from strongly task-related regions exhibit very different behavior from the behavior of voxels in unrelated regions, with elevated levels during periods of flashing checkerboard and depressed levels during the periods of fixation. The weakly task-related voxels on the edge of occipital cortex are similar to the unrelated voxels, although they do show some signs of the characteristic behavior of interest.

The other three panels of Figure 5.1 show close-ups of the three types of time courses; from this closer look it is apparent that the voxels on the edge of the relevant region are in fact responding to the visual task, but at much reduced levels compared to the strongly task-related voxels. By contrast, the levels of activation for the voxels in an irrelevant region of the cortex show no relation at all to the presentation of the stimulus.

These two broad patterns, one for responsive, or "active," voxels and one for unresponsive, or "inactive," voxels suggest some simple ways for proceeding with the data analysis on a voxel by voxel basis. For simplicity, assume that there is only one task of interest, so that the experiment alternates between blocks of task and blocks of control. Perhaps the easiest analysis in this setting is to ignore the temporal aspect of the data and to calculate the value of the t statistic for comparing the mean levels of activation during the task blocks and the control blocks, at each voxel. Hence, at voxel i the amount of activation is summarized by the statistic

$$t_i = \frac{\bar{X}_{\text{task}} - \bar{X}_{\text{control}}}{se},$$

where $\bar{X}_{\text{task}}$ is the average level of activation in voxel i over all times during which the task was being performed (that is, aggregating over all the task blocks), $\bar{X}_{\text{control}}$ is the average level of activation in voxel i over all times during which the subject was at rest or performing the control, and se is the standard error of the difference; commonly one would use the *pooled estimate*

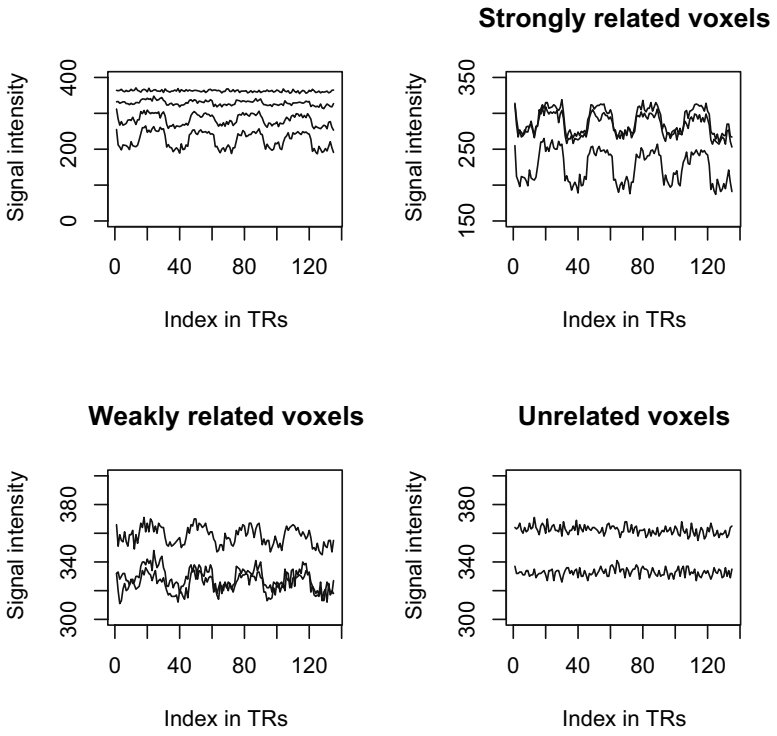


Fig. 5.1. Time courses for different voxels in a simple flashing checkerboard block design. Voxels that are strongly involved with the visual processing task show a characteristic boxcar behavior of increased activation during the stimulus presentation and relaxation during the fixation period. Voxels that are weakly involved in the task show a similar pattern, but not as strong. Unrelated voxels exhibit random "noise-like" activity throughout the course of the experiment. Data courtesy of Christine Krisky, Oregon Health and Science University.

of the variance,

$$s_p^2 = \frac{(n_{\text{task}} - 1)s_{\text{task}}^2 + (n_{\text{control}} - 1)s_{\text{control}}^2}{n_{\text{task}} + n_{\text{control}} - 2},$$

so that

$$se = s_p \sqrt{\frac{1}{n_{\text{task}}} + \frac{1}{n_{\text{control}}}}.$$

Here, n_{task} and n_{control} are the number of observations in the task and control conditions, respectively, and s_{task}^2 and s_{control}^2 are the sample variances of the activation in those two conditions, respectively, at voxel i.

After performing this analysis at each voxel for a given subject, the result can be plotted in a *statistical parametric map*, in this case a map of t values.

Typically, such a map will be generated for each subject, and these maps will in turn be subject to additional statistical processing, for instance *thresholding* to determine which voxels should be declared "active." See Chapter 10 for more on this issue. In most instances, practitioners will only be interested in those voxels for which the level of activation was higher in the task condition than in the control condition, i.e., large positive t values, leading to a one-sided hypothesis test for the difference being greater than zero. Voxels in which the level of activation decreases during the task condition relative to the control are termed *deactivated* and are usually of less scientific interest. Note that if there is more than one task condition, this basic approach can be extended to generate an F statistic at each voxel for comparing levels of activity over all conditions, leading to an F map for each subject, which in turn will be thresholded, and so forth.

This analysis makes certain strong assumptions, beyond the usual one of normality, namely:

1. *Independence of voxels.* Since the analysis is carried out for each voxel individually, resulting in a single test statistic at each point in the image, the implicit assumption is that those voxels are independent. This assumption of *spatial independence* is clearly not realistic and will be violated in all fMRI data sets.
2. *Independence in time.* Likewise, this analysis aggregates over time points, both within and across blocks of similar type (task or control), implicitly assuming that measurements collected at different points in time are independent. The assumption of *temporal independence* is also unrealistic and will be violated in all fMRI data sets.
3. *Onset of the hemodynamic response is immediate upon stimulus presentation and it likewise stops immediately when the stimulus is stopped.* In the simple version of the test described above, no consideration is taken of the physiological fact that the BOLD effect does not start instantaneously, but rather there is a delay of several seconds before the response begins, and several more seconds before the peak is reached. Similarly, once the stimulus is turned off in the control conditions, the hemodynamic response does not immediately drop back to baseline; rather the decline is gradual, again taking several seconds. These behaviors can be accommodated within the simple modeling framework, at the price of some loss of data, by deleting the first few images after each transition from one block type to another.
4. *Equal variances in task and control conditions.* Although this assumption may or may not be realistic, the two-sample t test is quite robust to violations of homoscedasticity, especially when the sample sizes are close to being equal (Miller, 1986), as they are in fMRI experiments run with block designs.

A recent study (Pavlicová et al., 2006) examined the assumptions of normality and equal variances in the rest and task conditions. Not surprisingly,

the study found that the data can exhibit departures from both of these assumptions. The authors proposed using, instead of the standard t test, modifications that allow for these departures: the Welch test (Welch, 1937)

$$t_{W_i} = \frac{\bar{X}_{\text{task}} - \bar{X}_{\text{control}}}{\sqrt{\frac{s^2_{\text{task}}}{n_{\text{task}}} + \frac{s^2_{\text{control}}}{n_{\text{control}}}}},$$

which permits the population variances to be different; and the Cressie-Whitford test (Cressie and Whitford, 1986), which has a complicated form and additionally allows for non-normality. Simulations show that the Cressie-Whitford test is more powerful than either the standard t or the t test modified according to Welch's correction.

In spite of the serious simplifying assumptions that are made, the simple t test analysis is common in the fMRI literature, since it is easy to implement and gives sensible scientific results for a wide variety of experimental conditions. Averaging over time (trials) also has the effect of increasing the signal to noise ratio, resulting in a more powerful analysis. Nonparametric versions of this statistic, which allow the researcher to drop the normality assumption, are also available, as they always are. These are not often used, however, perhaps because the degrees of freedom (as measured through the length of the time course, that is, the number of images) are generally large, and in this case nonparametric tests are approximately normal.

It is possible to refine the t test analysis and take some advantage of the temporal information in the data by carrying out a so-called *correlational analysis* (Bandettini et al., 1993) in which the stimulus time course or the predicted hemodynamic response is correlated with the voxel activation time course. That is, at voxel i calculate

$$r_i = r(\mathbf{S}, \mathbf{X}_i),$$

where S is the pattern of zeros and ones describing the block design stimulus presentation pattern, and X_i is the activation time course of voxel i; r represents Pearson's correlation coefficient,

$$r(\mathbf{Y}, \mathbf{Z}) = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})(Z_j - \bar{Z})}{\sqrt{\sum_{j=1}^{n}(Y_j - \bar{Y})^2 \sum_{j=1}^{n}(Z_j - \bar{Z})^2}}$$

between two vectors $\mathbf{Y}$ and $\mathbf{Z}$ of length n. Voxels with a high positive value of r_i are considered to be active, since there is a strong correlation in those voxels between the task (respectively, control) blocks and high (respectively, low) levels of activity. When there is no such correlation, voxels are inactive, and strong negative correlation indicates deactivated voxels, which, as mentioned above, are generally not of interest.

It can be shown, in fact, that when the number of images in the task condition is equal to the number of images in the control condition (that is,

a balanced block design), the two analyses coincide after transforming r_i by the usual

$$t_i = r_i \sqrt{\frac{n-2}{1-r_i^2}},$$

where in this case, with n as the total number of images we have $n/2$ images from the control condition and $n/2$ from the task condition. However, when the design isn't completely balanced, e.g., there is one more run of the task block than of the control block, the two analyses only approximately coincide.

An advantage of moving from the t test to the correlation framework is that we can consider different stimulus series other than the simple "boxcar" of the two-condition block design experiment. For instance, we can just as easily correlate fMRI time series with a time series of behavioral data collected in the scanner during the course of the experiment, to explore which voxels activate when a particular behavior is carried out. This may or may not correspond exactly with the series of stimuli as presented to the subject in the scanner. Or, we can correlate the fMRI time series with a smoothed version of the stimulus series, or some other suitable transformation. It is also easier to account for the delay in the hemodynamic response by considering correlations with a lagged stimulus series rather than the original one. Furthermore, designs which may not be amenable to a straightforward t- or F-type analysis, such as an experiment with more than two conditions that is not completely balanced, might lend themselves to a correlational analysis instead. In most applications in practice, of course, it will matter little which of the two approaches is taken. Some research groups tend to report their results in terms of the correlation analysis, and others in terms of the t analysis, but the difference seems mostly to be an issue of preference or habit.

Similar to the t test analysis, one can account for the gradual increase and decrease of the hemodynamic response at the transitions between task and control blocks by calculating a shifted or lagged correlation between the stimulus presentation series and the voxel time course. For longer lags more observations are lost.

5.3 Event-Related Designs: Basic Analysis

In the analysis of event-related studies, a major focus is estimation of the hemodynamic response function (HRF); estimated HRFs can in turn be used to identify areas of interesting activity. Beyond simple estimation of the HRF, event-related experiments allow for the investigation of a range of subtler effects as well, such as estimating the delay of response onset, exploring differences in brain behavior according to reaction to a stimulus (for instance, differences in the hemodynamic response when a correct answer is given to a question and when an incorrect answer is given), and so forth.

There are two main tools for the basic analysis of event-related designs: *trial averaging*, reminiscent of both the t test analysis for block designs and the analysis of event-related potential (ERP) studies of the brain (whence the name "event-related fMRI"); and *function estimation*.

By "trial averaging" we refer to the simple averaging together of like trial types (Dale and Buckner, 1997; Buckner, 1998). The trials in an experiment are sorted according to type and then the fMRI time courses of each type, at each voxel, are studied. In this way, we obtain the average fMRI time course for a particular type of trial, as well as the variance of the time course for that type of trial. Averaged time courses for voxels that are activated in reaction to a particular trial type should exhibit the stereotypical behavior of the BOLD response; for trial types that do not cause a reaction, the trial averaged fMRI series should not show the typical behavior.

The trial averaged time series can be further subjected to statistical analysis such as a t test or a correlational analysis, as described in the previous section, or to any of the more sophisticated approaches outlined in Section 10.6.

Dale and Buckner (1997) and Buckner (1998) show that this simple approach can be effective in eliciting meaningful responses when trials are spaced as little as 2 seconds apart; furthermore, when subjects are presented with more than one trial of the same type, in rapid succession, the BOLD response adds in a more or less linear fashion. Hence, trial averaging is not restricted to cases in which the interstimulus interval (see Chapter 2) is long enough for the hemodynamic response to return to baseline levels, as one might expect a priori.

Note that trial averaging yields an estimate of the HRF, albeit one that is not based on any model or assumptions. With the "function estimation" approach, we aim instead to obtain an estimate of the HRF that is model-based in some sense. The techniques can be roughly categorized as *parametric* and *nonparametric*.

5.3.1 Parametric Approaches to the Estimation of the HRF

As the name implies, the parametric approach to HRF estimation specifies some family of statistical distributions to model the shape of the BOLD response curve. Although the first model to be proposed was the Poisson (Friston et al., 1994), the family that is most often used for this purpose is the gamma (Lange and Zeger, 1997). Compared to the gamma family, the Poisson family is much less flexible, since it has only one parameter available to describe the HRF; furthermore, it is discrete, whereas the BOLD response evolves continuously over time.

Lange and Zeger (1997) model the HRF at time t and voxel i by a simple two-parameter gamma family:

$$h(t,i) = b_i(tb_i)^{a_i-1}\exp(-tb_i)/\Gamma(a_i),$$

where a_i and b_i are the shape and scale parameters, respectively, of the gamma distribution, which are allowed to vary from voxel to voxel, and $\Gamma(\cdot)$ is the normalizing constant for the density, the gamma function. The parameters are estimated iteratively, a potentially computer-intensive and time-consuming step. An example of such a function for $a = 3$ and $b = 4$ appears in Figure 5.2. Note that the shape of the curve for these choices of the parameters roughly parallels the fundamentals of the response: a slow rise to a peak value, followed by a (slower) decline back to baseline. Clearly the gamma family is flexible enough, through the choice of parameters to accommodate a wide range of functional shapes, only some of which will approximate the BOLD response.

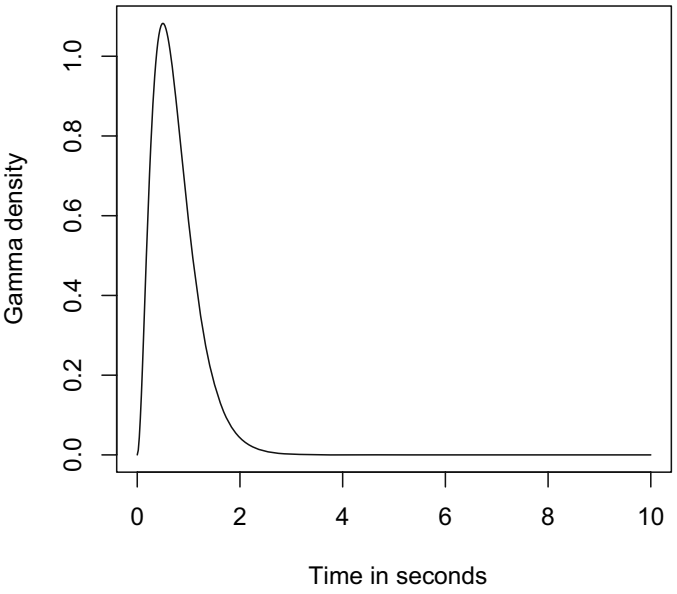


Fig. 5.2. Lange and Zeger model for the HRF based on a gamma distribution, here with shape parameter 3 and scale parameter 4.

Figure 5.2 shows that a single gamma component can describe the gross features of the BOLD response; the finer details, such as the initial delay prior to onset of the response and the undershoot before the recovery to baseline levels, are not so easily modeled with a single gamma. For this reason, Friston et al. (1998) modify the suggestion of Lange and Zeger by taking instead the difference of two gamma functions, with set parameter values as given by Glover (1999); this version is currently implemented in the SPM software package. Using this model, the slight underdip that is sometimes observed

prior to the return to baseline is accounted for:

$$h(t) = \left(\frac{t}{\tau_1}\right)^{\delta_1} \exp\left[-\frac{\delta_1}{\tau_1}(t-\tau_1)\right] - c\left(\frac{t}{\tau_2}\right)^{\delta_2} \exp\left[-\frac{\delta_2}{\tau_2}(t-\tau_2)\right],$$

with $\tau_j = 0.9\delta_j$, $\delta_1 = 6$, $\delta_2 = 12$, and $c = 0.35$. The two values τ_1 and τ_2 represent the time to peak and the time to undershoot peak, respectively. Using other values of the parameters in the function $h(t)$ besides those given here corresponds to different assumptions regarding the time to peak and peak undershoot, which may indeed vary by task and by region of the brain (Glover, 1999). The function, plotted in Figure 5.3, is more flexible than the simple gamma model, and more realistically describes behavior observed in studies of the HRF.

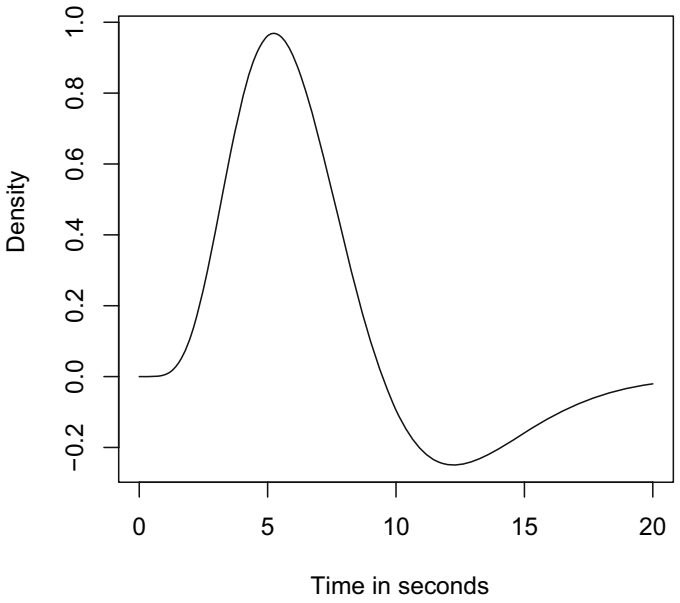


Fig. 5.3. The difference of two gammas, using parameter estimates from Glover, 1999, to model the HRF.

Estimating the parameters for the gamma-type model can be difficult, requiring iterative procedures, hence the tendency to use a "canonical" set of parameter values whether they are appropriate to the particular task and brain region under study or not. Alternatively, Rajapske et al. (1998) propose a simple Gaussian model for the HRF. In contrast to the other parametric

models, the authors claim that the parameters of the Gaussian model can be easily estimated, or approximately so, making it feasible to provide voxel- and task-dependent fits. In their comparison of the three model types, Rajapske et al. define two characteristics of the HRF: the *lag* (equated to the mean of the model function) and the *dispersion* (equated to the variance). Clearly, for the Gaussian family these are independent, whereas for both the Poisson and the gamma families they are not: for the Poisson the mean and the variance are the same, whereas for the gamma, the mean and the variance are related. Empirically, the authors find no evidence of a relationship between the lag and dispersion of the hemodynamic response in active regions over a variety of tasks. On the other hand, the symmetry of the Gaussian curve does not correspond to the supposed response, nor can it account for the often-observed undershoot.

These various approaches are subsumed into one general, nonlinear model by Kruggel and von Cramon (1999). Concentrating only on the subset of voxels declared significant by some other method, they consider models of the form

$$y(s,t) = g(t,\beta) + \epsilon(s,t),$$

where s indexes the voxel and t indexes time, and $g(t,\beta)$ can be taken as any of the existing parametric forms for the HRF described in this section. Across s and t, the vector $\boldsymbol{\epsilon}$ is assumed to be multivariate normally distributed with mean zero and unknown variance-covariance matrix Σ, whose elements must be estimated.

The β parameters of the model are estimated using maximum likelihood. Estimation of the variance-covariance matrix is more complicated, since without any further restrictions the number of parameters is large; hence some constraints must be imposed. The authors accomplish this by separating the variance matrix into temporal and spatial parts and assuming each of these has an AR(1) structure:

$$\Sigma = \gamma_2^2(\Sigma_S \otimes \Sigma_T),$$

where γ_2^2 is a variance term and $\otimes$ is the Kronecker product. Σ_S^{-1} and Σ_T^{-1} are modeled and estimated separately, and then used to build Σ^{-1} via

$$\Sigma^{-1} = 1/\gamma_2^2(\Sigma_S^{-1} \otimes \Sigma_T^{-1}).$$

5.3.2 Nonparametric Approaches to the Estimation of the HRF

The parametric approach to estimating the HRF has two main, intertwined, drawbacks. First, for simplicity it is usually assumed that the form of the HRF is the same at each voxel, under all conditions, and for all subjects. This is the implication of fitting the difference of two gammas with the canonical parameters as described in the previous section. Yet there is increasing evidence,

some of which we shall see in this section, that the HRF varies spatially, and across individuals. The need to fit the HRF with a single model regardless of the circumstances is in part a consequence of the second drawback, namely that many of the methods proposed in the previous section, as we saw, are computationally intense. This is a problem that plagues fMRI data analysis in general due to the highly complex and high dimensional nature of the data. It is common therefore to make the simplifying assumption that one model fits at each and every voxel, although this is clearly not realistic.

While the nonparametric approaches do not rid one of the computational burden (indeed, many of the methods we will see in this section are very computationally intensive), they do have the benefit of not imposing a priori and somewhat arbitrarily a prespecified shape to the HRF, such as a Poisson or a gamma. Under the rubric of nonparametric estimation we include techniques that use basis functions or otherwise avoid families of distributions to fit the shape of the response curve. Genovese (2000), for example, fits the HRF as one component of a more general Bayesian time course model, using cubic splines, with knots located at certain "critical points" of the curve. In this way, separate pieces may be fit for the initial delay, for the increase up to peak, for the decay, for the underdip, and for the recovery back to baseline levels. Genovese combines this spline model with terms for the baseline signal and the drift, as well as a series of priors, to develop a full-blown, computationally intensive, Bayesian approach. The need to specify priors, together with the computational complexity of this method, has limited its applicability in practice. We discuss Genovese's model in more detail in Chapter 9, but simply note here the use of a nonparametric method to estimate the HRF itself.

Gibbons et al. (2004) take as a starting point time-averaged curves for an event-related design with a "pseudo-block structure" (as in Dale and Buckner 1997, and others discussed above). Specifically, the design presents a single quick stimulus, followed by rest condition long enough to allow return to baseline; this same pattern of single quick stimulus and long rest is repeated many times. The time-averaged curve is obtained by averaging the responses at each corresponding time point in the pseudo-block (i.e., first images in each block are averaged, as are second images in each block, and so on). To these curves they then fit a simple cubic polynomial, which captures many of the main features of the hemodynamic response, without fixing a parametric form. The authors use a random effect polynomial regression with one effect for each voxel; the random effects represent the deviations of the voxels from the overall "average" cubic polynomial curve. The model can also incorporate fixed effects for particular covariates such as the brain region in which the voxel is located:

$$y_i = W_i\alpha + X_i\beta_i + \epsilon_i,$$

for $i = 1, \ldots, N$ voxels; y_i is the time averaged time course for the ith voxel, W_i is the known design matrix for the fixed effects (if there are any in the

model), X_i is the known design matrix for the random effects, and α, β_i are the unknown parameters. The random effects are assumed to be normally distributed with mean zero and unknown covariance Σ_β; the errors ϵ_i are also normal and assumed to have covariance $\sigma_\epsilon^2 \Omega_i$.

Random effect coefficients in this model are estimated using an empirical Bayes approach, which has the advantage that estimates of the polynomial coefficients at a particular voxel "borrow strength" from the information available in the other voxels in the analysis. This yields the estimates

$$\tilde{\beta}_i = [X_i^T (\sigma_\epsilon^2 \Omega_i)^{-1} X_i + \Sigma_\beta^{-1}]^{-1} X_i^T (\sigma_\epsilon^2 \Omega_i)^{-1} (y_i - W_i \alpha),$$

the empirical Bayes estimators for the voxel-level coefficients, and

$$\Sigma_{\beta|y_i} = [X_i^T (\sigma_\epsilon^2 \Omega_i)^{-1} X_i + \Sigma_\beta^{-1}]^{-1},$$

the posterior covariance matrix.

The fixed effect coefficients are estimated from the maximum marginal likelihood as follows:

$$\hat{\Sigma}_\beta = \frac{1}{N} \sum_{i=1}^{N} \tilde{\beta}_i \tilde{\beta}_i^T + \Sigma_{\beta|y_i}$$

gives an estimate of the population covariance,

$$\hat{\alpha} = [\sum_{i=1}^{N} W_i^T \Omega_i^{-1} W_i]^{-1} [\sum_{i=1}^{N} W_i^T \Omega_i^{-1} (y_i - X_i \tilde{\beta}_i)]$$

gives the estimates of the fixed effects, and the scale factor of the error term, σ_ϵ^2, is estimated by

$$\hat{\sigma}_\epsilon^2 = (\sum_{i=1}^{N} n_i)^{-1} \sum_{i=1}^{N} \mathrm{tr}[\Omega_i^{-1} (\hat{\epsilon}_i \hat{\epsilon}_i^T + X_i \Sigma_{\beta|y_i} X_i^T)],$$

where n_i are the number of time points observed on voxel i. Iterative procedures such as EM algorithm or Fisher scoring are needed to obtain final parameter estimates. A further step in the analysis classifies the voxels according to their estimated coefficients, to detect clusters of activation; see Section 10.6. Roy et al. (2005) generalize this work so that the first step, namely time averaging over the trials, is not needed. Instead, the entire time course is fit all at once using cubic splines, with knots in predetermined locations.

Burock and Dale (2000) avoid both trial averaging and specifying a form for the hemodynamic response altogether, preferring instead to estimate the HRF directly as the "parameter" in a linear model. First, they write the observed fMRI signal for a given voxel at time t (where time is considered to be discrete) as

$$y(t) = x_1 h_1(t) + x_2 h_2(t) + \ldots + x_k h_k(t) + \epsilon(t),$$

where x_i is a dummy variable representing the trial type, h_i is the hemodynamic response associated with the ith trial type, and ϵ is noise, normally distributed but with arbitrary covariance matrix (i.e., independence across time is not assumed). Noting that this can be written in the usual matrix formulation, now with the vector h being unknown, maximum likelihood methods can be used to estimate h without having to assume any particular shape for the hemodynamic responses of the different trial types. Since the covariance in the noise term is not known, this needs to be estimated as well, and the authors propose a procedure for modifying the simple ordinary least squares (OLS) estimates to take this into account.

Their method proceeds roughly as follows. First, obtain the ordinary least squares estimate of h, ignoring the temporal correlation,

$$\hat{h}_{\mathrm{OLS}} = (X^T X)^{-1} X^T y$$

and calculate the residuals $e = y - X\hat{h}_{\mathrm{OLS}}$. The residuals are then used to get an estimate of the unknown covariance matrix of ϵ, call it Σ_ϵ. This can either be a "global" estimate in the sense that the covariance matrix is assumed to be the same at every voxel in a slice, or "local" and varying from voxel to voxel. Certain assumptions also need to be made about the noise process; Burock and Dale assume a mixture of white (uncorrelated) Gaussian noise and a correlated (autoregressive) component. A component of the covariance matrix at lag k is modeled as

$$C(k) = \sigma^2[\lambda\delta(k) + (1-\lambda)\rho^{|k|}],$$

for parameters λ and ρ which need to be estimated, and σ^2 estimated using the residuals from the OLS fit above. The estimated covariance matrix $\hat{\Sigma}_\epsilon$ is

$$\hat{\Sigma}_\epsilon = \begin{pmatrix} \hat{C}(0) & \hat{C}(1) & \cdots & \hat{C}(j) \\ \hat{C}(1) & \hat{C}(0) & & \vdots \\ & \ddots & & \\ \vdots & & \hat{C}(0) & \hat{C}(1) \\ \hat{C}(j) & \cdots & \hat{C}(1) & \hat{C}(0) \end{pmatrix}$$

Here, j represents the biggest time lag considered, and $\hat{C}(k)$ is the estimate of $C(k)$ with λ and ρ replaced by their estimates; as mentioned above, these may be either global or local. With this in hand, $\hat{\Sigma}_\epsilon$ can be calculated, and the final estimate of the hemodynamic response is

$$\hat{h} = (X^T \hat{\Sigma}_\epsilon^{-1} X)^{-1} X^T \hat{\Sigma}_\epsilon^{-1} y.$$

A Bayesian extension of this approach is given by Marrelec et al. (2003), who model the response at time n ($1 \le n \le N$) as

$$y_n = h_0 x_n + h_1 x_{n-1} + \ldots + h_K x_{n-K} + \lambda_1 d_{1,n} + \ldots + \lambda_M d_{M,n} + \epsilon_n,$$

where $n = K+1, \ldots, N$, the h elements make up the hemodynamic response that is the target of estimation, the d elements are possible other nuisance effects such as signal drift, and the components of ϵ are assumed to be independent, identically distributed $N(0, \sigma^2)$. As in Burock and Dale, the goal is to estimate h from this linear model. Simple prior information on h is incorporated into the analysis in the form of two postulates:

1. *The hemodynamic response function starts and ends at zero.* This represents the fact that before any stimulus is presented, the subject is in a resting state, so voxel activity is at a baseline level, and that after the stimulus is turned off, in an event-related study, the voxel activity returns to baseline. This part of the prior requires setting the first and last points of the HRF to zero, which reduces by two the number of h parameters that need to be estimated.
2. *The hemodynamic response function is smooth.* This represents the observed behavior that there is a gradual rise from baseline to peak after stimulus presentation, followed by a gradual decay back down to baseline when the stimulus is stopped. There are no sudden jumps or jolts in the response; rather it evolves slowly over time. This part of the prior is expressed by putting a normal distribution on the norm of the second derivative of the HRF.

Interest then focuses on the marginal posterior distribution of h given the data y, which does not have a simple closed form solution. Instead of appealing to standard Bayesian simulation techniques (Markov chain Monte Carlo) to obtain the posterior distribution of h in its entirety, Marrelec et al. only evaluate the posterior mean of h given the data and the estimated value of a tuning parameter on the second (Gaussian) component of their prior specification, since this has a known distribution. Hence they avoid computational complexity, at the cost of not taking full advantage of the strength of the Bayesian paradigm. Although their approach does allow for inference in the form of detecting active voxels, it is likely that a more informative analysis would result if the entire marginal posterior distribution were calculated.

5.3.3 Methods for Estimating the Delay of the Hemodynamic Response

Several authors have addressed the specific problem of estimating the delay in the BOLD response, since the length of that delay may be of intrinsic interest (Saad et al., 2001). As discussed by Saad et al. (2001), variation in the response delay has been observed empirically, both for different voxels and for repeated measurements on the same voxel. The reasons for the variation in delay are not clearly understood; the basis could be physiological (delays in the flow of oxygenated blood through the large veins near the areas of neuronal activity,

heterogeneity in neuronal function or anatomy) or due to other sources, such as fMRI noise.

To explore some of these questions, Saad et al. collected data on a simple visual processing task (flickering rings, paired with a series of easy behavioral tasks designed to focus the attention) for seven subjects. The main foci of the statistical analysis were to detect which voxels were activated by the stimulus, and, for those voxels, to estimate the delay in the response relative to the stimulus presentation. This was accomplished in a single step by calculating the lagged correlation coefficients between the fMRI time course (observed response) and a reference time course (the ideal fMRI response). More specifically, the authors defined the reference time series $r(t)$ as

$$r(t) = x(t) + e_1(t),$$

where $x(t)$ is an "ideal" response (with no delay) and $e_1(t)$ is noise. For the purposes of the analysis, the reference time series was taken to be a sinusoidal wave with frequency that matched the frequency of the stimulus presentation, so that $e_1(t)$ was null and $r(t) = x(t)$.

For the observed time course the authors specified the lagged model

$$f(t) = ax(t - \Delta t) + e_2(t),$$

with a being a scaling factor, and $e_2(t)$ noise. Under the assumptions that $e_1(t)$ and $e_2(t)$ have mean zero, are uncorrelated with each other, and both are uncorrelated with $x(t)$, the correlation coefficient between $r(t)$ and $f(t)$ as defined above at lag γ is

$$r_{r,f}(\gamma) = ar_{x,x}(\gamma - \Delta t),$$

(Saad et al., 2003a, p. 495). Here $r_{x,x}(\gamma)$ is the autocorrelation of $x(t)$, which is maximal for $\gamma = 0$, so that $r_{x,x}(\gamma - \Delta t)$ is maximal when $\gamma = \Delta t$. Thus, one can calculate the lagged correlations, finding the lag for which the correlation is maximized, and thus obtain an estimate for the delay in the hemodynamic response. As a natural by-product of this procedure, the value of the maximal coefficient is also obtained. A voxel was considered activated if the maximal correlation was over 0.5. Computational aspects of this algorithm and further theoretical detail are developed in Saad et al. (2003a).

The active voxels were subjected to further analysis. Voxels were classified as "positive" or "negative," according to whether the measured signal increased or decreased with the onset of the stimulus; the authors then compared the estimated delay in these two types of active voxels. The voxels were also classified according to whether or not they were related to the large blood vessels in the brain, and the estimated delays for the two groups were compared. Finally, the authors investigated heterogeneity in the delay both across and within voxels.

All subjects possessed both positive and negative voxels, with the vast majority being of the former type, in general. Overall, the mean response delay

was found to be about 8.5 seconds; across voxels the variance of response delay was estimated at approximately 4.4 sec^2, with within-voxel variance estimated at 1.44 sec^2. There was no statistically significant difference in the average delay of response between the two voxel types. In a more detailed analysis of a subset of the subjects, the authors found that the spatial distribution of the positive and negative delayed voxels differed, and depended on the configuration of the stimulus (several configurations were tried over the course of the experiment). These voxels were located, as would be expected, in parts of the brain involved in visual processing. The average delays of response differed across the different visual areas, suggesting that in fact disparate brain regions react idiosyncratically to the same stimulus. Onset delays for voxels related to the blood vessels did not differ from onset delays for voxels unrelated to the blood vessels, although it should be noted that because of difficulties in classifying voxels along this criterion, there was much variability across subjects and across classification thresholds. There was some evidence for slightly longer delay (on the order of one or two seconds) for the blood vessel voxels, however, in all cases there was significant overlap in the distributions of the two types.

Estimation of the delay is also taken up by Liao et al. (2002),who note several drawbacks to the approach of Saad and colleagues, primarily the potential for computational complexity and the lack of theoretical standard errors for the estimate of the delay. Liao et al. aim to rectify these potential problems by use of a linearized model. They start with a reference HRF, such as the gamma model, or the difference of two gammas, described previously, denoted $h_0(t)$. The shifted model from the base is then described by a location family, that is $h(\gamma, t) = h_0(t - \gamma)$, where the lag γ is again permitted to vary. This function is approximated using as basis functions the first two components of the singular value decomposition of the matrix H defined by evaluating $h(\gamma, t)$ as both γ and t are varied along a regular grid of values. The result is a linearized model for the fMRI response at time i:

$$Y_i \approx x_0(t_i)w_0(\gamma)\beta + x_1(t_i)w_1(\gamma)\beta + \epsilon_i;$$

see Liao et al. (2002) for details.

The method of moments approach is used to get estimators for γ and β. Since everything is based on a linear model, the parameters are easily estimated, without need to resort to iterative or other time-consuming procedures. The authors also give an expression for the standard error of the estimated delay, making it feasible to test proposed values for the amount of delay using a t test, and also to compare the lags at two different voxels. Henson et al. (2002)propose a similar method, using Taylor series basis functions instead of the basis functions used here. As noted by Liao et al., the practical differences between the two, as evidenced by simulations, appear to be minimal.

Looking at a number of different reference functions and several variations on their basic methods for a tactile stimulation task the authors found the length of the delay to be approximately 6 seconds, compared to the canonical

value of 5.4 seconds in the difference of gamma model, and the estimated value of 8.5 seconds in the study described by Saad and colleagues. Furthermore, the amount of the delay did not vary considerably across the brain, again in contrast with the findings of Saad et al.

A much different idea is advocated by Calhoun et al. (2000). They model the signal in a voxel as

$$y = X_{\Delta t}\beta + \epsilon,$$

where Δt represents the delay in the start of the response following the stimulation, and the errors are assumed to be independent, identically distributed with mean zero. The main innovation of the method proposed by Calhoun and colleagues is to use a weighting function w to modify the errors, so that different amounts of importance can be assigned to various parts of the estimation problem. Where we are interested in focusing on the delay, or latency, of the response, for instance, we would assign higher weight to the beginning of the time course. The problem then becomes one of weighted least squares and it is straightforward to obtain estimates of the β parameters for a fixed value of Δt. Since the delay is itself unknown, in practice this calculation needs to be repeated for each Δt on a grid of values within a "reasonable" range (say 3-10 seconds following stimulus onset), and the choice of optimal delay is essentially a question of model comparison. As with some of the other methods discussed in this section, the activation status of voxels is determined by imposing a criterion on the value of the estimated latency.

The efficacy of this approach was tested on a simple visual motor task; subjects were presented with a flashing checkerboard and had to press a button upon seeing the pattern. Delays in the hemodynamic response were found to differ in the visual and motor areas, with longer delays occurring in the motor cortex. Compared to a least squares fit without weighting, estimates of the delay with weighting decreased in the visual areas, and increased in the motor areas. These results need to be interpreted with caution, since they are based on only a single subject.

5.4 The General Linear Model

Much of what has been described in the previous two sections can be subsumed under, or used in conjunction with, the *general linear model*, which is at the foundation of most traditional statistical analysis of fMRI data. Indeed, some of the methods from the previous section (Burock and Dale, 2000; Gibbons et al., 2004) utilize the general linear model as a means to an end, for instance, estimating the shape of the HRF for an event-related study. More prevalent is to have the focus of the analysis be the output of the general linear model itself. Due to its ubiquity in the fMRI literature, it behooves us to study this model more closely.

The form of the general linear model is of course familiar to statisticians, written in matrix form as:

$$Y = X\beta + \epsilon,$$

where Y is the response, X the matrix of predictors, β the unknown coefficients of the predictors, and ϵ the error, usually assumed to be normally distributed with mean zero and variance σ^2. In the context of fMRI, Y will generally be a matrix representing the time courses of all the voxels (hence, it will be of dimension t rows and v columns, say, one column for each voxel, and one row for each time point), X will be a design matrix reflecting the stimuli presented at each point in time, and ϵ may have constant or nonconstant variance, as well as nonzero covariance terms. The design matrix is often presented graphically. In the most basic expression of the general linear model, each voxel and each time point is assumed to be independent of the others, and σ^2 is assumed to be constant, so estimates of β can be obtained via ordinary least squares. With this formulation the general linear model subsumes the t test and the correlation analysis described in Section 5.2. More realistic assumptions allow for spatial and temporal correlation, a topic and class of models that we take up in Chapter 6.

The design matrix can, and usually will, incorporate a variety of different types of covariates of interest. First are factors that describe the experimental design, which are simple binary variables for the elementary block design, and more complicated categorical variables for more extended block designs (multiple tasks) and event-related designs. Usually the X matrix will also include predicted hemodynamic responses, which are obtained by convolving the stimulus time course with a model for the HRF (typically a simple gamma or Poisson model); this convolution takes into consideration the fact that the BOLD response does not start immediately upon presentation of the stimulus, and also brings in to the analysis other pertinent aspects of the hemodynamic response, such as the undershoot before return to baseline. These covariates are defined as

$$x(t) = \int_0^\infty h(u)s(t-u)du,$$

where $h(\cdot)$ is the model for the HRF and $s(t-\cdot)$ the stimulus time series. The value of the covariate at the ith scan is $x(t_i)$, with t_i the time of the image acquisition. See Figure 5.4. Note that the convolution has the effect of "smearing" the time course of the stimulus presentation, so that transitions from baseline to task (in this simple two-condition example) are smooth, rather than abrupt.

Finally, the model is also flexible enough to account for other categorical covariates besides those related to the design, such as subject demographics, group membership, and so forth. In this way, the general linear model provides a framework not only for the analysis of individual subject data, which has been our focus up to this point, but also more generally of data of groups of subjects, including comparisons across groups.

The model in its most fundamental form makes many of the same assumptions as are made for the t test analysis. This is not surprising, since the

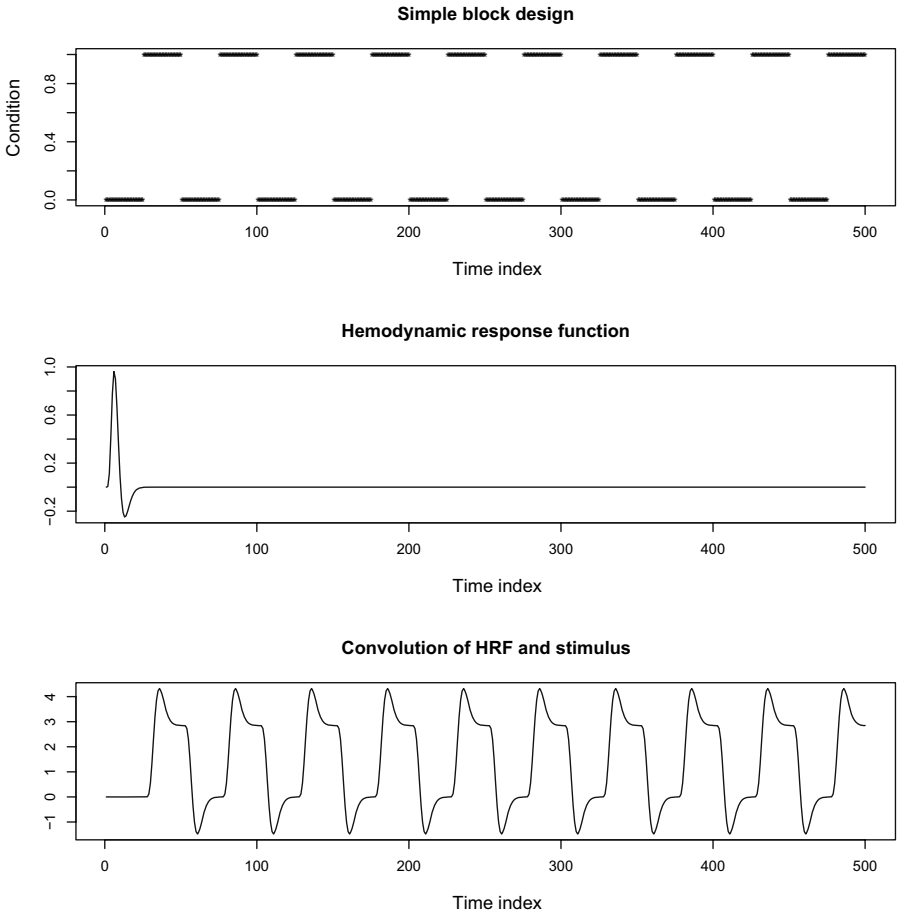


Fig. 5.4. The top panel shows the stimulus time course of a simple, two condition block design, with 0 representing the baseline and 1 the task. The middle panel gives the hemodynamic response function modeled as the difference between two gammas, using the parameters from Glover (1999), as described in Section 5.3.1. The bottom panel presents the convolution of the HRF with the stimulus time course; this is used as an independent variable in the general linear model.

general linear model is an extension of the simpler analysis to allow for the inclusion of additional covariates. In particular, when using the basic general linear model makes these assumptions:

1. *Voxels are independent.* The model makes no use of spatial information, whereas it is reasonable to assume that neighboring voxels will, to some extent at least, have similar behaviors. Hence, fMRI data analysis tends to be univariate in nature, even though the data themselves are massively

multivariate. However, implementing a truly multivariate linear model is generally infeasible, since the number of voxels is always much larger than the length of the time course, leading to problems of identifiability and parameter estimation.

2. *Time points are independent.* In practice this is a completely unrealistic assumption, as has already been discussed. While it is not clear how much of an effect the response at some time t_1 has on subsequent responses, nor how much the response at time t_1 is affected by past responses, such effects are surely present.
3. *The error variance at each time point (experimental condition) is the same.* This assumption precludes the possibility that some conditions may have more residual noise than others, which may or may not be realistic.
4. *The same model, as given by the design matrix, is appropriate for every voxel in the brain.* While the β estimates will differ from location to location, since they are derived independently at each voxel, the assumption is that one model fits everywhere. On the one hand this is a pragmatic stance to take; given the large number of voxels and the potential complexity of the models involved, it is easier to fit a single model to the whole brain. Interpretation might also be easier. On the other hand, from physiological and other considerations, one might find it more believable to fit different models at different locations, or at least to allow for such via a process of model selection. We take this issue up further in Section 11.3.

Since these common assumptions are almost surely unrealistic and hence violated in practice, much of the statistical research in fMRI has centered on ways of improving and extending the general linear model, or it has focused on alternate analysis paths. We explore these ideas in subsequent chapters. In the rest of this section, we consider in more detail three issues that are relevant to the immediate implementation of the basic general linear model: modifying the predictor variables to improve inference; fixed, random or mixed effect models; and analysis in "real time".

5.4.1 Some Implementation Issues

The basic model includes terms representing the hemodynamic response, specifically a convolution of the modeled HRF with the stimulus time course(s). This convolution is subsampled at the slice acquisition times to create the actual predictor as used in the general linear model. Delays between the real data and the HRF model are not uncommon; Friston, (Friston et al., 1998) for example, has suggested incorporating the first temporal derivative of the convolved HRF into the model as an additional predictor in order to account for these potential mismatches. Motivating the inclusion of the first derivative is the idea that this will allow for different response latencies and for misspecification of the stimulus onset relative to when the data were acquired in

practice (since the two do not occur simultaneously). Figure 1 in Friston et al. (1998) shows that the hemodynamic response model and its derivative exhibit different time and amplitude of peak, as well as different time and amplitude of poststimulus dip, thereby admitting a wider range of possible behaviors at the voxel level.

The model for a single voxel at time i therefore becomes

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \frac{\partial x_i}{\partial i} + \epsilon_i.$$

The effect of interest is still β_1, which is tested using the usual t test.

Calhoun et al. (2004) take this suggestion one step further and provide a rigorous basis for testing the effect of β_1 while directly accounting for the effect of the temporal derivative as well. The test they propose is

$$\frac{\text{sgn}(\hat{\beta}_1)\sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}}{e} \geq \tau,$$

where e is the residual error from fitting the model above, and τ is the critical value of the new t test. This test is derived under the very weak, and easily imposed, conditions that the vectors x and $\partial x/\partial i$ are orthonormal. As the authors show, it is also easy to extend their method to include additional derivatives, which in turn help capture additional spatially-varying features of the data. When there is a delay in the onset of the hemodynamic response, as is typically the case, the test that includes both terms outperforms the usual t test, and the corresponding model fits significantly better than a model that includes only the convolved HRF. As the delay increases (in simulated data), the ability of the nonderivative term to capture important features in the data decreases; specifically, the estimate of the amplitude of the response is dampened when only the first term is included, resulting in severe lack of fit. When the derivative term is included as well, the estimate of the effect amplitude is very close to the true value.

Including additional covariates beyond those having to do with the experimental design itself allows researchers to investigate the differences across multiple runs (scans) performed on a single subject over the course of a scanning session, or across multiple sessions. Furthermore, by adding indicators for group membership, one can also examine differences between experimental groups (men versus women, healthy controls versus clinical populations, and so forth). When such terms are included in the model as predictors, a natural question is whether they should be considered as fixed or random effects. This question has generated much discussion in the fMRI literature and seems particularly relevant for assessing the outcome of multisession and multiple subject studies. In the former we encounter again the question of reliability across scanning encounters (see Section 4.3.2); in the latter, we are concerned with inference from a given group of subjects to the more general population. In both cases there is a distinction that is made between within-session

(subject) effects and between-session (subject) effects; how these are treated is critical to the statistical inference as it is implemented and interpreted. In fMRI it has become standard to treat certain effects, such as sessions or subjects, as random (Friston et al., 2005), while others, such as the effect of the experimental manipulation, are treated as fixed, hence the general model is what would usually be considered a *mixed effect model* by statisticians.

The rationale for treating sessions and subjects (for example) as random is clear from the neuroimaging perspective, since this allows researchers to generalize their conclusions beyond the particular subjects that participated in an experiment to the broader population. Still, as we shall see in Section 5.5.2, the statistical question of how to treat subjects in the linear model is not completely resolved. Treating them as random effects results in a conservative testing procedure, in which patterns of activation need to be consistent across subjects, a rather stringent requirement for many fMRI experiments. The exploratory nature of many neuroimaging studies argues in favor of the more limited, but statistically more "generous," fixed effect approach to subjects. This question is discussed further in the next section. Also, see Friston et al. (2005) for more on the use of the mixed effect paradigm in fMRI, and Friston et al. (1999a) on the choice between treating subjects as fixed or random in the general linear model.

The procedures for estimating the parameters in the general linear model can be cumbersome, particularly for the mixed effect model where there are also multiple components of variance to estimate; the latter is usually accomplished by standard statistical approaches for such problems, including restricted maximum likelihood (REML), or iterative algorithms such as EM. Depending on the complexity of the model and the number of subjects, these calculations can be very time consuming. Yet there is continued (although sporadic) interest in the fMRI community for real-time analysis of the data as it is acquired in the scanner. One proposal for such an analysis is given by Bagarinao et al. (2003). As the authors note, an advantage of real-time analysis is that the quality of the data can be assessed as it is being collected, and changes can be made to the experimental paradigm "on the fly," if needed.

In the first step of the algorithm, the original linear model is transformed via a Gram-Schmidt orthogonalization, so that the response at voxel i at time j is given by

$$y_{ij} = \alpha_{i1}\phi_1(t_j) + \ldots + \alpha_{kp}\phi_p(t_j) + \epsilon_{ij},$$

where the ϕ terms are the orthogonalized versions of the original predictors. We are then interested in estimating the α coefficients instead of the βs from the original model. Bagarinao and colleagues demonstrate that the components required to obtain the estimates of the α (and hence the β) parameters at time j can be calculated based on those from time $j-1$ plus known quantities related to the data collected at time j. Hence it is possible in principle to update the parameter estimates and test their significance at each voxel after every volume acquisition (scan of the complete brain or whatever portion

thereof is of interest). Since the method updates the parameter estimates and the test statistics at every time point, one natural output is a "trace plot" of these quantities, which allows for monitoring over the course of the experiment; the statistical parametric map of activation can also be plotted at each time point, so that the effect of acquiring additional data can be assessed.

5.5 Methods for Combining Subjects

As a final topic before moving on to more advanced statistical analysis methods, we include a discussion of combining subjects from an fMRI study. Although notionally the topic itself is rather sophisticated, it is included here with the basic fMRI data analysis because the methods that are introduced in this section rely on the output of a basic analysis such as a t test, or are arrived at as part of a general linear model analysis.

Section 4.4 briefly outlined the two fundamental questions that arise when a researcher wishes to combine the statistical maps of multiple subjects to create a summary, or group, map. The first question involves finding a common space on which to consider the maps. The reason for this is that each and every brain is different, in size and shape at the very least, and possibly in other aspects as well. Hence, if the images of different subjects are to be combined into a single group map, account must first be made of these individual variations in anatomy. Although this is not a statistical question as such, the problem is key to much of the statistical endeavor in fMRI research, and so is given a brief survey in this chapter.

The second question can truly only be approached after some solution has been found to the first. This question involves finding a statistically optimal way of combining the data from the subjects in the study. What precisely is meant by "statistically optimal" is open to consideration, but given the nature of the data some possible desiderata include: computational efficiency and speed; good small sample properties; effective use of the data; and robustness.

5.5.1 The Anatomical Question

As already described, any effort to create a group map that summarizes subject activation, be it in the form of common patterns of activity or of an overall level that is characteristic of the group as a whole, requires a "warping" transformation to bring all of the brain images onto a common frame of reference. Such operations are often used in image processing prior to, or together with, statistical analysis. In fMRI the normalization step is almost always carried out as a separate preprocessing step, and the transformation has most often been to the standard *Talairach* coordinate system (Talairach and Tournoux, 1988). The Talairach and Tournoux system is a very detailed atlas of a single human brain that was obtained post-mortem; semi-automated software,

such as in the AFNI fMRI package (see Appendix A), allows users to easily transform images obtained from individual subjects into this coordinate space. Human intervention is required as a prelude to the transformation, since certain landmarks in the brain must be identified by hand and their coordinates (in original image space) fed into the program. Often, the warping transformation is carried out using simple linear interpolations, although more sophisticated schemes have also been developed, ones involving various forms of nonlinear interpolation.

When the data are shifted into "Talairach space," voxels are also resized, generally to 1 mm along each side. If a typical voxel is 3 mm along each side, then 1 "original space" voxel is transformed into 27 Talairach space voxels, thereby introducing extra correlation and potentially also blurring of the signal, depending on the warping algorithm. Still, in spite of these problems, the Talairach system, and more generally the warping of each individual brain onto a common atlas (of which several standards exist, beyond that of Talairach and Tournoux), remains one of the more popular approaches for handling the anatomical aspect of the group map question.

Another popular atlas is the Montreal Neurological Institute, or MNI, brain (Evans et al., 1993). By contrast with the Talairach system, the MNI atlas is an average based on a large number of living brains, over 300 MRI scans of healthy individuals. A template based on the MNI atlas has been adopted by the International Consortium of Brain Mapping (ICBM) as its standard (Chau and McIntosh, 2005). The ICBM has also developed a probabilistic atlas (both the Talairach and the MNI systems are deterministic) to account for individual differences (Mazziotta et al., 1995). However, since the Talairach system remains the standard for reporting in the literature, even when group maps or other analyses are calculated in these other coordinate systems, it is often necessary to transform the transformed data into Talairach space. This introduces further uncertainty into the results that appear in the neuroimaging literature, since the same point in MNI space for two different subjects might correspond to different points in Talairach space, and vice versa (Chau and McIntosh, 2005). It thus seems critical to establish a single standard that could be used by all neuroimaging scientists for research purposes. The development of a standard atlas, as well as the assessment of existing systems and the methods that are used to transform an individual brain image into the template space, continue to be topics of active research in the community.

A comparison of different warping procedures is given by Crivello et al. (2002), in which all brains are warped to a common brain (itself shifted into Talairach coordinates), but this is accomplished using a variety of methods: a 12-parameter linear (affine – involving rotation, translation, and rescaling operations) transformation; a fifth order nonlinear polynomial algorithm involving 168 parameters; a combined "linear and nonlinear" procedure; and a multigrid technique based on continuum mechanics.

The standard affine transformation is applied as a first warping step for all of the other methods. In the polynomial algorithm the initial warping is refined by the use of the fifth order nonlinear model, whereas in the "linear and nonlinear" combined method, it is improved by the use of discrete cosine basis functions. This is a very highly parameterized approach, as is the multigrid technique (requiring thousands and millions of parameters, respectively).

The authors consider the effect of using each type of warping procedure on both the structural (anatomical) and functional images. For the structural images they look at such criteria as how the different tissues in the brain are segmented (white matter, gray matter, cerebrospinal fluid); for the functional images, extent and patterns of activation following statistical analysis are relevant. In terms of the anatomical criteria, Crivello et al. find no differences among the four procedures for the percentage of voxels classified as gray matter after warping of the images; however the techniques do differ in the volumes of the white matter and cerebrospinal fluid. The multigrid fluid mechanics approach tends to give volume estimates of all tissues that are closest to the template brain used in this study. A more detailed examination of the voxels in each tissue type reveals that the linear transformation is outperformed by the three nonlinear methods. This is not at all surprising, given that the nonlinear methods are both more complex and much more heavily parameterized.

Turning to the effect of the spatial normalization procedures on the functional images, at low resolution (highly smoothed data) all three nonlinear methods result in slightly enhanced activation regions compared to the affine transformation, but in general the four methods produce similar activation maps. At higher resolution (less smoothing of the data) this changes, with the activation maps obtained after each warping procedure showing very little overlap, indicating that each method results in the detection of different areas and patterns of activation.

Although their analysis is not conclusive – the authors do not make any strong recommendation for one type of warping method over another – it does seem apparent that the advantages to be gained by the use of simple, moderately parameterized nonlinear methods can make them worthwhile when compared to linear warping with an extremely small number of parameters. The complexity of the brain and the differences among human brains are not likely to be captured with only affine transformations, but it is likewise doubtful that a normalization that uses millions of parameters and requires several hours to implement (per subject) will be practical in many real data analysis situations.

As pointed out by Kochunov et al. (1999), normalization by the use of landmarks and (affine) transformations are what might be called *global*. That is, the brain as a whole is described by a small set of parameters (landmarks), which need to be selected in such a way that simple operations such as scaling and rotating can align the image of interest with the target, or template image. Another possibility is the use of *regional* normalization, which matches

more localized features at a much more refined level of resolution, often using image intensity instead of landmarks. Such an approach is evidently computationally intensive, using heavily parameterized systems as we saw above; hence Kochunov et al. propose a method for regional normalization that is both quick (on the order of minutes instead of hours to align a single image) and accurate (comparable to the more expensive regional procedures), which they call *octree spatial normalization* (OSN).

OSN proceeds by successively splitting the volume of interest into octants; thus, the whole image is first split into octants, then each of the first level octants is further subdivided into octants, and so on. In each octant, at each level, a similarity measure (such as a correlation, or sum of squared differences) is calculated between the image of interest and the target, and the octant of interest is then suitably deformed to match the target octant as closely as possible (maximize correlation, or minimize sum of squared differences). Thus this procedure, unlike methods that warp to Talairach coordinates or some other template brain, is not landmark based. Any empty octants (those consisting entirely of air voxels) are removed from the processing stream; since brain voxels make up only about half of a typical image, this step alone can result in considerable computational savings. The authors also suggest other ways of optimizing computation.

Note that since at each step the volume is split into 8 equal cubes, the x, y, and z dimensions must all be the same size (same number of voxels), which is not typical for fMRI experiments, where x and y might be 64 each, and z only 20. Also, the length of each dimension must be 2^n, for n a positive integer; n will also give the number of processing steps in the algorithm. These restrictions are potentially limiting to the use of OSN, since they don't correspond to the way that the data are obtained in practice. One solution might be to "pad" the images, so that a volume that was originally $64 \times 64 \times 20$ would be padded with zeros in the z dimension; since empty octants are removed, this operation presumably would have little effect on the algorithm, although this point isn't addressed in Kochunov et al. (1999). Further modifications of the basic technique, which make it more applicable to brain images, are reported in Kochunov et al. (2000). Even with these modifications, some of which increase processing time, the OSN method is considerably faster than other heavily parameterized normalization regimes and achieves comparable accuracy. Furthermore, as would be expected, regional normalization outperforms its global counterpart; indeed, while apparently not strictly necessary, the latter is often taken as a preprocessing step for OSN and other region-based methods (see the examples in Kochunov et al. 2000).

The observation that registration techniques can be based on landmarks or on image intensity, pointed out above, forms the motivation for the work of Magnotta et al. (2003), who consider combining different types of information to create a more reliable atlas and normalization procedure. Specifically, they propose a method that uses anatomical landmarks, brain segmentation (of tissue or of regions), and image intensity information to minimize the distance

between a given image and the target image. First, a set of landmarks is manually identified, as an initialization for the general algorithm. This gives a global normalization in the sense discussed previously. Following this initial normalization, the algorithm proceeds iteratively to refine the match between the two images, using segmentation (into different tissue types and different brain regions) and intensity information. Iterations of the algorithm update transformed images of interest to the target, as well as transformed targets to the image of interest.

This algorithm is compared to a rigid normalization, which uses only a limited number of landmarks together with affine transformations (rotation and translation); to the piecewise linear normalization that characterizes the warping into Talairach space; and to a simpler version of the current procedure, which uses only intensity information to minimize the distance between two images (still utilizing both "forward" and "backward" transformations, as in the full algorithm). Not surprisingly, the two methods that use only landmark information, namely rigid and Talairach normalization, do not perform as well as the other two methods. For both of these the average amount of overlap with the target is reduced and the variability of overlap across subjects is greater when we look at specific regions or at the whole brain. The introduction of intensity information results in a significant improvement in most (but, interestingly, not all) regions: average overlap increases and the variability decreases. Finally, the strongest results are obtained with the full algorithm, which combines landmark and intensity information, although it should be noted that the improvement is not consistent; for some brain regions the difference between the two implementations is minimal (though still statistically significant), as it is also for the entire brain.

Readers who are interested in learning more about the issues involved in creating template systems and in translating an individual brain image to a target image are referred to the edited volume by Toga (1998).

5.5.2 The Statistical Question

Once we have normalized the individual subject images methods using any of the techniques from Section 5.5.1, we can consider the more purely statistical aspect of the problem, namely, creation of the group map from the subject maps. This is, in essence, a question of *combining information from independent sources*, where in our context each subject acts as an "independent source." Seen in this light, the statistical question is an old one, dating back to at least the 1930s and work by Tippett and Fisher. Ideas from the meta-analysis literature are also relevant, with subjects taking the place of published studies of a phenomenon of interest.

We take as a starting point work by Lazar et al. (2002), who carry out a survey of combining techniques from the statistics and psychology literatures and apply a selection of the multitude of available procedures to the formation of fMRI group maps. Following Hedges (1992), the authors distinguish

between two types of combining procedures: *combining tests* and *combined estimation*. The first type comprises methods that are based on individual test statistics, which in the fMRI setting are the individual statistical parametric maps obtained for each subject; for example, the output of the simple t analysis from Section 5.2. For these methods a group map is created by calculating, at each voxel, some function of the test statistic at that voxel over subjects (hence the need for a normalization procedure, so that voxels in the same location for different subjects will have an equivalent interpretation). Since the functions that are calculated are often based on the p-values of the test statistics rather than the test statistics themselves, these methods are also called *p-value techniques*.

Many p-value techniques have appeared in the statistics literature; among the most popular is one due to Fisher (1950),

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

where there are k independent tests of the particular null hypothesis in question, and p_i is the p-value associated with the ith test. In our context, the "k independent tests" are k subjects in an fMRI experiment, and the T_F statistic is calculated at each voxel independently. Fisher demonstrates that T_F is distributed as a χ^2 with $2k$ degrees of freedom under the null hypothesis of no effect (no activation in imaging studies). Hence, large values of T_F, when calibrated against the appropriate χ^2 distribution, lead to rejection of the null hypothesis.

Another early suggestion is found in Tippett (1931), namely to find at each voxel $T_T = \min_{1\le i\le k} p_i$, the minimum p-value over the k subjects. The null hypothesis is rejected at a given voxel if $T_T < 1-(1-\alpha)^{1/k}$ for a test of level α. Generalizing Tippett's idea, Wilkinson (1951) proposes looking at the rth smallest p-value and rejecting the null if this is smaller than a constant that depends on k, r, and α. Also in the order statistic family is the conjunction test of Worsley and Friston (2000), $T_W = \max_{1\le i\le k} p_i$, which rejects H_0 if T_W is less than $\alpha^{1/k}$. Note that this test requires the p-values at a given voxel be small for *every* subject in order to reject the null.

Although many other p-value methods are available (see Lancaster 1961, for a comprehensive review and Lazar et al. 2002, for specific suggestions in the fMRI setting), we mention here only one other, simple ad hoc procedure that is commonly used in the neuroimaging community, namely averaging the t statistics computed for the individual subjects,

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

where T_i is the value of the t statistic calculated for subject i at a particular voxel. The null hypothesis is rejected for large values of T_A as compared to percentiles of the standard normal distribution.

These statistics are all easily calculated, even over many hundred of thousands of voxels and with a large number of subjects, a definite advantage when dealing with neuroimaging data. On the other hand, p-value methods may be unduly affected by the outcome of a single study; specifically, H_0 at a given voxel may be rejected, or alternately fail to be rejected, based on the activation map of a sole, aberrant subject. See McNamee and Lazar (2004) for a detailed discussion of this point.

The second approach to making group maps, combined estimation, refers to procedures that come directly from the meta-analysis literature. As such, they are model-based, specifically relying on the linear model that is already prevalent in the analysis of fMRI data, and have been proposed in one form or another by several authors. Two models are relevant here, the *fixed effect* model and the *random effect* model. For a theoretical discussion of these models in the setting of group maps for fMRI see, as a recent comprehensive example, Beckmann et al. (2003).

When the various studies are homogeneous in design, so that one may reasonably assume that they are measuring the same phenomenon, the fixed effect model, which specifies

$$y_i = \theta + \epsilon_i,$$

where y_i is the *effect* observed in the ith study, θ is the common mean, and ϵ_i is the error. Here, the observed effects in all studies (subjects) are assumed to vary around the common mean θ, with ϵ terms that are usually taken to be independent, $\epsilon_i \sim N(0, V_i)$. The normality assumption is probably not warranted; see, for example, Thirion et al. (2007).

Defining the weight w_i to be inversely proportional to the variance in the ith study, θ is estimated by

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

$\hat{\theta}$ is unbiased for θ, has estimated variance $1/\sum w_i$, and is approximately normally distributed. Therefore, a test for the hypothesis that θ, the common mean, is zero, is given by

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

the null hypothesis is rejected for large values of T_X. As with the p-value methods, T_X is calculated at each voxel separately.

Note that this combining statistic is based on the effect y_i observed at a given voxel for subject i, and not on the value of the t or F (or other) statistic from the parametric brain map. For a simple block design with two conditions, y_i would be the difference in the average response. Note too that the fixed effect model constrains generalizability of results to the sample of observed subjects.

Often, however, researchers are interested in extrapolating beyond the particular sample of subjects to a broader population from which those subjects were drawn. More generally one may wish to pool the information available from heterogeneous studies of the same hypothesis. Even when studies are homogeneous, as required for the fixed effect model, intersubject variability is often much greater than intrasubject variability, and it is desirable to take account of both components of variance. For all of these reasons, the random effect model, which takes into consideration these various objectives, is preferred in many cases to the fixed effect model. Indeed, as we have seen above, it is one of the principal models traditionally used for the analysis of fMRI data.

The random effect model in the meta-analysis context on which we draw has the form

$$y_i = \theta_i + \epsilon_i$$
$$\theta_i = \theta + e_i,$$

where usually we assume $\epsilon_i \sim N(0, V_i)$, $e_i \sim N(0, \sigma_\theta^2)$ and all the e_i, ϵ_i are independent of each other. Each subject has a unique mean effect θ_i, and those θ_i are in turn drawn from a hyperdistribution with mean θ. The observed subject effect thus has two sources of variability: of the subject effect around its mean, and of the effect mean around the overall common mean.

As with the fixed effect model, we can define an estimator of the overall common mean θ as a weighted average of the y_i:

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

where $w_i^* = 1/(s_i^2 + \hat{\sigma}_\theta^2)$, with s_i^2 and $\hat{\sigma}_\theta^2$ estimating the two components of variance. The first one can be obtained simply from the data available on the ith study (subject), but estimating σ_θ^2 is harder, often requiring computationally intensive methods which make the random effect model time-consuming and difficult to implement on fMRI data, where the test statistic has to be calculated separately at each voxel. Some of the computational aspects of this model for fMRI are addressed by Beckmann et al. (2003) and Friston et al. (2005), who explore the use of a multistage approximation to the full analysis.

Once the variance components are estimated, it is then straightforward to define a test statistic similar to the one given above for the fixed effect model,

$$T_R = \frac{\hat{\theta}^*}{\sqrt{1/\sum w_i^*}},$$

rejecting the null hypothesis of no activation at a particular voxel for large values of T_R.

Note that the standard errors of the fixed effect estimate $\hat{\theta}$ tend to be smaller than those for the random effect estimate $\hat{\theta}^*$, since the latter takes

into account variability across subjects, which is ignored by the former. As a consequence, the random effect model requires much more computational effort to fit, compared to the fixed effect model; however, the ability to generalize beyond the sample to the population at large is considered to be of greater importance, and hence the random effect model is almost always preferred in fMRI data analysis. When σ_θ^2 is zero, the two models coincide, so the fixed effect model may be thought of as an "optimistic scenario" relative to the random effect model. On the other hand, there is some evidence (Lazar et al., 2002) that the random effect model, in particular in fMRI studies with a small number of subjects, is too conservative, making it very difficult to detect any interesting patterns of activation. Recent work (Thirion et al., 2007) indicates that a sample of 25 to 30 subjects may sufficiently alleviate the lack of sensitivity of the random effect model.

Many other statistical questions besides that of how to pool the available information from the different subjects also arise. In the rest of this section, we survey some of them.

An issue that comes up in much statistical analysis of fMRI data is that of spatial smoothing, namely, "Should one smooth, and if so, how much?" White et al. (2001) address this question in the context of creating group maps. As noted by those authors, when smoothing the data prior to combining the maps from multiple subjects, "... the goal is to utilize a filter that spreads the activation regions sufficiently to allow for an overlap of homologous activation sites during coregistration of group images, while preserving spatial resolution" (p. 579). The idea is therefore to smooth to help alleviate the anatomical problem of differential brain configuration, which may persist even after warping onto a common atlas. On the other hand, one mustn't smooth so much that all fine detail is lost. White et al. (2001) explore the effectiveness of applying different sizes of a Hanning filter (different amounts of spatial smoothing) prior to combining the individual subject maps. Their experiment is carried out for a simple block design finger tapping scenario, with six test subjects.

As might be expected, White et al. (2001) find different results depending on the size of the filter used on the individual images. When no smoothing is performed, the group map picks up little or no activation; the activation that is found tends to consist of one or two isolated regions comprising a small number of voxels. This may be misleading to the extent that the isolated sites are really part of the same functional region. Introducing even a small amount of smoothing has several effects: first, the discrete regions merge; second, activation on the dominant side of the brain is mirrored by activation in the other hemisphere (something that is not detected when no smoothing is applied); third, the intensity of the activation in the discovered region increases. When the filter size is increased to have a full width half maximum of greater than approximately 8 mm, anatomically distant and functionally distinct sites also begin to merge, highlighting some of the possible dangers of oversmoothing in this setting.

While the choice of filter size to be used in the preprocessing step of a group analysis is thus not an easy one, it is worth noting that much of the difficulty can be bypassed by careful choice of combination technique. The work described in Lazar et al. (2002), for instance, demonstrates a wide variety of regional contiguity across combining methods when applied to the same data. Where a random effect meta-analysis type model discovers only a few isolated active voxels in the group map, as one example, a map created using Fisher's method and the T_F statistic described above finds functionally meaningful, and connected, regions of activation, without the need for presmoothing.

In creating a group map, no matter what the method utilized, an implicit assumption is that the subjects are homogeneous enough that combining them is warranted. Both the fixed and random effect models, for example, suppose a common underlying mean from which all sample images are drawn. Although this assumption is not as obvious, nor seemingly as critical, for some of the p-value methods, it still isn't negligible. It has been noted in the statistical literature that T_F is sensitive to "outlying" studies (in particular, to large studies with an effect in the opposite direction from the effects in the rest of the studies), and there, too, a presumption of "sufficient homogeneity" should improve performance. Kherif et al. (2003) propose a method to assess the homogeneity assumption based on distance measures.

The authors start with calculating the *RV coefficient* (Robert and Escoufier, 1976) between two matrices Y_1 of dimension $p \times n$ and Y_2 of dimension $q \times n$, corresponding to the data observed on two subjects. Thus there are n observations on each subject, on p variables for the first and q for the second. Next, after mean correcting the raw data, summary matrices $Z_{11} = Y_1^T Y_1$ and $Z_{22} = Y_2^T Y_2$ are calculated, and the RV coefficient defined as

$$RV(Y_1, Y_2) = \frac{\text{tr}(Z_{11} Z_{22})}{\sqrt{\text{tr}(Z_{11} Z_{11})}\sqrt{\text{tr}(Z_{22} Z_{22})}},$$

a multivariate extension of the usual correlation coefficient. When the matrices for the two subjects are similar (linearly related), the value of the RV coefficient will be close to 1; when the matrices differ greatly, the coefficient will be close to zero. Using both $RV(Y_1, Y_2)$ and the similarity matrices Z_{11} and Z_{22}, various distance measures can also be defined.

When $p = q$, as will typically be the case for fMRI studies (we will have both the same number of time points and the same number of voxels for each subject, after warping to a common coordinate space), it is also possible to consider $RV(Y_1^T, Y_2^T)$. Kherif et al. (2003) note that $RV(Y_1, Y_2)$ can be used to assess *spatial similarity* between two subjects, while $RV(Y_1^T, Y_2^T)$ can be used to assess *temporal similarity*. In the first instance, the focus is on the similarity of the sites of activation in the brain, regardless of the temporal evolution of the hemodynamic response; in the second instance, the focus is on similarity in the fMRI time courses, regardless of the spatial distribution of activity.

The authors also extend the RV coefficient through allowing the use of covariates by embedding the problem in a general linear model framework. If X represents the matrix of regressors, the extension replaces Y_1 and Y_2 with $X^T Y_1$ and $X^T Y_2$, respectively, and then proceeds as previously. As a final modification they also introduce two corrections, one each for the spatial and temporal correlations inherent in the fMRI data. They then calculate the RV coefficient on the transformed data, and use the distance metric

$$D(Y_1^*, Y_2^*) = || \frac{Z_{11}^*}{\sqrt{\text{tr}(Z_{11}^{2*})}} - \frac{Z_{22}^*}{\sqrt{\text{tr}(Z_{22}^{2*})}} ||,$$

where Z_{ii}^* is the summary matrix for the matrix Y_i^*, which is Y_i after accounting for possible covariates of interest and correcting for temporal and spatial correlations in the data at the individual subject level. The distance measure is calculated for both the spatial and the temporal modes of analysis, as described in the previous paragraph.

Combined with techniques from multidimensional scaling and outlier detection in regression, a variety of graphical and numerical methods can be implemented to explore the similarities among the subjects in a study, to find unusual or possibly outlying individuals, and so forth. Kherif and colleagues demonstrate these procedures on a study of a mental calculation task performed by a group of 10 subjects. They are able to identify a small number of subjects whose temporal or spatial patterns of behavior are different from those of the rest of the sample. For instance, whereas most of the subjects in the study show task-related activation on both sides of the brain, some of the subjects exhibit this only on one side; these subjects are detected as being "far from the rest" in the spatial distribution of activity. Group maps created using a random effect model differ qualitatively when based on the entire sample versus discarding subjects who are influential in the temporal or spatial domain (or both).

McNamee and Lazar (2004) also address the problem of subject similarity and its impact on group maps by assessing the sensitivity of several of the methods studied by Lazar et al. (2002). They use a jackknife approach, (Quenouille, 1949; Tukey, 1958) treating the group map as the test statistic of interest. Deleting each subject in turn, the maps are recalculated based on a sample of size $n - 1$ rather than the original n, and discrepancy measures defined for the difference between the complete group map and the leave-one-out maps. The discrepancy measures are defined on the binary maps that are obtained after testing each voxel for significance and include a count of the number of significant voxels that are added by deleting a given subject, a count of the number of significant voxels that are removed by deleting a subject, and the percent overlap between a subject's individual map and the group map. Strictly speaking, this last is not a measure of the deviation of the jackknifed map from the complete group map, but it does give another indication of how important are the contributions of individual subjects to the overall picture. Note that unlike the work by Kherif and colleagues, where

the main interest lies in identifying unusual subjects, this is only a by-product of the approach taken by McNamee and Lazar, who concentrate instead on evaluating the combining methods.

Working on a test set of 11 subjects on a simple visual task, McNamee and Lazar (2004) demonstrate that of the procedures they consider, the random effect model is least affected by the deletion of any one given subject; in other words, it is a very robust method. This is not surprising, but it is apparently also an artifact of two other results: first, the group maps created by this combining method are extremely sparse, since the random effect model is a conservative approach to combining information due to having to account for two separate sources of variation; second, the amount of overlap exhibited by any single subject with the random effect group map is very small, indeed most subjects in this particular study have no overlap between their individual maps and the group map. On the other hand, and what also has been seen in the statistical literature many times, the group map created by Fisher's method of combining p-values may be highly affected by a single subject with unusual activation. This is manifested in the test set used by McNamee and Lazar, in the form of one subject with extremely strong activation in a region of the brain where no other subject has activity at all. The group map that includes this subject exhibits the pattern, even though it is attributable to only one individual. The authors hypothesize that the activity exhibited by that one subject is in fact a reflection of head motion, emphasizing the importance, also pointed out by Kherif et al. (2003), of carefully examining both the group maps and the maps of individual subjects when interpreting this type of analysis. Most other methods for combining individual maps fall somewhere between T_F and T_R in terms of their values on the discrepancy measures, and hence of their sensitivity to unusual (not necessarily outlying) subjects.

The choice of combining method has not been extensively explored outside of the work by Lazar and colleagues; the ad hoc procedure of averaging the individual maps together and the random effect model have been most popular, the former because of its ease of implementation and the latter owing to the general popularity of the random (or mixed) effect model in fMRI data analysis. All methods proposed to date have both strengths and weaknesses. The p-value methods, being relatively new to the fMRI community, have not yet gained widespread acceptance, in spite of the fact that they are easy to calculate (even for large numbers of subjects) and tend to be more liberal in admitting activation, which would seem to be ideal for the exploratory nature of much fMRI research. On the other hand, they are often more influenced by individual subject behavior than is the popular random effect model, and it may be difficult to interpret the results of a p-value-based approach as a consequence, without further exploration of the individual level data. Also, the plethora of p-value methods may be an impediment to their further adoption, unless a strong recommendation can be given in favor of one over the others. Finally, the p-value methods, for the most part, do not have obvious

extensions to the problem of comparing groups of subjects, which is often one of the scientific aims of a neuroimaging study.

In contrast with the p-value methods, the model-based combined estimation methods require the 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). Thus, both in terms of data storage and computation, these procedures, and in particular the random effect model, are more demanding. On the other hand, it is of course relatively easy to use either the fixed or the random effect model to perform group comparisons, simply by introducing "group" as a covariate. In choosing between treating subjects as fixed or random, both McNamee and Lazar, and Friston and colleagues, have pointed out that the fixed effect model can be usefully implemented even on a small number of subjects, whereas the random effect approach to subjects requires much larger sample sizes, which may not always be feasible in fMRI studies. Correspondingly, the conclusions that one can draw from treating the subjects as random effects are stronger.

As a final consideration, it is worth mentioning that the theoretical properties of various combining methods, especially the p-value techniques, have received a fair amount of attention in the statistical literature over the years. Some of these methods enjoy certain statistical optimality properties not shared by the others, and in particular Fisher's method would seem to be a natural recommendation, in spite of its extreme liberality, since it has been shown to be an asymptotically optimal procedure for combining information from independent sources,; see Littell and Folks (1973) and Berk and Cohen (1979).

6

Temporal, Spatial, and Spatiotemporal Models

The methods considered in the previous chapter ignore, for the most part, the temporal and spatial correlations in the fMRI data; at the very least they underutilize the rich information available in the time courses, as well as failing to take full advantage of spatial patterns. Ignoring this structure can lead to loss of efficiency, bias, and misstatement of type I error (Worsley and Friston, 1995; Purdon and Weisskoff, 1998). In this chapter we examine the more sophisticated models that exploit temporal correlation, spatial correlation, or both. From a statistical perspective these models are closer to being useful than the univariate, independence-based models of Chapter 5, although even those provide practitioners with good (or good enough) summaries of the manifested patterns of brain activity, as evidenced by the popularity of the basic general linear model. The tradeoff, especially for the spatiotemporal models that will be the focus of Section 6.3, is in more complexity of model, translating in turn to higher computational costs. For data sets as large as those that are typical in fMRI, this tradeoff is a nontrivial concern.

6.1 Temporal Models

We first examine models for the time course; we have already encountered some such models in Chapter 5, in the context of estimating and classifying the hemodynamic response function. The emphasis in this section is rather on the time series nature of the voxel observations themselves and ways of exploiting this information directly. There are two ways of analyzing time series data: in the time domain itself, and in the frequency domain following transformation of the data. Both have been used for fMRI time series.

A basic model for the time series can be obtained by extending the linear model from Chapter 5 so that the error term is a stochastic process, such as an autoregressive model of small order (Marchini and Ripley, 2000); thus at time t the response Y_t is modeled as

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_6,

$$Y_t = X_t\beta + Z_t,$$

where X_t is the design matrix, as before, but now $\boldsymbol{\epsilon}$, which was assumed previously to be, say, normal with mean zero and variance $\sigma^2 I$, is replaced with the process Z_t of mean zero and unknown covariance structure. Instead of the identity matrix I, there is a general matrix V whose elements depend on the autocovariance function between two time points. As noted by Marchini and Ripley (2000), these types of models can be analyzed in the original time domain, or the data can be Fourier transformed and analyzed in the frequency domain.

Writing the covariance in the new model as $\sigma^2 V$ with V written as KK^T, a general class of procedures is obtained by premultiplying by a matrix D, yielding $DY = DX\beta + DZ$, where the dependence on t has been suppressed (Marchini and Smith, 2003); ordinary least squares estimates of β follow standard theory. If V is known, the best linear unbiased estimator of β is achieved when $D = K^{-1}$. In that case, premultiplying by D is called *prewhitening*, since application of the transformation removes the correlation structure, resulting in errors that are white noise (Gaussian). Of course, in practice V is not known, and so it must be estimated. A wide variety of methods for estimating V have been proposed in the fMRI literature. In the next sections we examine some of the these, in both the time and frequency domains.

6.1.1 Time Domain Analysis

As noted above, it is in some sense natural to extend the basic linear model analysis, which collapses over the time dimension, to an analysis of the voxel time series themselves. The error structure of the model will change to accommodate the temporal correlation. This approach has a rich and long history in fMRI data analysis, going back to work of Friston and others in the early 1990s.

Worsley and Friston (1995) present an early and accurate accounting of the time series analysis. Readers should be aware that two prior presentations by Friston and colleagues (Friston et al., 1994; Friston et al., 1995) contained various mathematical and statistical errors that are corrected in Worsley and Friston (1995). The goal of Friston's original work in this area was to estimate the parameter β in a linear model $Y = X\beta + \mathbf{e}$, where Y is the unsmoothed time series, and the error vector $\mathbf{e}$ contains components that are independently identically distributed normal with mean zero and variance σ^2.

Framed in this way, the problem is simply one of allowing for serial correlation in a regression setting, and solutions exist. Although an optimal estimator is based on applying the Gauss-Markov theorem to unsmooth and decorrelated data, Friston and colleagues choose instead to smooth the time series, noting that good performance of the former is "... very sensitive to the correct specification of K" (where K is the smoothing matrix). Upon smoothing the data first, the least squares estimator of β becomes $\hat{\beta} = (X_1^T X_1)^{-1} X_1^T KY$,

with $X_1 = KX$. Worsley and Friston note that, because of the smoothing, this estimator isn't fully optimal, but it is unbiased, and in many situations the loss of efficiency is not great. The usual least squares theory then produces estimates of the variance and a test statistic that can be used to assess the behavior at each voxel. By contrast, Wicker and Fonlupt (2003) carry out an analysis of this type of model using generalized least squares (GLS) and an empirically determined correlation matrix.

There is more than one way to model the time course directly in the time domain. One such alternative route is taken by Bullmore et al. (1996b), who use trigonometric basis functions, namely sines and cosines, to capture the frequency information from the time series of signal intensities. For signal Y_t at time t, the fitted model is

$$\begin{aligned} Y_t = {} & \gamma \sin(\omega t) + \delta \cos(\omega t) + \gamma' \sin(2\omega t) + \delta' \cos(2\omega t) \\ & + \gamma'' \sin(3\omega t) + \delta'' \cos(3\omega t) + \alpha + \beta t + \rho_t. \end{aligned}$$

In this expression, ω is the fundamental frequency for the data (collected in a periodic experimental paradigm); the first three pairs of terms represent sine waves at the fundamental frequency and the first two harmonics. The term $\alpha + \beta t$ is a linear trend, and ρ_t is the error at the time point t. Since the residual errors are correlated, Bullmore et al. estimate the parameters via pseudogeneralized least squares.

To identify active voxels a two-stage approach, using temporal information only in the first stage and spatial in the second, is used. More specifically, at the first stage the authors calculate the *fundamental power quotient* at voxel i, defined as

$$\mathrm{FPQ}_i = \frac{\hat{\gamma}_i^2 + \hat{\delta}_i^2}{\sqrt{2(se(\hat{\gamma}_i)^4 + se(\hat{\delta}_i)^4)}}.$$

They then find the significantly active voxels by the use of a permutation test (see Section 10.3). In this way, distributional assumptions are avoided, at the cost of having a computationally more expensive procedure. For testing at the second stage, it is assumed that all voxels found in the first stage are false positives. A measure N_{vox} is then defined, which counts the number of voxels in each 8-connected cluster. Truly false positives should, in theory, be isolated, whereas truly activated voxels should cluster together. Only voxels that pass a threshold for both FPQ and N_{vox} jointly are considered active. As noted by the authors, the two measures are not independent, and so looking at them jointly serves merely to locate voxels of potential interest.

From the methods proposed by Bullmore et al. (1996b) additional detail regarding the timing of activation can be extracted. Noting that there is information also in the signs of $\hat{\gamma}$ and $\hat{\delta}$, they split the significant voxels into four groups according to whether each of the two estimated parameters are positive or negative. In their interpretation the sign of $\hat{\gamma}$ is related to the condition of the experiment to which the voxel is responding (positive for task or

negative for rest) and the sign of $\hat{\delta}$ is related to the timing of activation (positive for anticipatory or negative for delayed); hence it is possible to classify a voxel, for instance, as being active in delayed response to the task condition, or in anticipation of rest. For the data set they examine in this study (a single subject performing a simple visual and linguistic task), most of the active voxels show increased signal in delayed response to either the task (the majority fall in this category) or rest conditions, and very few voxels show an anticipatory reaction. Presumably for different types of tasks, one would find different breakdowns of voxels into the four groups.

Locascio et al. (1997) use traditional time series methods, namely autoregressive moving average (ARMA) models for the fMRI time course, on a voxel by voxel basis. At time t the signal intensity Y_t is modeled as

$$Y_t = \alpha_0 + \sum \alpha_i C_{it} + \beta_1 \text{time} + \beta_2 \text{time}^2 + \frac{\theta(B)}{\phi(B)}\epsilon_t,$$

where $\sum \alpha_i C_{it}$ is a term representing contrasts of interest between the experimental and baseline conditions; "time" counts the order of successive images; B is the backshift operator $BX_t = X_{t-1}$; $\theta(B)$ is the moving average operator $\theta(B) = 1 - \theta_1(B) - \cdots - \theta_q(B)^q$ for a moving average component of order q; $\phi(B)$ is the autoregressive operator $\phi(B) = 1 - \phi_1(B) - \cdots - \phi_p(B)^p$ for an autoregressive component of order p; ϵ_t is white noise at time t. Since the model incorporates both contrast and ARMA components, Locascio and colleagues term this a "CARMA" model.

The CARMA models are fit to each voxel individually. At each voxel, an essentially stepwise procedure is performed to search for the best AR (autoregressive) or MA (moving average) model of up to order 3. Higher orders could of course be considered, as could mixtures of AR and MA components (Locascio et al., 1997). The authors prefer to use AR over MA when both models fit equally well or nearly so, since the former is more easily interpreted. Finally, the residuals from the fitted model at each voxel are subjected to a test of white noise to determine if relevant time trends and autocorrelations have been accounted for. Voxels must pass this test as well in order to be declared significant. Significance in general is assessed for the contrasts of interest, on the voxels that pass the white noise test, using permutation methods (see Section 10.3).

A distinct advantage of this approach is that it allows for a different model to be fit at each voxel, since AR and MA components are tested separately and the best fitting order of the appropriate component is taken. These may differ from voxel to voxel and recognition of this fact is desirable; most often, for computational and conceptual ease the same model is fit at each voxel. Furthermore, unlike many of the temporal models, that of Locascio et al. (1997) can accommodate arbitrary experimental designs; it is not necessary for them to be periodic or have any other special structure. However, this method is purely temporal, with no attempt to borrow strength from neighboring voxels and fit similar models to nearby physical locations, and it is not immediately

clear how one could easily extend their procedure to have a spatial component and still keep to the spirit of the analysis.

Other proposed models for the temporal correlation include AR(p) (Bullmore et al., 1996a) and AR(1) with added white noise (Purdon and Weisskoff, 1998); estimation then proceeds, as in Locascio et al. (1997), under the assumption that the prespecified model holds.

6.1.2 Frequency Domain Analysis

Now we move to the analysis in the frequency domain, as espoused for instance by Lange and Zeger (1997) and Marchini and Ripley (2000). Let ω_j denote the Fourier frequencies, $\omega_j = j\delta/n$, where n is the length of the time course, δ is the sampling interval, and $j = 0, 1, \ldots, \lfloor n/2 \rfloor$. Then the Fourier transform of a series w is given by

$$d_w(\omega_j) = \frac{1}{n} \sum_{k=0}^{n-1} w_k \exp(-i2\pi\omega_j \delta k).$$

Hence in the frequency domain, the model for the time course can be represented as

$$d_Y(\omega_j) = d_X(\omega_j)^T \beta + d_Z(\omega_j).$$

For large n, the "error terms" $d_Z(\omega_j)$ are approximately uncorrelated (Marchini and Ripley, 2000).

For periodic stimulus designs, such as the standard block design traditionally used in fMRI, analysis in the frequency domain is simpler than that in the time domain, since the model will simplify considerably, relying only on the Fourier frequencies that correspond to the period of the block design, the rest being zero. Therefore, if we take the discrete Fourier transform of each voxel time series, most of the frequencies will not, in fact, contain information about the signal. Specifically, Lange and Zeger, and Marchini and Ripley point out that the fundamental frequency of activation in the spectral domain contains most of the information relevant for inference; additional information is found in the harmonics. Parametric (Lange and Zeger, 1997) or nonparametric (Marchini and Ripley, 2000) methods can then concentrate on the estimation of the few relevant components, as opposed to the entire spectrum. In a simple two-condition block design consisting of c repetitions of "control-stimulus," the relevant Fourier frequencies are $\Omega = \{\omega_j : j \in (c, 2c, \ldots, \lfloor n/2 \rfloor)\}$.

Lange and Zeger (1997) start with the time domain model $Y(t,i) = X(t,\theta_i)\beta_i + Z(t,i)$ and

$$X(t,\theta_i) = \sum_{0 \le s, t-s \le T-1} \lambda(s,\theta_i) x(t-s),$$

for location i, time $t = 1, \ldots, T$, $\lambda(\cdot,\cdot)$ the two-parameter gamma family described in Section 5.3.1, and $Z(t,i)$ mean zero random error. Upon applying the discrete Fourier transform to this model, it becomes $d_Y(\omega_j, i) =$

$d_X(\omega_j, \theta_i)\beta_i + d_Z(\omega_j, i)$ and $d_X(\omega_j, \theta_i) = d_\lambda(\omega_j, \theta_i)d_x(\omega_j)$. Using an iterative algorithm of complex least squares, estimates are obtained for the β and θ parameters at each spatial location separately.

As noted by some of the commenters on the paper by Lange and Zeger, the approach has several technical drawbacks: the two-parameter gamma model may not be flexible enough to capture the behavior of the hemodynamic response function, although it is no doubt more flexible than some of the other parametric models that have been proposed; there may in addition be issues of parameter identifiability and convergence of the complex least squares algorithm. Furthermore, their approach is only appropriate for periodic experimental designs. Particularly as more and more researchers are moving to the use of event-related studies, this is a serious limitation of the spectral domain analysis.

In spite of the inherent limitations of analysis in the frequency domain, it has continued to hold attractions for methodological researchers, who have built on Lange and Zeger (1997) and extended their approach in various directions.

Marchini and Ripley (2000) start with the time domain model $Y_t = X_t\beta + Z_t$, and assume that the time series has also been preprocessed to remove trends and other confounding factors. Then the terms $d_X(\omega_j)$ vanish except at the fundamental frequency and its harmonics, thereby reducing the model to

$$d_Y(\omega_j) = \begin{cases} d_X(\omega_j)^T\beta + d_Z(\omega_j) & j \in \Omega \\ d_Z(\omega_j) & \text{otherwise} \end{cases}$$

The authors note that if one takes the discrete Fourier transform of the time series at each voxel, most of the frequencies will contain information only about the underlying correlation structure of the stochastic process at that voxel. Now, the periodogram at frequency ω_j is exactly given by $I(\omega_j) = n|d_Y(\omega_j)|^2$, and so by studying the periodogram it is possible to learn about the response to the stimulus at each voxel, or, more precisely, which frequencies of the signal are evidence of response. Since much of the variance is explained in the fundamental frequency for active voxels, this is where Marchini and Ripley focus their inferential efforts. In particular, they demonstrate that the value of the periodogram at the fundamental frequency is related to the optimal estimator of β in the model for $d_Y(\omega_j)$, and hence tests for significance of the response to a *periodic* stimulus are based on this value.

The test statistic they define is

$$R_j = \frac{I(\omega_j)}{g(\omega_j)},$$

for $\omega_j = j/\delta n$ and $g(\cdot)$ a smoothed version of the periodogram which is used as an estimator of the spectral density. For periodic designs one need consider

R_j only at the fundamental frequency and its harmonics, as described above. The authors recommend using nonparametric methods such as smoothing splines to get the estimate $g(\cdot)$ of the spectral density, which is asymptotically unbiased. Under the null hypothesis of no activation, R_j at the fundamental frequency is asymptotically standard exponential; in fact this is true at other frequencies as well, save the edges, and so if inference on some of the harmonics is also of interest, the same result can be applied.

An interesting aspect of the proposed method stems from the observation that only at the fundamental and first few harmonic frequencies is there expected to be any response to the periodic stimulus. Hence, the values of R_j at the other frequencies can be considered as drawn from the null hypothesis; these therefore provide a large sample from the null to use as an empirical distribution against which to compare the values of R_j at the frequencies most likely to be of interest. This obviates the need to hew to the theoretical exponential distribution, if, for example, it is not a good fit for a given data set. The empirical distribution can be used instead for calibration of the test statistic.

Müller et al. (2001), also working in the spectral domain, consider instead a multivariate approach, in the hope of teasing out, in addition to regions of activation, the functional connectivities among such regions. This is a delicate question in the analysis of fMRI data (see also Sections 4.3.1 and 11.2). As in Marchini and Ripley (2000), Müller et al. (2001) assume a periodic experimental design and focus on the fundamental frequency. Using multivariate time series methods in the spectral domain, they estimate two key parameters for understanding the temporal connections between pairs of voxels: the coherence and the phase lead.

Letting $f_{jk}(\lambda)$ be the cross-spectral density function at frequency λ, the coherence is defined to be

$$\rho_{jk}(\lambda) = \frac{|f_{jk}(\lambda)|}{\sqrt{f_{jj}(\lambda) f_{kk}(\lambda)}}$$

and the phase lead is the function $\nu_{jk}(\lambda)$ in the expression

$$f_{jk}(\lambda) = |f_{jk}(\lambda)| e^{i\nu_{jk}(\lambda)}.$$

Coherence in this context is analogous to correlation, namely it is a measure of linear association between two time series at a particular frequency. Voxels that have high coherence with each other are "correlated" in this sense, and the expectation would be that coactivating voxels would form clusters with high coherence. After the clusters with high coherence are identified, the method proposed by Müller and colleagues computes the phase lead, again pairwise between voxels, again for selected frequencies that are related to the periodic experimental design. The authors describe the phase lead as a measure of the amount of temporal displacement in the BOLD response for one region relative to another. Thus, examination of the phase lead in theory can shed

light on some aspects of connectivity, such as which regions activate earlier and which later in reaction to a particular stimulus. Note that, while this analysis may describe the temporal sequence in which different regions become active, it does not indicate causality; simply because voxels in one region activate before those in another, one cannot of course conclude that activity in the former leads to activation in the latter.

According to Müller et al. (2001),working with a simple visual task, their method performs comparably to the standard general linear model analysis, picking out similar regions of activation (in both location and extent). A purported advantage of their procedure is that some understanding of the network, via the different lags in BOLD response for different regions, is obtained. As mentioned above, however, it is still not possible from this approach to infer causality in the network. Also, because the authors assume weak stationarity, the experimental design needs to be periodic (that is, a block design experiment) or nearly so. Hence it won't be an appropriate analysis path for more advanced or complex experimental designs.

A somewhat different approach to the modeling of the time series is given by Gonzalez Andino et al. (2000), who start from the assumption that time series for voxels that are related to "signal" should look different from those that are related to "noise" (or "nonsignal," more generally). In particular, the time series should be differentiable according to their complexity, with "signal" voxels having less complex patterns made up of a few temporal components (Gonzalez Andino et al., 2000). The measure they propose to use for the purpose of distinguishing signal from noise time series is the *Renyi entropy*, which makes minimal assumptions about how the signal is generated. There is no need to assume normality or stationarity, and the HRF is not estimated. It is only assumed that the characteristics of noise and signal time series differ.

A concept that is basic to their approach is the *time frequency representation*, or TFR; this is a two-dimensional plot showing how the frequency of a series varies over time. Time series that contain organized signal will have a few "hot spots" in the TFR, whereas those that are essentially noise will have many such spots scattered at random. Thus the number of hot spots in the TFR can be taken as a measure of the complexity of the time series, with the rationale that many components are needed to describe a noise series and only a few are needed to describe a series with a clear pattern.

To formally measure the complexity of a signal, the authors use the following definition of Renyi entropy:

$$H_\alpha(C_s) = \frac{1}{1-\alpha} \log_2 \int \int \left(\frac{C_s(t,f)dtdf}{\int \int C_s(t,f)dtdf} \right)^\alpha ,$$

where α represents the order of the entropy, and $C_s(t,f)$ are the coefficients of the TFR of the time series s. Based on earlier empirical studies by various researchers, the value $\alpha = 3$ is chosen. When the number of components in the TFR is small (organized signal), this entropy measure will also be small; for a large number of diffuse components (noise), the entropy will be high.

On a simple motor task, Gonazalez Andino and colleagues find that values of the Renyi entropy are clearly separated for voxels declared active and those declared inactive by another method (correlation analysis). While this is only one study, and based apparently on a single subject, the results are indicative of the potential power of the approach. As is evident from the expression for the Renyi entropy itself, there is no need to estimate the HRF, nor to assume a reference vector (such as for the correlation method), meaning that this type of analysis can, in principle, be applied also to event-related experiments of arbitrary complexity. On the other hand, users do need to choose the time frequency representation, of which there are many possibilities, and the order α; it is not clear how sensitive conclusions are to these choices. Furthermore, the authors do not offer a formal way of distinguishing between voxels to be declared active and those to be declared inactive based on values of $H_\alpha(C_s)$ when there is not a natural separation between the two groups. One could presumably bring existing statistical methods to bear on this problem.

6.1.3 Effect of Ignoring Temporal Correlation

Purdon and Weisskoff (1998) report a simulation study that explores the importance for precise statistical inference of accounting for the temporal correlation in the fMRI time series. They look at two block design paradigms, one of low frequency (blocks of 40 seconds) and one of high (blocks of 20 seconds), and three statistical procedures: the nonparametric Kolmogorov-Smirnov test, the t test, and a Fourier-based F test.

For each combination of design and test, the authors calculate the *false positive characteristic*, or FPC; this is a plot of the false positive rate in simulated "null" data (data with no activation, representing noise or resting brain) against the assumed level of significance. If a data set meets the assumptions of a particular test, the FPC line should be straight with a slope of 1, since in this case the proportion of false discoveries will match the preset significance level. Deviations from the test assumptions will manifest themselves in a nonlinear FPC curve. Purdon and Weisskoff (1998) find that in the low frequency design, for all three tests the FPC line curves upward, indicating that more significant voxels are detected than is warranted by the declared α level. For smaller values of α, the bias is worse than for larger values. In the high frequency design, there are fewer false discoveries than in the low frequency paradigm, which under some circumstances results in bias in the opposite direction, that is, fewer significant results than would be expected for a given level. The TR also influences the amount of bias in the results.

In short, as one might expect, if there are indeed temporal autocorrelations present in the fMRI data, ignoring them introduces bias in the assumed significance levels, resulting in tests that may be conservative or liberal in direction, depending on the TR and the experimental design. Purdon and Weisskoff also suggest a way of modifying their analysis and that of Worsley and Friston (1995) to account for temporal autocorrelation.

Woolrich et al. (2001) examine the effectiveness of several statistical methods for directly handling autocorrelation in the fMRI time series, namely coloring with a low-pass filter, correcting the variance, and prewhitening. All of these are variations on the theme of premultiplying the general linear model by a matrix D that will simplify calculations or the data; they differ in the choice of matrix. In addition, the autocorrelation may be estimated using a variety of approaches from the time series literature: parametric or nonparametric techniques, or windowing/tapering. Finally, Woolrich and colleagues explore the effect of different experimental designs: a regular block design, and event-related studies with fixed, jittered, and random interstimulus intervals.

For all experimental designs prewhitening is the most efficient method of handling the autocorrelation, followed by variance correction and coloring. For the block design experiment the differences in efficiency are minimal, indicating that as long as the autocorrelation is accounted for, the particular way in which this is done is not very important. For the event-related designs with either fixed or jittered ISI, the differences between prewhitening and variance correction are small, but coloring is clearly inferior, with the latter achieving 70% of the efficiency of the former two; in the random ISI design, coloring is only 20% as efficient as prewhitening, and variance correction is 80% as efficient.

Although prewhitening is the most efficient method according to this research, it does require a robust estimator of the autocorrelation (Woolrich et al., 2001). Hence in the second part of their study, Woolrich and colleagues compare the estimation techniques mentioned above. They find that a simple windowing works best, and, when applied locally to small neighborhoods, is computationally efficient as well. However, even for the most effective method, there is considerable bias far in the tail, in the regions of the distributions that are of most interest for statistical inference, and especially when corrections are made for multiple tests. This bias can be offset to some extent, as evidenced in the study, by applying a small amount of spatial smoothing. See also Marchini and Smith (2003) for additional simulations and a theoretical discussion which together further highlight the usefulness of smoothing in this setting.

6.2 Spatial Models

In contrast to the large amount of work that has been done to model the temporal aspect of fMRI data, the purely spatial aspect has seen much less development (Hartvig and Jensen, 2000). As discussed in previous chapters, the traditional approaches to the spatial correlation problem have been to ignore it (assume independence) or to spatially smooth, perhaps in conjunction with another analysis. In part this is because the nature of the spatial dependence, which clearly must exist, is much harder to decipher and hence

to propose models for, than the temporal correlation structure. Physical location alone is not enough to describe the spatial dependence; for example, Broca's area and Wernicke's area are involved in language processing and may be present in both hemispheres of the brain (although more developed on one side than the other), hence it is possible, and even likely, that voxels that are not in spatial proximity to each other would still show high correlation. Yet the large number of voxels precludes calculating all pairwise correlations and assessing them for patterns. A different approach is needed beyond simple autoregressive spatial models.

Another possible reason for the relative dearth of purely spatial models is that, unlike for temporal models where an analysis can be performed on the time courses directly assuming independence of voxels, a spatial analysis that ignores the temporal element still requires some prior processing. One could, for example, construct a model concerned only with spatial relationships on the basis of a linear model analysis of the time courses, or some other summary of the temporal information. That is, the time course is analyzed before any spatial model is built. The alternative, in analogy to the temporal analysis, would be to fit a spatial model at each time point independently, and view the results as a movie. While feasible computationally, this is hard to interpret and I am not aware of this approach being put forth in the literature. By contrast, it is not strictly necessary to smooth or execute any other spatial processing before constructing a temporal model.

Interestingly, much of the work that has been done on purely spatial modeling is Bayesian in nature. The reason for this is most likely that if one is not going to smooth, and one still wants to take advantage of the neighboring voxels, the Bayesian framework is the obvious choice for borrowing strength. We will see other uses of the Bayesian approach for fMRI data analysis in Chapter 9. Another possibility is to apply clustering techniques to the statistical map. Both approaches have been proposed and we take them up in the following sections.

6.2.1 Bayesian Spatial Models

An example of a (Bayesian) spatial model built on a statistical map that summarizes over the time dimension is Hartvig and Jensen (2000), who suggest a variety of spatial models based on mixtures. They write that "[t]hough the model may be used as the spatial part of a spatio-temporal model, we will only consider the problem of estimating the activation pattern based on a single summary image (or volume) of voxel-wise activation estimates, also known as a statistical parametric map (SPM)" (p. 234). Their analysis starts with the intuitively pleasing idea that active voxels will tend to cluster together (see also Forman et al. 1995, and the discussion in Chapter 10). Hence it makes sense to consider clusters of voxels, or, more specifically, the activation status in clusters or neighborhoods of voxels. For a given voxel i, then, we want to borrow strength from its 8 immediate neighbors (all those in the same

slice, with voxel i in the center of the square), or its 26 immediate neighbors (all those in the same slice, or in the two adjacent slices, again with voxel i in the middle of the cube); voxels that are in an "active neighborhood" are themselves more likely to be active.

Denote the activation status of voxel i by A, where $A = 1$ indicates that i is active and $A = 0$ indicates that it is not. Also, the activation statuses of voxel i's k neighbors ($k = 8$ or $k = 26$ for the two cases described above; one could of course define neighborhoods with different characteristics, but these have the advantage of simplicity) are denoted by Hartvig and Jensen as $A^1, \ldots, A^k$. None of these A values are observed; rather the authors take a Bayesian approach and estimate the posterior probabilities of each being 1, based on a model for the prior and a likelihood function given the state of activation.

What is observed in their scenario is the interim statistical map, for example, the output from a simple t test performed at each voxel. The first step is to specify a likelihood for this observed value, called x, given the activation status A of the voxel. Hartvig and Jensen suggest a normal distribution with mean zero when $A = 0$, and either a normal distribution with mean $\mu \neq 0$ or a gamma distribution when $A = 1$. This determination of the likelihood expresses the mixture nature of the problem. Next, the prior is specified, and this is the main focus of Hartvig and Jensen (2000). Once these two are determined Bayes rule gives the posterior probability of a pattern of activation and of a particular voxel being active (regardless of the neighbors). Letting subscript C denote properties of the cluster configuration (i.e., which voxels are active), the posterior probability for the entire pattern of activation is given by

$$P(A_C = a_C|x_C) \propto f(x_C|a_C)P(A_C = a_C)$$

and the posterior probability of voxel i being active is

$$P(A = a|x_C) \propto \sum_{a^1=0,1} \cdots \sum_{a^k=0,1} P(A_C = a_C|x_C).$$

Now, in general this latter expression will be hard to calculate, since it requires summing over all the possible activation patterns of the neighbors of voxel i, therefore Hartvig and Jensen propose priors that result in a closed form expression for the posterior probability. All of these priors are applied to small, local neighborhoods, and aim to capture the notion that truly active voxels should tend to "clump together." Let S be the number of 1s in the cluster under consideration. The three priors are:

1.

$$P(A_C = a_C) = \begin{cases} q_0 & S = 0 \\ q_1 & S > 0 \end{cases}$$

2.

$$P(A_C = a_C) = \begin{cases} q_0 & S = 0 \\ \alpha\gamma^{S-1} & S > 0 \end{cases}$$

3.

$$P(A_C = a_C) = \begin{cases} q_0 & S = 0 \\ \alpha_1\gamma_1^{S-1} + \alpha_2\gamma_2^{S-k} & 1 \le S \le k \\ q_1 & S = k+1 \end{cases}$$

The first prior has in effect only one parameter, and is thus particularly easy to work with. It represents the prior belief that the active clusters are of intermediate size. Single voxels are not believable, but neither are neighborhoods that are very large.

The parameter γ in the second prior is a measure of correlation among neighboring voxels. The other free parameter can be rewritten in terms of the probability of a voxel being activated. The third prior induces symmetry in the way active and nonactive voxels are treated. It can be written in terms of the probability of a voxel being activated, plus four parameters that describe the correlation across voxels.

Based on simulations and a real data analysis of a visual processing task performed by a single subject, the authors recommend the second of their three priors, applied to a small neighborhood (3×3 for a slice, or $3 \times 3 \times 3$ for a volume). This combination of model and neighborhood performs the best, in terms of power and of minimizing classification error. The mixture model is also comparable to nonparametric spatial smoothing methods (see their Figure 5) in the appearance of the activation maps, although the latter does result in somewhat smoother clusters. Note that the procedures described here are all local; that is, the activation probability of a given voxel depends only on the behavior of its immediate neighbors. However, since the model is applied at every voxel, contiguous regions that span the brain, for instance bilaterally, can be formed. The local fitting reduces the computational burden, as do the closed form expressions that the authors derive for the posterior probabilities.

A very different Bayesian analysis is implemented by Smith et al. (2003) (see also Smith and Fahrmeir 2007).Their point of departure is the basic linear model, as described in the previous chapter. In their analysis the time course at voxel i is modeled as the sum of a baseline trend (which is not of direct interest), an "activation profile," and error. The second term, the activation profile, is the focus of the analysis. Smith and colleagues assume a latent variable, γ_i underlying voxel i, so that $\gamma_i = 1$ if voxel i is active and $\gamma_i = 0$ otherwise. The regression parameter for the activation profile in the linear model then represents the amplitude of activity, being nonzero only if $\gamma_i = 1$.

In this configuration, the vector γ that summarizes the activation pattern is the parameter of interest. The authors suggest imposing spatial correlation and incorporating information (for instance, anatomical) via a prior that has the *Ising form*; this is a common prior in spatial statistics (Besag, 1986; Besag et al., 1991). The Ising prior for γ is

$$\pi(\gamma) \propto \exp\left\{\sum_{i=1}^{n} \delta_i \gamma_i + \theta \sum_{i \sim j} \omega_{ij} I(\gamma_i = \gamma_j)\right\}.$$

Here, the first term in the sum is called the *external field*; anatomical or expert prior information enters the model through this part of the prior. The second term in the sum models the spatial correlation: the sum over $i \sim j$ is over the neighbors of voxel i; ω_{ij} are weights for the interaction between neighboring voxels. Finally, the parameter θ is used to control the amount of spatial smoothing. The components of the γ vector are independent when $\theta = 0$ and become more spatially correlated as θ increases. Specification of the model is completed by setting priors on the other elements of the linear model and on the activation amplitudes given $\gamma_i = 1$. For each voxel, Markov chain Monte Carlo is used to obtain the posterior probability of activation, and the posterior distribution of the activation magnitude.

Based on simulation and analysis of a real data set, the authors note that their Bayesian procedure seems to find more isolated voxels than a comparable frequentist (linear model-based) analysis. On the other hand, they find increased sensitivity, apparently due to the use of anatomical prior information, to details of activated brain structures. This feature of the method makes it potentially well-suited for mapping of the brain before surgery. In that case, as the authors point out, it is important to have precise, individually tailored inference, especially as regards the regions that are involved in particular cognitive tasks.

6.2.2 Clustering for Spatial Modeling

The use of clustering to localize and characterize spatial patterns of activation can blur the distinction between purely spatial and spatiotemporal models. Approaches that cluster the fMRI time course (Baumgartner et al., 2001) or models for the hemodynamic response (Gibbons et al., 2004), for example, are really more spatiotemporal than spatial. Since clustering methods require a dimension along which to measure closeness or similarity of behavior, this is to some extent unavoidable in functional neuroimaging problems: if one is looking for voxels in the brain that cluster together, it is natural to consider the behavior of those voxels over the course of an experiment. This leads, in turn, to clustering over the time course. Hence, the most statistically natural way of applying clustering techniques in the fMRI context gives a spatiotemporal, rather than a spatial, analysis. Such analyses will be taken up in Section 6.3.

One might then ask if there is an informative way of clustering that leads to a more purely spatial model. From the previous discussion, it is evident that to answer in the affirmative we require a summary of the experiment-wise behavior of each voxel, much as in the Bayesian approach of Hartvig and Jensen (2000). That is, we wish to cluster on features of the data other than the time course, but those features must be such that they capture the activation patterns at the voxel level.

Given the difficulty of building spatial models that don't also have a strong temporal component, it is not surprising that this approach has received little attention in the literature. Indeed, only recently (Bowman and Patel, 2004) has an attempt to tackle the problem appeared. Although the methodology of Bowman and Patel (2004) was developed and tested on positron emission tomography (PET) data – another imaging modality – many of the statistical issues are the same; the authors furthermore indicate that the analysis can be used for fMRI data as well. The essence of their approach is to cluster parameter estimates obtained from a general linear model, or contrasts based on those estimates. They additionally suggest a "multiple classification approach" whereby many clustering algorithms are evaluated simultaneously. Following the evaluation, either the algorithm that produces the single best classification is chosen, or, if there is no such algorithm, a conglomeration of several is used instead.

Among the algorithms that come under test are hierarchical procedures (single linkage, complete linkage, and so forth), K means, and fuzzy clustering. In order to evaluate the performance of the various algorithms, Bowman and Patel propose a new measure, which they call the *relative information* associated with a particular partition, or RI. One rationale for the new measure is that most absolute measures of cluster performance will improve simply by the addition of more clusters. RI, by contrast, attaches a penalty to the number of clusters. It is assessed relative to a "reference clustering" solution, the procedure which yields the least probable partition of the data. The relative performance of the possible classifications is evaluated using RI, which leads either to the choice of a particular scheme, or a pooled procedure, with RI providing the weights. If the latter is chosen, a voxel is not uniquely (discretely) identified as belonging to a particular cluster; rather it is summarized by a weighted average of its plausible class memberships.

Not unexpectedly, the multiple classification approach produces good results, since it combines the outcomes of the individual algorithms, or picks the best among them. However, it should be noted that clustering algorithms are in general computationally intensive, especially on large data sets such as fMRI, even collapsing along the time dimension. Use of the multiple classification procedure adds a level of computational complexity. The authors also find in some of their simulations that RI still tends to favor a large number of clusters for some clustering methods; this despite the attempt to incorporate a bias towards parsimony.

6.3 Spatiotemporal Models

Spatiotemporal models for fMRI data aim at incorporating both time and space effects. In terms of building statistically valid and realistic models, these spatiotemporal approaches are the most natural way to handle functional neuroimaging data. However, both computational and conceptual barriers

have historically prevented the development and widespread application of this idea. Computationally, the task of fitting a full spatiotemporal model to fMRI data is a formidable one, involving hundreds of thousands of voxels over hundreds of time points. Estimation of model parameters and their standard errors can be challenging in this scenario. Conceptually, as we have already seen, the spatial correlation in particular is difficult to summarize in a form that admits a simple statistical model.

With advances in computing power, simultaneous models for the spatial and temporal elements are becoming more feasible. Two schools of thought are commonly found: clustering of time series (as introduced in Section 6.2.2) and "direct modeling." Time series clustering is nonmodel based in that no parametric model is specified for the spatial relations; rather, these are elucidated by the detected clusters. By contrast, with direct modeling one attempts to fit spatial models, often with the aid of prior information provided by neuroscientific rationale or previous experiments. The two approaches are thus fundamentally different in perspective and assumptions about the spatial component in particular.

6.3.1 Clustering fMRI Time Series

A relatively straightforward way to incorporate time and space into the statistical analysis is to apply clustering techniques to the time series data. In this application there are various questions that need to be addressed, including: What should be clustered – the raw time series or some function of these? What clustering algorithm or family of algorithms should be used? How many clusters are needed and how should this be decided? All of these have received attention in the fMRI clustering literature.

Regarding the first question, there are two main perspectives. One, as exemplified by the work of Baumgartner and colleagues (Baumgartner et al., 1997; Baumgartner et al., 1998) clusters the time series themselves, looking for similarities in behavior. This seems to be the dominant approach. Other authors (Goutte et al., 1999) claim that clustering the time series is unstable, and that the resultant clusters will not necessarily reflect similarity of responses to the stimulus. Goutte et al. (1999), for example, recommend clustering instead on the correlation function of the series with the experimental paradigm. Also on this topic, many authors have noted that since most voxels are inactive, if all of the voxels in the brain are tossed at a clustering algorithm, the clusters that come out will not be able to easily distinguish active from inactive locations. In other words, the algorithms will not be likely to find clusters that are made up of solely active voxels, or even more generally, clusters of activation. Hence most researchers in this area recommend doing some sort of screening first, to eliminate voxels that are clearly not active. This is done in different ways and taking different features of the data into account, as we will see in more detail below.

Choosing the "best" or "right" number of clusters when this is not known a priori is also a problem that warrants attention. In fMRI data it is reasonable to assume that most of the identified clusters will be made up of inactive voxels, even if some screening procedure is implemented before clustering. If the number of clusters is too small, active voxels are likely to be clumped together with inactive ones, and the clusters that result will not be easily interpretable. If the number of clusters is too large, the active voxels may be split across several clusters, which may or may not have a physiological intepretation. In general, this is a delicate question; we consider several potential solutions in the sequel.

We first briefly survey works in which the fMRI time courses are clustered directly. These investigations differ mainly in the details of implementation: what clustering algorithm is used; how the number of clusters is decided; and whether or not there is prescreening to remove inactive voxels from the analysis.

1. *Clustering Algorithm.* The most popular clustering algorithms in fMRI data analysis are K means (for example, Balslev et al. 2002) and fuzzy clustering (for example, Baumgartner et al. 1998; Fadili et al. 2000), although hierarchical methods (Stanberry et al., 2003) have also been considered. Filzmoser et al. (1999) use a combination of K means and hierarchical clustering; first they use a hierarchical approach to narrow down the number of clusters (hence avoiding in part the need to specify this in advance) and then take those identified clusters as the starting point for a K means clustering.
2. *Number of Clusters.* Hierarchical clustering methods such as average, single, or complete linkage do not require prior specification of the number of clusters, although a clustering threshold must be picked. Fuzzy clustering and K means require that the number of clusters be chosen ahead of time; thus researchers need in any case to confront this question. Aside from the two-stage approach of Filzmoser et al. (1999) mentioned immediately above, a posteriori validation of detected clusters by statistical testing (Baumgartner et al., 1998), cross-validation (Balslev et al., 2002), and iterative unsupervised learning with a fuzzy clustering algorithm (Fadili et al., 2000) are some possibilities that have been suggested in the literature. But the problem remains inherently difficult.
3. *Reduction of Brain Volume.* Most authors recommend reducing the set of voxels on which the algorithms are applied, due to the fact that the proportion of active voxels in the brain is relatively small. Without a prescreening step there is the concern that even the active voxels will get clustered among the noise, rather than forming clusters of their own; see, however, Gibbons et al. (2004) for one example of clustering where only air voxels are masked out, yet meaningful clusters of activation are detected. Reduction of the voxel set is achieved by segmentation to strip away the white matter and cerebrospinal fluid, so that only voxels in the gray matter

are clustered (Fadili et al., 2000) or crude thresholding according to values of a simple test statistic (Goutte et al., 1999; Fadili et al., 2000; Balslev et al., 2002).

Minimal spanning trees (MST) (Hartigan, 1975)) have also been used for analyzing fMRI time courses (Baumgartner et al., 2001), thereby providing another spatiotemporal approach based on clustering ideas. The MST is a multidimensional generalization of an ordered list; Baumgartner et al. (2001) suggest that it can thus serve as a means of investigating the temporal evolution and connectivity among groups of spatially clustered voxels. The starting point of such an investigation is therefore voxel time series that have already been clustered by some other method.

Once the clusters are identified, their MST algorithm proceeds by combining all clusters together and ordering all of the voxels according to Euclidean distance from a root node; the root node has depth of zero by definition, and the depths of the other time courses are defined by their distances from the root. Voxels can then be ranked according to distances and interest centers on whether or not the time courses from different clusters are distinguishable. This is determined by examining the *runs structure* of the MST, where a run is a consecutive sequence of voxels from the same cluster. The total number of runs and the length of the longest run give information about the separability of the clusters. For instance, if there are many short runs, this means that the observations from different clusters have similar temporal behavior, since their distances from the root time course are similar, and the clusters aren't separable. By contrast, a small number of long runs indicates that the observations from different clusters have different depths, i.e., different temporal behavior. Ideally, the voxels from different clusters would completely separate into one run for each cluster. Baumgartner et al. (2001) demonstrate this separability on several simulated and real data sets.

Coactivation is inferred when no such separation results from the MST ranking of the time courses. In this instance the temporal behavior of different regions or clusters of voxels cannot be distinguished: voxels that are physically distant from each other exhibit similar response to the stimulus. In terms of the characteristics of the relevant tree, time courses from one region are ranked near time courses from other regions. No separability is possible. When the ranks are plotted back onto a brain image, we expect to see in this case an intermingling of voxels from different regions.

Figure 6.1 shows the minimal spanning tree constructed from voxel time courses from two previously identified clusters; one cluster, containing $n_1 = 19$ voxels, is believed by the researcher to be related to the experimental task, whereas the other, containing $n_2 = 12$ voxels, is believed to be noise. As can be seen, the two clusters are completely separated in the tree. The complete cluster separation is also apparent in the ordered index plot (Figure 6.2). One conclusion that can be drawn from this representation of the data is that the

interpretation of the clusters as containing voxels with different behaviors is justified.

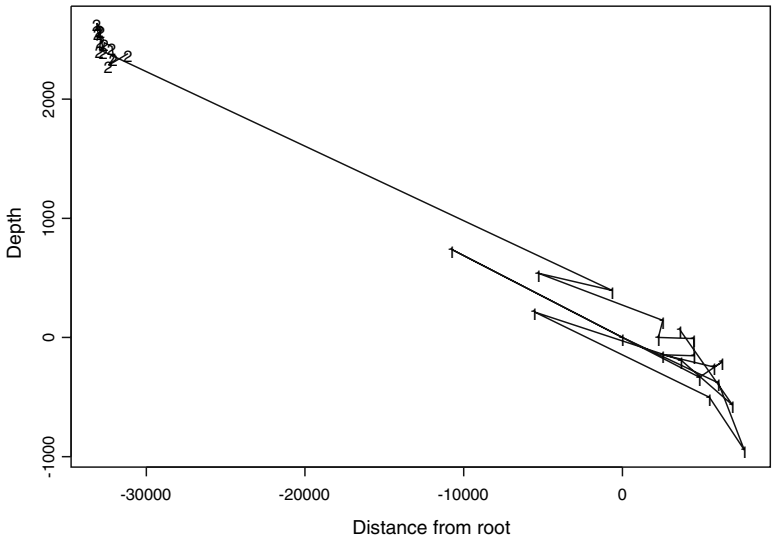


Fig. 6.1. Minimal spanning tree built from voxels belonging to two pre-identified clusters. The first cluster contains 19 voxels and is thought by the researcher to be related to the task. The second cluster contains 12 voxels and is believed to be noise. While the complete separation of the clusters apparent in the tree cannot validate these claims, it does confirm that the time courses of the voxels in the two clusters exhibit very different behaviors.

Finally, Figure 6.3 shows the results of applying three hierarchical clustering algorithms – complete linkage, average linkage, and single linkage – to the combined data set from the two clusters. All methods identify the two clusters correctly, although they differ slightly in the details of the structure (which voxels within a cluster are deemed "closest" varies from algorithm to algorithm).

Stanberry et al. (2003) also use the idea of the minimal spanning tree, through its connection with the single linkage hierarchical analysis, and "dendrogram sharpening." Dendrogram sharpening is a way of reducing the data that are input to the clustering algorithm in order to produce more distinct clusters. The data that are discarded during the sharpening stage are then classified into one of the clusters identified by the single linkage algorithm. Sharpening involves looking at every parent node in an initial dendrogram based on all of the data, starting at the root node. Any branch of the

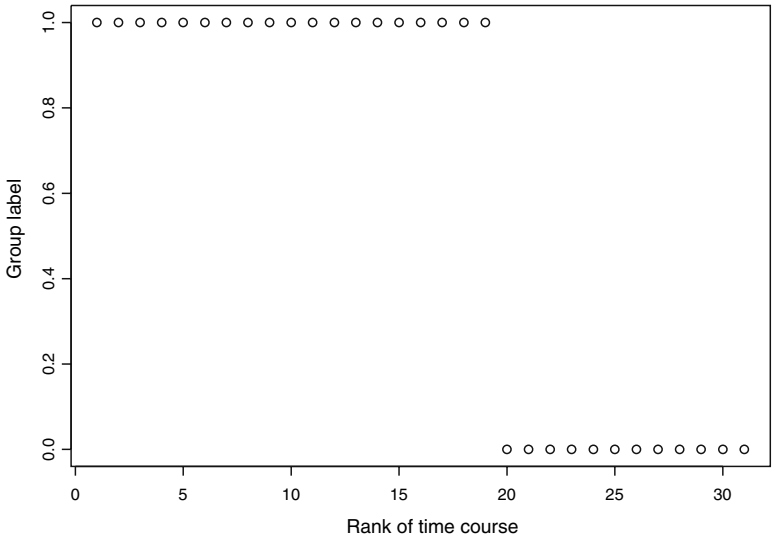


Fig. 6.2. Ordered index plot, based on the minimal spanning tree for the two clusters. The ordering is based on distances between time courses. The voxels in the two clusters are completely separated from each other.

dendrogram that has a minimum number of nodes is a candidate for sharpening; child nodes that are smaller than a preset size are eliminated. Hence the amount of data reduction is governed by two tuning parameters – the minimal size of a parent node to be a candidate for sharpening (denoted n_{core}), and the maximal size of the child nodes (denoted n_{fluff}). For example, with $n_{\text{core}} = 10$ and $n_{\text{fluff}} = 3$, all the children of size 3 or smaller will be discarded from every node of size 10 or more. Size refers to the total number of descendants of a node. The algorithm proceeds from the root up, discarding as it goes. Additional data reduction is achieved by a prescreening step; the distance measure in their algorithm is the correlation between time courses and any voxel that doesn't have a high correlation (greater than 0.5) with at least four other voxels is discarded even prior to the sharpening. In the examples shown by Stanberry et al., vast reductions in the size of the data set are achieved by these two tools, and single linkage clustering applied to the time courses of the survivors often reveals clear structure.

To demonstrate some of the ideas behind this approach, consider a simple simulated data set made up of $n = 15$ observations, 8 of which are drawn from a bivariate normal distribution with mean $(0, 0)$ and covariance matrix I, and 7 from a bivariate normal with mean $(1.5, 1.5)$ and covariance matrix I. Hence there are two clusters but with some overlap; see Figure 6.4. The dendrogram

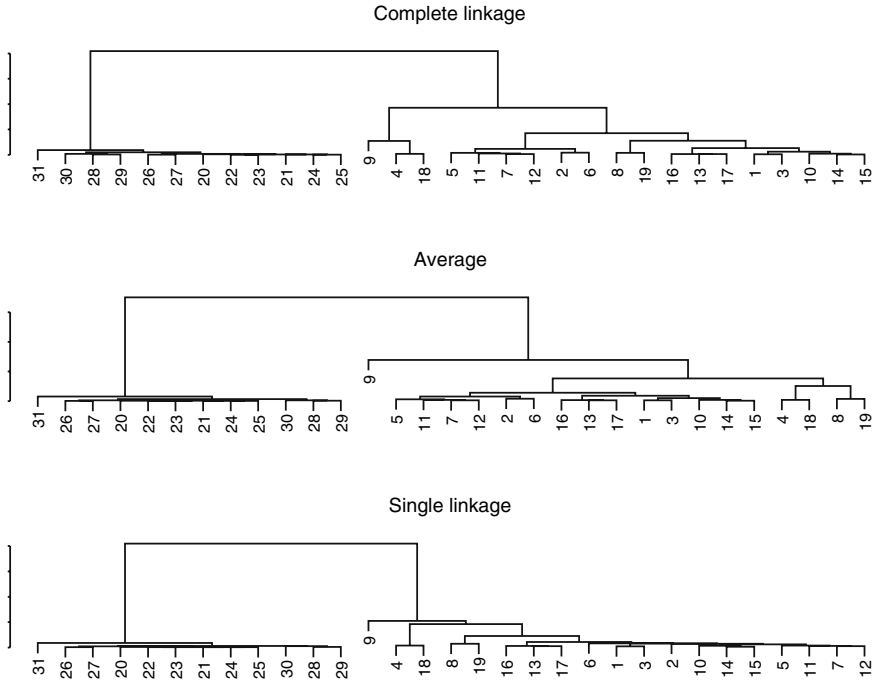


Fig. 6.3. Hierarchical clustering applied to the combined data set of $n = 31$ voxels. All three methods clearly and correctly identify the two clusters from which the voxels are drawn.

for this data set based on the single linkage algorithm is in Figure 6.5. There are two clear clusters identified in the figure, but some of the observations from the second distribution (observations 9 and 15) are misclassified as coming from the first.

For this example we set the parameters $n_{\text{fluff}} = 2$ and $n_{\text{core}} = 5$. The root node is of size 15, so it will be analyzed. It has two children, the left of size 10 and the right of size 5. Both are greater than 2, so will be further considered. The right child is of size not greater than 5, so it will be retained in its entirety. The left child is subject to sharpening. It has children of size 1 (left) and 9 (right). The left child is of size smaller than 2, so it is discarded. The right child has children of size 2 (left) and 7 (right), so again the left child is discarded. The right child is a candidate for additional sharpening. Its children are of size 3 (left) and 4 (right); both are greater than 2, but less than 5, and so are kept. The three observations $\{6, 7, 8\}$ are deleted; these are denoted in Figure 6.4 as open circles with dots inside of them.

The dendrogram for the sharpened data set is given in Figure 6.6. Although observations 9 and 15, from the second distribution, are still misclassified as

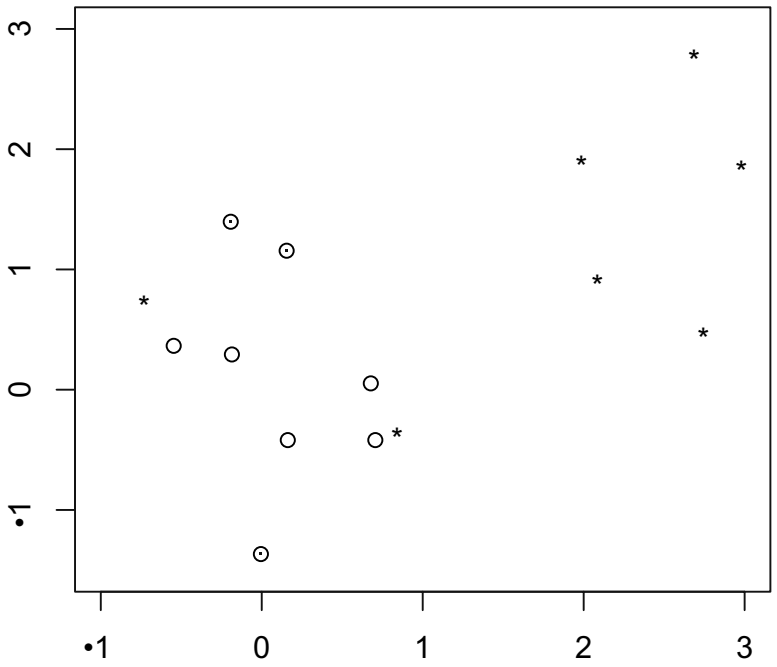


Fig. 6.4. Scatterplot of simulated data; 8 points (open circles) are taken from the standard bivariate normal and 7 (stars) from the bivariate normal with mean $(1.5, 1.5)$ and covariance matrix I. Two clusters are discernible in the data, with some amount of overlap. The three open circles with dots inside of them are points that are discarded by the sharpening algorithm.

coming from the first cluster (they were not discarded by the sharpening), the two identified clusters are now more distinct than they were previously.

Goutte et al. (1999) is a good example of clustering on something other than the time courses themselves. As noted above, these authors suggest that clustering on the correlation function of the fMRI time series with the experimental paradigm can yield more meaningful groupings. Let T denote the length of a time course and let y_j be the measured time series at voxel j. Then the correlation function, which is used by Goutte and colleagues as the metric for the clustering algorithms that they evaluate, is defined as

$$x_j(t) = \frac{1}{T} \sum_{s=1}^{T} y_j(s) p(s-t),$$

with $p(\cdot)$ the stimulus series (for instance, the boxcar typical of a block design). This is the usual convolution of the observed series with the stimulus, now evaluated at each time point instead of being summarized into a correlation coefficient. The correlation function is also used as the screening device to rid the data of "clearly inactive" voxels prior to clustering. In this instance the

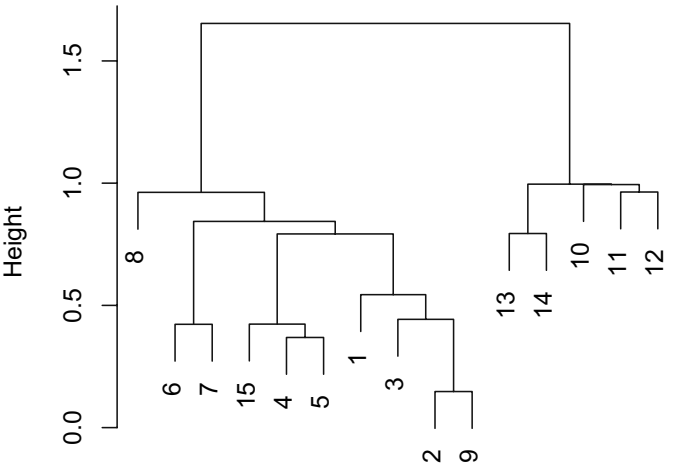


Fig. 6.5. Dendrogram for simulated data, using single linkage algorithm. Two clusters are identified; however, not all observations are classified correctly.

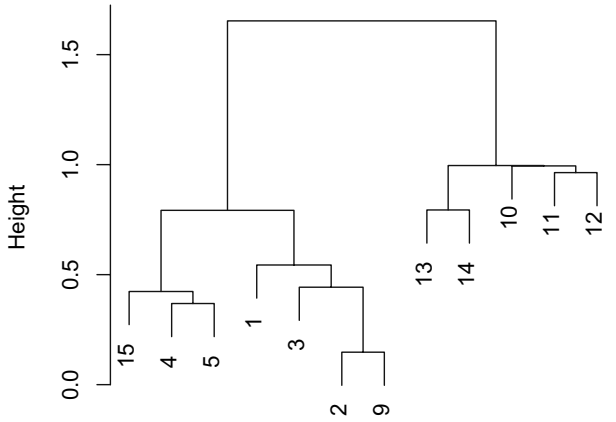


Fig. 6.6. Dendrogram of sharpened data set. The observations are numbered according to their original indices.

maximum value of the function is considered; if that *maximum* isn't large, the voxel is discarded. Finally, the clustering algorithms (K means and the hierarchical group-average agglomerative method, or "Ward's method") are applied on the reduced data sets.

A drawback of the K means algorithm, as we have already seen, is that it requires that the number of classes be known, or at least specifiable, a priori. This will not usually be the case in fMRI studies, in particular the exploratory ones. Ward's method does not have this problem, as hierarchical algorithms in general allow the user to consider different numbers of clusters; however, a choice as to the "best" number still needs to be made, and this choice is perforce subjective in nature, although one may try to make it more objective. Goutte et al. (1999) propose one way of doing this, by considering changes in the "within-class inertia," that is, the average distance of a voxel from the center of its cluster. As this inertia decreases, the clusters become more homogeneous. But, the measure also decreases simply because the number of clusters increases. Hence, on its own the within-class inertia is not sufficient; one could look at a plot of the inertia against the number of clusters and use this like a scree plot, seeking a point where the decline levels off, however this would again be a subjective way of picking the number of clusters. The authors suggest instead that one look at the curvature (second derivative) and find "bumps" or anomalously high values, an indication that, when moving from k to $k-1$ clusters, a dramatic change in homogeneity occurrs.

Using this combination of ideas, Goutte et al. (1999) succeed in finding physiologically and functionally meaningful clusters on data collected from a single subject, with the very evocative flashing checkerboard paradigm. While this result is encouraging for the potential of simple spatiotemporal analyses to produce reasonable models of the working brain, a more convincing validation would be carried out using tasks that are inferentially challenging. Of course, as we have seen already – and will continue to see – new statistical methodologies tend to be tested first using stimuli that evoke a strong and well-localized response; unfortunately the next step, namely testing them on more complicated tasks, is often overlooked.

Lu et al. (2003) introduce a "region growing method" for the problem; these methods are popular in the image segmentation literature and hold some advantages for fMRI data as well, since they exploit the idea that true activation tends to occur in clustered regions. In contrast with clustering techniques, however, the number of clusters does not need to be known in advance; rather this is one output of the algorithm.

The method described by Lu et al. (2003) has two steps: region growing and region selection. In the first step, region growing, each voxel acts as a "seed" voxel around which a region is grown. Thus, there are initially as many regions, or clusters, as there are voxels, and obviously any particular voxel may belong to more than one region. Voxels that neighbor the seed are added to its region according to a homogeneity criterion, which in this case is taken to be the ordinary Pearson correlation between the voxel time series and

the average time course of the voxels already in that region; if the correlation is above a specified threshold, the new voxel is added to the region, otherwise it isn't. For each seed this process is continued until no more voxels can be added to its defined region.

In the region selection step the grown regions are pared down successively. First, the largest region (in terms of number of voxels) is located; regions that are derived from the voxels in this region are eliminated. Then, one continues finding the current largest region and discarding regions that are grown from its constituent voxels until all remaining regions are above a specified threshold in size.

Finally, some postprocessing of the regions left after the second step may be necessary, for example, merging overlapping regions into one bigger cluster. The authors also suggest removing regions that do not seem to be task-related, either due to anatomical considerations or lack of correlation with the experimental paradigm, but this seems rather dubious in that such an approach will bias the resultant statistical map.

6.3.2 Direct Modeling

The second approach to full-blown spatiotemporal fitting of fMRI data is direct modeling. Here, both temporal and spatial components are modeled explicitly. Due to the computational complexity of this approach, it has only become feasible relatively recently. It still is not the most prevalent analysis even though from a statistical perspective it is, perhaps, the most complete and correct. Within the rubric of "direct models" are included models based on regression and wavelets, and Bayesian models, among others. In short, a wide variety of analyses are available.

Purdon et al. (2001) attack the analysis of fMRI data as a spatiotemporal system identification problem, seeking the relationship between the input (a sensory or cognitive stimulus) and the observed output (the measured fMRI response to that stimulus). Their model is "physiologically inspired" (p. 912), based on BOLD fMRI studies of animals and on simulations, and it comprises three components: the hemodynamic response, the noise (which itself is made up of physiologic noise and scanner noise), and a drift. Model-fitting uses local spatial regularization on the parameters so that estimation is not done on a voxel by voxel basis.

The *hemodynamic response* part of the model follows known or assumed patterns of change in the blood oxygenation following stimulus presentation, as described in previous chapters. In particular, Purdon et al. (2001) allow for the characteristic delay, peak, decline, and undershoot of the BOLD response.

The *noise* part of the model has two components: white for the scanner noise and AR(1) for the low-frequency physiological noise. Equivalently, as the authors note, this formulation of the noise can be thought of as ARMA(1,1).

The *drift* part of the model is linear in time, and accounts for slow drifts in the external field, as well as small amounts of motion that were not corrected in any motion correction step.

At each voxel there are noise parameters and signal parameters that need to be estimated. Denote the entire vector of parameters at voxel v by $\boldsymbol{\theta}_v$. The authors set an overall fitting criterion that is separable, i.e.,

$$J(\boldsymbol{\theta}) = \sum_{v=1}^{V} J_v(\boldsymbol{\theta}_v);$$

the criterion at a given voxel v is a spatially locally weighted log-likelihood,

$$J_v(\boldsymbol{\theta}_v) = \sum_{q \in N_v} K_{v-q}^h L_q(\boldsymbol{\theta}_v).$$

Here, $L_q(\boldsymbol{\theta}_v)$ is the Gaussian log-likelihood based on the time series at voxel q, K is a kernel function, h controls the size of the neighborhood on which the kernel function is concentrated, N_v is the neighborhood of voxel v. Finally, estimation proceeds iteratively, alternating between the noise and signal parameters. Thus these are separately spatially regularized.

An interesting feature of the approach is that, since noise and signal parameters are regularized separately, one can choose to focus on subspaces of the whole parameter space and hence improve estimation of those subspaces by borrowing strength from neighbors. It might be computationally expensive, and time-consuming, to regularize fully on all parameters, and the iterative procedure affords the user some measure of control. In the article Purdon et al. (2001) concentrate on regularizing the AR(1) noise parameters, for instance, but they note that their algorithm can be arbitrarily partitioned, "allowing different parameter subsets to be regularized with different degrees of spatial smoothing" (p. 915).

Compared to the commonly used approach of spatially smoothing the fMRI data prior to statistical analysis, the authors show on both simulated and real data that local regularization is able to better estimate both the signal and the noise. In particular, estimates of noise are smoother under the analysis of Purdon and colleagues; at the same time, estimates of the signal aren't blurred, as they tend to be with presmoothing. Note that the estimation of the noise is improved relative to either presmoothing or no regularization due to the regularization on the noise parameters. In later work (Long et al., 2004) the spatial component is incorporated via wavelets.

A similar idea of local spatial regularization is proposed by Katanoda et al. (2002). These authors extend the basic linear model described in the previous chapter to involve a multiple regression. In this multiple regression the model for each voxel involves the time series of its neighbors, as well as its own time series. The model is built as follows. Assume for simplicity a simple "on-off" block design experimental paradigm. Recall the simple linear model for the response for voxel v at time t:

$$y_{v,t} = \beta_v x(t) + \epsilon_{v,t},$$

where $x(t)$ is 0 or 1 depending on whether the task was "off" or "on" at time t. Now, consider instead the multiple regression for voxel v that includes also neighbors $v_1, \ldots, v_p$:

$$\begin{pmatrix} y_{v,t} \\ \vdots \\ y_{v_p,t} \end{pmatrix} = \beta_v \begin{pmatrix} x(t) \\ \vdots \\ x(t) \end{pmatrix} + \begin{pmatrix} \epsilon_{v,t} \\ \vdots \\ \epsilon_{v_p,t} \end{pmatrix}$$

Note that this model includes as dependent variables the time courses for the voxel v of interest, and also of its neighbors. Since each voxel has different neighbors, each model is different and the estimated β coefficients are different as well. Hence the spatial aspect of the data is (at least partially) accounted for by using local neighborhood structure to fit the coefficient at each voxel. In addition, Katanoda et al. posit a separable model for the spatial and temporal correlation, namely $\mathrm{Cov}(\epsilon_{v_a,t}, \epsilon_{v_b,s}) = \sigma_a \sigma_b \rho_1(a,b)\rho_2(t-s)$, where σ_a^2 is the variance of $\epsilon_{v_a,t}$ (and likewise σ_b^2); $\rho_1(a,b)$ is the correlation between $\epsilon_{v_a,t}$ and $\epsilon_{v_b,t}$; and $\rho_2(t-s)$ is the temporal autocorrelation of ϵ at lag $t-s$. In this formulation, the temporal and spatial components are separately modeled, which is a considerable simplification.

Estimation of model parameters is carried out in the frequency domain, i.e., the data are subjected to a Fourier transform and then analyzed. This manipulation, as we have seen previously, can simplify some of the calculations required for estimating the various parameters. Generalized least squares (GLS), taking account of neighboring structure, is used for obtaining the parameter estimates.

The authors compare their "neighborhood GLS" to ordinary least squares (OLS) and GLS carried out on a voxel by voxel basis, as well as a "neighborhood OLS" method. On simulated data both neighborhood methods perform similarly. When areas of activation are large, the neighborhood algorithms tend to have better power than the voxel-wise methods. The pattern of activation (spherical or cubic in the simulations) is also, more surprisingly, relevant to the performance of the different algorithms. Interestingly, the neighborhood methods, which in theory should be borrowing strength across voxels, do not always have better power than voxel-wise methods. And incorporating the spatial and temporal autocorrelations via the GLS analysis does not always lead to improvement over OLS.

On real data from a simple finger tapping experiment, the neighborhood GLS model does seem to give better results than the single voxel methods, in the sense of discovering larger, more coherent areas of activation, and fewer scattered, small areas (which are often assumed by the scientists to be spurious). It also seems to outperform the neighborhood OLS, from the opposite direction, namely finding tighter, more focused regions of activation. By contrast, the regions detected by the neighborhood OLS approach appear quite diffuse (see, for example, their Figure 4).

Continuing in the least squares vein, McIntosh et al. (2004) propose the use of partial least squares specifically for event-related fMRI experiments, although their method is apparently suitable for block designs as well, with some modifications. Partial least squares is a multivariate extension of multiple linear regression. "Partial" refers to computing the best least squares fit but only to part of a covariance (correlation) matrix. The "part" in question is specified to be related to the experimental design or to subject behavior, depending on the goals of the study and the analysis.

Consider a multiple linear regression in general form, $Y = X\beta + \epsilon$. Partial least squares is related to other multivariate extensions of this basic model, such as discriminant analysis, principal components analysis (PCA), and canonical correlation analysis (CCA), as follows. All search for so-called *prediction functions*, functions of the dependent or independent variables that reveal multivariate structures in the data. Whereas most of these methods extract factors from the Y^TY or X^TX matrices only, and not from the cross-product matrices, with partial least squares the factors are extracted from Y^TXX^TY, i.e., they involve simultaneously the dependent and independent variables. This allows for many more factors than when either set of variables is taken on its own.

The implementation of the spatiotemporal partial least squares (ST-PLS) advocated by McIntosh and colleagues for fMRI has two core elements: (i) rearranging the data array into a matrix to reflect the multivariate nature of the PLS approach; (ii) singular value decomposition on the rearranged matrix, or some transformation of it, to extract the factors. For simplicity, suppose that there is a single subject, an experiment with c conditions and k trials per condition. In the data-rearranging step, a matrix is created that has a row for each of the $c \times k$ combinations of condition and trial. The columns are the measured signal at each voxel and each time point. Starting with the first voxel, the first t columns make up the time series for that voxel; the next t columns are the time series for the second voxel; and so on. Hence with v voxels and t time points, there are $v \times t$ columns in the matrix. This data matrix contains both spatial and temporal information in the columns and information about the experimental design in the rows.

Two versions of ST-PLS are presented. In the first the data matrix is mean-centered and factors are extracted by applying a singular value decomposition to this new matrix. In the second the original data matrix is transformed by a set of orthonormal contrasts representing effects of interest. The covariance matrix of these contrasts is then calculated and the singular value decomposition is applied to that matrix instead.

Taking the first approach as exemplar (there is in practice little difference between the two versions of ST-PLS), the result of the singular value decomposition is a set of factors, sometimes called "latent variables." These factors relate brain activity and experimental design due to the layout of the rearranged data matrix. Two sets of weights identify (i) groups of voxels that are most related to the effects expressed by the different factors, and (ii) the

degree to which different tasks are related to patterns of BOLD response. In addition, the analysis yields "brain scores" and "design scores" for each factor. Brain scores indicate the strength with which different subjects (in a multiple subject study) express the patterns detected by each factor, while design scores do the same for tasks.

One of the difficulties of factor analysis, principal component analysis, and the like, is the choice of number of factors to retain. This is also, of course, an issue for the ST-PLS algorithm. The authors address the problem via permutation testing, to decide on the number of significant factors, and bootstrap to evaluate the significance of weights on the significant factors.

McIntosh et al. (2004) validate their approach on a multiple subject study involving two types of task, one of visual processing and one of auditory processing. They conduct two types of ST-PLS: task analysis to detect spatiotemporal patterns in the stimulus response; and behavioral analysis to examine the spatiotemporal structure of brain behavior and reaction time on the tasks. From the first analysis they find two significant factors. The first is attributed to the main effect of task versus rest; the temporal pattern indicates peak activity 6-10 seconds after stimulus presentation, consistent with what has been found in many other studies; the spatial pattern reveals those areas of the brain that are most similar to the expressed temporal trend of the hemodynamic response. The second significant factor yields the interaction between type of stimulus (auditory or visual) and condition (task versus baseline). Again, it is possible to interpret this factor in terms of the spatiotemporal patterns in the data.

From the second analysis, the behavioral ST-PLS, only one significant factor is discovered; the authors interpret this factor as the overall correlation of reaction time with brain activation in both tasks. They find clear temporal fluctuations in the correlation pattern. Also coherent areas of slower or faster reaction time, as reflected in the factor weights, can be seen.

The potential of multivariate methods, in particular principal and independent components analyses, which have been heavily used with fMRI data, will be explored more fully in Chapter 7.

Lastly, we turn to Bayesian spatiotemporal inference. Gössl et al. (2001) introduce a series of hierarchical Bayesian models that can account for spatial and temporal effects individually or simultaneously. As their base likelihoods they consider either the usual linear model for the response as a function of baseline drift and the stimulus convolved with the hemodynamic response function, or a state space model (Gössl et al., 2000). We'll consider the latter model as an example of their general approach.

The state space model is written as

$$y_{it} = a_{it} + z_{it} b_{it} + \epsilon_{it}$$

for voxel i at time t, where a_{it} is the baseline trend, z_{it} is the stimulus convolved with the HRF, b_{it} is the activation effect; $\epsilon_{it} \sim N(0, \sigma_i^2)$ and

$$a_{it} = 2a_{it-1} - a_{it-2} + \zeta_{it}, \quad \zeta_{it} \sim N(0, \sigma^2_{\zeta_i})$$

$$b_{it} = 2b_{it-1} - b_{it-2} + \eta_{it}, \quad \eta_{it} \sim N(0, \sigma^2_{\eta_i}).$$

Once the models have been specified, it remains to set the prior distributions. Of course there are many ways to do this and still incorporate spatiotemporal structure. For relative ease of computation, Gössl et al. recommend imposing spatial or spatiotemporal Markov random field priors for the parameters at the second stage of the hierarchical model. Take for instance the stimulus effect parameters b_{it} and let $\mathbf{b}_i$ be the vector attributed to voxel i. Under the assumptions outlined above, it is possible to write

$$\pi(\mathbf{b}_i|\lambda_i) \propto \exp\left(-\frac{1}{2}\lambda_i \mathbf{b}_i^T Q \mathbf{b}_i\right),$$

where λ_i is the precision. The Q matrix imposes smoothness on the $\mathbf{b}_i$ vector over time. These are both derived from the second-order random walk model for b_{it}. The result from this step of the analysis is coefficients at voxel i that vary slowly and smoothly over time.

For the spatiotemporal model, Gössl et al. (2001) introduce both additive and nonadditive approaches. In the additive model, the effect b_{it} is written as

$$b_{it} = \alpha_i + \beta_{it},$$

where α_i is constant over time but not location, and β_{it} varies over both time and location. The prior for α_i is a spatial smoothness prior of the form

$$\pi(\alpha_i|\lambda) \propto \exp\left\{-\frac{1}{2}\lambda \sum_{i\sim j}(\alpha_i - \alpha_j)^2\right\},$$

where $i \sim j$ denotes the neighbors of voxel i. For β_{it} the temporal random walk prior $\pi(\mathbf{b}_i|\lambda_i)$ defined above is used, with suitable modifications.

As noted by Gössl et al., this still models on an individual voxel basis, hence it does not take full advantage of the strengths of the Bayesian framework. These are more fully realized in the nonadditive model, in which the spatial and temporal random field priors are combined via the Kronecker product of their respective precision matrices. This yields

$$\pi(b|\lambda) \propto \exp\left\{-\frac{1}{2}\lambda \sum_{i\sim j}\sum_t (\Delta^2 b_{it} - \Delta^2 b_{jt})^2\right\};$$

here $\Delta^2 b_{it}$ is the second differences of b_{it}, or $\Delta^2 b_{it} = b_{it} - 2b_{it-1} + b_{it-2}$. This prior allows both spatial and temporal smoothness to be enforced, so that there are not too sudden transitions in either dimension. Further fine tuning of the priors is suggested by Gössl and colleagues to enhance smoothness

in space or time over that which occurs naturally through the random field specification.

Finally, whatever prior model is used, hyperpriors on the parameters λ_i and σ_i^2 are set; as is common practice, these are taken to be conjugate but relatively diffuse, while still ensuring propriety of the resultant posterior distributions. The use of conjugate priors means that the full conditional distributions have simple forms, and Gibbs sampling will give the required posteriors.

Penny et al. (2005) formulate a similar Bayesian model to that of Gössl et al. (2001). At voxel i they assume a general linear model with autoregressive errors, i.e.,

$$y_i = Xw_i + \epsilon_i$$

and

$$\epsilon_i = E_i a_i + z_i.$$

In these models, the regression coefficients are given by w_i and a_i, X is a matrix of covariates, E_i is a matrix of lags, z_i is normal with mean zero and precision λ_i.

The prior for the regression coefficients w is normal with mean zero and (spatial) precision that is unique for each covariate; furthermore, the prior factors over the different regressors, so that covariates may have differing amounts of smoothness. The priors on the spatial precisions and the λs also factor, with each component being a diffuse gamma. Note that this last factorization therefore implicitly assumes that the variances of neighboring voxels are independent. While this is most likely an unrealistic assumption, it underlies the standard general linear model analysis, as we have seen and as the authors also point out. Lastly, a factorized (again over voxels) diffuse normal prior is placed on the autoregressive parameters.

The major difference between the work of Penny et al. and the work of Gössl et al. is in the form of the priors. The random field priors proposed by the latter result in relatively computationally intensive sampling procedures, even though the straightforward Gibbs sampler can be used. By contrast, the factorized priors of Penny et al. allow for the development of approximate posterior distributions via the variational Bayes theory, from which the relevant parameters can be more efficiently sampled. The drawback of course is that we are then sampling from an approximation to the posterior, rather than the true posterior itself. It is not clear how important this is in practice; results reported by Penny et al. on simulated and real data indicate that their procedure is effective at picking out true (in the case of simulated data) or reasonable (in the case of the real data) areas of activation.

There is still much scope for the application of Bayesian methods in fMRI. We return to this topic in Chapter 9.

6.4 Software Issues

Unlike the basic linear model analysis described in the previous chapter, which is featured prominently in all of the major software packages (see Appendix A), temporal, spatial, and spatiotemporal models are still mostly the output of particular laboratories and researchers. As such, they have largely not yet been codified into any of the standard analysis packages. Instead, individual researchers tend to write their own code, usually in MATLAB (Mathworks Inc.) or C, to implement the algorithms they have developed. Due to the collaborative nature of fMRI work, this code is often freely available from the authors. Readers are encouraged to seek out the programs that are of interest to them directly from the investigators.

7

Multivariate Approaches

In this chapter we look at fMRI data from the multivariate perspectives of component and correlation analyses. The former include principal components analysis (PCA) and independent components analysis (ICA); the latter include canonical correlation analysis and maximum correlation analysis. ICA is by far the most popular of these methods. All of the procedures of this chapter share the feature that they are "data driven" rather than "model" or "hypothesis driven." The implication is that the researcher does not need to specify a priori all the possible effects and behaviors of interest; indeed, the components that are produced as a result of the various decompositions will often lend themselves to unexpected interpretations. For instance, in addition to components that are task-related, with associated time courses that follow the experimental paradigm (and which could be predicted in advance), the methods can discover components that correspond to transient effects, and even some that don't relate specifically to the task, but are consistently found across subjects or cluster spatially, indicating their "veracity" as elements of interest (see, for instance, Calhoun et al. 2001a). This seems to be especially true of ICA, and is perhaps one explanation for its popularity (Hu et al., 2005).

In contrast to the methods we explored in the previous chapters, which were essentially voxel-based even when spatial structure was taken into account, the techniques of the current chapter aim to find or characterize the multivariate nature of the data, seeking out subspaces or high dimensional directions of common behavior. These directions may be in space, time or both, depending on how the analysis is performed. For example, a multivariate analysis might uncover regions of the brain (clusters of voxels) with similar temporal behavior; such temporal behavior may be related to the experimental task, but could also arise from noise or other artifacts. This is, evidently, another way of describing the spatial and temporal dependencies in the brain. As such, these multivariate methods are a potentially useful alternative to the models of the previous chapter.

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_7,

7.1 Description of Methods

7.1.1 Principal Components Analysis

Principal components analysis (PCA) is a classical multivariate statistical tool, developed primarily by Hotelling (1933). The goal of PCA is to find linear combinations of the original variables that parsimoniously describe the dependence structure of the data. Given random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ drawn from a multivariate distribution with mean the vector $\boldsymbol{\mu}$ and variance the matrix Σ of rank $r \leq p$, the usual estimator of Σ is the sample covariance matrix S. The first principal component is the linear combination $\mathbf{a}_1^T\mathbf{X}$ with maximal sample variance $\mathbf{a}_1^T S\mathbf{a}_1$ among all coefficient vectors with length 1 (i.e., such that $\mathbf{a}_1^T\mathbf{a}_1 = 1$). It is easy to show, using a Lagrange multiplier argument, that $\mathbf{a}_1$ is the eigenvector corresponding to the largest eigenvalue of S (Morrison, 1978). Furthermore, that largest eigenvalue is the variance $\mathbf{a}_1^T S\mathbf{a}_1$ of the first principal component.

The second principal component is the linear combination $\mathbf{a}_2^T\mathbf{X}$ with maximal variance subject to the constraints that $\mathbf{a}_2^T\mathbf{a}_2 = 1$ and $\mathbf{a}_1^T\mathbf{a}_2 = 0$, that is, the coefficient vector has length 1 and is orthogonal to the coefficient vector of the first principal component. Similar to the results for the first component, it can be shown that the second component is the eigenvector corresponding to the second largest eigenvalue of S, and that the sample variance of the second component is that eigenvalue. Subsequent principal components are obtained in this way, such that the coefficients of each successive linear combination are orthogonal with all the others.

A consequence of the derivation is that the first principal component explains that largest proportion of the variance in the sample, the second principal component explains a smaller proportion than the first but the largest among all remaining components, and so forth. That is, each successive component explains a smaller portion of the total sample variance. Another, geometric, interpretation of the principal components is also commonly exploited. Under this interpretation we think of the first principal component as being the principal axis of the p-dimensional scatter cloud of the data (the "longest" direction of the cloud, that is, the one with most variability); this defines a rotation of the data from their original orientation. The second principal component is chosen from the remaining $p - 1$ minor axes to be orthogonal to this principal axis, and it is the longest of the remaining axes in the $p - 1$-dimensional subspace. Consecutive components are chosen in similar fashion, as orthogonal axes to those already defined. Once all have been picked the data have been rotated into a new set of coordinates, given by the values in the appropriate eigenvectors (see Morrison 1978, for more detail). Figure 7.1 shows 500 points drawn from a bivariate normal distribution, with vector mean zero, variance 1 in each direction and correlation 0.7. The solid line is the first principal axis, capturing the direction of greatest variability. The dashed line is the second principal axis, which is orthogonal to the first and captures the direction of greatest remaining variability.

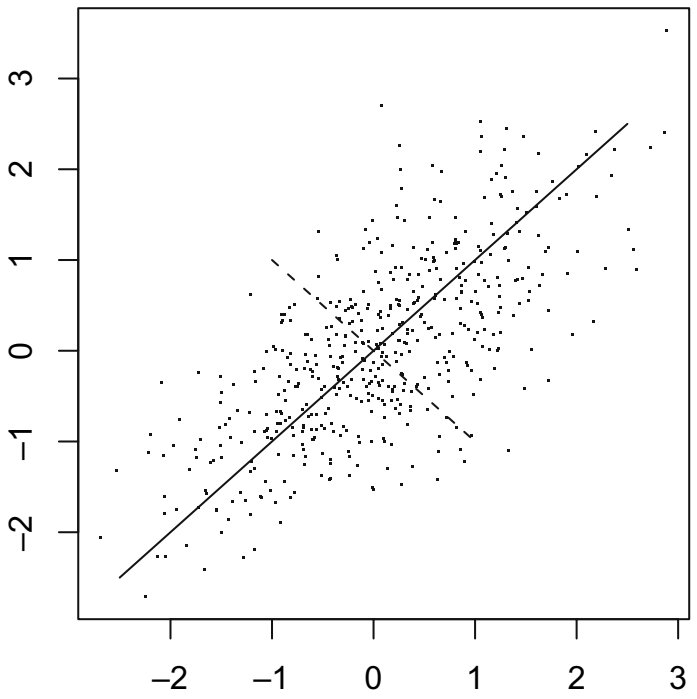


Fig. 7.1. 500 points drawn at random from bivariate normal with correlation 0.7. The solid line shows the first principal axis and the dashed line shows the second.

7.1.2 Independent Components Analysis

Independent components analysis (ICA) is a modern multivariate technique that has become very popular in recent years. The prototypical ICA problem is the so-called "cocktail party problem" (Hyvärinen and Oja, 2000), in which one wishes to take the general hubbub of a party (for instance, the various people who are speaking all at once) and identify its independent sources (that is, separate out the speech of each of the guests from that of the others), using microphones placed around the room. More generally, ICA looks for linear combinations of the original data, assumed to be non-normal, that are maximally independent. In this, it is similar to PCA, but with a number of important differences. First, PCA is based on the covariance of the data, whereas ICA uses also information available in higher moments (hence it doesn't assume normality, a typical PCA assumption). Second, an explicit goal of PCA is dimension reduction; this is not an aim of ICA and indeed the dimension may be increased if the number of "sources" identified is greater than the dimension of the original data.

A common approach for ICA is the so-called "noise-free model" for the random vector $\mathbf{X}$. Here, one wishes to estimate the model

$$\mathbf{X} = A\mathbf{s},$$

where $\mathbf{s}$ contains independent latent variables and A is a mixing matrix that defines how the latent variables combine to make up the observed vector $\mathbf{X}$. Note that only $\mathbf{X}$ is observed; both $\mathbf{s}$ and A need to be estimated. Often, the number of mixtures (the observed $\mathbf{X}$ values) is the same as the number of independent components (although this isn't necessary), so that A is a square matrix. In the cocktail party example, this would mean that the number of microphones in the room and the number of guests at the party are the same.

On the face of it, this problem doesn't admit a solution, since neither the latent variables $\mathbf{s}$ nor the mixing weights in A are observed or known. But, in fact, the noise-free model is identifiable if (i) the independent components are non-normal (one of the components may be normal but the rest may not); (ii) the dimension of the data $\mathbf{X}$ is at least as large as the number of independent components; (iii) the matrix A is of full column rank. Additionally, for uniqueness of results it is usually assumed that $\mathbf{X}$ and $\mathbf{s}$ are centered and that $\mathbf{s}$ has variance 1. In contrast to PCA the resultant independent components are not naturally ordered, although it is possible to order them (Hyvärinen and Oja, 2000).

The assumption of non-normality is key, as shown in Hyvärinen and Oja (2000). For, if the independent components are normally distributed and the mixing matrix is orthogonal, then the joint density of the components of $\mathbf{X}$ is a symmetric (spherical) normal and A is not uniquely identifiable. Estimation of A (or its inverse) is based on maximizing the non-normality of the independent components that would result; different measures of non-normality (for instance, kurtosis) can be used toward this goal. In general terms, an ICA algorithm proceeds as follows: We are looking for a linear combination of the $\mathbf{X}_i$, call it $\mathbf{y} = \mathbf{w}^T\mathbf{X}$. If $\mathbf{w}$ were a row of A^{-1}, then $\mathbf{y}$ would actually be one of the independent components. Defining $\mathbf{z} = A^T\mathbf{w}$ it is easy to show that $\mathbf{y} = \mathbf{z}^T\mathbf{s}$, so that $\mathbf{y}$ can also be thought of as a linear combination of the latent variables. By the Central Limit Theorem, $\mathbf{y}$ is "more normal" than any of the $\mathbf{s}_i$ (as a linear combination of independent random variables); $\mathbf{y}$ is "least normal" therefore when it exactly equals one of the $\mathbf{s}_i$ and $\mathbf{z}$ has only one nonzero element. The goal becomes to find the vector $\mathbf{w}$ that maximizes the non-normality of $\mathbf{y} = \mathbf{w}^T\mathbf{X}$, and this in turn gives the first independent component. Other components are found by maximizing the non-normality in successive (uncorrelated) subspaces.

Application of ICA involves two main preprocessing steps: data reduction and whitening. PCA is often used for the data reduction step in such a way that the majority of the variability in the data is captured; the rationale is that even if the number of required components is large, it will still be smaller than either the number of time points or the number of voxels in a typical fMRI study. PCA is also used to prewhiten the data. Whitening transforms the search space to be orthogonal. A variety of algorithms, which constrain the results to be uncorrelated and take advantage of the higher order features of the data, are then available to actually perform the ICA.

7.1.3 Canonical Correlation Analysis

Canonical correlation analysis, or CCA (Hotelling, 1936), is a way of quantifying the correlation between sets of variables. More specifically, suppose we have two random vectors, $\mathbf{X}$ and $\mathbf{Y}$. With canonical correlation analysis, we seek the linear combinations $\mathbf{a}_1^T\mathbf{X}$ and $\mathbf{b}_1^T\mathbf{Y}$ so that the correlation $\text{cor}(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y})$ is maximized. The random vectors $\mathbf{a}_1^T\mathbf{X}$ and $\mathbf{b}_1^T\mathbf{Y}$ are the first pair of canonical variables. We then seek linear combinations $\mathbf{a}_2^T\mathbf{X}$, $\mathbf{b}_2^T\mathbf{Y}$ maximizing the correlation subject to the constraint that they be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. The procedure is iterated, with each successive pair of canonical variables being uncorrelated with the previous pairs, up to k pairs, where k is the dimension of the smaller of $\mathbf{X}$ and $\mathbf{Y}$.

Let the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ be

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

in the obvious notation for the submatrices. Then it can be shown that the coefficients of the jth pair are given by

$$(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - c_j\Sigma_{XX})\mathbf{a}_j = 0$$

and

$$(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - c_j\Sigma_{YY})\mathbf{b}_j = 0,$$

where c_j is the jth largest root (the eigenvalue) of the equations $|\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \lambda\Sigma_{XX}| = 0$ or $|\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \lambda\Sigma_{YY}| = 0$.

If the eigenvalues are unique, then clearly the coefficient vectors $\mathbf{a}_j$ and $\mathbf{b}_j$ will be as well, and the corresponding linear combinations will be uncorrelated with other canonical variates. The method of canonical correlation analysis is a useful exploratory tool and has been used as such in the behavioral and social sciences, historically. An advantage of the approach is that the transformation of the original variables into the new scale reveals the correlation structure between the sets; this would not always be apparent if the simple pairwise correlations between components of the two sets were calculated instead.

7.2 Multivariate Analyses

For the multivariate analyses described above, an important and relevant feature of fMRI data is that they can be described in two computationally and mathematically equivalent ways: considering the time course of each voxel, or considering the level of activation in the brain at each time point (Andersen et al., 1999). The distinction revolves around whether the multivariate analysis focuses on the common temporal or spatial structure in the data. The number of voxels in either a two-dimensional slice or a three-dimensional

volume is much larger than the number of time points at which data are collected and observed, and it is the latter that determines the amount of independent information available for detecting common structures in the data. Hence, considerable computational savings can be achieved by calculating, for instance, principal components on the temporal scale and then transforming to the spatial (Andersen et al., 1999). This computational duality between the spatial and the temporal domains implies that the large number of voxels does not present as much of a challenge to inference as it does for standard univariate methods. In the typical univariate approach calculations are carried out on a voxel by voxel basis (or, perhaps, over conglomerations of voxels, but still the number of locations is large); in the multivariate approach, calculations can be carried out using time, rather than space, as the variable of interest, and hence the number of voxels is not as important in determining the computational burden.

7.2.1 Principal Components

Much of the research on the use of PCA for functional neuroimaging has been done in the setting of positron emission tomography (PET), rather than fMRI. See, for example, Friston et al. (1993),in which principal components analysis is performed on the voxel time series, thereby elucidating spatial regions with common patterns of temporal behavior. This simple use of PCA sheds light on functional connectivities in the data (Friston et al., 2000b), a question of much interest in current fMRI research. Since most research on PCA for fMRI data has been rather specialized, the rest of this section is devoted to those topics: nonlinear PCA, functional PCA, kernel PCA, and the choice of the number of components.

Nonlinear PCA

In this section we consider a nonlinear version of the most basic principal components analysis, introduced in the fMRI context by Friston and colleagues (Friston et al., 1999b; Friston et al., 2000b).

The motivation for the extension to the usual PCA proposed by Friston and colleagues was the observation that the conventional analysis imposes biologically implausible constraints on the solutions. The first constraint is that the decomposition be into linearly separable components; the second constraint is that the components be orthogonal and account successively for the greatest amount of remaining variance. Friston et al. (2000b)note that the first constraint, of linearity, is the more severe, since it precludes the possibility of interactions among brain systems. By contrast, cognitive neuroscientists broadly believe that brain systems do interact with each other, and in quite complex ways at that. Hence the desirability of a decomposition that allows for interactions and other more complex relationships among the components.

In essence, nonlinear PCA takes the idea of finding the linear combination through the data that explains the most unexplained variability, while being orthogonal to already-detected directions, and replaces it with a general curve; that is, instead of looking for a linear combination of the original variables, one seeks a general function such that the average distance of points from the curve is minimized. Friston and colleagues consider only a particularly simple extension, namely, the linearization up to second order – via Taylor expansion – of the nonlinear function.

Let $f(\cdot)$ denote the general function. Assume there is a small number J of "input sources" contributing to the observed data, and let there be in total n voxels. The observed data at voxel i over time is taken to be a function of the sources, $y_i(t) = f_i[s(t)]$. Taylor expand this up to second order around some "expected value" $\bar{s}(t)$, to obtain

$$y_i(t) \approx f_i(\bar{s}) + \sum_j \frac{\partial f_i}{\partial u_j} u_j + \sum_{j,k} \frac{\partial^2 f_i}{\partial u_j \partial u_k} u_j u_k,$$

where $u(t) = s(t) - \bar{s}(t)$. This system is solved in a neural network framework, casting it as a general linear model.

The "first order modes" correspond to the typical components found by a PCA. The "second order modes" allow for nonadditive relationships between components. Friston et al. (1999b) demonstrate the method on a task that combines motion and color processing. They find two main principal components, one that is mostly an effect of motion (accentuated by the presence of color cues) and one that is mostly an effect of color processing. These components in turn are localized to the appropriate anatomical regions for the respective types of processing. The second order mode shows where the two types of processing interact.

Functional PCA

Viviani et al. (2005) take the perspective that fMRI data can be considered as functional data in the sense of Ramsay and Silverman (1997), that is, the voxel time courses are taken as functions that evolve over time, such that each time course is a continuous function, an integral whole, which is sampled at discrete TR intervals. These are usually estimated by fitting with a set of basis functions. Many standard statistical techniques, such as analysis of variance, have functional counterparts; for analysis of variance, as an example, the data are curves observed at each level of the covariates, rather than single observations. Likewise with functional PCA, the analysis is carried out in a way that treats the data at each voxel as a continuous function of time, and the eigenanalysis is performed on these functions. Viviani et al. (2005) examine the usefulness of functional PCA as an exploratory tool.

Functional PCA therefore requires two steps:

1. *Smoothing (estimating) the time series.* Here, each voxel time series is estimated or smoothed via a series of basis functions. As in all such applications, there are decisions to be made regarding the choice of basis function (Fourier, spline, B-spline, and so on) and value of the smoothing parameter. Advantages and disadvantages of some of these choices specifically for functional PCA of fMRI data are discussed in Viviani et al. (2005). Regarding smoothing, if no smoothing is applied, there will be little or no difference between ordinary PCA and functional PCA, since the time series functions will be simple interpolation. On the other hand, oversmoothing will eliminate the signals of interest. Hence it is crucial to pick the amount of smoothing carefully. The choice of basis function, as usual, is less critical.
2. *PCA on the estimated functions.* Once the voxel time series have been estimated or smoothed, PCA is then applied on the resultant functions. More specifically, each voxel time series is expressed as a linear combination of a small number of basis functions; the PCA decomposition is carried out in the usual way on these linear combinations.

The authors demonstrate the capabilities of functional PCA on three case studies, one of working memory (block design), one of episodic memory (block design) and one of finger tapping (event-related design). In all three cases the first component of the functional PCA captures important features of the data related to the experimental design, that is, the signals of interest. Ordinary PCA picks these up in the second component, if at all; the first component mostly seems to reflect noise, or at any rate is not easily interpretable. Most of the variance is explained by one component in the functional approach; the cutoff for the number of components in the ordinary PCA is much less clear. Finally, the patterns of activation revealed by functional PCA, in all three cases, more closely match the results of previous examinations of the tasks in question, giving them more scientific plausibility. These case studies highlight the potential of the functional data analysis approach for fMRI more broadly; this is an area that is still in need of development.

Kernel PCA

As noted by Thirion and Faugeras (2003) ordinary PCA assumes that the underlying structures of interest are uncorrelated both spatially and temporally. This assumption is not likely to hold for fMRI data. Hence they propose a modification of the basic PCA procedure, in which as a first step each voxel time series is analyzed univariately (for instance, using the general linear model), resulting in a temporal characterization of that voxel's behavior. Next, the voxel-based models are subjected to a multivariate analysis, and specifically in this case "kernel PCA." The goal of the kernel PCA is to preserve the temporal patterns extracted in the first modeling step, something that is not attainable with ordinary PCA because of the assumption that components are uncorrelated.

Let the original data be represented by n voxel time courses, each of length T, denoted $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ (so that each $\mathbf{X}_i$ is in fact a vector). After model-fitting at each voxel univariately, let the resultant data be $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$. Based on either of these representations it is trivial to calculate the covariance or correlation matrix. For a given pair of time series, the correlation will be 1 if they are positively linearly correlated, it will be -1 if they are negatively linearly correlated, and it will be zero if they are not correlated. As the correlation approaches zero from either above or below, the strength of the linear relationship between the two time series decreases.

The kernel PCA approach adopted by Thirion and Faugeras introduces nonlinearity by penalizing values of the correlation coefficient that are far from 1. If $\mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j)$ is the ordinary covariance matrix and $\mathrm{Cor}(\mathbf{X}_i, \mathbf{X}_j)$ is the correlation matrix, then the modified covariance matrix, call it Cov^*, is obtained as $\mathrm{Cov}^*(\mathbf{X}_i, \mathbf{X}_j) = \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j) * \phi(\mathrm{Cor}(\mathbf{X}_i, \mathbf{X}_j))$. Different forms of the function $\phi(\cdot)$ result in different types of penalty. A scaling factor determines after what level of decorrelation will two time courses be considered uncorrelated. Note that in this configuration, the values 1 and -1 need not be treated symmetrically, as they usually are in discussions of correlation. For instance, in their implementation the authors wish to treat strong negative correlations as indicating time courses with distinct patterns of behavior; that is, if the correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$ is close to -1, those time courses exhibit opposite reactions to the stimulus and so should be treated as orthogonal. Principal components analysis is carried out on the new covariance matrix, Cov^*.

With the choices of using a prior temporal model or not, and of performing ordinary PCA or kernel PCA on the time courses, the authors have four methods to compare: No temporal modeling and ordinary PCA corresponds to the standard PCA of fMRI data; temporal modeling with kernel PCA corresponds to the complete analysis path advocated by Thirion and Faugeras. Various challenges arise in applying any of the four methods: the number of components needs to be determined; for the kernel methods the scaling factor needs to be picked or estimated; the methods are computationally intensive, especially the kernel approaches, unless some data screening is first carried out, retaining only voxels that have the most interesting signal (in which case the question remains of how many voxels to keep for the subsequent analysis). Based on simulations, furthermore, it appears that the kernel method may perform poorly if the temporal modeling step is left out. This is an interesting finding that probably deserves more attention.

The complete analysis path of Thirion and Faugeras (2003) bears some similarity to statistical techniques such as "factor analysis regression." There, a dimension reduction is effected via factor analysis, which results in the creation of new variables. These are then used as the explanatory variables in a regression model. We will see this idea again in the next section, under the rubric of "ICA regression." For both of these the decomposition comes first,

followed by the linear model; in the approach of Thirion and Faugeras, the steps are reversed – first the linear model, and then the decomposition.

Picking the Number of Principal Components

Deciding the appropriate number of components in a PCA is a difficult problem in any setting and is still a question of active research in the general statisics literature. Hansen et al. (1999) discuss the issue in the context of fMRI. When PCA is being used as a dimension reduction technique, one would wish to have as much variance as possible contained in as few components as possible. Likewise for exploratory purposes it might be desirable to reduce the data to a small number of principal components, as this will generally be easier to interpret. However, for other uses of PCA it is not clear that having a small number of components is appropriate. Anecdotal evidence indicates that there are situations where one might wish to retain a relatively large number of components, as even in the lower components there is often important signal.

In the general statistical literature there have been many proposals for choosing the number of principal components. Some of these are ad hoc, for example, examining a scree plot of the proportion of variance explained by each successive component and deciding where it "levels off"(Jolliffe, 2002). Others have more of a theoretical basis, for instance, using the bootstrap to understand the PC distribution. As with smoothing for functional PCA, here too there is an essential tension between fidelity to the data (leading to less smoothing or retaining more principal components) and generalizability or interpretability (leading to more smoothing or retaining fewer components). Hansen et al. (1999) take generalizability as the criterion to help guide the choice of the number of principal components to retain.

Given data x from a model parameterized by θ (possibly vector-valued), training and generalization errors are defined as (Hansen et al., 1999)

$$E(\theta) = \frac{1}{N} \sum_{j=1}^{N} \epsilon(x_j|\theta),$$

and

$$G(\theta) = \int d(x)p(x)\epsilon(x|\theta),$$

respectively, where N is the size of the training set, $\epsilon(x|\theta)$ is the cost or error function associated with a particular procedure, and $p(x)$ is the probability density of x. Minimizing the generalization error $G(x)$ yields the desired procedure, in this instance, the number of principal components. Note that $G(x)$ is not observable and hence it must be estimated.

The authors propose two ways of estimating $G(x)$, one which they term "analytical" and the other "empirical." The analytical estimate is based on the relationship

$$\hat{G} = E + \frac{\dim(\theta)}{N},$$

which holds for large training sets (Hansen and Larsen, 1996). The training error is estimated from the training set in a particular split of the data. The empirical estimate is obtained directly from the testing set in a particular split. See Hansen et al. (1999) for explicit formulae using the negative log-likelihood as the cost function and normality assumptions on the data.

Based on two case studies, a finger tapping experiment and a flashing checkerboard study, the authors find that the analytical estimate is too generous, suggesting retention of a relatively large number of components compared to the empirical method. Of the components that are retained by the empirical estimator, some reflect signal that is related to the experimental paradigm, as would be expected.

7.2.2 Independent Components

As noted above, ICA is perhaps the most popular of the multivariate methods currently being used for the analysis of fMRI data. It was introduced in the late 1990s by McKeown et al. (1998), and both temporal and spatial versions have appeared in the literature (see, for example Biswal and Ulmer, 1999, for temporal ICA; Calhoun et al., 2001b, for spatial and temporal ICAs; Calhoun et al., 2003b, for temporal ICA; McKeown et al., 1998, for spatial ICA), although the latter is by far the dominant mode of analysis (Calhoun et al., 2003a).

Suppose there are n voxels, each with a time course of length T. For spatial ICA, $\mathbf{X}$ is a $T \times n$ matrix, strung into a vector, and the signals are the n voxels; hence there are T instances of each signal. The decomposition described above, $\mathbf{X} = A\mathbf{s}$, now indicates that A is a $T \times T$ mixing matrix, and $\mathbf{s}$ is a $T \times n$ matrix that contains the T independent components. The rows of $\mathbf{s}$ are spatially independent images, and the columns of A are spatially independent time courses associated with those images. Recall that both of these are estimated by the ICA algorithm.

For the temporal ICA, everything "flips." Hence, $\mathbf{X}$ is now $n \times T$, the transpose of the matrix used for the spatial version. A is the $n \times n$ mixing matrix and $\mathbf{s}$ is the $n \times T$ matrix of n independent components. Temporally independent time courses are in the rows of $\mathbf{s}$ and their associated temporally independent images are in the columns of A. Since n, the number of voxels, is much larger than T, the length of the fMRI time courses, the temporal ICA is much more computationally intensive.

Although in many circumstances the two approaches will yield similar results, in terms of the extracted time courses and spatial maps of activation, Calhoun et al. (2001b) show that it is possible for them to diverge. They designed four visual paradigms, each consisting of two spatiotemporal components. The two components were either spatially dependent, temporally dependent, spatially and temporally dependent, or spatially and temporally

uncorrelated. On both simulated and real fMRI data, the ICA algorithms perform as would be expected a priori; namely, with strong spatial dependence in the signal, spatial ICA cannot recover the original sources. Likewise, with strong temporal dependence, temporal ICA fails. The algorithms succeed at separating out the sources when there is no dependence in the appropriate dimension. Thus, where the underlying signal has dependence of one type or another (or both), the respective ICAs will both differ from each other and fail to give correct results (when compared to a linear model analysis, for example).

While this result is not at all surprising, it is a useful demonstration of both the strengths and weaknesses of the ICA approach. In particular, the choice of whether to perform the analysis in the spatial or temporal domain should rely on prior knowledge about the tasks performed in a given experiment. For instance, if the areas activated by the tasks are assumed to be spatially independent, spatial ICA should be suitable and informative. It shouldn't be used if the activated areas are assumed to overlap (as one example of possible induced dependence). Of course, researchers won't always have even this level of knowledge when performing an initial statistical analysis, a rather severe limitation of ICA.

Validation of ICA Results

Independent component analysis really refers to a system of different algorithms with the same goal of separating the signal into its independent sources. A disturbing side effect of this fact is that different algorithms could potentially yield different components, and hence differing interpretations of the same data. Furthermore, since most of the algorithms have a stochastic element, different runs of the same algorithm will give different results (Himberg et al., 2004).

Two recent studies have addressed the question of validating ICA for neuroimaging studies. In the first, Esposito et al. (2002) examine two of the popular algorithms used for spatial ICA – Infomax and fixed-point based. Both algorithms attempt to minimize the mutual information in the components of $\mathbf{s}$, i.e., to make those components as independent as possible. They accomplish this in different ways. With Infomax, the output entropy of a neural network with as many outputs as the number of components is adaptively maximized. The outputs are nonlinear functions which must be suited to the particular application at hand; for fMRI data, sigmoidal functions have been found to be effective (McKeown et al., 1998).

The fixed-point algorithm uses the concept of normalized differential entropy, or *negentropy*. One interpretation of negentropy is as a measure of non-normality. Hence maximizing the negentropy finds directions of maximal non-normality in the data, one of the objectives of ICA. Negentropy can be used to estimate the independent components one at a time, or simultaneously.

The algorithm is non-adaptive, using most or all of the available observations at each updating step of the unmixing matrix.

The authors find, using simulated data and real data from simple motor and visual tasks, that both algorithms perform well in the sense of identifying components that reflect true (for simulated data) or expected (for real data) patterns. However, neither algorithm dominates the other; depending on what criterion is used for comparison, either may give "better" results.

Another form of validation has to do with statistical reproducibility and consistency of results. As noted above, there is a stochastic element of the ICA algorithms, which implies that different runs on the same data will give different results. Himberg et al. (2004) use computational and graphical techniques to help evaluate the "algorithmic and statistical reliability" of ICA. By "algorithmic reliability" they mean the consistency with which an algorithm converges to the same solution; this they assess by starting each run with different initial values. By "statistical reliability" they mean an assessment of the significance of the output and this they accomplish via bootstrap.

Visualization plays a role in the assessment since the results from each run are clustered and the clusters plotted using a software called *Icasso* developed by the authors for this purpose (see Himberg et al. 2004, for details). Projection methods allow each estimated component from each run to be represented as a point; a reliably estimated component should produce a cluster of points over the repeated runs. The assumption is that a reliable cluster corresponds to a real component. A real component should therefore be represented by a small, tight cluster that is relatively isolated from other clusters. On the other hand, points that do not correspond to real independent components should be scattered throughout the space, and do not belong to any particular cluster. Himberg et al. (2004) introduce a cluster quality index to help users of *Icasso* locate potentially interesting clusters for further examination. This is defined as the difference between the average intracluster similarities and the average intercluster similarities. As a cluster becomes more diffuse (less compact and isolated), the value of the index decreases; an "ideal cluster" is a single point, and it has an index value of 1.

The authors give two examples of the *Icasso* software methods, one a magnetoencephalogram (MEG) of the whole brain, where the measurements of the brain were disrupted by outside signals such as eye blinks or other muscular activity; the other an fMRI study of finger tapping with the dominant hand. In the MEG study the most reliable components detected by the analysis correspond to sources such as eye movement, biting, and so forth – artifacts that are known to disturb the signal of interest. In the fMRI study the most reliable components correspond to the task, to head motion, and to vascular activity. Finally, both analyses find reliable components that don't have any immediate interpretation, but appear (based on the associated time courses or localization in the brain) to be real (and not just consistently estimated), and hence worthy of further investigation.

ICA Regression

As noted by McKeown (2000) and Hu et al. (2005), and described briefly in the introduction to this chapter, an advantage of ICA is that it can automatically find components that correspond to effects of interest, without these needing to be prespecified. This is useful in particular for experimental studies or clinical groups of patients, where the scientist does not know what to expect in advance, and hence creating the appropriate covariates for the general linear model approach is difficult, if not impossible (Hu et al., 2005). The advantage is not without a price, however, namely that ICA does not have a framework within which to assess the significance of results. We have already seen one way of handling this issue, i.e., validation studies such as that of Himberg et al. (2004). In this section we consider a different idea – combining ICA with the standard fMRI general linear model.

McKeown (2000) exploits the structural similarity between the simple general linear model $X = G\beta + \epsilon$, where G is the design matrix and ϵ are independent, identically distributed normal errors, and the ICA specification $\mathbf{X} = A\mathbf{s}$, where A is the mixing matrix and $\mathbf{s}$ are the (latent) independent components, to write $X = G\beta$. Formally, this is equivalent to the linear model without the error term, but the interpretation and solution differ. Now, instead of being fixed by the experimenter, G corresponds to the mixing matrix A, and hence is found by the ICA algorithm. Furthermore, equating β with $\mathbf{s}$, β is not estimated but instead is obtained from $\mathbf{s} = A^{-1}\mathbf{X}$.

The ICA model as just described derives from the noise-free ICA, implying that there is no residual error. Thus, framing the model in this way does not, in fact, allow for the assessment of statistical significance. To get around this problem, McKeown considers two possibilities. The first is to simply discard some of the independent components as predictors and treat them as random noise. McKeown observes that small components behave "more normally" and so this approach might make sense. However, if the eliminated components contain important non-normal features, then the model would be misspecified. The second idea he proposes is therefore to keep all of the independent components, but to group them into two classes. One class consists of components that are task-related; these are combined into a single regressor, thereby releasing degrees of freedom for error. The other class contains the other components, which are the data-driven confounds.

A critical question in this analysis then becomes how many task-related components need to be included in the composite regressor? At one extreme, all of the components except for one are confounds; this is ordinary noise-free ICA. At the other extreme, all of the components are combined into the task-related effect; this is the standard general linear model. In between are the various models that have some task-related effects and some confounds. The question of the number of components in the combined regressor is therefore one of model selection. McKeown suggests choosing the optimal number of task-related components via a modified version of the PRESS criterion, which

is a leave-one-out predictive summary statistic. Clearly, though, other model selection criteria could be employed for this task.

Some of these same issues are explored in Hu et al. (2005), who note a number of potential weaknesses of McKeown's hybrid ICA. First, the ICA regression in McKeown (2000) is built around spatial ICA. Hu and colleagues suggest that temporal ICA is more natural, since with tICA the components are time courses, and this fits easily into the general linear model framework. Next, they point out that hybrid ICA models task-related activity with one static image and one task-related time course (since the task-related components are combined into one in order to create degrees of freedom), which may miss subtleties in the response. Finally, the linear model used in McKeown's hybrid ICA is the simplest one, ignoring spatial and temporal dependencies in the response.

As an alternative, Hu et al. propose what they call a "unified approach," built on the same general principles as hybrid ICA but differing in some of the details of implementation. Their idea, as in McKeown (2000), is to partition the components into signal and noise subspaces and thereby create regressors for the general linear model. At each voxel a statistical test based on fitting the linear model with the particular signal components as explanatory variables determines whether or not the signal is significantly expressed. Note that the assumption of ICA is that the components are all "true causes" of the observed signal and hence globally they are all expressed in the data. In any given region, by contrast, they may or may not be expressed. One would expect, for example, that task-related components would express mostly, or exclusively, in task-related regions.

However, instead of spatial ICA, as in McKeown (2000), they use temporal ICA. The order of the model is chosen by BIC rather than by the PRESS criterion, and the regressors are those components that have the highest correlation with the experimental paradigm (rather than combining the chosen components optimally into a single regressor). Finally, since the inputs to the model are time courses, the result of the general linear model analysis is to identify spatial regions in which the independent components are expressed.

Hu et al. (2005) compare their unified approach to a standard general linear model analysis. On simulated data they find that under most conditions that they examine the ICA linear model is more powerful than the classical model. For small type I error, the two approaches are comparable, with the usual general linear model performing better (smaller false positive rate) when the levels increase. Analysis of real data confirms the higher sensitivity of the unified approach. Unfortunately, the other interesting comparison, with McKeown's hybrid ICA, is not performed.

Group Analysis with ICA

Moving beyond single-subject analysis, Calhoun et al. (2001a) propose a model for group inferences using ICA. As we have seen already in Section 5.5,

combining individual subject results into a group map presents some interesting statistical challenges. The computational aspect is often problematic, due to the large amounts of data that are involved in creating a group map. Hence an important part of the Calhoun et al. procedure is data reduction. In fact, there are two data reduction steps. The first is part of the preprocessing performed on the individual subjects; here, any necessary cleaning of the data (such as motion correction) is carried out, the data are translated into Talairach coordinates, and an initial dimension reduction via PCA is used to decrease the number of time points in the images, while preserving as much of the variability as possible.

The reduced individual data sets are then concatenated into a $n \times (k \times l)$ matrix, where n is the number of voxels in each image, k is the number of subjects, and l is the number of time points after PCA reduction in the temporal dimension. l is the same for all subjects. Next, the second dimension reduction is carried out on the combined data matrix; a model selection criterion such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC) is used to determine the number of sources (components) in the grouped data. Again using PCA the aggregated data are reduced down to this new dimension. ICA is performed on the reduced concatenated data. Finally, time courses and spatial maps at the group and individual levels are reconstructed, and the spatial maps are thresholded.

Calhoun et al. show that the mixing matrix in their group ICA is "approximately separable across subjects" (p. 143). The implication of this separability is that the time courses for different subjects are distinct. That is, there is no assumption of a common underlying temporal pattern, hence different subjects can have different activation time courses. However, this approach does impose a common space of observations (hence the need to transform the images into Talairach coordinates prior to analysis). In subsequent work (Calhoun et al., 2003b), the authors extend their methodology to allow for different temporal delays (latencies of activation onset) in each source.

Apparently independently of Calhoun and colleagues, and at approximately the same time, Svensén et al. (2002) also proposed an extension of ICA from individual to group settings. In contrast to Calhoun et al., Svensén et al. perform a spatial ICA, concatenating the data from individual subjects into an $n \times (l_1 + l_2 + \cdots + l_k)$ matrix, where n is the number of time points collected for each subject over the course of the experiment and l_i is the number of voxels for subject i after masking out the air. Masking out air voxels is an effective way of reducing the dimension of the data, since for many subjects as much as 50% of an image is made up of voxels outside of the brain. Furthermore, as noted by Svensén et al., with their approach there is no need to transform the images into Talairach coordinates.

The analysis of Svensén and colleagues results in a single set of time courses that is common to the group as a whole (that is, the mixing matrix is common to all subjects, rather than being separable), and a set of individual spatial maps. This is the "flip" of the output of the Calhoun et al. analysis.

There are several consequences of this. First, there is the implicit underlying assumption that common time courses for the entire group are appropriate. Svensén et al. justify this by noting that subjects who are combined follow the same experimental paradigm, and so it makes sense to have common temporal components. On the other hand, this means that components with different temporal behavior across subjects will not be extracted. Some of this will be noise, but it is possible that real, subtle, differences among subjects will be suppressed. By contrast, since the spatial patterns are individual, the analysis can pick out components in this dimension that are characteristic of only a subset of the subjects, or even of single subjects.

A more formal comparison of these two approaches is given by Schmithorst and Holland (2004). They use the term "subject-wise concatenation" to describe the approach of Calhoun and colleagues, "row-wise concatenation (across time courses)" to describe the method of Svensén et al., and they consider in addition across-subject averaging prior to performing ICA, as a computationally simple alternative to the other two. The three group methods are tested on a series of simulated data, with the number of "subjects" expressing each of 20 sources varying from 1 to 20, and one simulation with 100 "subjects." The first simulation includes a set of runs with only the 20 common sources, and a set of runs with these plus individual sources representing differences among subjects (subjects having components that are unique to them, due for example to head motion); the second simulation is run on the set of components with the unique sources, and common components being present in from 1 to 20 subjects out of the 100.

For the simulations without any individual components, all three methods are found to perform more or less comparably with respect to estimation of the time course. Subject-wise concatenation offers a distinct advantage over the other two methods with respect to the accuracy with which sources are estimated, when the number of subjects in which a component appears is small (fewer than 10). When the unique sources are added in, row-wise concatenation suffers a serious degradation in accuracy in terms of estimating both the sources (even those present in all subjects) and the associated time course. The subject-wise concatenation method is not strongly affected by the presence of unique components, performing almost as well as in the first set of simulation runs. Across-subject averaging followed by ICA falls somewhere in between the other two. When the number of subjects expressing a component is small, and there are in addition components that are uniquely expressed, this method performs poorly, at the level of row-wise concatenation. However, for large number of subjects expressing the common sources, across-subject averaging is comparable to subject-wise concatenation. Likewise, when (i) the total number of subjects is large (100), (ii) there are unique components, and (iii) a common component is expressed in 1 to 20 of the 100 of the subjects, subject-wise concatenation remains quite robust, unless the number of subjects with the source is very small. Averaging across subjects has more

difficulty with this particular scenario, unless the common component is present in a relatively large number of subjects.

Based on their simulation results, Schmithorst and Holland (2004) conclude that row-wise concatenation is not a feasible method for group inference, since it is both computationally expensive and its performance is dominated by that of the other two procedures. Between those two, subject-wise concatenation is the more accurate, but it is also more computationally intense. Furthermore, for large studies, across-subject averaging followed by ICA performs almost as well if the common sources are present in a sufficiently significant fraction of the subjects, and it is less demanding of computing time and resources. It thus appears to represent a viable alternative, although there are clearly many more parameters that could be manipulated to test the accuracy of all three group analysis techniques.

Esposito et al. (2005) approach the problem via similarity measures to study the commonalities among the independent components calculated for each of the subjects; components can then be clustered using any of a number of standard techniques.

In more detail, Esposito et al. (2005) first carry out ICA on each subject individually. They define a flexible and general similarity measure between components i and j as

$$\mathrm{SM}(i,j) = \lambda R_s(i,j) + (1-\lambda) R_t(i,j),$$

where $R_s(i,j)$ is the spatial correlation coefficient between components s_i and s_j, $R_t(i,j)$ is the temporal correlation between the time courses associated with s_i and s_j, and λ is a weight between 0 and 1, which allows the researcher to control the emphasis that is placed on spatial or temporal similarity. The matrix of similarity values is transformed into a matrix of dissimilarities, or distances, via $\mathrm{DM}(i,j) = \sqrt{1-\mathrm{SM}(i,j)}$; this is the input for the clustering algorithm.

In the initial implementation of their method, Esposito et al. use a "supervised hierarchical clustering algorithm, linking the components to each other only when differently labeled (i.e., belonging to different subjects)" (p. 197). A new cluster is created when the within-cluster distances are below the current value of a threshold, and the cluster is "representative" of a group of subjects according to a minimum group size specified by the user (this is done instead of fixing the number of clusters). Researchers thus have a fair amount of discretion in guiding the algorithm, via an interactive visualization tool developed by the authors, the *SOM toolbox*. A final step that is incorporated into the graphical representation is a projection of the similarity/dissimilarity matrix onto a two-dimensional space using methods related to multidimensional scaling.

The main features of the "self-organizing ICA" are tested on two real studies of simple visual stimulation (flashing checkerboards presented either to the entire visual field at once, or alternately to the left and right sides, in both instances interspersed with periods of rest) to six healthy subjects.

For both experiments the self-organizing ICA detects components related to the paradigm – a single, simple boxcar time course component in the first study, and two time course components in the second study, corresponding to stimulus presentations to the left and right visual field. Other components, representing transient effects, noise, and so forth, are also found, as is typical with ICA.

Group inference using ICA is an active area of research. In particular, extensions to multigroup settings for the comparison of several groups of subjects are still needed; see Beckmann and Smith (2005) for a first step in this direction.

Modifications to the Basic ICA Approaches

Given that ICA is becoming such a popular tool in the analysis of fMRI data, it should come as no surprise that researchers continue to propose modifications to the basic spatial and temporal methods, in order to better take advantage of the characteristics of neuroimages. In this section we survey a few of these modifications, but readers who are interested in ICA as an analysis path should consult the current literature for new ideas.

Stone et al. (2002) introduce two new ICA methods, *spatiotemporal ICA* (or stICA), and *skew-ICA*. Their motivation for the first modification is the observation that carrying out the ICA in the spatial dimension or the temporal dimension, as is standard practice, results in physically impossible forms for the "dual" dimension, in order to achieve independence in the extracted signals. For instance, in temporal ICA we seek the set of independent time courses and the corresponding set of images (the "dual" in the terminology of Stone et al.) is unconstrained. The lack of constraints on the set of images means that tICA might yield components that are meaningless in terms of the underlying science. Likewise for spatial ICA. Hence they propose the stICA procedure to maximize independence over space and time simultaneously, without necessarily achieving independence in either dimension individually. Such an approach acknowledges that there may be small amounts of spatial or temporal dependence in the sources.

The second modification, skew-ICA, is motivated by the observation that "source images are likely to consist of spatially localized features surrounded by an homogeneous background" (p. 408). The corresponding density function that describes such sources is a skewed distribution. Typical analyses use a source distribution with heavy tails (high kurtosis); the authors surmise that long tails might be more realistic. The goal of skew-ICA is to allow for this possibility.

In all, three new variants are proposed: stICA, skew-sICA, and skew-stICA. As before, let the number of voxels in an image be n and the length of the time course be T. Furthermore, suppose that the dimension of the decomposition is k, a number much smaller than either n or T. All three methods work with this reduced-dimension data set, call it $\tilde{X}$ in the notation of the

authors, and it is the result of applying a preliminary PCA to the original data, X. That is, $X \approx \tilde{X} = UDV^T$, and they write the right hand side of the last equality as $\tilde{U}\tilde{V}^T$.

For the basic stICA procedure the underlying assumption is that each eigenimage in $\tilde{U}$ is a linear combination of k spatially independent images S, and each eigentime-course in $\tilde{V}$ is a linear combination of k temporally independent sequences T. There are then two unmixing matrices, call them W_S and W_T, such that $S = \tilde{U}W_S$, $T = \tilde{V}W_T$, which can be found by simultaneously maximizing a function of the spatial and temporal entropy of the signals. Stone and colleagues suggest a function of the form

$$h_{ST} = \alpha h_S + (1 - \alpha)h_T;$$

this allows the preferential weighting of either the spatial or the temporal component.

Skew-sICA replaces the high-kurtosis probability distribution function characteristic of ICA with a skewed distribution for the spatial sources. Stone et al. (2002) use an exponential decay model. The rest of the ICA procedure is as for the standard sICA. Finally, skew-stICA combines the two above ideas, maximizing the weighted entropy function, with h_S replaced by the skewed distribution model.

On both real and simulated data, stICA alone is shown to suffer from some of the same defects as other analyses considered by the authors (general linear model, PCA, sICA, tICA), namely, it extracts time courses that are inconsistent with the known experimental paradigm and does not correctly identify the spatial sources. Skew-sICA is somewhat better at extracting the time courses and much improved at identifying the spatial components. Skew-stICA further improves over skew-sICA, both spatially and temporally.

Most analyses of fMRI data consider the magnitude information primarily or exclusively. However, there may also be activation-relevant information in the phase images (Calhoun et al., 2002). To account for this, Calhoun et al. (2002) suggest ICA in the complex domain, with three variants: allowing the time courses to be complex-valued and the images to be real, allowing the images to be complex-valued and the time courses real, and allowing both the time courses and the images to be complex-valued.

Notably, they continue to use standard (real-valued) algorithms, rather than algorithms already existing in the ICA literature for handling complex data. They achieve this by a suitable organization of the data. For example, when both time courses and images are complex, they write the fundamental noise-free ICA model $\mathbf{X} = A\mathbf{s}$ as

$$\begin{bmatrix} X_{Re} & X_{Im} \\ X_{Im} & -X_{Re} \end{bmatrix} = \begin{bmatrix} A_{Re} \\ A_{Im} \end{bmatrix} [s_{Re} s_{Im}]$$

The real and imaginary parts are separated into distinct (scalar-valued) submatrices, and ordinary ICA algorithms can be applied. When only one dimension (spatial or temporal) is complex, the model is similarly rewritten.

Compared to a magnitude-only ICA, Calhoun et al. show that ICA in the complex domain, in any of the three variants, detects 10% to 25% more contiguous voxels above a set significance threshold. The time courses extracted from the magnitude-only ICA are smoother than those from the complex domain analyses, apparently due to the fact that voxels with task-relevant phase changes will be detected as active, even if the changes in magnitude are weak. The fully complex variant poses some challenges of interpretability, since phase information will get mixed among the images and time courses, therefore the authors suggest an analysis with one dimension being complex, to capture the phase information, and the other real.

7.2.3 Canonical Correlation

An early use of CCA for fMRI data is found in Friman et al. (2001), for block design experiments. As $\mathbf{X}$ they take a local neighborhood of nine voxels: the voxel of interest and its eight adjacent voxels in the same slice. Thus the dimension of X is 9. As $\mathbf{Y}$ they take a sinusoidal basis function set comprising the six elements $(\sin(\omega t), \cos(\omega t), \sin(3\omega t), \cos(3\omega t), \sin(5\omega t), \cos(5\omega t))$, where $t = 1, \ldots, T$ with T the length of the time series at each voxel and $\omega = 2\pi/m$, where m is the length of a block, assumed to be the same for rest and task conditions. Their specific choice of $\mathbf{Y}$ is the Fourier series expansion (truncated) of the square wave describing a block design. The dimension of $\mathbf{Y}$ is therefore 6 and there are six canonical correlations that can be calculated; the authors concentrate on the largest only. This largest canonical correlation is assigned to the center voxel in each neighborhood, yielding a statistical map of correlation values.

To complete the analysis it then remains to determine which voxels are active. For the particular choice of $\mathbf{X}$ and $\mathbf{Y}$ given here, the largest canonical correlation measures how well the voxel time series in each 3×3 neighborhood matches with the idealized square wave of the block design. If the correlation is large, that neighborhood is similar to the experimental paradigm, and the central voxel (to which the largest canonical correlation is assigned) would be declared active. This is one criterion, namely to conclude that a voxel is active if the largest canonical correlation is higher than some threshold. Note however that incorrect inferences could be reached as a result of the convention to assign the highest correlation to the center voxel of each neighborhood (Nandy and Cordes, 2004a). The problem is exacerbated by the fact that assessing statistical significance of the largest canonical correlation is difficult (the distribution is not tractable even under assumptions of normality and independence, both of which are unrealistic for fMRI data; see also the discussion in Nandy and Cordes 2003a).

Friman and colleagues therefore define two additional criteria for determining activation status of voxels, using the coefficients from the CCA for the sinusoidal basis set. These are *shape* and *response delay*. Shape captures the notion that active voxels in a block design experiment should exhibit a

repeating "on-off" pattern that is roughly synchronized with the timing of the task and rest blocks. Voxels that don't show this behavior are probably not activating in response to the stimulus presentation. Response delay is the maximum allowed delay between presentation of the stimulus and the start of the hemodynamic response. As we have seen in previous chapters, there is typically a small delay, on the order of 2-6 seconds, between the stimulus and the onset of the measured response in the brain. So some delay is expected in an active voxel, but it should not be too long. Voxels that do not meet the criteria for shape and delay are eliminated from future consideration.

When compared to standard univariate analyses (t or F tests) on a single subject performing a simple finger motion task, the results of the CCA appear smoother and have fewer isolated active voxels (which most likely are spurious). Furthermore, in this one example the canonical correlation analysis reveals subtleties that are not apparent in the other maps. On the other hand, the authors discuss the possibility that the detected regions are too large, an artifact of assigning the largest canonical correlation value to the center voxel in a neighborhood; this problem, they theorize, will be especially serious near blood vessels. Nandy and Cordes (2004a) propose an adaptive assignment scheme instead, and show via simulation and a real data example that indeed the convention adopted by Friman et al. can result in regions that are artificially large.

In other work, Nandy and Cordes (2003a) address the intractability of the distribution of the maximum canonical correlation. First, they suggest using a test statistic that is a function of all the canonical correlations, instead of just the largest. The likelihood ratio test for all of the canonical correlations being zero (hence that $\mathbf{X}$ and $\mathbf{Y}$ are independent, under an assumption of normality) is Wilks' Lambda:

$$\Lambda = \prod_{i=1}^{k} (1 - r_i^2),$$

where $r_1 \geq r_2 \geq \cdots \geq r_k$ are the canonical correlations calculated on the sample. In fact one typically works with a function of Λ that is asympotically χ^2 in distribution. Using just the maximum, one can achieve only a bound on the significance level by setting $r_2, \ldots, r_k$ to zero. Nandy and Cordes use the same test statistic as do Friman et al., without setting the smaller canonical correlations to be zero.

Their second modification is to perform a nonparametric analysis, noting that most of the assumptions underlying the parametric analysis of Wilks' Lambda, in particular the assumption of temporal independence of observations within a voxel, are violated. To carry out the nonparametric analysis, they collect a so-called "null" data set, in which the subject is resting in the scanner with eyes shut (awake, but not performing any task). The test statistic is calculated at each voxel in the null data set, resulting in an empirical distribution against which the statistics calculated for the task data can be

calibrated. Convergence of the empirical distribution function to the true distribution relies on independent samples, and the voxels in the brain are not spatially independent; it is necessary therefore to select subsets of the voxels from the null data that may reasonably be assumed to be independent, for instance, voxels that are far away from each other.

With these modifications to the simple CCA, Nandy and Cordes show that, especially for studies with weak activation (for instance, tasks that do not elicit a robust response), the multivariate methods outperform the standard univariate analyses. However, the differences are not as great when the elicited response is strong and consistent and may be in favor of either approach. This is not surprising, since one would expect that any reasonable statistical analysis should be able to detect clear, strong activation patterns. Also, not unexpectedly, they find that the distribution of the test statistic in Friman et al. (2001) differs greatly from theirs; neither distribution coincides with the asymptotic distribution, although that of Nandy and Cordes is much closer.

Maximum Correlation Analysis

Friman et al. (2002a) develop a technique related to CCA, which they call *maximum correlation modeling*, or MCM. Around each voxel, consider the eight neighbors (so that each voxel is at the center of its own 3×3 neighborhood). Out of those nine voxels, one creates five new time series: $\mathbf{x}_1(t)$ is the time course of the center voxel; $\mathbf{x}_2(t)$ is the average of the voxels to the immediate left and right of the center; $\mathbf{x}_3(t)$ is the average of the voxels in the upper left and lower right corners; $\mathbf{x}_4(t)$ is the average of the voxels above and below the center; and $\mathbf{x}_5(t)$ is the average of the voxels in the upper right and lower left corners. These five new time courses are combined linearly via weights w_1, w_2, w_3, w_4, w_5 which are non-negative, sum to 1, and such that w_1 (the weight applied to the center voxel) is larger than the others. Finally, the correlation of this spatially smoothed time course with the convolution of the stimulus trail (block design) and the difference of two gamma model for the HRF, is calculated. One seeks the values of the weights and the parameters of the model such that the correlation is maximized.

Recall from the discussion in Section 5.3.1 that the difference of two gamma model has several parameters; to make the estimation problem computationally feasible, Friman et al. fix some of the parameters and impose constraints on others, so that the optimization is performed only over the weights w_i and two of the parameters in the difference of gammas model, namely the delay in the hemodynamic response and the size of the undershoot (poststimulus dip). The parameter estimates that result from solving the optimization problem are interpreted as yielding the "optimal" spatial filter (via the weights that are assigned to the center voxel and its neighbors) and hemodynamic response model (via the delay and undershoot of the HRF).

Friman et al. (2002b) also propose an exploratory use of CCA to separately detect temporal and spatial directions of maximal autocorrelation. The

motivation for this approach is the twofold observation that (1) in the time domain, "interesting" (that is, potentially active) voxels should exhibit autocorrelation, whereas inactive voxels are more likely to be white noise and (2) in the spatial domain, active voxels should appear in clusters whereas singleton active voxels are more likely to be spurious. Hence for the temporal analysis, $\mathbf{X}$ in the CCA is taken as a time course and $\mathbf{Y}$ is taken as that same time course shifted by a lag of 1; for the spatial analysis, X is taken as a voxel and Y is taken as the sum of its four immediate adjacent neighbors (to the left, to the right, above, and below). Compared to both spatial and temporal versions of PCA and ICA, the CCA method is computationally efficient and seems to find directions of high "interestingness." Interpreting the components that are revealed by the analysis can be challenging, a well-known problem of multivariate methods such as PCA or Factor Analysis. In spite of this, in the example Friman and colleagues present they are able to interpret the first two or three temporal components and likewise with the spatial, relating them directly to aspects of the experimental design in the first case (e.g., the boxcar function of the block design experiment) and to brain structure in the second (e.g., areas involved in the experimental task). Later components do not have such obvious meaning attached to them.

7.3 Software Issues

Since the methods discussed in this chapter are relatively new to neuroimaging, they are not fully integrated into any of the major software packages. Many researchers, particularly those developing modifications on the standard approaches, also write their own code, which is usually available on their websites for free download. Much of this code is written in MATLAB, R/Splus or Fortran, and hence should be relatively accessible. Readers who are interested in specific techniques are urged to explore the appropriate sites for additional information. Links are not included here as the web addresses are prone to change without notice.

8

Basis Function Approaches

The methods examined in this chapter have in common that they model the response of interest via some set of basis functions. The most popular approach by far is to use wavelets (Chui, 1992; Daubechies, 1992; Vidakovic, 1999). Wavelets have found a variety of applications in the fMRI literature and these are surveyed here together. Other basis functions, such as sets informed by anatomical considerations, and typical families such as splines, have also been explored by fMRI researchers, but in a much more limited capacity. Finally, polynomial and trigonometric basis functions are sometimes used as additional predictor variables in the general linear model.

This chapter focuses on wavelets and anatomically informed basis functions. The former are of interest because they have wide applicability to fMRI data beyond being an extension of the basic linear model; indeed, wavelets have been used for creating activation maps, as a resampling technique, for data compression, and for modeling. The latter are of specific interest because they represent attempts to directly use prior anatomical information, and hence to derive a set of functions that have intrinsic physiological meaning and interpretation (as opposed to trigonometric or spline functions, for example).

8.1 Wavelets

The attractiveness of wavelets for the analysis of fMRI data has been persuasively argued by Bullmore et al. (2003), who enumerate the following advantages: wavelets are multiresolutional, i.e., they can model phenomena at different scales; they are adaptive to nonstationary or local features; the wavelet transform has a decorrelating ("whitening") effect and this may be statistically convenient for modeling purposes; the wavelet transform is useful for data compression and denoising; the discrete wavelet transform is very fast computationally, even compared to the fast Fourier transform; the brain has a fractal nature, and wavelets are an effective way of modeling such processes.

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_8,

The last point would seem to argue for the use of two- or three-dimensional wavelet decompositions for the spatial part of the spatiotemporal problem; however, in practice, wavelets have most often been used in fMRI to model the temporal structure of the data, i.e., long-range temporal dependence.

Wavelets are a family of orthonormal basis functions obtained by translation and dilation of a "mother" wavelet ψ with $\int \psi(t)dt = 0$ and a "father" wavelet or scaling function ϕ with $\int \phi(t)dt = 1$ in the following manner:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - 2^j k}{2^j}\right)$$

and

$$\phi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\phi\left(\frac{t - 2^j k}{2^j}\right).$$

Here, $j = 1, 2, \ldots, J$ indexes the scale (dilation) $S_j = 2^j$ and $k = 1, 2, \ldots K$ indexes the location (translation). Wavelets are additionally characterized by their smoothness, or the number R of vanishing moments (of the mother wavelet). Wavelets come in many forms, or families. Two simple examples of mother wavelets are the Haar and "Mexican hat," shown in Figure 8.1.

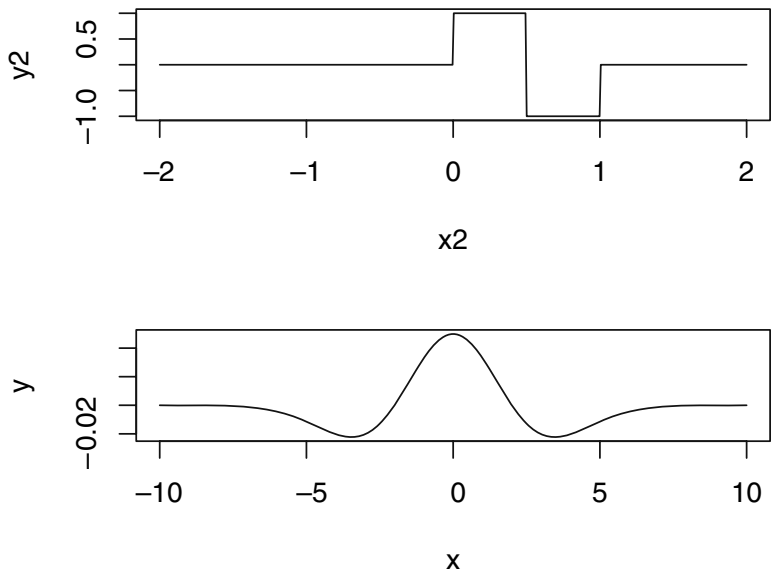


Fig. 8.1. Haar (top panel) and Mexican hat (bottom panel) mother wavelets.

The successive translations and dilations of the mother and father wavelets yield a set of basis functions. At each scale S_j the data are then decomposed

into "detail coefficients" $d_{j,k}$ and "approximation coefficients" $a_{j,k}$, which are orthogonal to each other; they are found by taking the inner product of the data and the scaled and translated mother and father wavelets respectively. The detail coefficients summarize variability in the data at the given scale j, while the approximation coefficients are residuals after the information at scale j and all finer scales has been removed. The original data can be reconstructed without loss by adding the approximation at the coarsest scale J and the details at all scales up to and including J, that is,

$$X = \sum_{j=1}^{J} \sum_{k} d_{j,k}\psi_{j,k} + \sum_{k} a_{J,k}\phi_{J,k}.$$

Often, many of the detail coefficients are small and therefore we might in fact want to set them to zero for modeling purposes as their contribution is minimal. In this manner wavelets encourage a sparse representation of the data. Since they are localized in both time and frequency, wavelets can model abrupt changes in the signal (for instance, local spikes) even with a small number of coefficients. Wavelet thresholding has been extensively studied by Donoho and Johnstone (1994).

Extending wavelet analysis from one dimension to two, or more, is straightforward (Ruttimann et al., 1998) by the use of tensor product basis functions. This involves applying the one-dimensional composition separately to each direction of the multidimensional data. It is therefore easy in principle to perform a wavelet analysis for fMRI data in space, time, or both.

The set of detail coefficients has other uses in wavelet analysis as well. For instance, the correlation between coefficients at different or the same scales can be shown to decay exponentially if the number R of vanishing moments satisfies $R > 2H+1$, where $0 < H < 1$ is the Hurst parameter, a measure of the long-range dependence in the data. This proves the decorrelating effect of the wavelet transform. Furthermore, the sample variance of the detail coefficients at the jth scale can be used to get a simple estimate of H (Park et al., 2007). See Bullmore et al. (2003) for quick derivations of these two results.

8.1.1 Creating Activation Maps with Wavelets

Analysis of fMRI data with wavelets dates back to the late 1990s (Ruttimann et al., 1998). In this work Ruttimann et al. (1998) use wavelets to address the problem of multiple hypothesis testing; Chapter 10 contains an in-depth discussion of the statistical issues and some of the varied solutions that are available in the fMRI literature. Here we thus focus on the wavelet aspect of the application.

The authors start with the (possibly unrealistic) assumption that the noise in an image is independently, identically distributed (i.i.d.) normal, and hence the wavelet coefficients of the noise are also i.i.d. normal. Let the common variance be denoted σ^2. The method proceeds in two steps. In the first step each

level and direction of the (now two-dimensional) wavelet transform undergoes an overall test of significance using the assumption under the null that

$$\sum_{k=1}^{K_j} \left(\frac{d_{j,k}}{\sigma} \right)^2 \sim \chi^2 K_j,$$

with K_j being the number of coefficients at level j of the wavelet decomposition. If the null hypothesis is not rejected at a given level, that level is not investigated further. For the levels at which the first null hypothesis is rejected, in the second step each coefficient is then subjected to an individual hypothesis test. Finally, an activation map can be constructed by applying the inverse wavelet transform on the coefficients that survive both levels of hypothesis testing. The two-step procedure reduces the total number of tests that needs to be carried out, and the authors control type I error by a Bonferroni correction at each stage. As we will see in Chapter 10, in general the Bonferroni correction is too conservative for fMRI data, due to the extremely large number of voxels that must be tested. However, performing the analysis in the wavelet domain and screening out some coefficients in the first stage of the Ruttimann et al. approach, mitigates this problem to a certain extent.

A disadvantage of performing the entire analysis in the wavelet domain is lack of easy interpretability, since the effects that are detected are not in the original image space (Van De Ville et al., 2004). Hence, Van De Ville et al. advocate splitting the procedure back into two parts, approximation and detection, carried out in the two domains, wavelet and spatial, respectively. They propose what they term an "integrated framework" in which the links between the two domains are incorporated as constraints to limit the solution set. As such, there are two threshold values, one for the wavelet domain and one for the spatial domain; these are chosen to control the overall type I error rate and to minimize the difference between an unprocessed reconstruction (that is, reconstruction of the raw wavelet coefficients without any preprocessing) and the thresholded reconstruction. The first step of processing is in the wavelet domain, where the linear model is applied to the wavelet transformed data (see Section 8.1.2). The standardized wavelet coefficients are thresholded according to the value determined for this step and the data reconstructed back into image space. A wavelet-reconstructed standard error map is also built for image space, and the second threshold is applied to the standardized values in image space.

The authors present some initial results on simulated, null, and activation data, and note that the approach tends to be conservative. Regions of activation that are detected are somewhat sparser in extent and smaller in amplitude than regions found by a linear model with some presmoothing. By considering two different wavelet bases, they also note that the effectiveness of the method does depend quite crucially on the choice of basis. This question is left relatively unexplored in the initial work, however. The choice of wavelet order is also not examined in this paper, but, as we shall see below, this is

a relevant factor in other wavelet analyses and hence warrants exploration in this context as well.

8.1.2 Wavelets for Modeling

A more traditional application of wavelets would have them used as part of the modeling procedure for fMRI data. Several authors have considered this possibility. As noted by Fadili and Bullmore (2002), among others, there is experimental evidence that the error structure in fMRI data might contain long-term dependencies and a model that accounts for this more complicated behavior will result in improved inference. Here, wavelets are applied to model the temporal structure of the data, rather than the spatial. Since certain types of long-term dependence exhibit so-called "self-similar" (or fractal) properties, and these can be "whitened" by applying a discrete wavelet transform, wavelets provide a natural basis for characterizing and modeling these processes (Fadili and Bullmore, 2002). Fractional Gaussian noise and fractional ARIMA models are examples of models for long-range dependence that are amenable to a wavelet representation. Park et al. (2007) have shown that fractional Gaussian noise is a realistic model for long fMRI time series of resting data, bolstering the relevance of the wavelet modeling approach.

Based on these ideas, Fadili and Bullmore (2002) propose a wavelet-generalized least squares algorithm. They start with the basic voxel-level linear regression model

$$Y = X\beta + \boldsymbol{\epsilon},$$

where X, as before, is a matrix of predictors (for example the convolved hemodynamic response, and relevant covariates) but now $\boldsymbol{\epsilon}$ is assumed to have a long-term dependence structure such as fractional Gaussian noise. This model is then shifted to the wavelet domain by taking the discrete wavelet transform of both sides of the equation, resulting in

$$Y_w = X_w\beta + \boldsymbol{\epsilon}_w.$$

The advantage of transforming the data into wavelet domain is that the variance-covariance matrix of the errors is approximately diagonal, which facilitates inference. A novel feature of the Fadili and Bullmore analysis is that the parameters of the long-range dependency process (for instance, the Hurst parameter of fractional Gaussian noise) are estimated simultaneously with the parameters of the linear model. The procedure they suggest is iterative, cycling between estimation of the β components and estimation of the parameter that characterizes the error structure. All estimation is done within a maximum likelihood framework.

Using simulated data, Fadili and Bullmore show that their wavelet-generalized least squares estimates possess desirable and theoretically derived properties such as unbiasedness and asymptotic normality. Compared to ordinary least squares, the new method has the advantage of yielding, as part

of the procedure, an estimate of the long-range dependence parameter. Its behavior on the simulated data is also better. Compared to generalized least squares, which requires prior knowledge or specification of the entire variance-covariance matrix of the errors, the wavelet least squares approach is simpler, since it makes the assumption that the off-diagonal elements are zero. On the simulated cases presented by Fadili and Bullmore, this assumption does not introduce serious bias compared to the GLS estimates. Wavelet-generalized least squares is more robust than OLS to changes in the value of the long-range dependence parameter (not surprisingly) and controls the type I error in hypothesis tests of the regression parameter β at nominal levels, whereas OLS is led astray by the presence of long-range dependence in the error structure (again, not surprisingly). An autoregressive model of short-term dependence is intermediate in performance between OLS and the models that assume long-range memory behavior.

8.1.3 Wavelet Resampling

Another use of wavelets in fMRI has been for resampling, or *wavestrapping* (Bullmore et al., 2001; Breakspear et al., 2004),in order to obtain valid statistical inferences for activation detection. The wavelet resampling method, which again exploits the whitening characteristic of the discrete wavelet transform, was introduced by Bullmore et al. (2001). In fact, their procedure is remarkably simple. First, one takes the discrete wavelet transform of an fMRI time series, adjusted to have mean zero. Next, one resamples the coefficients at each level of detail; since this is done *without* replacement, one obtains a permutation of the wavelet coefficients. Finally, one executes the inverse wavelet transform of the permuted coefficients to get a reconstructed time course that has the same second-order properties (variances and covariances) as the original.

Note that one cannot merely permute or resample the components of the time series, as this will destroy the correlation structure (Friman and Westin, 2005). Hence, one must either carry out the resampling in such a way as to preserve the structure (for instance, block resampling), or decorrelate the data first (for instance, model prewhitening or transforming into a domain – wavelet or Fourier – in which elements become exchangeable).

Bullmore et al. report that, on null and simulated data, this procedure controls adequately for type I error, although it tends to be somewhat conservative. Looking at activation data acquired in both block and event-related paradigms, and on magnets of differing field strength (1.5T and 3T), they also find that the wavelet resampling approach is robust to a variety of imaging conditions. This is not the case for another resampling method also considered in the study (permuting the time courses after they have been prewhitened by autoregressive models of low or moderate order), which is shown to be highly sensitive to field strength and experimental paradigm.

Validity of wavestrapping as an inferential tool depends heavily on the lack of correlation among coefficients at the same level of detail. But, if the original data are highly correlated and the time courses are not long, application of the wavelet transform may not sufficiently decorrelate the detail coefficients (Breakspear et al., 2004). In that case the simple resampling scheme needs to be modified, for instance, resampling can be done in blocks (this is similar to modifications of the basic bootstrap that are necessary when handling time series data). Furthermore, if one wishes to account for spatial structure, as in fMRI data, one should not necessarily resample each voxel independently, as done in Bullmore et al. (2001). The question then arises, How best to resample the coefficients in each direction at each level, independently or together? These issues are addressed in Breakspear et al. (2004), an extension of the basic ideas formulated in Bullmore et al. (2001).

Breakspear et al. (2004) build a two-step wavestrapping procedure that aims to preserve both spatial and temporal second-order structure present in the original data. In the first step each slice at each point in time is spatially permuted. The same resampling (permutation) at the same scale is applied to each slice and time point. At different scales the resampling is independently performed. This part of the algorithm may involve certain spatial constraints (for instance, separation of brain voxels from air). In the second step the time series from each voxel is temporally permuted. Again, at a given scale the same resampling is applied at each voxel. All resampling is done in the wavelet domain, so before the first step a two-dimensional decomposition is carried out, and before the second step, a one-dimensional decomposition.

Although the basic idea behind wavestrapping is relatively easy and straightforward, as with any resampling scheme "the devil is in the details" and many technical issues need to be resolved, or at least explored, prior to implementation of this approach. These include, but are not limited to, the choice of wavelet family, the order of the wavelet, the number of vanishing moments to impose, resampling in multiple directions jointly or independently, dealing with edge effects. Detailed discussions of these points and more can be found in Bullmore et al. (2001) and especially in Breakspear et al. (2004). See also Bullmore et al. (2003) for an overview of wavestrapping fMRI data.

As noted above, transforming to the wavelet domain in order to get exchangeable elements to resample is not the only way to preserve relevant structure in the original data. One can also transform into the Fourier domain; at each frequency the components are approximately independent and so the same rationale as for wavestrapping holds. However, Breakspear et al. (2004) point out that the multiresolution characteristics of the wavelet basis make it, perhaps, better suited for data such as fMRI, where small local changes exist alongside larger-scale variations. A more formal, if limited, comparison of these two resampling domains is presented by Laird et al. (2004). The authors look at two variations on Fourier resampling, and three wavelet resampling schemes. The Fourier methods exactly preserve the temporal and spatial autocorrelation structure of simulated data; wavelet resampling using

either the same permutation at each voxel or different permutations at each voxel does well at preserving the temporal structure, but only the former effectively preserves spatial structure. A two-dimensional wavelet resampling scheme does not preserve spatial or temporal structure as effectively as any of the other methods. When wavelet resampling is used, somewhat larger areas of activation in a real data set are detected, compared to Fourier resampling. Finally, based on analysis of sensitivity and specificity of wavelet and Fourier resampling, Laird et al. (2004) conclude that the former is superior as a method of statistical inference.

Friman and Westin (2005) compare the bias in the thresholds obtained by three resampling methods: prewhitening with an autoregressive model and then resampling, wavelet resampling, and Fourier resampling. These three resampling paradigms are applied to block and event-related designs, and the resamples carried out on the original or residual (after regressing out the BOLD response) data. They find that the Fourier method is the most biased of the three, and the whitening method the least, with wavelet resampling in between. The bias is worse for block designs than for event-related; wavelet and Fourier methods are biased upward when applied to the original data (significance thresholds are too high) and downward when applied to the residual data (significance thresholds are too low). Although this would seem to indicate the overall superiority of prewhitening resampling and event-related designs, we should note that the resampling method is sensitive to the choice of model; in particular, if the data are prewhitened with an incorrect model, this will again introduce bias.

These two works are only initial studies of the relative strengths and weaknesses of wavestrapping. More research in this area, and a deeper exploration of wavelet resampling in general, is warranted.

8.1.4 Assessing Wavelet Methods

Some of the technical issues involved in using wavelets for the analysis of fMRI data have been explored by Desco et al. (2001), who test different families, orders, and levels on a computer-generated phantom. This phantom allows them to manipulate characteristics of the data, such as extent and strength of activation, amount of noise, and effect of presmoothing. Wavelet packets from the Gabor, Daubechies, Lemarie, and Symlet families are evaluated, and compared as well to a more traditional t test analysis.

Several comparisons are made. First, the authors consider the ability of each method to recover the known true activation pattern when various amounts of noise (ranging from 5% to 20%) are added in. Without presmoothing, the performance of the t test deteriorates severely as more noise is added; at the highest levels of noise, the recovered image (voxels deemed significantly active) is made up of random, isolated voxels. The wavelet methods all fare better, even in the presence of noise and without presmoothing. Presmoothing

improves the quality of the recovered images in all cases, although high levels of noise continue to pose difficulties for the standard analysis.

A second comparison involves looking at receiver operating characteristic (ROC) curves, which plot sensitivity (true discoveries) against specificity (true nondiscoveries) at different levels of significance. Area under the ROC curve is a standard way of assessing the relative quality of methods. Using this criterion, all the wavelet methods outperform the t test, no matter the extent or level of activation, or the amount of noise added to the data. Without presmoothing, the differences are more noticeable; presmoothing brings the various methods into much closer agreement. Indeed, when the data are presmoothed, the areas under the ROC curves are similar for low and high amounts of noise (5% versus 20%), regardless of method of analysis.

Looking at specificity, sensitivity, and ROC curves within a family for different orders of wavelet, wavelets of lower order tend to give better results than those of higher order. As for the choice of family, the Gabor family of wavelets is found to be optimal, but this might be due to the shape of the simulated areas of activation (all circular). Notably, true brain regions of activity are unlikely to be round, and so this family might not dominate others on real fMRI data.

Different wavelet methods for testing the hypothesis of activation at each voxel are compared in Fadili and Bullmore (2004). They look at the problem as one of shrinkage of the wavelet coefficients and consider three algorithms for effecting this shrinkage: frequentist shrinkage with control of the false discovery rate (FDR) (Benjamini and Hochberg, 1995); frequentist shrinkage with recursive testing (Ogden and Parzen, 1996); and Bayesian shrinkage. Although the three algorithms obviously differ in their details of implementation, all have the same goal, namely to identify a subset of the wavelet coefficients that contains the important signal and to threshold out the rest. The retained coefficients are then used to reconstruct the brain image via the inverse wavelet transform.

The three shrinkage algorithms are compared along a number of dimensions on simulated and real data; the real data consist of a resting data set (representing the "null" hypothesis in some loose sense) and an event-related finger movement experiment. Type I error control is evaluated using the null data. All three algorithms give good control of the type I error, with the FDR approach being somewhat conservative and the Bayesian somewhat liberal. Sensitivity is assessed using simulated data, with area under the ROC curve the criterion for comparison. Bayesian shrinkage is the most sensitive, although the advantage is slight. FDR is more sensitive than recursive shrinkage, in general, except at the smallest amplitudes. As the amplitude of the signal increases, the sensitivity of all the algorithms increases, as would be expected, and the ordering $Bayes > FDR > recursive$ is preserved.

When implementing the wavelet transform, users can determine the amount of smoothness or regularity via the number of vanishing moments and the level to which the decomposition will be carried out. Fadili and Bullmore also

explore the effects of these choices. For fixed level of decomposition and significance level α, the false positive fraction (the proportion of wavelet coefficients that are falsely retained) observed in the null data decreases as the number of vanishing moments increases, for all three methods. Desco et al. (2001) report this same conclusion. Keeping the number of vanishing moments fixed and letting the level of decomposition vary, Fadili and Bullmore find that the observed false positive fraction decreases as the decomposition level increases. As noted by the authors, as the level of the decomposition decreases, the number of wavelet coefficients increases, i.e., more hypothesis tests are carried out, and so such a finding is expected.

Finally, on the activation data set, all three thresholding algorithms yield broadly similar results. Bayesian shrinkage is somewhat more sensitive, providing, in the words of Fadili and Bullmore, a "richer or more sensitive characterisation of the cerebral response... " (p. 1123); this reflects also the higher sensitivity of the Bayesian method on the simulated data. Also consistent with the simulated data results, FDR control yields a slightly more conservative picture on the activation data, compared to the other two approaches.

8.2 Basis Functions Informed by Anatomy and Physiology

Whereas wavelets offer a general set of basis functions capable of effectively modeling most data curves, other researchers have proposed instead to use families of basis functions that are driven by the anatomy of the brain. In this section we consider two of these approaches. Neither is yet a standard analysis stream for fMRI data.

The first focused effort to use physiological information to create basis functions for the analysis of fMRI data is due to Kiebel (Kiebel et al., 2000). An important component of this approach is the use of anatomical images together with the functional ones. The anatomical images are preprocessed to reconstruct the cortical surfaces on a so-called "flattened" representation. With this representation the original "folded" cortex is first inflated, so that all of the sulci and gyri are on the surface (we can think of this as taking the crumpled handkerchief of Chapter 1 and blowing air into it in such a way that the relative positioning of the valleys and peaks is preserved). The inflated surface is then pulled out flat, again preserving the basic structure of the folded cortex. The transformations from one reconstruction to another are accomplished by discretizing the maps and manipulating the vertices that define each element of the partition. See Kiebel et al. (2000) for more detail and some references on this procedure.

Letting Y denote the functional observation in voxel space, and V_G the vertices of the reconstructed gray matter in the flat map, the goal of the authors is to estimate a smooth distribution that best explains the observed data. They write this in general model form as

$$Y = g[f(V_G)] + \epsilon.$$

The smooth distribution $f(\cdot)$ is estimated by a linear combination of *local, smooth, partially overlapping* basis functions defined on V_G. g is an operator that integrates the basis functions, transforming them from the space of vertices back into voxel space. The basis functions are defined on the cortical flat map with a two-dimensional coordinate system as follows: the basis function b_F^j with center at coordinate (x_j, y_j) is

$$b_F^j = c_1 \exp\left\{\frac{-[(x-x_j)^2 + (y-y_j)^2]}{2w^2}\right\},$$

a circular Gaussian. c_1 is a normalizing constant and w is a user-specified window, the width of the Gaussian basis function in the x and y directions. The basis functions are patterned in a hexagonal layout; with (x_j, y_j) the center of b_F^j, the centers of the six neighbors are given by $(x + d/2, y + d_0)$, $(x + d, y)$, $(x + d/2, y - d_0)$, $(x - d/2, y + d_0)$, $(x - d, y)$, and $(x - d/2, y - d_0)$, with d the fixed distance between basis function centers and $d_0 = d \sin 60^\circ$. Figure 8.2 gives an example with $d = 4$, and hence $d_0 = 2\sqrt{3}$.

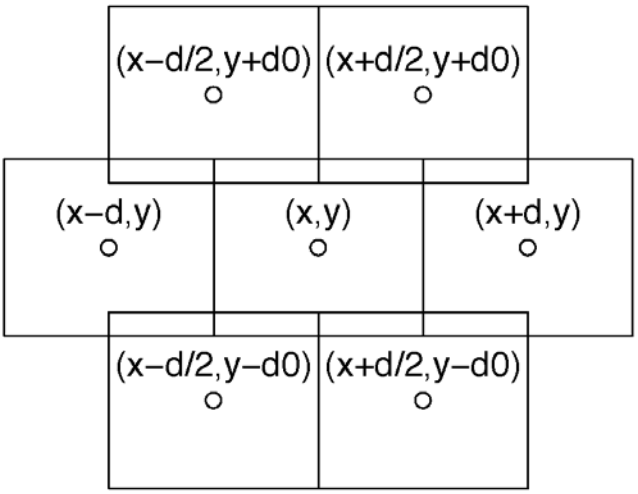


Fig. 8.2. "Honeycomb" neighborhood for anatomically informed basis functions. Here $d = 4$ and $d_0 = 2\sqrt{3}$. Starting with the function centered at the point labeled (x, y), the hexagonal neighborhood is defined. Note that there is a slight overlap of neighbors. Successive neighborhoods are built from the neighbors of this first point in the same fashion.

Suppose then that n_B basis functions are defined in this manner. Via the integration operator g described above, the b_F^j are tranformed into voxel space,

and it is this set of basis functions that is used to model the functional data summarized in Y. Let the number of functional voxels be n_V (this is also the number of basis functions in voxel space after the transformation). Then the model that is actually fit is

$$Y = A\beta + \epsilon,$$

a linear model. In practice, Kiebel et al. implement a ridge regression approach to estimate β, i.e., $\hat{\beta} = (A^T A + \lambda I)^{-1} A^T Y$ which is necessary if the basis functions have substantial overlap. If the overlap is small or none, λ is set to zero.

Note that all of the above applies to a single vector of functional data. The next steps of the analysis consist of assembling the estimated parameter vectors for each Y_i into one matrix, which is then subjected to further analysis, either univariate or multivariate, to make inference about response patterns over time. The method therefore consists of fitting two linear models to the data, the first in the spatial domain and the second in the temporal. Fitting the analysis into the linear model framework means that the extension to multiple subjects is relatively straightforward. A *canonical cortical surface* that is representative of all subjects is used instead of the individual flat maps. The rest of the analysis proceeds as before (Kiebel and Friston, 2002).

More recently, Harms and Melcher (2003) implement a general linear model with a set of basis functions that are meant to reflect temporal features of the BOLD response. They call this basis set *OSORU*, for "onset, sustained, offset, ramp, undershoot" – the main features of the hemodynamic response as we have seen it described in previous chapters; see Figure 8.3 for a graphical depiction of these basis functions. These basis functions make up the design matrix in the general linear model. Compared to other basis functions that can be used with the general linear model, such as sinusoidal (discussed in Chapter 5) and wavelets (discussed in the previous section), the OSORU set admits a direct interpretation.

The *onset* component represents the initial reaction to the stimulus, peaking at around 6 seconds after presentation, and returning more or less to baseline by 14-16 seconds. Harms and Melcher characterize this as a transient response, in contrast with the *sustained* component, which is a convolution of the onset with the stimulus stream (coded as 0-1 for a block design, for instance). The *offset* component is a similarly transient response to the termination of the stimulus, represented as a time-shifted version of the onset. The *ramp* is an approximately linear response observed for some stimuli, and characterizes "signal recovery" following the decline of the onset component. The *undershoot* is a third transient component. In the basis function representation it is defined, like the other functions, as a positive deviation from baseline. Hence, its manifestation in the model when a significant undershoot exists is expected to be via a negative coefficient.

The particular choice of basis function derives from the observed phenomenon that for some types of stimuli different temporal patterns of

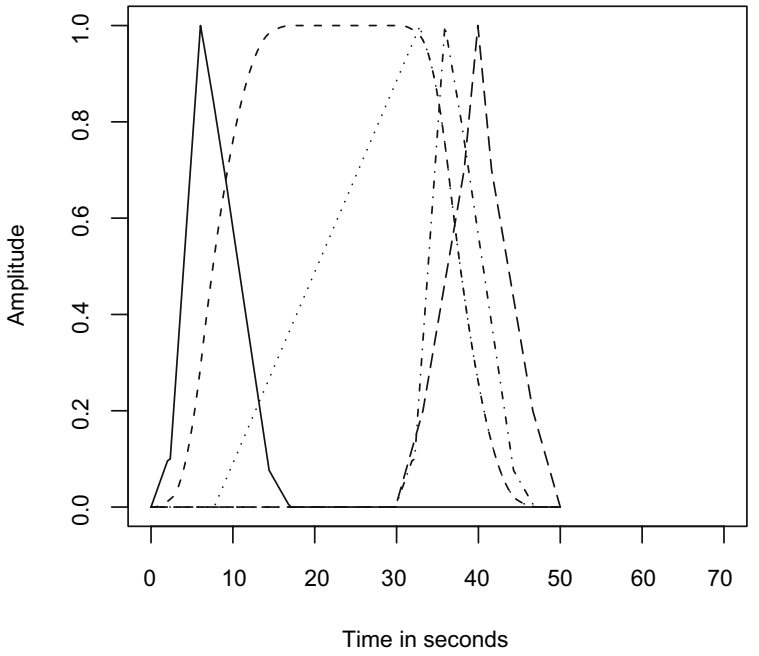


Fig. 8.3. The OSORU basis functions; the solid line is *onset* component, the short dashed line is *sustained* component, the dashed dotted line is *offset* component, the dotted line is *ramp* component, and the long dashed line is *undershoot* component. The last component, when it plays a role, is expected to have a negative coefficient in the model due to its positive representation in the basis set. In this example, the stimulus is on for the first 30 seconds. Adapted from Harms and Melcher (2003).

activation may be manifested. Harms and Melcher, for example, note that for repeated noise bursts in a block design experiment, when the repetition rate is low the response is sustained, but when it is high, the response is "phasic" (that is, shows peaks at onset and offset) in the *same* cortical areas. Thus it is desirable to have a set of basis functions that is flexible enough to capture both of these sorts of behavior. Incorporating the OSORU basis into the general linear model framework gives this flexibility, since the regression coefficients will reflect the relative importance of each component of the basis set, at each voxel, and so different shapes of hemodynamic response can thereby be modeled. Obviously wavelets or other functions have this flexibility as well, but Harms and Melcher emphasize ease of interpretation and physiological meaning. This is because they are not interested primarily in

modeling the hemodynamic response, as in other approaches we have already considered; rather their focus is on trying to understand the neural activity behind the fMRI response. For this purpose it is advantageous to have basis functions that are directly interpretable. On the other hand, the specific set of functions is largely paradigm-dependent; while the general shapes might be appropriate for a wide range of studies, the details of latency times, plateau, and so forth will differ from experiment to experiment. Indeed, Harms and Melcher use a subset of their data to precisely define the OSORU basis functions. The need to tailor the basis functions to the experiment is a potentially serious drawback of this method.

In addition to carrying out the general linear model with the OSORU functions as elements of the design matrix, Harms and Melcher also define a "waveshape index" (WI) to characterize temporal patterns as *sustained*, *phasic*, or intermediate between the two extremes:

$$\mathrm{WI} = \frac{1}{2}\left(\frac{\mathrm{On} + \mathrm{Off}}{\mathrm{Mid} + \max(\mathrm{On}, \mathrm{Off})}\right);$$

$\mathrm{WI} \in [0, 1]$. In this equation, *On* and *Off* are the estimated coefficients of the onset and offset components from the linear model, and *Mid* is the magnitude of the sustained component plus one-half the coefficient of the ramp component. Large values of WI indicate that the onset and offset are of comparable size, and are, in addition, big relative to the midpoint value. This will be the case for phasic responses. Values of WI near zero indicate a response that is dominated by the sustained or ramp component.

For the noise burst data that motivated this work, Harms and Melcher find that the OSORU basis detects more activity in the relevant cortical areas than does a general linear model with only the *sustained* component, which is roughly equivalent to a standard t test. It is comparable to the sinusoidal basis consisting of the first through fourth harmonics, detecting slightly more activation under some circumstances. The "sustained only" analysis misses voxels that have a phasic response, as would be expected. Both the OSORU and the sinusoidal basis sets are capable of detecting these voxels, again as would be expected. Using essentially a partial F test to assess the relative importance of the different components in the basis function set, they find that all five are important for fitting the responses in a preponderance of the active voxels in auditory cortex.

8.3 Summary

Basis functions, and in particular wavelets, have found a variety of uses in fMRI data analysis. Although some authors have attempted to compare the merits of the different technical choices relating to the implementation of these approaches (for instance, the family or order of wavelets), there is still scope

for exploration and development. Other basis function families have not found as wide expression in fMRI data, even though they have proven useful in other areas of application. Splines, for example, have been applied in preprocessing steps, as we saw in Chapter 3 (Tanabe et al., 2002), but only rarely for modeling. Orthogonal polynomials are another rich family that has not been exploited by the fMRI community, despite effectiveness in other neuroimaging modalities (but see Clark 2002, for an application of orthogonal polynomial regression to event-related fMRI data). Further developing these relatively unexamined methods and incorporating anatomical knowledge as in Kiebel et al. (2000) or Harms and Melcher (2003) remain research directions of some promise.

9

Bayesian Methods in fMRI

In some ways the Bayesian framework is ideal for the analysis of functional MRI data. As we have seen in previous chapters, the data are often described in a hierarchical manner, with voxel-level models being embedded in subject-level models, which in turn may be nested in a group-level model. The hierarchical nature of the standard general linear model approach fits well into the Bayesian setting (Friston et al., 2002). Spatiotemporal models are another class that well describe functional neuroimaging data, and these too lend themselves quite naturally to a Bayesian analysis. Indeed, in Chapter 6 we saw several examples of Bayesian spatial or spatiotemporal analyses (for example, Gössl et al. 2000; Hartvig and Jensen 2000; Gössl et al. 2001; Smith et al. 2003). In addition, the basis function approaches that incorporate anatomical information, as discussed in Section 8.2, have a distinctly Bayesian "flavor" even if they aren't explicitly Bayes methods.

The well-known criticisms of classical significance testing – the non-intuitive meaning of a p-value, the lack of symmetry between the null and alternative hypotheses (such that the null can never be accepted), increasing sensitivity with sample size so that a "statistically significant" result can always be found given enough observations – have lately come under discussion in the neuroimaging community as well (Friston et al., 2002). As is true for many applied fields, the result of frequentist hypothesis testing (namely, declaring voxels to be significantly active or not) is not necessarily the output that is most easily interpretable or even of interest to neuroscientists. Instead, the posterior probability of the effect being over a specified threshold is possibly a more intuitive measure (Friston and Penny, 2003).

Finally, over the past decade and a half, as fMRI has become more widespread, as well as other imaging techniques, neuroscientists have started to accumulate a wealth of information about brain activation in general. For instance:

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_9,

1. Various robust experimental paradigms such as flashing checkerboards, finger tapping, eye movement tasks (saccades), are known to elicit consistent and reliable activation in particular brain regions.
2. Activation is unlikely to be localized to single, isolated voxels. Instead, true activation should be found in contiguous clusters.
3. Individual differences in the extent and amplitude of activation within an affected region are to be expected, as are fluctuations for a given individual at different scanning sessions.
4. The magnitude of the response is typically about 2% - 5% of the intensity at rest (baseline).
5. The parameters guiding the hemodynamic response may vary from voxel to voxel, or from brain region to brain region.

These pieces of knowledge can be used to build up prior distributions for a Bayesian analysis, in a classic application of the principle that one obtains posterior inference by updating prior knowledge (in this instance, the data gathered over years of fMRI studies) in light of new data. Use of prior information has the added benefit that it constrains the high dimensional fMRI parameter space (Genovese, 2000).

Despite the apparent suitability of Bayesian methods to fMRI data, this line of research has only recently begun to take root in the literature. Part of the issue, not unexpectedly, is computation. The large quantities of data and the complex relationships within and among voxel time series, have precluded widespread implementation of Bayesian analysis, due to the complicated forms of the likelihood and priors. The works we saw in earlier chapters are distinguished by the fact that they all attempt to make simplifying assumptions or find computational shortcuts that would make the analysis feasible. Where this has not been done (for instance, Genovese 2000), in an attempt to develop a fully Bayes analysis, the resultant methods have not been adopted by the neuroimaging community.

A compromise position, advocated for example by Friston (Friston et al., 2002; Friston and Penny, 2003) is the use of empirical Bayes methods, whereby the parameters of the priors are estimated from the data, rather than themselves being subject to prior specification as in a fully Bayes model. Friston and Penny (2003) show how to embed the empirical Bayes approach in the hierarchical general linear model framework, as implemented for example in the software SPM (see Appendix A) and solved using standard statistical tools such as restricted maximum likelihood and the EM algorithm.

Clearly, Bayesian methodology holds promise for fMRI, and equally clearly, there is still much that statisticians can contribute in this area. In this chapter we look at some of the ways in which researchers have tried to tackle the various components inherent in a Bayesian analysis: specification of fully Bayes models and of priors, and computational shortcuts or approximations to ease the burdens imposed by the large scale of the data.

9.1 Fully Bayes Models

The first fully Bayesian model for fMRI data appears to be Genovese (2000) (Worsley, 2000). Almost contemporaneously, however, Kershaw et al. (1999) propose a very similar model; they use noninformative (Jeffreys) priors for all parameters, thereby avoiding the computational issues. By choosing a prior of convenience, rather than one that allows for a true summary of prior information and beliefs, one could argue that this work is not fully Bayes. Significantly, from an historical perspective, these works only address the voxel time course, with voxels taken to be independent (no spatial prior). Within a year or two of Genovese's paper, several models incorporating spatial information already had appeared.

At every voxel, Genovese models the observed signal at time t as

$$Y(t) = \mu + d(t) + a(t; \mu, \gamma, \theta) + \sigma\epsilon(t),$$

where

1. μ is the baseline signal for the voxel, or, the average signal over time in the absence of activation and noise; the prior for μ is a scaled t_1 centered at a fixed value, μ_0.
2. $d(t)$ is the drift in the measured signal over time; it is modeled as a spline, where the number of knots may be large or small, and their location fixed or random; when the number of knots is not known, this is given a Poisson prior, and the locations have a prior that is derived from the Dirichlet.
3. $a(t; \mu, \gamma, \theta)$ is the activation profile, which captures the shape of the hemodynamic response; the hemodynamic response function may include up to eight components: lag-on, attack, rise, lag-off, decay, fall, dip, and skew (see Figure 9.1); θ determines the shape of the response to a single stimulus presentation and γ specifies the amplitude of the signal change relative to baseline; the θ parameters for attack, decay, lag (on and off), and dip have independent gamma distributions, and the priors for rise, fall, and skew are uniform over their respective ranges; the γ parameters are constrained to be non-negative and have a gamma prior when strictly positive.
4. $\epsilon(t)$ is the noise, characterized as white; the prior on σ is inverse-gamma.

An advantage of this modeling framework is that it allows researchers to ask, and answer, questions beyond those of localization ("Where are the active voxels?"). Genovese, for example, studies monotonicity of response, tackling the question of whether or not the strength of the response increases with the difficulty of the task. This is a very hard question to answer using frequentist techniques, but of course one of the strengths of the Bayesian logic is that once the posterior distributions of the parameters are obtained they can be manipulated to answer a wide range of questions about the parameters. Kershaw et al. (2001) likewise use Bayesian methods to test linearity of the hemodynamic response in event-related studies. In addition, the model parameters are

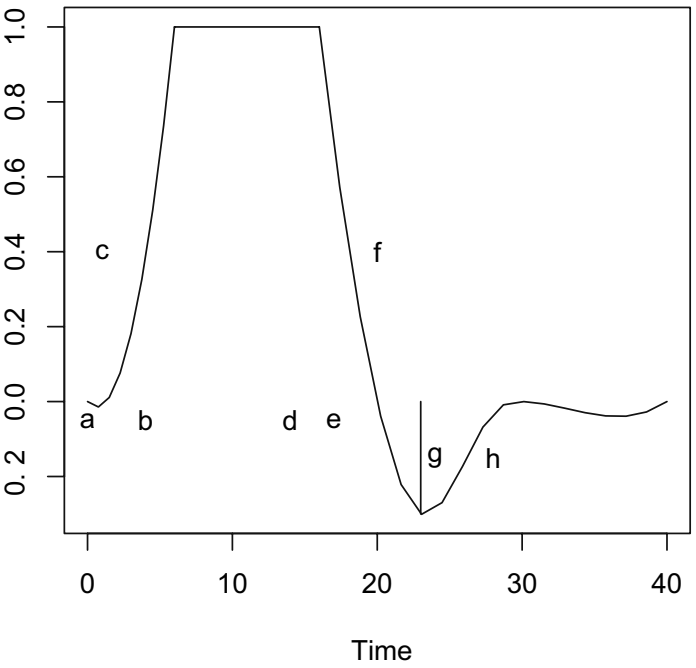


Fig. 9.1. Schematic of hemodynamic response function as given in Genovese (2000). The task is performed from the start of "a" to the start of "d." The parameters of the model, some of which may be omitted for a given analysis, are the time of lag-on ("a") and the length of attack ("b") following stimulus presentation, the strength of the rise ("c"), the time of lag-off ("d") and the length of the decay ("e") after the stimulus or task has ended, the strength of the fall ("f"), the size of the dip ("g"), and the shape of the skew ("h").

easily interpreted in terms of the physiology of the hemodynamic response. Compared to the fixed models of the HRF (for instance, the canonical difference of two gammas), Genovese's approach is more flexible, allowing features such as the poststimulus undershoot to be present in some voxels but not others. This feature is also shared by the basis function approaches from the previous chapter.

On the other hand, no spatial information is incorporated in this analysis, as commented on by several of the discussants to this paper (some of whom suggested simple ways of including spatial priors). And the method is extremely computationally expensive. Genovese (2000) points out (p. 702) that it required time on the order of one day to process the data from a single subject. This is clearly a prohibitive barrier for neuroscience researchers.

Advances in computing memory and power only partially solve the problem; writing four years after Genovese, Woolrich et al. (2004c) cite a running

time of six hours to process a single slice of data using their fully Bayes model. This model, like that of Genovese, posits a model for the hemodynamic response function, but, unlike that of Genovese, also introduces a spatiotemporal noise component.

The noise process is decomposed into two parts: nonstationary deterministic (large-scale variation) and stationary short-scale stochastic variation. The former is removed via high-pass filtering and is not considered in the model. The latter is modeled as a multivariate normal, with mean zero and variance-covariance matrix Σ. Since Σ is over all time points and all voxels, it is very large, hence infeasible to estimate without some simplifying assumptions. Woolrich et al. use a family of *space-time simultaneously specified autoregressive* (STSAR) models, combining a temporally fixed spatial AR(1) with a spatially varying general order temporal AR. That is, they model the short-scale variation at voxel i and time point t as

$$s_{it} = \sum_{j \in \mathcal{N}_i} \beta_{ij} s_{j,t-1} + \sum_{p=1}^{P} \alpha_{pi} s_{i,t-p} + \epsilon_{it},$$

where $\mathcal{N}_i$ is the neighborhood of voxel i, β_{ij} is the spatial autocorrelation between voxel i and voxel j at a time lag of 1, α_{pi} is the temporal autocorrelation between time point t and time point $t-p$ at voxel i, and ϵ_{it} is normally distributed with mean zero and variance that depends on the voxel, but is fixed over time. Through the β parameter, this model can accommodate spatial stationarity or nonstationarity.

For the prior specification on this part of the model, Woolrich et al. (2004c) assume prior independence among the parameters α, β, and ϕ (the precisions of the ϵs). Several options for the distribution of α and β are considered; these include the diffuse normal (to give a noninformative prior), the Markov random field, and the *automatic relevance determination* priors. The latter is a way of automatically adjusting the order of the autoregressive process at each voxel (that is, they allow the order of the autoregression to differ from voxel to voxel) that avoids the computational burden of methods such as reversible jump MCMC (Green, 1995). The precisions of the ϵs, as well as the precisions for the Markov random field and automatic relevance determination priors are gammas with set hyperparameters.

The model for the signal is separated into the height of the response and the assumed shape of the HRF. For the HRF, Woolrich and colleagues propose to add four half-period cosines, resulting in the schematic form shown in Figure 9.2. This model has six parameters: four for the periods of the cosines, the ratio of the height of the poststimulus undershoot to the height of the main peak, and the ratio of the height of the initial dip to the height of the main peak. The priors on the half-periods are taken to be uniform with relaxed time ranges to represent the relative uncertainty about when the peaks, dips, and so forth will take place (across voxels and experimental paradigms, as we have already seen, there is a good deal of variability in the timing of these

events). Since it isn't clear that the initial dip or the poststimulus undershoot will occur, the authors apply the automatic relevance determination prior here as well, on the parameters that govern the sizes of these phenomena. Finally, for the purposes of activation height modeling, all voxels are treated independently.

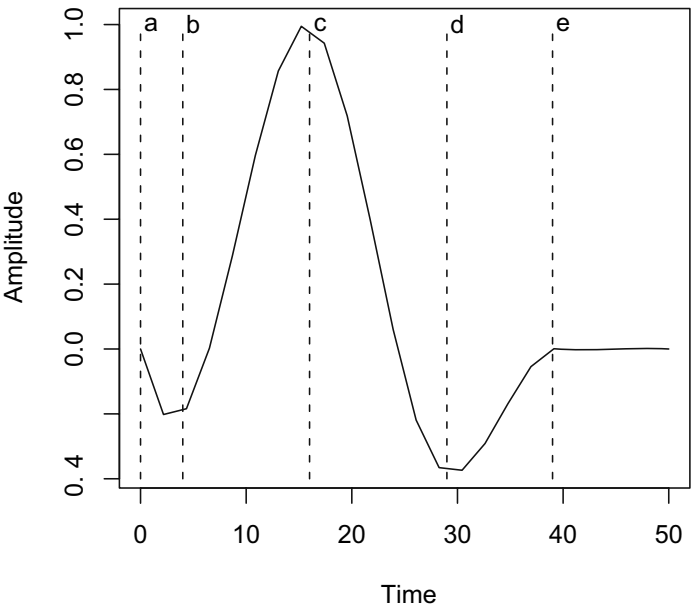


Fig. 9.2. Schematic of HRF used by Woolrich et al. (2004c). The parameters in this model are m_1 (the distance between the vertical dashed lines marked "a" and "b"); m_2 (the distance between the vertical dashed lines marked "b" and "c"); m_3 (the distance between the vertical dashed lines marked "c" and "d"); m_4 (the distance between the vertical dashed lines marked "d" and "e"); c_1 (the ratio of the size of the initial dip below baseline in the period between "a" and "b" to the size of the main peak); c_2 (the ratio of the size of the undershoot below baseline in the period between "d" and "e" to the size of the main peak).

Advantages of this model over that of Genovese are the incorporation of spatial information and a modeling framework that allows for model comparison (an issue that isn't directly addressed by Genovese), for example in the choice of autoregressive order. Also, the formulation of the HRF is appropriate for different experimental designs, whereas the model given by Genovese is aimed at block design experiments. However, as already noted, this approach

is also computationally intensive, even with the simplifications such as filtering out the large scale temporal variation. Further simplifications of the model, including the use of variational Bayes approximations to speed up computation (see Section 9.3) are given in Woolrich et al. (2004b).

A different perspective on the Bayesian analysis of fMRI data is presented in Hartvig (2002), who uses marked point processes to describe the spatial activation pattern. His general model for the signal measured in voxel i at time t (after preprocessing) is

$$Y_{it} = (A_i(X) + \eta_i)\phi_t + \epsilon_{it},$$

where η and ϵ are governed by Gaussian processes, $A(X)$ is the level of activation, parameterized by the marked point process X, and ϕ is temporal variation related to the BOLD response.

Consider for ease of exposition the case where data are on a two-dimensional slice (the three-dimensional volume is treated similarly). The interpretation of the points X_j of the marked process is as centers of activation, with location μ_j and marks that describe the magnitude and shape of the center. More specifically, if the marked point process is $\{X_1, X_2, \ldots, X_n\}$, then the activation pattern $\{A_i(X)\}$ is assumed to be a sum of Gaussian functions $A_i(X) = \sum_{j=1}^{n} h(i; X_j)$, where

$$h(i; X_j) = a_j \exp\left\{-\frac{\pi \log 2}{d_j}\left(\frac{l_1^2}{r_j/(1-r_j)} + \frac{l_2^2}{(1-r_j)/r_j}\right)\right\}.$$

$(l_1, l_2) = R(-\theta_j)(i - \mu_j)$ for $R(\theta)$ a rotation with angle θ; $(a_j, d_j, r_j, \theta_j)$ are the parameters of the process at location μ_j, with a_j being the height of the normal density at center μ_j, d_j the area of the contour ellipse at half height (the two-dimensional version of full width at half maximum), $r_j \in (0, 1)$ a measure of the eccentricity of the ellipse, and $\theta_j \in [-\pi/4, \pi/4]$ the orientation of the ellipse.

The mark parameters a, d, and r are assumed to be independent a priori, with the first two given inverse-gamma priors and the last one a Beta prior. These are also a priori independent of μ, which is characterized by an intensity function β. Taking the product of this factorized prior over the points in the process X gives the prior density for the spatial part.

Choosing to emphasize the spatial aspect in his initial analysis and presentation, Hartvig takes a simple model for the time component, ϕ:

$$\phi_t = \sum_i \pi_{t-i} \frac{\mathrm{TR}}{\sqrt{2\pi}3} \exp\left(-\frac{(i \times \mathrm{TR} - 6)^2}{18}\right),$$

where $\pi_t = 1$ during periods of stimulation presentation and zero else, and TR is the repetition time. This corresponds to the typical convolution of the stimulus trail with a function for the hemodynamic response, taken here to be normal.

Simulations from the posterior distribution of the activation centers here is somewhat complex, requiring the use of a modified reversible jump Markov chain Monte Carlo algorithm (Green, 1995) (note that the number of centers is a parameter of the model). The reversible jump steps are inserting a new point, removing an existing point, and changing the position or marks of an existing point. Within these steps the other parameters may also need to be updated, and this is accomplished sequentially, as it is not feasible to sample directly from the full posterior. There are obviously other subtleties and details of the complete implementation, and Hartvig also proposes several modifications to his basic algorithm, including more complicated (and realistic) models for ϕ.

9.2 Priors for fMRI Data

When we set priors for the Bayesian analysis of neuroimaging data, a crucial point is to account for the knowledge that true activity is at least locally homogeneous and occurs in regions that are spatially connected (Penny et al., 2005). This points to a certain lack of symmetry in the treatment of the two components – spatial and temporal – from the Bayesian perspective, with the former demanding more attention than the latter. We have already seen the use of Markov random fields (Gössl et al., 2000; Gössl et al., 2001) and mixture models (Hartvig and Jensen, 2000) as examples of incorporating prior spatial knowledge. Here we consider two more recent efforts in this direction.

First, da Rocha Amaral et al. (2004) define what they term a "multigrid prior." This is really a hierarchy of priors, from the finest scale of individual voxels to the coarsest scale of an entire region of interest (ROI). To move from scale to scale, voxels are gathered into nonoverlapping square neighborhoods of size $2^p \times 2^p$. So, the finest level treats each voxel; at the next level, voxels are grouped into 2×2 neighborhoods that cover the whole region (which therefore must itself be a square of an appropriate size); at the next level these 2×2 neighborhoods are combined into 4×4 neighborhoods that cover the whole region, and so on. Figure 9.3 shows the gridding procedure for three levels. Within each neighborhood, the response is defined as the average response of voxels in that neighborhood, and the model is the average of the models at the individual voxels. The overall global prior is taken to be the proportion of active voxels in the entire region and is uniform across the region. Priors at finer levels of the hierarchy are obtained sequentially from the posterior at the previous level; that is, for a given level k, the posterior probability that a region r_k is active is simply

$$P(A_{r_k}|D_{r_k}) = \frac{P(A_{r_k})P(D_{r_k}|A_{r_k})}{P(D_{r_k})},$$

where A_{r_k} denotes the event that the region r_k is active, and D_{r_k} is the average response in the region r_k. The prior at the finer level $k-1$ is the posterior of the coarse level k; in this way, working down from the coarsest level, priors

and posteriors can be recursively calculated. With the finest level of the grid being individual voxels, the method therefore provides the posterior probability of activity for each voxel, and this incorporates both local neighborhood information (through the immediately closest voxels) and information from neighbors at coarser levels. da Rocha Amaral et al. (2004) claim that their method for specifying the prior can be used with any model, and hence they do not focus on any one in particular.

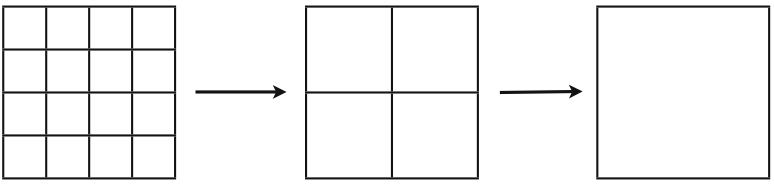


Fig. 9.3. The multigrid procedure. At the finest level, each of the 16 voxels in this example is its own neighborhood. At the second level, the voxels are grouped into square neighborhoods of size 4. At the last level, the entire region is treated as one neighborhood.

By contrast, Penny et al. (2005) set out a detailed Bayesian model with an autoregressive AR(p) error structure built on to the general linear model. The coefficients of the model are taken to be a priori normal with mean zero and variance $\alpha_k^{-1}(S^TS)^{-1}$, where α_k is the spatial precision for variable k (determines the amount of smoothness) and S is a spatial kernel matrix. Furthermore, the overall prior for the regression coefficients factors as the product of the individual priors, enabling different amounts of smoothness (via α_k, which is estimated from the data), although S is shared by all of the regression coefficients (and hence dependencies are allowed even though the prior factors). The α_k are taken to be a priori independent with diffuse gamma hyperpriors, as are the precisions of the observations at each voxel. Finally, the autoregressive parameters are a priori independent across voxels, each with a mean zero Gaussian distribution. Readers will recognize these as conjugate priors for the normal likelihood.

Prior spatial information is quantified in this model via the matrix S. Penny et al. (2005) use the Laplacian operator, depicted graphically in Figure 9.4. As is evident from the picture, by placing negative weights on the cardinal directions this prior penalizes differences among neighbors.

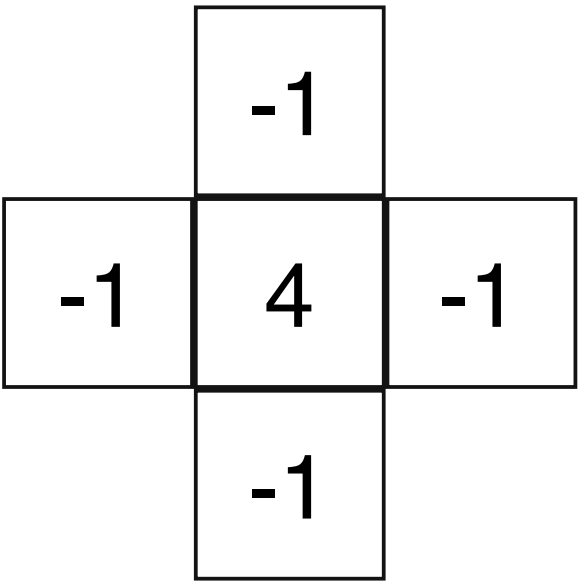


Fig. 9.4. The Laplace operator, used as a spatial prior by Penny et al. (2005). Differences among neighboring voxels are penalized by negative weights.

9.3 Computation

Although Bayesian methods hold great promise for fMRI data analysis, computation of posterior distributions has been an impediment to more widespread implementation within the neuroimaging community, as we have already seen. Simplifying assumptions in the model or the likelihood are often made to enable computation, weakening the attraction and power of the approach. Recognizing this, attempts have been put forth, again in recent years, to alleviate the computational burden in different ways.

Variational Bayes, for instance, is gaining popularity within some segments of the fMRI world. This is a family of techniques for approximating the required integrals and densities, thereby avoiding the need to perform computationally expensive Markov chain Monte Carlo simulations. It appears to have been first considered in the fMRI context by Penny et al. (2003), and is implemented as well in Penny et al. (2005).

Consider a general posterior density, $p(\theta|y)$ in standard notation. The logarithm of the marginal likelihood $p(y)$ can be written as

$$\log p(y) = \int q(\theta|y) \log p(y) d\theta,$$

for $q(\theta|y)$ an arbitrary density. Using the definition of conditional density, and then multiplying and dividing by $q(\theta|y)$, it is easy to see that

$$\log p(y) = \int q(\theta|y) \log \left[\frac{q(\theta|y)p(y,\theta)}{p(\theta|y)q(\theta|y)} \right] d\theta.$$

Define

$$F = \int q(\theta|y) \log \left[\frac{p(y,\theta)}{q(\theta|y)} \right] d\theta$$

and

$$KL = \int q(\theta|y) \log \left[\frac{q(\theta|y)}{p(\theta|y)} \right] d\theta;$$

then

$$p(y) = F + KL.$$

KL is the Kullback-Leibler distance between $q(\theta|y)$ and the target $p(\theta|y)$.

Now, KL is non-negative, attaining the value of zero when $q(\theta|y)$ is equal to the true posterior. Therefore, F (the "variational free energy") is a lower bound on the marginal density of y, and is equal to $p(y)$ when KL=0. The density $q(\theta|y)$ is viewed in this setting as an approximate posterior. Variational Bayes aims to maximize F, which brings the approximate and true posterior densities close to each other. Under the assumption that the approximate density factorizes in an appropriate way, Penny et al. (2003) and Penny et al. (2005) give explicit forms for the (approximate) posterior densities, which are of conjugate form due to the choice of priors (but note that the dependence structure included for the regression coefficients means that a naïve MCMC approach would not be suitable).

Several authors have also discussed the possibility of simplifying the Bayesian inferential procedure by breaking the analysis into steps, only some of which involve the calculation of posterior quantities, or by imposing assumptions such as independence on some levels of the hierarchy but not others (Neumann and Lohmann, 2003; Woolrich et al., 2004a; Chen, 2004). We consider Neumann and Lohmann (2003) as an example.

Neumann and Lohmann focus on what they call "second level" analysis, that is, group inference. Their method is a hybrid of frequentist and Bayesian approaches in that for the first (subject) level analysis they use the ordinary general linear model without priors, and only incorporate a Bayesian scheme at the second level. For each subject, the model $Y = X\beta + \epsilon$ (or some variant of it) is fit by generalized least squares to obtain the parameter estimates $\hat{\beta}$. An effect of interest is given by $c\hat{\beta}$, for c a vector of weights summing to zero that define contrasts (note, however, that if a single effect is of interest, c will be 1). Suppose in addition that there are k subjects. Under typical normality assumptions on the likelihood and the prior, i.e., contrasts $c\hat{\beta}_i$ are normal with variance σ_i^2, the posterior distribution of the combined contrast is also normal, with mean

$$\frac{\sum_{i=1}^{k} \sigma_i^{-2} c\hat{\beta}_i}{\sum_{i=1}^{k} \sigma_i^{-2}}$$

and variance

$$\frac{1}{\sum_{i=1}^{k} \sigma_i^{-2}}.$$

These will be recognized as the usual Bayesian updates for normal likelihoods with conjugate priors, or, indeed, simply as the estimates from a fixed effect model (see Section 5.5.2).

One can then carry out standard Bayesian manipulations to assess posterior probabilities of activation across the group (simple in this case since the density is normal), to compare subgroups, to compare effects of covariates within a subject, and so forth, via the use of appropriate contrasts.

A similar approach is suggested by Woolrich et al. (2004a). The main difference between the two is the choice of prior; whereas Neumann and Lohmann stay within the conjugate framework, Woolrich and colleagues introduce reference priors as a way of modeling ignorance (Bernardo and Smith, 2000). This necessitates evaluation of complicated integrals; Woolrich et al. (2004a) explore a fast approximation and a slower Markov chain Monte Carlo approach. In both Neumann and Lohmann (2003) and Woolrich et al. (2004a), summaries from the first level analysis, which may in fact be frequentist or Bayesian, are used as input to the higher levels in the hierarchy. Both sets of authors indicate that their methodology can be extended to more than two levels (for instance, scan within a session, session within a subject, subject within a group).

As with many of these developments, the statistical novelty represented here is minimal. The importance of such works lies rather in bringing to the attention of the neuroimaging community both the existence of Bayesian methods and the proof of the feasibility of their application. "Shortcuts" such as that suggested by Neumann and Lohmann (2003) are well-known in the statistical literature; in order for Bayesian techniques to penetrate into this new domain and become accepted, users need to know that they can still carry out their analyses in a reasonable amount of time, which is a major concern with some of the "fully Bayes" procedures that have been put forth. The fact that several groups of researchers have arrived independently and approximately simultaneously at similar solutions (in "flavor" if not in all details of implementation) is evidence that there is a recognition among neuroimaging scientists of the usefulness of Bayesian techniques, as well as a desire to put these into practice. Devising computational and other tools to facilitate Bayesian analysis of large, complex data sets is a challenge that statisticians will need to face more fully.

9.4 Conclusion

The last decade has seen a growing awareness among fMRI researchers that the logic and philosophy of the Bayesian paradigm could prove advantageous to the goal of building and drawing inference from the "correct" spatiotemporal models ("correct" in the sense that they more accurately reflect the known or surmised spatial and temporal correlation structures, than the voxelwise linear model in any of its guises; in particular this is true of the spatial structure). At the same time, advances in computing capacity and memory, and the development of new statistical methodologies for handling large, complicated data sets, have made the application of Bayesian procedures more feasible in general. The potential of Bayesian fMRI data analysis has not yet been fully exploited. I expect this to continue to be an area of active and fruitful collaboration between statisticians and neuroimaging scientists.

10

Multiple Testing in fMRI: The Problem of "Thresholding"

Regardless of which specific methods are used to analyze the data and to create a statistical parametric map of the brain, one cannot entirely avoid the question "Which voxels show significant levels of activation compared to the control?" While it is possible to argue that a simple dichotomizing of each individual voxel into active/not active does not truly answer the question that is of scientific interest (Jernigan et al., 2003) – *this* would more likely focus on the behavior of regions – the reality is that this is the question that often, in practice, *is* asked, as well as answered.

From a statistical perspective, it is clear that the task of classifying each of the tens, or hundreds, of thousands of voxels in a typical study as "significantly active" or not is a formidable problem of multiplicity. In any statistical test, a binary decision (significant/not significant) is made. The true state of nature is also binary (significant/not significant). 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 10.1).

	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

Table 10.1. 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 – leads each voxel in the brain to fall in one of four categories, as shown above. Interest generally focuses on attaining control over the voxels falsely declared active.

In total, there are m voxels tested (typically in the hundreds of thousands). Of these, $m_{.1}$ are declared active (null is rejected), and $m_{.0} = m - m_{.1}$ are declared inactive (null not rejected). In reality, $m_{0.}$ voxels are inactive (the

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_10,

null hypothesis is true) and $m_{1\cdot}$ are active. We are interested mainly in the m_{01} voxels for which the null is rejected, when it is in fact true. These are the *false positives* or *false discoveries*. 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 over the ensemble of tests. More recently, methods that control the *false discovery rate* (FDR), or the (expected) proportion of voxels erroneously declared to be active, out of all voxels declared active, have become popular in statistics. This is due, in part, no doubt to the proliferation of large data sets, not only in functional neuroimaging, but also in fields such as microarray analysis, where control of FWER may be inappropriate and is, in any case, too conservative.

Psychologists, accustomed to controlling for multiple testing by the use of the Bonferroni correction for FWER, for example, early on found that this method was much too conservative for their purposes in fMRI. By this, they meant that the Bonferroni-adjusted significance level of α/m, where m is the number of tests, α is the overall significance level, and m is very large, results in a criterion that is so strict, that even in areas where activation "should take place" (according to theory or prior knowledge attained using other techniques), nothing can be detected.

The extreme conservativeness of the Bonferroni method, coupled with its inability to take into consideration the particular features of fMRI data (such as the dependence among voxels), requires other techniques for error control. In this chapter, we discuss some of the approaches currently in use in the fMRI community, as well as some that have been proposed but not widely adopted. Sections 10.1 through 10.5 describe 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 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 instead of the familywise error rate (as is done, for instance, by the Bonferroni correction), which offer advantages in terms of power, ease of use, adaptability, and interpretability; and an ad hoc method, which involves setting the threshold by eye, based on the practitioner's experience and knowledge. The first three of these can be seen as variations on a single theme. We then examine some recent proposals for detecting regions of activity by directly using the properties of estimated hemodynamic response functions from an event-related study, and more generally by working with the fMRI time series. The survey concludes with a look at various other methods that have been suggested in the literature. The last two sections of this chapter look at two recent comparisons of thresholding techniques used in fMRI data analysis, and explore some other issues of relevance.

10.1 Cluster Thresholds

An early attempt to grapple with the multiplicity issue is found in work by Forman et al. (1995). The approach here is to note that true activation is likely to be spread over several contiguous voxels, since these are themselves simply rather arbitrary divisions of the brain, with no intrinsic physiological meaning. Indeed, most researchers consider isolated single active voxels to be spurious discoveries, and so tend to ignore them. On the other hand, in a null map containing no true activation, individual voxels that do attain significance are more likely to be scattered throughout the brain rather than form clusters.

It is easy to confirm these conjectures via a small-scale simulation. Here, 200 16 x 16 grids of values uniform on [0,1] are generated. Any value under 0.05 is declared "significant" and the following summaries found for each simulation: the total number of clusters (where a cluster is defined as by Forman and colleagues, to be made up of neighboring pixels, a "neighbor" being any of the eight pixels bordering a given pixel); the sizes of the clusters; and the total number of significant pixels. Over the 200 simulations the average number of clusters is 10.215 and the average number of significant pixels is 12.285 (close to the nominal 0.05 level). Out of the 2043 total clusters, 1699 of them, or just over 83%, are singletons, made up of a single pixel; 289 clusters (14%) are of size 2; 37 clusters (2%) are of size 3; 16 clusters (just under 1%) are of size 4; and 2 clusters (less than one-tenth of one percent) are of size 5. There are no clusters over size 5, but it is noteworthy that clusters of "significant pixels" will of course be found even in null data, by the definition of type I error.

Under this specification of the thresholding problem, there are two elements that determine the probability of making a false discovery (that is, declaring an inactive voxel to be active): (i) the criterion for rejecting the null hypothesis for a given voxel at level α, call this $C(\alpha)$, and (ii) S, the size of a contiguous cluster of active voxels. In order to be considered active, a voxel must cross the value $C(\alpha)$, and enough of its touching neighbors must as well, to form a contiguous cluster of size S. If these two conditions are not both met, significance is not attained. In particular, if the cluster size threshold is set at S, no active clusters smaller than S in size will be found. But the user has the ability to consider tradeoffs between the size of the cluster and the threshold for declaring voxels to be active. For example, if $C(\alpha)$ is fixed, as S increases the probability of detecting false positives decreases. The reason for this is that the probability of v voxels all exceeding $C(\alpha)$ and being contiguous with each other is lower than the simple probability of those v voxels exceeding the threshold. The extra condition of contiguity has real implications in terms of the relevant probability calculations. By combining the two thresholds, $C(\alpha)$ and S, we lose the ability to detect small (smaller than S) areas of activation, but we gain power to detect larger areas. Under the assumption that real activation will indeed spread over several voxels, this tradeoff can be expected to work to the benefit of the researcher.

Forman et al. explore combinations of S and voxel-level values of α in the typical range of 0.005 to 0.2, for both uncorrelated and correlated data. The method is entirely simulation-based, empirically finding the expected number of clusters of a given size, for a given level of voxel significance, under the null situation of no truly active voxels. From these simulated maps it is then possible to calculate the probability of finding a false positive voxel, on a per voxel basis (as opposed to over the entire cluster), as a function of S and α. These in turn provide the requisite $C(\alpha)$ values to be applied to each voxel in the cluster. Correlation is induced in the simulated data by using Gaussian filters of varying widths, with small amounts of smoothing representing less correlation and large amounts of smoothing representing more correlation. Implementations of this thresholding technique in fMRI software such as AFNI (see Appendix A) allow the user to estimate and specify the amount of smoothing suitable for a given data set when calculating the voxel significance thresholds for a particular S threshold and overall level of significance.

Figures 10.1 and 10.2 show how the probability of a false positive voxel, per voxel, changes as a function of the cluster size threshold, S, for uncorrelated and correlated voxels respectively. Each line in Figure 10.1 represents an overall α level, for $\alpha = 0.05, 0.03, 0.025, 0.01, 0.005$, from the highest line down. As can be seen in the figure, increasing the cluster size for a fixed α results in decreasing probability of a false positive voxel, per voxel, as would be expected. Indeed, for clusters as small as 5, for the various overall levels of significance, the probability per voxel for voxels in those clusters is small enough to be indistinguishable from zero.

Similar conclusions are reached from examination of Figure 10.2. Each panel in the figure represents a different overall level of significance; within each panel, the probability of a false positive voxel, per voxel, is plotted as a function of cluster size for different amounts of smoothing. The highest line in each panel corresponds to the greatest amount of smoothing, that is, the highest amount of induced spatial correlation. Again, as the size of the cluster increases, for fixed α, the probability per voxel in the cluster of a false positive declines to zero. For larger amounts of spatial correlation, the decline is slower. This is also not surprising, as we would expect that voxels that are spatially correlated would be clustered together, and if one is a false positive, its neighbors in the cluster might be as well. As the amount of correlation decreases, so should the effect of any given voxel on its immediate neighbors, and the clusters will be made up of more "nearly independent" voxels.

As demonstrated by Forman et al., the use of cluster thresholding methods leads to increased power to detect true activity, particularly for uncorrelated data, but also in the correlated case. Based on their results, the authors recommend taking $7 \leq S \leq 9$ and α of between 0.02 and 0.03 to gain most of the power benefits of the method, while still being able to detect reasonably small clusters of activity.

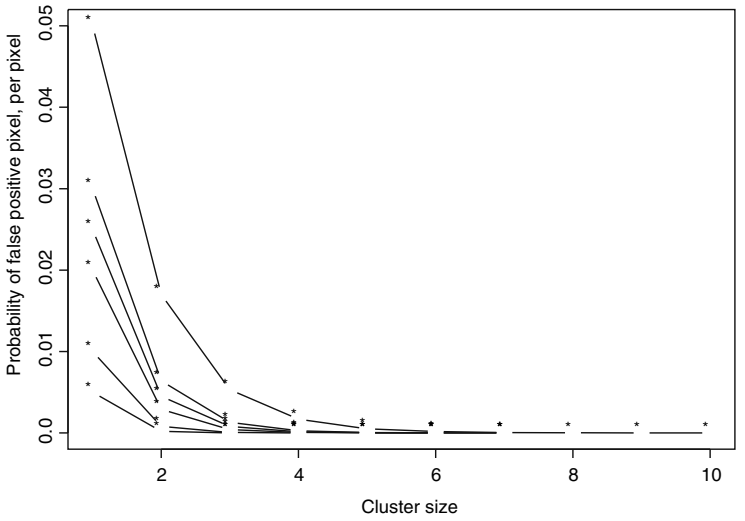


Fig. 10.1. Probability of a false positive voxel, per voxel, for overall α levels ranging from 0.005 to 0.05, as a function of the size of the cluster threshold. Voxels are assumed to be uncorrelated. Adapted from Table 1 in Forman et al., 1995.

10.2 Random Field Theory

Like the cluster threshold method of the previous section, both random field theory, the topic of this section, and permutation thresholds, the topic of the next, aim at detecting contiguous groups of active voxels. Random field and permutation thresholds achieve this, as pointed out by Nichols and Hayasaka (2003), by considering the distribution of the maximum of the statistical map, or equivalently of the minimum p-value, to account for dependence in the data and hence to encourage detection of clusters of activity.

The random field approach to thresholding relies on a quantity called the *Euler characteristic* of the *excursion set* (Worsley, 2003). The excursion set describes those voxels that are above the specified threshold. For high thresholds, the Euler characteristic of that set roughly corresponds to the number of local maxima (or clusters). For lower thresholds, the Euler characteristic counts the number of connected clusters minus the number of "holes;" when the threshold is high enough, the holes disappear and only the clusters remain. As the threshold increases to approach the maximum statistic in the map, the Euler characteristic will take on the value 1 if the maximum is above the threshold, and zero otherwise (Worsley, 1996). That is, the Euler characteristic is, for high enough thresholds, approximately an indicator function

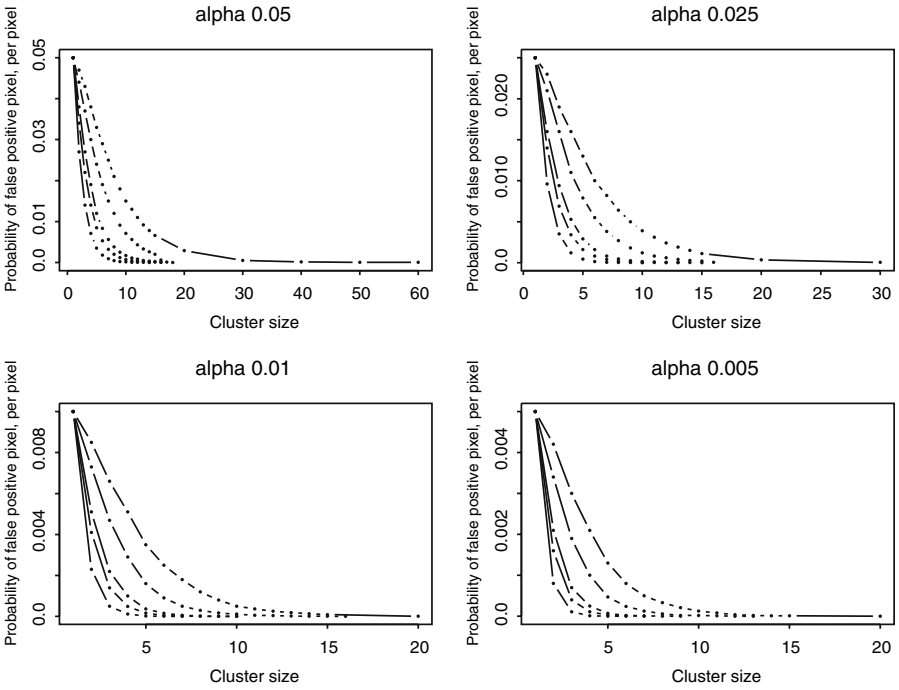


Fig. 10.2. Probability of a false positive voxel, per voxel, for overall α levels of 0.005, 0.01, 0.025, and 0.05, as a function of the size of the cluster threshold. Voxels are assumed to be correlated, with differing amounts of correlation being induced by kernel smoothers of differing window widths. Adapted from Table 2 in Forman et al., 1995.

for having any voxel above threshold. This establishes the connection to the control of FWER.

Although the theory governing random fields is complicated, the procedure has gained widespread popularity in the fMRI community, partly due to the fact that relatively simple and easily interpretable formulae for the quantities in question have been derived by Worsley and his colleagues. This has allowed the random field methodology to be implemented in fMRI software packages such as SPM (see Appendix A). For example, Worsley (2003) describes the approximate probability that the maximal statistic $t_{\max}$ is above the threshold t by

$$P(t_{\max} \geq t) \approx \sum_{d=0}^{D} \mu_d(S)\rho_d(t),$$

where S refers to the search space, in our case the brain, $\mu_d(S)$ is the "d-dimensional intrinsic volume" of the search space, and $\rho_d(t)$ is the

"d-dimensional Euler characteristic density" of $t(s)$, the statistic calculated at the point $s \in S$. The last contribution in the sum, for $d = D$, tends to be the most important; the other terms are boundary corrections.

When $D = 3$ and S is a ball of radius r, the intrinsic volumes are given by $\mu_0(S) = 1$ (Euler characteristic), $\mu_1(S) = 4r$, $\mu_2(S) = 2\pi r^2$ (0.5 times the surface area), and $\mu_3(S) = (4/3)\pi r^3$ (volume). The Euler characteristic density depends on the type of test statistic as well as the threshold, hence it has to be worked out individually for each type of random field. For Gaussian random fields see Adler (1981); for χ^2, t, F random fields see Worsley (1994). A parameter that appears in the expression for the density terms is related to the roughness of the field, and this can be estimated from the data.

The random field approach is attractive in large part because of its generality; it can be applied to situations as diverse as brain imaging and studies of the structure of the galaxy, since the theory is based on performing inference for a random image. In spite of the difficult mathematics that lie at the foundation of the method, peaks and excursion sets are rather intuitive and lend themselves to natural scientific interpretations in the brain imaging context.

10.3 Thresholds Obtained via Permutation

In contrast to the theoretical random fields approach, which relies on various rather strong assumptions such as having images that are smooth, and uniformly so, as well as having a high enough threshold that the EC approximation holds, the permutation method for obtaining cluster thresholds (Holmes et al., 1996; Bullmore et al., 1996a; Nichols and Holmes, 2001; Friman and Westin, 2005) relies solely on exchangeability under the null hypothesis (Hayasaka and Nichols, 2003).

The basic idea, even in the neuroimaging context, is a familiar one. If there is no difference between experimental conditions (that is, under the null hypothesis), then the labels "rest" and "task" (for example, in a simple block design study) can be thought of as arbitrary in the sense that any observation arising from the "task" condition could just as readily have been an observation from the "rest" condition, 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 "rest" and "task" among the observations. For each such permutation the relevant test statistic (say, a t test) is computed, and the observed value of the statistic is then compared to this permutation distribution.

To control for multiple testing and 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. Nichols and Holmes (2001) describe a way of doing this through the use of a *maximal statistic*, thereby also forging the connection with Worsley's random field theory from Section 10.2. Essentially, the maximal statistic is as its name implies, the maximum voxel value in an

image, which acts as a summary of the image as a whole. Nichols and Holmes consider two types of threshold: a *single threshold test* and a *suprathreshold cluster test*.

The single threshold test thresholds the image at a given critical value. If all voxels are below that critical value, then the overall null hypothesis of no activation cannot be rejected. If even one voxel is above the threshold, then a rejection occurs. Hence, the maximum of the image becomes relevant for this type of test, and more specifically its distribution, which is obtained via the permutation method. For a fixed overall level of significance α and number of permutations N, the critical threshold is calculated as $c + 1$, where c is αN rounded down to the nearest integer. This represents the $100(1 - \alpha)$ percentile of the permutation distribution of the maximum. The null hypothesis is rejected at *any* voxel with value exceeding the critical threshold based on the distribution of the maximum. Note that this test is still performed at the voxel level. Holmes et al. (1996) prove that this permutation test has strong control over the type I error, experiment-wise.

The second type of test considered by Nichols and Holmes, the suprathreshold cluster test, relates more formally to the methods outlined in Section 10.1. Here, the aim is to discover and assess the significance of clusters of connected voxels that are all above some predetermined threshold. Only contiguous regions larger than a specified size are considered to be active. Now what is needed is the distribution of the largest cluster above the initial threshold, under the null hypothesis, which again is determined by permutation methods. Although this approach is more powerful, as we have already seen, a main drawback pointed out by Nichols and Holmes is that no tests are conducted for individual voxels, but rather only for clusters of connected voxels. Therefore, one cannot say that an individual voxel is significant, only that it belongs to a significant cluster. In practice, this is most likely the statement that researchers would prefer to make, thus the inability to declare significance at the voxel level does not seem to me to be a serious flaw in the procedure.

Another use of the permutation test together with ideas from cluster thresholding can be found in Hayasaka and Nichols (2004), which deals with evaluating the extent of cluster activation using permutation methods to assess significance. They suggest combining information on voxel intensity with information on cluster extent using a variety of simple *combining functions*. Let P_i^I be the (corrected) p-value for the peak intensity of the ith cluster and P_i^S the (corrected) p-value for the size of the ith cluster; then one can define the combination of these two effects in various ways, for example, considering $T_i^T = \min(P_i^I, P_i^S)$ (the superscript T denoting that this minimum statistic was suggested as a combining metric by Tippett, as early as 1931) or $T_i^F = -2(\log P_i^I + \log P_i S)$ (Fisher, 1950). See Lazar et al. (2002) for some other suggestions of combining functions. Instead of evaluating the significance of these combined statistics according to their theoretical distributions, which have been worked out in many instances, Hayasaka and Nichols propose a multilevel permutation approach, first deriving corrected p-values for

P_i^I and P_i^S by the usual permutation tactic; then doing the same for the combined statistics, and finally, if desired, defining and assessing the significance of a *meta-combining statistic*, which compares the individual combining functions. Another feature of the Hayasaka and Nichols approach is that it is easy to give more weight to one or the other of the cluster attributes, either size or peak intensity; whereas evaluating significance of these modified statistics might be difficult using theory-driven results, it is of course trivial to do this via permutations.

As with many applications of permutation tests, in many types of neuroimaging experiments the number of possible permutations will tend to be very large, making it infeasible to enumerate them all. In this case it is acceptable to take a random subsample of the set of all possible permutations. Care should be taken to make this random sample large enough that desired levels of significance can, in principle, be attained. If one wishes to set $\alpha = 0.001$, for example, at least 1000 permutations should be considered. Any less than that and it will be impossible to achieve the significance level, even if the observed data configuration is more extreme than any of the other permutations contributing to the empirical distribution.

10.4 Control of the False Discovery Rate

Whereas the Bonferroni correction controls for the familywise error rate, that is, the probability of erroneously identifying even a single null value as significant, other types of error control are possible. In the context of fMRI specifically, it is not reasonable to control the familywise rate, since scientists care about the overall picture of activation, and not any one particular voxel. The lack of power and conservative nature of Bonferroni 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, and was suggested by Benjamini and Hochberg in 1995. It quickly became apparent that this powerful, intuitive, and easy to implement procedure would have widespread applicability for the analysis of large datasets. Control of the FDR was introduced to the neuroimaging community by Genovese et al. in 2002 (Genovese et al., 2002).

A convenient step-up procedure for controlling the false discovery rate at level q, under rather weak assumptions, is as follows (see Figure 10.3):

1. Order the m p-values in increasing order,

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m-1)} \leq p_{(m)},$$

where the hypothesis corresponding to $p_{(i)}$ is denoted $H_{(i)}$.

2. Let r be the largest i such that

$$p_{(i)} \leq q\frac{i}{m}.$$

3. Reject the null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(r)}$.

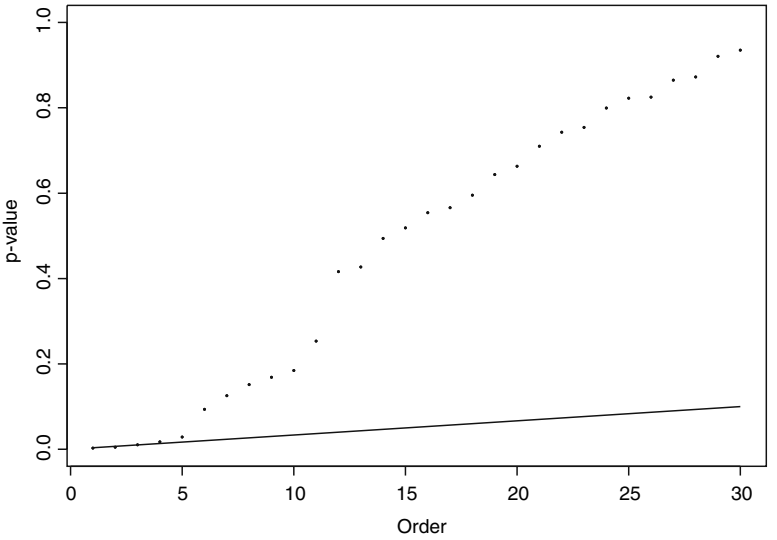


Fig. 10.3. Calculating the threshold to control the false discovery rate using the Benjamini and Hochberg procedure with $q = 0.1$. The p-values are ordered from smallest to largest. The ith p-value is compared to $q(i/m)$, which is the solid line in the plot. The largest p-value to fall below that line is the cutoff point for significance.

If no hypotheses are rejected, that is, the criterion in the second step of the procedure is not met for any i, then the false discovery rate is defined to be zero.

The parameter q can be interpreted in this context as the proportion of false discoveries in the data that the researcher is willing to tolerate. As such, it has a more objective, or at least understandable, interpretation than does a p-value. While there is a tendency to set q values similar to standard p-values, such as 0.05, there is in fact no practical reason for doing so. Values of q as high as 0.15 or 0.2 have been found to work well in some settings (Benjamini, personal communication); in much of my own work on fMRI data, I have found more conservative q values, even as low as $q = 0.01$, to be appropriate, 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. Furthermore, the q value is adaptive to the levels of activation in an fMRI statistical map, meaning that the same value of q will give different "equivalent thresholds" on the scale of the test statistic, for each subject in a study. This obviates the need to set one threshold that provides "good looking" results for all subjects.

This is demonstrated in Figure 10.4, which shows the ordered p-values for two simulated subjects, one a "high activator" and the other a "low activator." For the "low activator," 200 of the 1000 simulated p-values are taken from a uniform distribution on $(0, 0.2)$ and the other 800 on a uniform $(0, 1)$. For the "high activator," 250 of the simulated p-values are taken from a uniform distribution on $(0, 0.01)$, 250 on a uniform $(0.01, 0.05)$, 250 on a uniform $(0.05, 0.15)$, and 250 on a uniform $(0.15, 1)$. Note that the curves of the ordered p-values look very different, so that the value at which the FDR criterion line crosses the p-value curve is different for both subjects. Even though q is taken to be the same, 0.05, for the two subjects, the p-value thresholds that are determined by the FDR procedure are different. This in turn translates into different thresholds on the scales of the original test statistics, be they t, F, or something else. Using the Bonferroni adjustment, the same p-value threshold would have applied in both cases. For the "low activator" the procedure to control FDR in this simulated example picks out 6 active voxels, whereas the Bonferroni correction to control FWER does not pick out any. For the "high activator," by contrast, the procedure to control FDR finds 255 active voxels; using the Bonferroni correction, even on this simulated subject who was designed to have a concentration of small p-values (which would often be indicative of significant activation), not a single active voxel is detected.

Finally, in their original paper Benjamini and Hochberg (1995) show that the step-up method for the control of FDR is more powerful than the Bonferroni correction, making it especially attractive in settings where there are vast numbers of tests to be carried out, a scenario that is increasingly prevalent in many scientific disciplines. As a result there has been a flourishing of research, both theoretical and applied, on FDR methods in the past five years especially. It is outside the scope of this chapter to survey all of the recent research on FDR and related procedures, but we return to some aspects of this work that seem particularly relevant for the analysis of fMRI data in Section 10.9.

10.5 An Ad Hoc Method

We briefly mention also what might best be termed an ad hoc thresholding method. This is based, not on any statistical reasoning or justification, but rather is driven by the practical need to decide which voxels are indeed active. The method is simple, namely, to arbitrarily set a threshold on the test

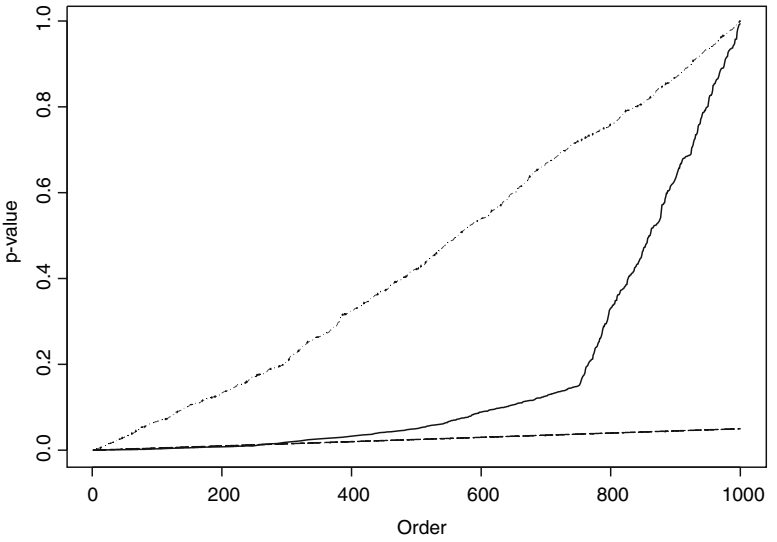


Fig. 10.4. A simulated "low activator" (dotted line) and a simulated "high activator" (solid line), together with the FDR criterion line (dashed-dotted line) using $q = 0.05$.

statistic according to what "looks right." With this approach, researchers rely on their knowledge and prior beliefs about what an experiment should yield in setting the threshold. For instance, for many data sets collected on typical 1.5T magnets, a t-map threshold of 3 or 4 has been found to work well by this standard. However, it does not provide any guidance on how to threshold data sets (from different subjects, different tasks, different types of magnet, or magnets of different strengths) for which the arbitrary threshold does not work. As circumstances change, for example as laboratories move from 1.5T scanners to 3T scanners, it is usually necessary to adjust the threshold, which involves searching (by eye) for the new "good performer." Clearly, such an approach is unsatisfactory from a statistical perspective, even if a given arbitrary threshold can be justified a posteriori by appeal to the Bonferroni correction, control of the FDR, or other means. Still, by definition, it gives "good" results from the practitioner's perspective.

10.6 Procedures Based on fMRI Time Series and the HRF

A more recent trend has been to use the fMRI time series itself, or estimates of the HRF obtained from an event-related experiment, to detect regions of

activation and hence to indirectly threshold the maps. For the most part, the procedures in this section take advantage of characteristics of the time courses to distinguish those voxels whose behavior is indicative of activity.

Clare et al. (1999)use an analysis of variance approach to detect voxels that exhibit a consistent pattern of response to the stimulus over time in event-related studies. When a voxel shows a repeatable response to a stimulus, the authors argue, averaging across trials of a given type will not diminish the variance of the trial time course, and indeed the response will be highlighted by gathering strength across like trials via the averaging operation. By contrast, trial averaging will tend to reduce both the signal and the variance of the trial time course for nonactivated voxels, since the pattern of response is essentially random. The method then compares the variance within time points to the variance between time points at each voxel, where both voxels and time points are treated as being independent. Under the null hypothesis of no activation, and following the usual ANOVA theory, the ratio of these two quantities will follow an F distribution with appropriate degrees of freedom. Once this F map is created, the authors still need to resort to thresholding methods described in previous sections, such as random field theory, but the interesting step in this procedure is to exploit the different characteristics of the time courses of active and nonactive voxels.

Fuzzy clustering is proposed by Fadili et al. (2000) as a way of both exploiting the features of the fMRI response curves and avoiding standard thresholding approaches. Instead of thresholding a statistical map and confronting the multiplicity problem, these authors suggest (fuzzy) clustering of the voxel time series to identify clusters of interest. Clusters of interest would be those in which the voxels show behaviors characteristic of active response to a stimulus. A first step in the algorithm is to reduce the number of voxels under consideration. The reason for this is that the number of active voxels is expected to be small relative to the total number of voxels in the brain, a fact which can produce difficulties for any clustering algorithm. Two types of voxel reduction are used: elimination of time courses that are statistically no different from white noise, and removal of voxels that are not in the gray matter tissue of the brain. Following this step the fuzzy clustering algorithm itself is applied.

Two crucial elements of the fuzzy clustering algorithm are the number of clusters, which needs to be predetermined, and the *fuzziness index*, which measures the "strength" of the partitioning. High values of the index indicate a "fuzzier" partitioning, whereas small values of the index indicate a "crisper" partition (in the language of Fadili et al.). Based on simulations, the authors recommend setting the fuzziness index in the range 1.5-2.5, with the value 2 being a popular choice in the fuzzy clustering literature. They also propose a procedure for determining the number of clusters prior to implementation of the clustering algorithm.

Gibbons et al. (2004) propose an approach that allows for simultaneous estimation and classification of the hemodynamic response in an event-related

fRMI study. It is the classification of the response curves that creates an activation map under their specification. The first step in the procedure is to average the trials; hence if the experiment consists of k repetitions of "present stimulus for one second, followed by t seconds of rest" there will be in total $k(t+1)$ observations prior to averaging, and only $t+1$ after averaging. A third degree polynomial is used to fit the HRF to the averaged signal, treating the coefficients of the curve as random effects. The authors estimate the random effects using an empirical Bayes approach, which allows the estimates in a given voxel to "borrow strength" from neighboring voxels.

Now, depending on the nature of the responses at a particular voxel, it may or may not exhibit the characteristics of the BOLD response, as described in previous chapters. More specifically, recall that when a region of the brain becomes active, changes in the hemodynamic response take place with a lag, or delay, relative to the stimulus onset. Once the response starts to form, it reaches a peak after a certain amount of time, depending on the individual and the task. In the absence of continued stimulation, the response then dies down, returning to baseline after a possible undershoot (see Figure 1.8). Therefore, an activated voxel should have roughly that behavior: a lag, followed by a maximum in the first half of the $t+1$ observations, followed by a minimum in the later part of the observation trail. In addition, the difference between the attained maximum at the peak, and the attained minimum prior to or at the return to baseline, must be large enough that activation can be considered to have taken place. These criteria can be written in terms of the *critical points* and *critical values* of the cubic polynomial, which are in turn one to one functions of the polynomial coefficients. However, unlike the coefficients of the polynomial, the critical points and values have biological meaning related directly to the nature of the BOLD response.

Finally, voxels are clustered using an algorithm such as K mediods and voxels or clusters that do not show the requisite behavior are dropped. Only clusters containing voxels that meet the criteria outlined above are retained. These clusters, and in practice there rarely seem to be more than one or two that qualify, give the active voxels, without any formal hypothesis testing or thresholding. On the other hand, the method relies on averaging across the $k(t+1)$ original observations, resulting in a significant loss of data, and ignoring potentially important differences among the k repetitions. These issues are addressed in Roy et al. (2005).

10.7 Other Techniques

Finally, we turn here to a brief survey of some other procedures that have been proposed in the recent literature for detecting activation in statistical neuroimaging maps. Although not all of these methods were developed for fMRI, the issues are general enough that there should be no question of applicability.

Turkheimer et al. (2000) develop a region of interest (ROI) analysis, with the goal of comparing multiple ROIs across experimental groups of subjects. This is a rather different focus than other methods considered thus far in this chapter, where the emphasis has been on detecting active voxels in individual maps. Nonetheless, this work by Turkheimer and colleagues deals directly with problems of multiplicity. For each ROI in each experimental group, one calculates the linear combination of the mean and median, $\lambda\bar{x} + (1-\lambda)\text{med}(x)$ for some $\lambda > 0$ and x being the observed data in the ROI. The choice of λ is via bootstrap-like methods, and thus may differ for each ROI and for each of the experimental groups, resulting in an adaptive procedure. To compare two groups of subjects at a given ROI, one compares the values of the linear combination, scaling by the pooled sample standard deviation.

The multiple testing problem across the ROIs is attacked using permutation methods, as described in Section 10.3, in conjunction with step-down procedures. For the step-down approach, the test statistics in the k ROIs are ordered from largest to smallest. The empirical distribution of the ordered statistics is obtained via permutation, and the largest observed test statistic is compared to the empirical distribution of the maximum. If the observed value is extreme in comparison to its distribution, the null hypothesis of no difference at the corresponding ROI is rejected. The procedure is then repeated on the next $n-1$ largest test statistics, with regions being rejected and eliminated from consideration until the hypothesis being tested cannot be rejected.

An advantage of the adaptive test statistic is that one need not assume that the distribution of the observations is the same in all ROIs and for all experimental groups, a restrictive assumption that is often made in practice. In their examples, both simulated and real, the authors seem to choose λ only once for each ROI, but in principle there does not appear to be any barrier to letting λ vary by group. Their simulations show that the adaptive statistic does indeed adapt to the underlying distribution of the data, with λ changing as the true distribution of the observation varies. They also demonstrate an increase in power over the usual t test with randomization, for moderate numbers of subjects in each group. On the real data, the correction for multiplicity does indeed result in increased sensitivity to detect differences between two groups of subjects in multiple ROIs.

As two final points, the authors note that their methodology can be extended to the more general problem of detecting active voxels in a statistical map; they also recommend against using their procedure when it is of interest to compare more than two groups of subjects (the traditional "multiple comparisons" problem in statistics).

Levin and Uftring (2001) combine a correlation approach with the cluster thresholding of Forman et al. (1995) to create a "model independent" method for detecting activation. Noting that traditional correlation analysis distinguishes between active and inactive voxels on the basis of the correlation of the voxel time series with a prespecified reference wave (for instance,

the boxcar representing a block design), Levin and Uftring instead advocate repeating each fMRI experiment at least twice on the same subject during the same session, and correlating each voxel with itself. Here, the time series in a given voxel from the first run would be correlated with the time series for that same voxel in the second run. The argument is that voxels that are truly active should show a high level of consistency in their behavior, whereas the time courses of voxels that are inactive will be mostly noise, which will not be repeatable from run to run. Levin and Uftring hence propose a two-stage criterion: first, the correlation between the time series from the two runs must be above a certain threshold, say $r \geq 0.6$; second, the voxel must be part of a cluster of voxels, all of which have correlation above the threshold. As in the work by Forman and colleagues described earlier, Levin and Uftring use simulation to study the expected number of clusters of a given size under the null model, although tables are not provided.

An advantage of this procedure, as demonstrated on a small number of experiments with a small number ($n = 2$) of real subjects is that activation can indeed be detected when no a priori assumptions are made regarding the model that drives the brain response. This is shown in the paper with simple hand clenching experiments, where activation is detected on both the left and right sides of the brain, in response to left and right hand clenches. By contrast, using the standard correlation method, voxels responding to left hand clenching are detected on one side of the brain when the time series are correlated with a block design that is "on" for left clenches and "off" otherwise, and similarly for right hand clenching. The standard correlation approach detects only one side of activity at a time, whereas the bilateral effect is found using the model independent analysis.

A potential drawback is the strong assumption that an active voxel will show a consistent enough behavior that it will attain a high level of correlation with itself in a repeat of the same experiment in a single session. As noted by the authors, this is a different assumption from the test-retest reliability research described in Section 4.3, in which the reliability was not for repeated runs in the same session, but rather across sessions separated in time by weeks or longer. Even with this crucial distinction, the method may still be questioned on the grounds that, particularly for simple tasks, subjects may become habituated to the stimulus, or lose focus in the second run. Furthermore, the requirement to repeat each experiment at least twice in the same session for each subject adds to the expense and burden of carrying out an fMRI study. Subjects would need to spend more time in the scanner, which could lead to problems in recruitment and retention.

10.8 Evaluation of Methods

Interestingly, in light of the severity of the multiplicity problem in fMRI data and of the centrality of the thresholding question, to date not much work has

been done to compare the performance of the most commonly used thresholding methods.

Nichols and Hayasaka (2003) compare techniques for controlling the familywise error rate (FWER). Their survey encompasses permutation/resampling methods, random fields and Bonferroni, as well as corrected versions of Bonferroni, as applied to Gaussian and t images. The choice of thresholding approaches was guided in part by the desire to include methods that could account for at least some of the spatial dependencies in the data. In all, 11 real data sets are examined by these authors, 9 from fMRI studies and 2 from positron emission tomography (or PET, another functional neuroimaging modality) studies. Results from all of the studies are reported here, since imaging modality does not appear to influence the main conclusions. Nichols and Hayasaka find that the threshold for declaring significance is always lowest using permutation methods, often markedly so. Differences among the methods are more pronounced for studies with small degrees of freedom. Compared to Bonferroni, adjusted "Bonferroni-like," and permutation methods, the random field thresholds become more conservative (that is, higher) as the degrees of freedom in the underlying maps decrease.

In terms of numbers of active voxels detected by each of the thresholding procedures considered by Nichols and Hayasaka (2003), the lower thresholds determined by the permutation method obviously lead to more active voxels. Interestingly, however, in 3 of the 11 real data sets in the study, even using the permutation method to set the threshold, no active voxels are detected. The random field theory's conservative thresholds imply that fewer active voxels are found overall; in four of the experiments, no activity is detected. Bonferroni and related methods are between these two extremes. In only one of the experiments is activation found using the permutation threshold, but not by any of the other techniques. These conclusions are confirmed in a series of simulated data examples. Random field thresholds tend to be (overly) conservative unless the data are sufficiently smoothed first and the degrees of freedom of the underlying map are sufficiently large. The permutation thresholds are the most liberal, in general. Similar results are reported in Nichols and Holmes (2001) and in Hayasaka and Nichols (2003).

Another comparative study (Logan and Rowe, 2004) considers the control of three error rates: the simple, voxelwise type-I error without accounting for multiple testing; the familywise error; and the false discovery rate. For each error rate the authors also consider different methods that can be used for their control, ranging from simple procedures carried out on an individual voxel basis, to techniques that account for possible spatial dependencies, as in the comparison carried out by Nichols and Hayasaka (2003). Working with a simulated data set designed to mimic a simple finger tapping experiment, the authors incorporate several different types of correlation structures to test the various thresholding methods. Logan and Rowe, like Nichols and Hayasaka, also consider the effect of smoothing, as this is a common preprocessing step, as we have already seen. Five thresholding methods are compared: unadjusted

for multiplicity; Bonferroni correction; permutation resampling to control the FWER; simple control of the FDR; and a resampling procedure to control the FDR.

Not surprisingly, in all cases the unadjusted method has high power to detect true activations, but at the steep price of many false discoveries. As noted by the authors, every voxel in the simulated image, which also included "air" outside the "brain," is declared active at least once, and usually multiple times, over the course of the simulations, when no correction is made for multiple tests. This is what one would expect.

When the voxels are assumed to be uncorrelated, there is little or no advantage to using a permutation approach to control the error rate, be it FWER or FDR. For both of these, the permutation versions give almost identical results as the simpler, nonpermutation-based thresholds. Again as would be expected from theory, the FDR methods have higher power than the procedures that control FWER, although the former also falsely declare more inactive voxels to be active, even though the rate is controlled at the desired level on average. The conclusions are the same with moderate amounts of spatial correlation: there is no apparent advantage to utilizing a permutation method, and FDR procedures are more powerful than FWER procedures, but with more false positives. The strength of the permutation-based thresholds becomes apparent when there is strong spatial correlation; in that case, the resampling versions for control of FWER and FDR perform better than the simple versions, having higher power to detect true activation. FDR procedures remain more powerful than their familywise counterparts. Finally, smoothing produces a correlation structure intermediate between uncorrelated and moderately correlated in the simulations performed by Logan and Rowe, indicating that there is little to be gained from a permutation resampling approach. The findings from the simulation are also replicated when the authors apply the five thresholding techniques to a real finger tapping experiment: little difference is found between the simple procedure for the control of an error rate and the more complicated, resampling version that incorporates spatial dependence; and FDR procedures discover more active voxels than do FWER procedures, since the former impose less stringent criteria.

Finally, Marchini and Presanis (2004) present a comparison of three thresholding approaches: control of FWER via random field theory, control of FDR, and Bayesian posterior probability thresholding. In terms of power, posterior probability thresholding has the best performance, followed by FDR control. Control of FWER via random fields is the least powerful of the methods considered here. For actual control of the type I error, Marchini and Presanis find that the random field method produces the lowest type I error, and has smaller standard errors of estimates as well. FDR control is again between the other two approaches. This comparison ends on the rather unsatisfactory conclusion that no one method dominates the others, and that the choice of thresholding technique will depend on the balance of type I and type II errors that an investigator is willing to tolerate.

The findings from these comparison studies are not themselves directly comparable, as they employ different thresholding methods. However, it is possible to make some practical recommendations. First, it is apparent that FDR controlling procedures are to be preferred over FWER controlling procedures for fMRI, both because they are more powerful and because they control a rate that seems to be more relevant and interesting to practitioners. Second, the added computational burden from permutation methods may not be worthwhile, unless there is an a priori reason to believe that there is strong correlation in the data. This may be the case, for instance, when considering a group map based on multiple subjects, or when the data have been heavily smoothed (not, admittedly, a common practice). Thresholds based on random field theory may be overly conservative, especially in studies of an exploratory nature. It remains to be seen how the other methods described in this chapter fare relative to those that have already been compared.

10.9 Other Issues

It is evident that the question of thresholding is central to fMRI, and to neuroimaging in general, and has provided the impetus for much new statistical work as well. In this section, we survey several general issues related to the thresholding question.

First, it is worth noting again that there is an alternative to whole-brain analyses, namely analysis of predefined regions of interest. ROIs are areas of the brain that are defined by anatomy, by function, or both, to be of specific interest to the researcher. In such cases, where the focus is a priori on a particular ROI, or perhaps several (for instance, a confirmatory rather than an exploratory study, or an experimental paradigm that is known to have a reliable effect on specific areas of the brain), statistical power can be gained and the multiple testing problem reduced by concentrating on just those areas. In this manner, power is not expended searching for activation in parts of the brain where it cannot, or should not occur. Although the number of voxels in an ROI will be smaller than the number in the entire image, that number may still be in the thousands. Furthermore, the ROI has to be defined on an individual basis for each subject in a study, an effort that can be very time consuming. And if ROIs across subjects are to be comparable, this work must be carried out in a common atlas space, such as Talairach space. Still, in spite of its difficulties, ROI analysis offers a viable alternative that is used, exclusively or in conjunction with a whole brain analysis, in many fMRI studies. When such an analysis is performed, the focus will often be on characterizing the activation patterns within ROIs, for example, the proportion of active voxels, or the extent of activity in the ROI. These measures are then compared across subjects or across experimental groups using standard statistical techniques, and the thresholding question is mitigated to a certain extent (see Lange 2003 for one example of this).

As an intermediate step between whole-brain analysis and ROI analysis, it is also possible, if not entirely common practice, to "strip away" parts of the image that are not valid for activation; the image can be *masked* so that only voxels inside the brain are included, and the brain itself can be further segmented into gray matter, white matter, and cerebrospinal fluid, with activation expected only in the first of these three categories. Both of these actions carry with them the risk of inducing additional error into the analysis, by potentially stripping away relevant voxels. I have often found that working with the whole image is useful for diagnostic purposes: if thresholds are set too low, for example, this can manifest itself in high levels of apparent activation outside of the brain.

Inference can also be sharpened, and power gained, by reducing the number of tested voxels in another way, that is, by first estimating the number of truly null hypotheses $m_{0.}$ and then basing any adjustment for multiplicity on $m - m_{0.}$ instead of on m (Benjamini and Hochberg, 2000; Turkheimer et al., 2001). Ideally, we would like to identify which voxels should be eliminated from the analysis altogether, but this will most likely introduce bias and error, since it is of course possible that a test with a high p-value was in fact generated from the alternative and not the null. Consider the situation in Figure 10.5. It is clear that if any null hypotheses are false, they are probably the ones that correspond to the smallest p-values. It is equally as clear that the null hypotheses corresponding to the highest p-values are probably truly null. What is not so clear is where the transition from "probably false nulls" to "probably true nulls" takes place. In addition, we know that under the null hypothesis, the p-values are uniformly distributed on the interval [0,1], so that if *all* tested hypotheses were truly null, the ordered p-values would be like the order statistics of a U[0,1] random variable, and would fall along the straight line of slope 1. Deviations from this line could then, in principle, be used to get an estimate of the number of "true nulls."

Benjamini and Hochberg (2000) suggest one way of doing this in the context of improving FDR control. Their method is essentially graphical in nature, and is based on the U[0,1] distribution of the p-values under the null hypothesis. When the number of true nulls $m_{0.}$ is less than the total number of voxels tested m, then as described above the p-values from the false nulls will tend to be smaller than those from the true nulls. We are looking for the "break" from the line denoting uniformity, that is, we need to find an estimate of $m_{0.}$, $\hat{m}_{0.}$, based on the largest p-values. Now, in the region of the plot of the ordered p-values where the null holds, the points should still be roughly linear (uniform), and the slope of that line will be $b = 1/(m_{0.} + 1)$. To get an estimate of $m_{0.}$ from this relationship, we simply need to draw a line through the largest p-values that passes through the point $(m + 1, 1)$ (the point in the upper right corner of the plot), with slope $\hat{b}$; then $\hat{m}_{0.} = 1/\hat{b}$.

The question still remains, how many of the largest p-values to include in this calculation? Benjamini and Hochberg propose the following simple method: fit the line using all m points and calculate the estimated slope;

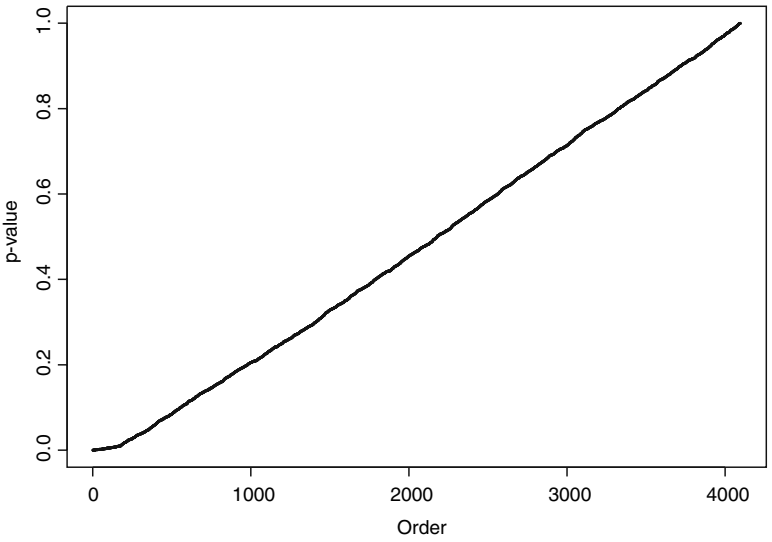


Fig. 10.5. An exaggerated plot of the ordered p-values from one slice including $64 \times 64 = 4096$ voxels.

then, fit the line using the $m-1$ largest p-values and calculate the estimated slope; continue in this manner, successively dropping the smallest p-values from the set, until the first time that the estimated slope decreases. Although many possibilities for calculating the slope exist, the authors advocate the simple *lowest slope estimator*, the slope of the line passing through the points $(m+1, 1)$ and $(i, p_{(i)})$, $S_i = (1-p_{(i)})/(m+1-i)$. The algorithm then specifies that as long as $S_i \geq S_{(i-1)}$, one should continue; stop for the first k such that $S_k < S_{k-1}$ and estimate the number of true nulls by $\hat{m}_{0.} = \min[(1/S_k+1), m]$, the smaller of m and the integer larger than the inverse of the slope S_k. The estimate $\hat{m}_{0.}$ is then used in the FDR-controlling procedure described in Section 10.4.

This is demonstrated in Figure 10.6. We first draw the line connecting $(m+1, 1)$ and $(1, p_{(1)})$, then the line connecting $(m+1, 1)$ and $(2, p_{(2)})$, and so on. Note that as we move through the first five ordered p-values, the slope of the line increases. When we pass the line connecting $(m+1, 1)$ and $(6, p_{(6)})$, the slope decreases for the first time. For the simulated data in this example, the slope for the line at the sixth ordered p-value is approximately 0.0605; since $1/0.0605 = 16.53$, we estimate the true number of nulls at 17, out of the 20 observations in the data set.

Turkheimer et al. (2001)use changepoint analysis together with the graphical presentation of the p-values to estimate $m_{0.}$. Their procedure is as follows:

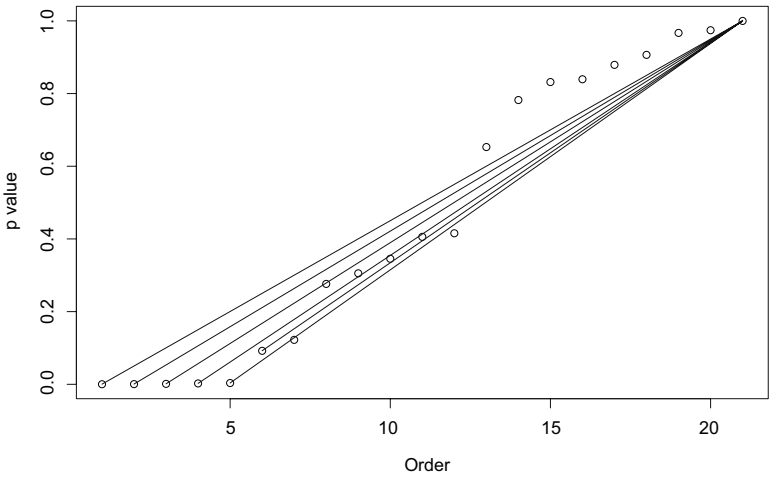


Fig. 10.6. Demonstration of the Benjamini and Hochberg (2000) adaptive procedure for estimating the number of true null hypotheses.

A test of uniformity is applied to sets of ordered p-values, in an iterative fashion. Starting with an initial set, if the test for uniformity is rejected, the smallest p-value in the set is discarded; the test is reapplied on the reduced set, and so forth, until the hypothesis of uniformity cannot be rejected. The slope of the plot of the ordered p-values is then calculated based on this final set, and $\hat{m}_{0.}$ determined as described above for the Benjamini and Hochberg technique. By comparison to Benjamini and Hochberg's adaptive procedure, that of Turkheimer et al. is somewhat more complicated. The test for uniformity is based on the differences between the observed $p_{(i)}$ and its expected value if there are in fact $m_{0.}$ true nulls (in which case those $m_{0.}$ p-values come from a uniform distribution). The maximum "residual" is taken as the test statistic, which then requires the use of simulation or tables to assess whether a set of p-values significantly differ in distribution. The slope estimation is also considerably more complicated, using a weighted least squares approach.

A model-based version of the graphical adaptive ideas is introduced by Pounds and Morris (2003). If one draws a histogram of the p-values, instead of plotting the ordered values, in the case where some tests are true nulls and some are not, the histogram will look something like the depiction in Figure 10.7. The left hand side of the histogram, representing the smallest p-values, will be made up of some small p-values coming from true nulls and some p-values coming from false nulls. As a result, there will be more small p-values than expected if all tests are truly null. The rest of the histogram will look roughly uniform. Again, the issue is where to draw the line across the

histogram, to delineate the true null, and hence uniformly distributed, part from the rest.

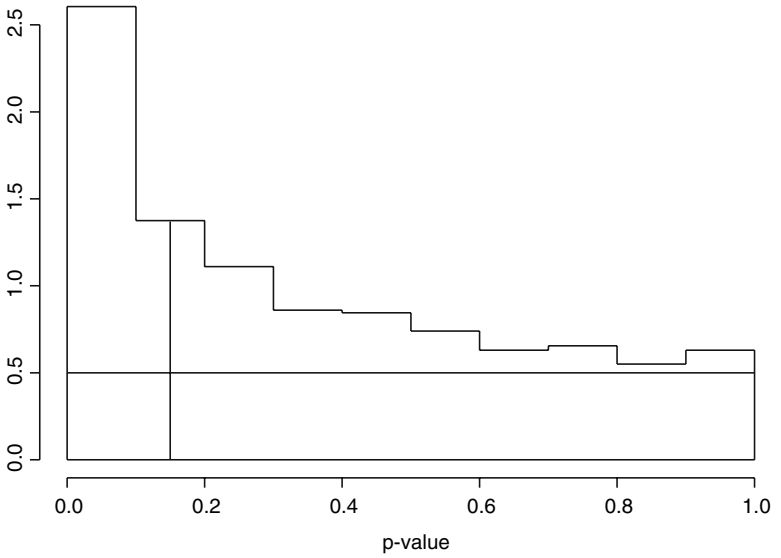


Fig. 10.7. Idealized histogram showing the distribution of the p-values as a mixture of two components. One component is $U(0,1)$, coming from those voxels that are truly null. The second component comes from voxels that are not truly null and can be modeled in a variety of ways. The vertical line shows the cutoff for determining significance; for purposes of demonstration it is drawn here at 0.15. All voxels to the left of this line are declared active and all voxels to the right of it are declared inactive. The horizontal line is the dividing line between the true null voxels and the true non-null voxels, that is, between the uniform component and the other component. All voxels below the line come from the uniform component (truly null) and all voxels above it come from the other component (truly non-null). In this way, the two lines divide the histogram, conceptually, into four parts: true positives (upper left), false positives (lower left), true negatives (lower right), false negatives (upper right).

As has been proposed by several authors, it is sensible to model the distribution of the p-values as a mixture of a U[0,1] and some other distribution. The Beta family is flexible and includes the uniform as a special case, making this a sensible choice. The model in Pounds and Morris (2003) takes the distribution of the p-values to have the form

$$f(p|\pi, a) = \pi + (1-\pi)ap^{a-1},$$

a mixture of Beta(a,1) and U[0,1]. Various ways of estimating the density represented in the histogram via this model are presented. The estimate of π is of particular interest, since it is the estimate of the proportion of true null hypotheses in the data. Based on estimates from this mixture model, one can also attain estimated upper bounds on the false discovery rate, which in turn leads to an adaptive threshold.

Pounds and Cheng (2004) explore the use of the *conditional FDR* (cFDR), which is simply the expected proportion of false discoveries given that r voxels are declared active, i.e., the number of false discoveries among the r most significant tests. Moving away from the model-based approach in Pounds and Morris, which may be too restrictive in some applications, Pounds and Cheng suggest smoothing the histogram of the p-values to obtain a more robust estimate of their density and from this derive an estimate of the proportion of true null hypotheses, π. In simulation studies, although not theoretically, this smoothed histogram method is shown to perform better than the model-based procedure of Pounds and Morris.

Of these various methods for estimating the number of true null hypotheses and carrying out the appropriate control for multiple testing, Benjamini and Hochberg's adaptive procedure has the virtue of simplicity. Many fMRI studies are exploratory in nature, so having a technique that is relatively easy to implement, even if it does result in a rather crude estimate of m_0, is probably sufficient.

10.10 Conclusion

The question of appropriate thresholding of statistical image maps, with the goal of determining which voxels or regions show significant amounts of activation, has taken a central place in the analysis and interpretation of fMRI data. Starting from the clearly unsatisfactory solution of applying a standard Bonferroni correction in order to control the FWER, research into the matter has spread in a variety of directions. Practical issues of implementation have inspired much of the current research, and will no doubt continue to do so, owing to the proliferation and indeed increasing prevalence of very large data sets for which multiplicity becomes a problem, such as DNA microarray analysis.

11

Additional Statistical Issues

In this penultimate chapter, we examine a number of critical statistical issues that can only be fully appreciated from within a broader understanding and knowledge of the analysis methods currently in use for fMRI data. As we have seen in the preceeding chapters, this rich data source, and the fascination of the human brain, have created an attitude among researchers of "I have a hammer, fMRI data look like nails.". Almost any statistical procedure that one can consider has been brought to bear on some aspect of the analysis stream, whether it be preprocessing, modeling, thresholding, or all three. This is not necessarily a bad thing, as it means that some very smart people have thought about a very hard problem. Still, it is worthwhile to take a step back from the minutiae of implementation, as outlined in the chapters devoted to particular techniques or approaches, and consider some general issues that can be gleaned from the more detailed discussions. The goal of this chapter is to provide that overview.

11.1 Whitening Versus Smoothing

One of the main obstacles to easily fitting models to fMRI data is the complicated correlation structure, particularly in the spatial dimension. As we have seen, one approach is to ignore this difficulty altogether, and thence to fit the model independently to each voxel and collapsing over time. This is the example of the simple t test or basic linear model. Although such a simplistic and unrealistic analysis does yield "good enough" results on robust experimental paradigms, one would obviously prefer either to fit the existing models with a more realistic error structure (the general linear model), or to fit more complicated models that directly incorporate the temporal and spatial correlations (spatiotemporal models).

The discussion of the general linear model in Chapter 5 alluded to the main alternatives available to researchers wishing to go this route and

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_11,

accommodate temporal structure other than independence. These are whitening and smoothing.

To understand the logic of these two approaches, we will rewrite the general linear model as

$$\mathbf{Y} = X\beta + W\boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is the observed time series, X is the design matrix, W describes the autocorrelation structure, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. Now the variance-covariance matrix of the data is $V = WW^T$.

Whitening means to remove the correlation structure altogether by premultiplication of both sides of the general linear model equation by a suitable matrix. We can see how to do this using the rewritten model above. Suppose W is known. Then if we premultiply by W^{-1}, we obtain

$$W^{-1}\mathbf{Y} = W^{-1}X\beta + \boldsymbol{\epsilon},$$

and now the errors are independent, identically distributed, and normal. β can be estimated using ordinary least squares, and the estimates are minimum variance unbiased. Of course, in practice W is not known, and so we cannot premultiply by W^{-1} precisely. These misspecifications result in biased estimates of the variance (Friston et al., 2000a). Friston et al. (2000a) also show that prewhitening models such as autoregressive of low order yield unacceptably large biases of the variances of the parameter estimates and their contrasts. Still, prewhitening is quite commonly used in fMRI data analysis.

In the same vein as whitening, but with the opposite goal or orientation, is *precoloring*. Here, the model is premultiplied with a coloring matrix C to give

$$C\mathbf{Y} = CX\beta + CW\boldsymbol{\epsilon}.$$

The effect of this transformation is to obtain a known autocorrelation structure for the errors, making it possible to use the theory of generalized least squares (GLS). That is, rather than eliminating the correlation structure, we aim to impose known structure prior to analysis. This approach is not as prevalent, and so is not discussed further.

An alternative method is *smoothing*, which refers to an initial smoothing or filtering of the fMRI time series as advocated by Carew et al. (2003). The idea now is to premultiply the general linear model by a smoothing matrix S, such that the assumed variance-covariance matrix SS^T is approximately equal to the true variance-covariance matrix under smoothing, SVS^T. Friston et al. (2000a) show that the bias can be controlled even when V isn't known, although direct minimization is difficult.

Carew et al. (2003) use spline smoothing to compute S, with the optimal degree of smoothing for each voxel time series chosen by generalized cross-validation (GCV). On a real fMRI data set, they find that most voxels require only a small amount of smoothing, with a few time series demanding large amounts. In general, the GCV spline approach smooths more than does SPM,

the default analysis path for many fMRI researchers. On simulated data the authors find that spline smoothing produces, on average, unbiased variance estimates for contrasts of interest, whereas both SPM and no smoothing are biased (the latter more so, obviously). The spline estimates are, however, less efficient.

Whitening is more efficient than smoothing (Friston et al., 2000a; Carew et al., 2003). But, as noted by both Friston and Carew, if the model used for whitening the errors is misspecified, the results can be very biased (in terms of standard error estimates, and hence inference as a whole). This is the typical bias-variance tradeoff, and should come as no surprise.

It is important to emphasize that the parameter estimates $\hat{\beta}$ are in any case unbiased. The discussion in the fMRI literature surrounding whitening or smoothing and the general linear model is in reference to the possible bias in the variance estimates. Bias in the variance estimates means that the test statistics are also biased, leading to a failure to detect truly active voxels or falsely detecting inactive voxels.

In the spatial dimension, smoothing is also often applied as a preprocessing step to induce spatial correlation prior to any statistical analysis, which can then be done on a voxel-by-voxel basis. The size of the smoothing kernel (the bandwidth) determines the strength of correlation, and ideally the data would be smoothed "just enough" to match local structure. Of course it is hard to assess the requisite amount of smoothing needed to achieve this matching. Alternatively, images may be analyzed voxelwise and the results then smoothed, but this is less common.

11.2 Functional and Effective Connectivity

A contentious and elusive concept (Horwitz, 2003) in the neuroimaging community – and one of great interest to researchers – is *connectivity*. Loosely speaking, connectivity refers to networks that model or explain relationships between brain regions. Discussions of connectivity thus allow neuroscientists to move away from the thresholding question ("Which voxels are active?") to the more general, and more scientifically relevant, question of how different areas of the brain interact to create thought. From a statistical standpoint, this is a difficult issue to explore, since it typically involves questions of causality; beyond that, however, the (temporal) resolution of the data may not support inference at this level of detail, and scientists don't always agree on what exactly is meant by this term, as noted by Horwitz (2003). fMRI researchers generally distinguish between functional and effective connectivity, although, again as noted by Horwitz (2003), different researchers use these same terms to mean different things, and furthermore the same concept is evaluated using different levels of data (for instance neuronal versus brain region) and measures.

Functional connectivity is usually taken to refer to the temporal correlation between spatially remote events, whereas effective connectivity is the influence that one neural system has over another. Clearly these two ideas are related and the demarcation between them is somewhat fuzzy, further exacerbating the conceptual and statistical difficulties accruing to the discussion of connectivity as a whole. Connectivity, in addition, is a fluid notion: it seems reasonable, on the face of it, that different networks come in to play when we solve a hard math problem than when we read *Macbeth*, for example. Hence, the search must be for general principles and procedures that can elucidate the connections between regions of the brain, even if specific networks are of interest. In spite of the various difficulties (conceptual, practical, statistical) surrounding connectivity, in this section I attempt to survey some of the work that has been carried out in this important area, and to draw some general conclusions.

In the current literature there are two predominant ways of trying to assess connectivity: correlation methods and structural equation models (SEM). The former looks at correlations between voxels or regions of interest in an effort to determine which areas of the brain coactivate. This is some measure of functional connectivity, perhaps, but not of effective connectivity, since it is well known that correlation does not imply causation. Structural equation models are a class of models arising mostly from social science and econometrics that do aim at attaching some notion of causality to the relationships that are found among the components. Hence the SEM approach is in fact an attempt to quantify effective connectivity. As with the related method of path analysis, there is some fair amount of controversy within statistics regarding the propriety of inference obtained via SEM, and so these results should be treated with caution.

11.2.1 The Use of Correlation to Assess Connectivity

It is intuitively plausible that correlation could be used to assess connectivity between voxels or regions. When the time courses of two voxels are highly positively correlated, one would tend to infer that they are connected with each other, if in no other sense than that the two voxels are active at the same times and inactive at the same times (more or less). Similarly, a high negative correlation would lead to the conclusion that when one voxel is active, the other is not; and again this is a type of connection between the activation patterns of the two. In this limited interpretation, similarity of function can be summarized by the simple correlation coefficients between voxel time courses.

A small example is presented in Figure 11.1. The correlation matrix shows pairwise correlations among 67 voxels that, according to a prior analysis, belong to three different categories: the first 26 are deemed to be noise, the next 22 are believed to be voxels that are exhibiting head motion, and the last 19 are evidently task-related. In this plot, lighter colors mean correlations

closer to 1. A number of conclusions can be quickly drawn, making this an attractive approach:

1. First, we can see that there are patches of higher correlation corresponding to the head motion voxels (in the middle of Figure 11.1) and the task-related voxels (in the upper right corner). This implies that all of the voxels in those two categories exhibit similar temporal behavior.
2. Second, the strength of the correlation among the noise voxels is weaker.
3. Third, some patching in the noise block is evident; some of this is an artifact, since the noise block is in fact made up of noise voxels from three different brain locations ($n = 8, 6, 12$) and interestingly the correlations pick this up.
4. Fourth, the last noise block, made up of 12 voxels according to the initial analysis, seems to exhibit at least two different types of temporal behavior.
5. Fifth, there is correlation among the voxels in the head motion class and the voxels in the task-related class. While weaker than the respective within-class correlations (as is to be expected), it is stronger, or on a level with, the within-class correlation for the noise voxels.

In spite of these various observations, it is still not clear what we can learn more precisely about the connections, if any, among the different classes of voxels in this example. The existence of moderate levels of correlation among the head motion and task-related classes is intriguing, but it is a far leap from noting the effect to concluding connectivity.

Computationally such a naïve approach is challenging since it requires the calculation and evaluation of all the pairwise correlations between time courses. And this is not even to consider simple extensions such as partial correlations to encompass sets of voxels or lagged correlations. One can ameliorate the computational burden by first screening the candidate set of voxels and retaining only those that show evidence of activation (Bullmore et al., 1996b). However, the inferences about connectivity from this implementation are still unsatisfactory.

As an example of this idea, in Bullmore et al. (1996b) task data are first analyzed to identify a subset of probably active voxels. The 170 voxels that remain after this screening belong to several regions that are related to the visual-linguistic task in the experiment. In this formulation the regions of the network are loosely identified through testing for activation at the individual voxel level; that is, active voxels are found and attributed to specific anatomical or functional areas. Looking at the correlations between pairs of time courses, the authors conclude that voxels in the same anatomical region tend to be positively correlated, as one would expect if the correlation approach is at all valid. Across regions, voxels may be positively or negatively correlated; again this is reasonable since it indicates that different areas of the brain are activating at different times. However, they also note deficiencies, in terms of what can be learned from the simple correlation matrix, and these are instructive more broadly. First, the correlations only give information about whether

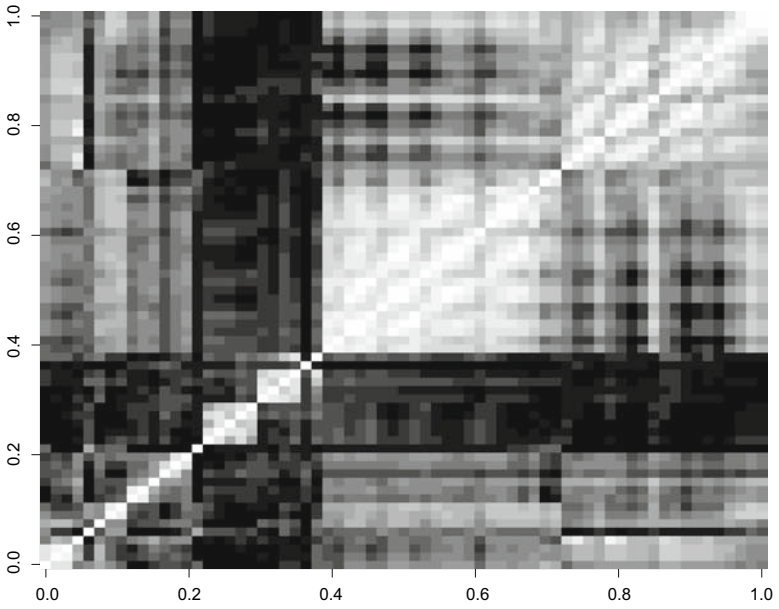


Fig. 11.1. Correlation matrix among 67 voxels belonging to three different categories: noise ($n = 26$), head motion ($n = 22$), and task-related ($n = 19$). Light colors indicate correlations closer to 1 and it is evident that the two blocks corresponding to head motion and task-related activation exhibit a temporal coherence that is not found in the noise block. Since these are simple correlation coefficients, it is not entirely clear what we learn about connectivity from such an analysis. Voxels extracted and classified by Jun Ye, from a data set provided by Rebecca McNamee, University of Pittsburgh.

or not two voxels are connected, but they do not inform on the patterns of temporal activity that are shared by functionally connected voxels. Second, it is hard to discern finer details of the functional relationships between regions, when they exist.

Bullmore and colleagues therefore suggest multivariate methods on the time courses, such as PCA (via singular value decomposition of the covariance matrix of the activated voxels) and canonical variate analysis, to help elucidate the connections between elements in the network. For example, the first principal component summarizes the most important temporal behavior in the data. Therefore, if two voxels have positive weights on that component, one can conclude that they are positively connected in terms of that dominating temporal pattern. This gives a more subtle picture about voxel coactivation than the pure correlation method. Bullmore et al. (1996b) provide a variety of analytical and graphical techniques to evaluate the

connections derived from the multivariate analysis. In large part conclusions from the simple correlation matrix are confirmed, and sharpened with additional detail.

Although motivated by the correlation approach, Bullmore et al. do not directly use the correlations across voxel or regional time courses in their analysis; other researchers have suggested instead refinements to the approach itself. Hampson et al. (2002), for example, introduce an interesting (if somewhat intricate) procedure that defines regions of interest and hypothesizes functional connections between them on one data set, then tests the connectivity hypotheses on another, independent data set obtained from the same subject. The roles of the two data sets are then reversed, and one can switch back and forth between the two sets multiple times to pick out additional ROIs in the network. The first data set is obtained from a block design, alternating between task and rest conditions. The second data set is obtained from "steady state" data of the same two conditions, i.e., a continuous stretch of rest data and a continuous stretch of task data.

The rationale for using two data sets is that, as discussed previously, functional connectivities are almost surely different for different tasks, and between task and rest. Connectivities uncovered during a resting state might give insight into neuronal networks that are in place regardless of the task. On the other hand, clearly the connections that arise during task performance are also of interest and they can be more realistically evaluated if compared to what is expected at rest.

In the first step, ROIs are defined from the set of data from the block design. The authors use a combination of voxel by voxel t testing (with generous thresholding) and anatomical information to identify the ROIs (call them A and B for illustrative purposes). A functional relationship is hypothesized between the two regions. To test the hypothesis, correlations with the signal from region A are calculated for the two steady state runs comprising the second data set: the time courses are filtered, the average time course for the run is calculated, the average time course of voxels in region A is calculated, and the partial correlation between the time course of each voxel and that of region A is found, after removing the overall average from the former. Since this procedure is followed separately for each of the two steady state runs, in the end one obtains a map of correlations between voxels at rest and region A, and a map of correlations between voxels during task and region A. These maps undergo some additional processing, but in the end the correlations to region A are evaluated within region B and a test performed for the significance of correlations between activation in the two regions.

This completes one iteration of the algorithm. In the next iteration, one uses the two steady state data sets to define region C, which is hypothesized to be functionally connected to, say, region A. The respective states of the block design data set are then brought to bear to evaluate the hypothesis of connectivity between region C and region A, similar to the fashion described

for the first iteration. One can continue in this manner, iteratively adding prospective regions to the network.

Clearly this technique is intensive both in terms of data collection (since extra data sets need to be acquired that wouldn't necessarily be of interest otherwise) and analysis (the processing within each iteration is extensive). Hampson et al. (2002) illustrate the method on a language task, with results that coincide with others in the literature and generally accepted networks. It would be interesting to see how their approach would handle cognitive networks that are less well understood. In particular, much of the efficacy of the algorithm appears to hinge on the ability to detect or define ROIs, which might be difficult for some systems. Of course, the same criticism can be leveled against any of the methods for studying connectivity that rely on "seed" voxels or ROIs, a very common starting point.

Beyond the defects of the correlation approach that we have already outlined, some researchers have raised questions about the use of correlation specifically as the measure of functional closeness (Lahaye et al., 2003), since this implies that only linear relationships with no time delay are of interest. To this end, Lahaye et al. (2003) explore both nonlinear relations and the possibility of lags. The gray matter of the brain is first divided into "parcels," or groups of voxels that are segmented according to levels of the measured signal. These are similar to ROIs, but are based on a predetermined number of seed voxels rather than any functional or anatomical constraints. As in Bullmore et al. (1996b), the motivation for this work is the perceived limitation of the pure correlation analysis, but here too, correlations are not directly used.

Instead, Lahaye et al. (2003) devise a hierarchical system of models for the functional relationship between parcels A and B of which the instantaneous linear is but the simplest. Noninstantaneous effects of one parcel on another are modeled by including lagged terms in a general linear model framework; nonlinear effects are modeled by polynomial terms. Models also include terms for the lagged effect of parcel A on itself, recognizing that what happens in a particular region at time t_i most likely influences behaviors in that region at later times $t_j, j = i + 1, \ldots, i + k$, for some k. The simplest model in this framework is thus the instantaneous linear effect with recent history of the parcel; the most complex includes as well nonlinear and lagged effect terms with another parcel.

As a final example of extensions to the basic correlation approach, although others exist and are no doubt being developed still, we consider Sun et al. (2004), who use coherence measures to assess functional connectivity. Coherence, the spectral analog of correlation, measures the linear time-invariant relationship between two time series. As such, it is another way of accounting for the possibility of lags in the effect of one voxel (or region) on another. If the time series in one voxel is broadly similar to that in another, but with a time delay, then the ordinary correlation between the two will be moderate or low (depending on the amount of lag); the coherence, by contrast, will be high

within the bandwidth of the hemodynamic response function (which, based on empirical evidence, is concentrated at low frequencies).

Let $\mathbf{x}$ and $\mathbf{y}$ denote two voxel time series. The power spectrum of $\mathbf{x}$ is

$$f_{xx}(\lambda) = \sum_t \mathrm{Var}_x[t] e^{-j\lambda t},$$

and likewise for $\mathbf{y}$. The cross-spectrum is

$$f_{xy}(\lambda) = \sum_t \mathrm{Cov}_{xy}[t] e^{-j\lambda t}.$$

The coherence is then defined to be

$$C_{xy}(\lambda) = \frac{|f_{xy}(\lambda)|^2}{f_{xx}(\lambda) f_{yy}(\lambda)}.$$

The rest of the analysis proceeds as is typical; ROIs are chosen for the task at hand, and within those ROIs seed voxels are picked to be the basis of comparison. Instead of correlations, coherences are calculated and mapped out. As discussed in previous chapters, many analyses and interpretations are more easily carried out in the spectral domain than in the time domain, and the approach of Sun and colleagues falls squarely into that camp.

11.2.2 Structural Equation Models

The second major approach for investigating connectivity is structural equation modeling. Again, the starting point is generally a set of regions of interest relevant to the task at hand; these are most often theory-driven. The researcher hypothesizes the functional connections between different ROIs, including direction (region A to region B, region B to region A, or both). Within the setting of SEM connections are hypothesized to be captured in the covariances between regions, and brain function results from changes in those covariances (Gonçalves and Hall, 2003). Structural equation models are related to general linear models, but the parameter estimates are taken to refer to the strength of connections between regions. For purposes of analysis, one does not use every voxel in the ROIs, as this would be very computationally burdensome. Instead, representative, or "seed," voxels are again chosen, one or several from each region, and these are included in the SEM. In fMRI data analysis, structural equation models are usually used in a confirmatory sense, rather than an exploratory one (but see Zhuang et al. 2005). That is, the researcher will specify a set of connections and then use SEM to confirm that the specified model fits the covariance structure of the data. Notably, these models do not use temporal information (Harrison et al., 2003).

An implementation issue for fMRI data analysis that does not arise in other uses of SEM is, as noted briefly above, that the entire region of interest is not entered into the analysis, but rather some representative voxels are

picked from each. Potentially, then, there is the difficulty that the results of fitting a structural model of this type, namely the postulated behavior of the network, could be influenced by the particular voxels that happen to be chosen. Often the seed voxel is the "most active" in the region, and perhaps its immediate neighbors as well, which might bias the outcome to look stronger than it truly is. Gonçalves and Hall (2003) examine the effect of choice of voxel(s) on the resultant SEM output and find that there is, indeed, some variability according to which voxels go into the model. However, overall the nature of the conclusions is similar; thus, while the choice of voxel might affect the estimated strength of a connection, generally it won't affect the existence of that connection. The authors caution though that the choice still should not be arbitrary (since one might then pick as an exemplar for an ROI a voxel that is actually inactive); rather, they advise using the peak (most highly activated) voxel or the average of voxels around the peak as the summaries of an ROI.

Although SEM, and inferring causal networks more generally, is a difficult question for statisticians, it is easy to see why such an approach would be attractive for neuroscientists interested in understanding effective connectivity in particular, since it provides a way of confirming theoretically derived networks of brain regions, and gives in addition estimates of the strength of pathways. Modules for performing SEM exist in both SPM and AFNI, two of the major statistical packages for the analysis of functional neuroimaging data, attesting to its popularity in the community.

An interesting use of structural equation modeling that seems to be more specific to the issues of neuroimaging is presented in Mechelli et al. (2002), who build a multisubject network, rather than individual networks for each subject, to look for underlying commonalities that would hint at effective connectivity. We have already seen that there is a fair amount of variability across subjects in terms of patterns of activation. Is the same true for the underlying networks that create that activation? On the one hand, it is definitely plausible to hypothesize that different people would (implicitly) call on different brain networks to perform the same task; this is akin to using different strategies and might help explain the observed variations in activity. On the other hand, there is also a great deal of similarity in the overall patterns of activation, in spite of the existing individual differences. This argues in favor of a common underlying effective connectivity across subjects. For a unified theory of brain, we would hope that there are broad commonalities in the networks, with perhaps several "subtypes" representing the various processing strategies.

Mechelli and colleagues note that studies of effective connectivity based on multiple subjects have generally taken one of two approaches. In the first instance, researchers treat the data from the different subjects as if they were all from the same subject, which assumes that connectivity patterns do not widely vary from person to person, and what variation there is, is random. Both of these assumptions are questionable. In the second instance, researchers perform subject-specific network analyses. The difficulty with this approach lies

in interpretation, since if subjects exhibit different networks, it isn't clear whether or not these correspond to scientifically important differences and there is apparently no easy way to test this.

Building a framework for the evaluation of group commonalities and differences in connectivity networks is the main motivation for the model developed by Mechelli et al. (2002). In this model, they posit m ROIs and n subjects. Within a subject, ROIs are connected as in a standard network model; regions from different subjects are not allowed to be connected, since, clearly, subjects act independently. Hypothesis testing is performed by comparing two formulations, one in which the effects (connections) are required to be constant across all subjects, and another in which these connections are allowed to vary from subject to subject. With m ROIs and n subjects there are a total of $m \times n$ regions in the model, and $n \times m \times (m-1)$ connections in the most complete network specification (bidirectional between each pair of regions). To construct the basic network, one still must rely on theory, of course. For example, Mechelli et al. (2002) look at a reading task in their study. The ROIs are chosen according to an anatomical model of reading developed by one of the researchers; in this study they include all possible connections between pairs of regions, in both directions, although this could be narrowed down according to additional theoretical considerations.

The main advantage of this approach is that it provides researchers with a way of objectively assessing the variability across subjects via model comparison. As noted by Mechelli and colleagues, statistically significant differences in connectivity may arise either as a result of the networks themselves differing from subject to subject, or from subjects sharing the network pattern but having differences in the strength of its expression. It's not clear that the multisubject analysis is sensitive enough to distinguish between these two cases, but it can at least serve as the springboard for a more detailed investigation.

11.2.3 Other Approaches

Various other techniques have also been applied to the problem of studying connectivity. Among these are self-organizing maps (Ngan and Hu, 1999; Peltier et al., 2003), spatial ICA (van de Ven et al., 2004), ranking time courses with spanning trees (Baumgartner et al., 2001), and multivariate autoregressive models (Harrison et al., 2003). Most of these efforts are relatively recent, and therefore don't have the history of the correlation method (which is derived from standard ways of thinking about fMRI data) or structural equation modeling (which was first introduced in the neuroimaging literature for the analysis of positron emission tomography data more than a decade ago). They do not appear to have yet been developed beyond the initial work that introduced them and it remains to be seen which of these directions, if any, will lead to additional insight into the thorny problems of understanding brain networks.

11.2.4 Conclusion

Functional and effective connectivity are at the core of the neuroimaging enterprise. Classification of voxels into "active" or "inactive" – long the default analysis – is a mere weak proxy for the true questions of interest. However, methodological limitations (at both the imaging level and the statistical) have made it difficult for scientists to address the problem head-on. The existing methods show promise, in that they have demonstrated ability to confirm theoretically posited networks or have replicated results found in other analyses. Generating previously unidentified or unsuspected networks of connections is a much more challenging prospect, yet it seems necessary to achieve this level of sophistication if we are ever truly to understand the inner workings of the brain. I suspect that this will continue to be one of the harder, and more interesting, questions to face statisticians working in fMRI data analysis.

11.3 Model Selection

Although model selection has a venerable history in statistics as a whole, has received much attention on both the theoretical and applied levels, and has been the focus of much discussion in the disciplines, this topic has not been at the forefront of the conversation about fMRI data analysis. Part of the reason for this no doubt has been that, until very recently, researchers have had to be so focused on the questions of devising and assessing adequate models that they have not had the luxury of considering alternative models on a large scale.

There are two levels at which model selection questions might be addressed in the context of fMRI. First there is the selection of a model to fit to the brain as a whole, from among a group of competing models. That is, given the panoply of models that have been suggested to date, which one fits "the best" in some (as yet to be defined) sense? The second level touches on a deeper, and computationally more difficult, problem, namely should the same model be fit at each voxel of the brain? It seems evident that the answer to this question must be "no," since there is no reason to expect any one single model to provide a reasonable explanation of the observed behavior at every voxel. Thus when the same model is used at each voxel, some will be underfit and some will be overfit (Razavi et al., 2003). Furthermore, many researchers have commented on the apparent differences in the characteristics of the measured signal in different regions of the brain, bolstering the intuition that models should selected on a voxel basis. This presents a computational challenge, but with ever-increasing speed and memory, not to mention the option of parallel processing (if each voxel is treated independently, they can be analyzed simultaneously on multiple processors), it becomes more feasible. For theoretical reasons one might still prefer to fit a single model to the whole brain; thresholding with the random field approach, for example, requires a single model (Kherif et al., 2002).

Model assessment has been similarly neglected, possibly leading to several drawbacks in the way that fMRI analyses are conducted (Razavi et al., 2003), including:

1. model misspecification;
2. improper model selection based on the number of active voxels (wherein a model is judged to be of higher quality if it detects more activation);
3. lack of systematic model evaluation, resulting in the use of potentially oversimplistic models;
4. use of the same (prespecified) model at each voxel in the brain;
5. lack of standards for formal model comparison.

Razavi et al. (2003) aim to rectify some of these shortcomings by introducing a formal framework for the assessment and comparison of fMRI models. They assign two dimensions of quality upon which to evaluate competing models, *validity* and *goodness of fit*. Other fMRI researchers (Kherif et al., 2002) note that a "good" model should also be *parsimonious* and *simple*, both traits that aid in interpretability and generalizability.

Validity is determined by how well the data fit the assumptions of the model. For example, the general linear model makes certain assumptions about the error term (normality, equality of variances, and independence). Independence is probably the most critical of these, and the one that is almost surely violated by fMRI data. Whitening, coloring, and smoothing have all been used to account for dependence structure, as noted above. If they are effective, the residuals from such models should not have much (if any) remaining correlation. The authors therefore suggest assessing the quality of a model by testing for the presence of temporal autocorrelation (of first order) in the residuals using the Durbin-Watson statistic:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Goodness of fit refers to the closeness of the fitted model to the data. The authors summarize this quality by the usual multiple regression R^2 (or its adjusted version). Among the valid models with good fit, those with higher values of the goodness of fit statistic should be preferred.

According to Razavi et al., this series of steps (assessing validity, assessing goodness of fit, comparing values of adjusted R^2) offers a more rigorous and formal way of comparing competing models than counting the number of detected activations. As they point out, a valid model with a high value of adjusted R^2 could yield a smaller number of active voxels than a valid model with a lower value of adjusted R^2. The former would still be preferred, since there is nothing intrinsically relating the quality of a model and the number of active voxels it detects, whereas adjusted R^2 does have statistical meaning as a measure of model quality.

The framework developed by these researchers is a promising one, since it also provides a context for model building and selection. One might quibble

with their choice of R^2, even in its adjusted version, as the measure of goodness of fit, realizing that there is a wealth of criteria for model comparison (AIC, BIC, and C_p, to name just a few; many others should be familiar to the reader). The importance of the study lies in opening the door for additional exploration of this critical topic, and not so much in the particular implementation choices made by the authors.

11.4 Evaluation of Competing Methods

Related to the questions of model building and selection is the evaluation and comparison of competing methods. This topic, too, has started to receive more attention in recent years, as a number of researchers have proposed practical frameworks within which these comparisons can be made. The predominant method of evaluation is to use receiver operating characteristic (ROC) curves (see, for example, Skudlarski et al. 1999, for an early application).

ROC curves compare methods based on the proportions of true and false discoveries. In more detail, following the discussion in Nandy and Cordes (2003b), the ROC curve can be defined as follows. For each voxel, when a statistical analysis is performed and a decision reached regarding the rejection of the null hypothesis, we know that there are four possible states: correctly declare a voxel to be active; incorrectly declare a voxel to be active; correctly declare a voxel to be inactive; incorrectly declare a voxel to be inactive. The decision (declare the voxel to be active or inactive) is based on the value of some test statistic and threshold. The ROC curve is a plot of the conditional probability of declaring a voxel to be active, given that it is truly active (or, the response contains signal), against the conditional probability of declaring a voxel to be active, given that it is truly inactive (or, the response is pure noise), for different levels of the test statistic. (Sometimes this is equivalently plotted for different levels of significance.) Area under the ROC curve is generally used as a measure of quality; methods with greater area under the curve are preferred.

A simple and highly stylized example is shown in Figure 11.2 where $n_1 = 100$ data points are generated from $N(0, 1)$ and $n_2 = 40$ from $N(2, 1)$. All observations below the value c are assumed to come from the first population (with mean zero) and all observations above c are assumed to come from the second (with mean 2). Obviously, any particular observation can be correctly or incorrectly classified. The ROC curve shows the false positive rate (observations incorrectly classified as belonging to the second population, i.e., they have a value above c although they come from the $N(0, 1)$ distribution) against the true positive rate (observations correctly classified as belonging to the second population, i.e., they have a value above c and come from the $N(2, 1)$ distribution).

As the threshold c increases we have fewer false positives; that is, we are better able to allocate the observations from the first distribution correctly.

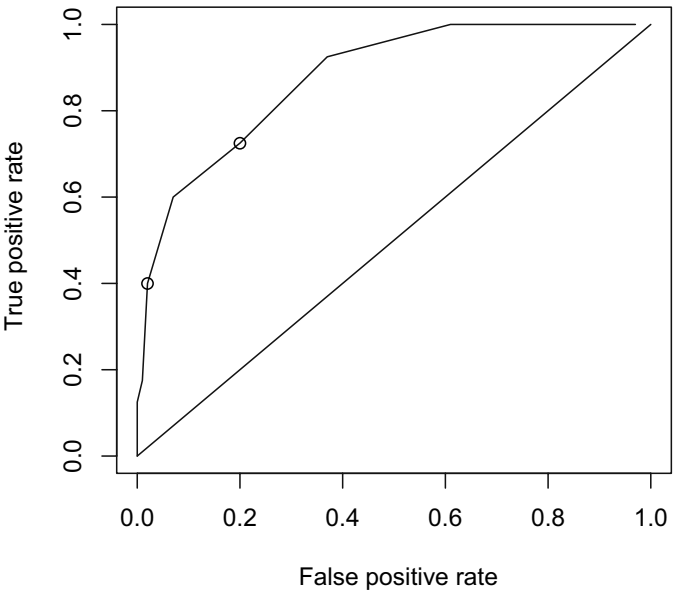


Fig. 11.2. ROC curve for a data set composed of $n_1 = 100$ observations from a $N(0, 1)$ distribution and $n_2 = 40$ observations from a $N(2, 1)$ distribution. The lower point on the plot corresponds to the characteristics when the cutoff for assigning an observation to the second population is $c = 2$; the upper point gives the characteristics when the cutoff is lowered to $c = 1$. With the lower threshold we are better able to winnow out the observations from the distribution with the lower mean, resulting in fewer false positives, but we are less able to correctly assign observations to the distribution with the higher mean, resulting in fewer true positives as well.

However, we also have fewer true positives. This makes sense, since as the threshold increases, there are fewer observations overall, from whichever population. When the threshold is very low we can correctly assign all of the observations from the $N(2, 1)$ distribution, but almost all of the observations from the $N(0, 1)$ distribution are assigned here as well. Hence there is a tradeoff between the two rates.

Now in practice, the conditional probabilities in question are not, and cannot, be known precisely for fMRI data, since investigators don't know which voxels are truly active and which are truly inactive. Hence the fractions need to be estimated; this can be accomplished in various ways, most frequently via simulation of some sort. There has been some discussion in the literature regarding the "best" standard for obtaining estimates of the ROC curve; a consensus seems to be forming among many researchers that real data is

preferable to completely artificial simulated data (Nandy and Cordes, 2003b; Nandy and Cordes, 2004b). Skudlarski et al. (1999) have subjects perform a standard cognitive processing task, then use slices that do not exhibit activation related to the task as the source for "noise" or null state; to this noise, they add artificial activation. By contrast, Nandy and Cordes (2003b) take the more conventional route and use ordinary resting data to represent the null state. Skudlarski et al. (1999) justify their choice by noting that the variance of resting data sets is different from the variance of data sets obtained while the subject performs a task. Nandy and Cordes (2003b) justify their choice as correctly incorporating the spatial correlation structure present in fMRI error.

Beyond this, the two studies are different in their areas of investigation. Skudlarski et al. (1999) use standard ROC curves to explore the effects of the various choices available to researchers at each stage of fMRI data analysis, ranging from experimental design, to preprocessing, to statistical analysis, and more. Based on their simulations, and the settings used for each of the stages they consider, the authors are able to make concrete recommendations for investigators. For example:

1. Temporal detrending at the level of individual voxels is useful, but smoothing the time course decreases the ability to detect true activation.
2. The ordinary t test, and its nonparametric Mann-Whitney analog, are very powerful. Paired t tests perform well for some experimental designs, but not others.
3. The optimal length of a task block is about 18 seconds for a simple block design.
4. Spatially smoothing raw fMRI images with a Gaussian filter of 1-2 voxels FWHM prior to analysis is generally beneficial.
5. Motion correction has little impact on the relative effectiveness of the various statistical tests.

Nandy and Cordes (2003b) propose a modified ROC curve to bypass the difficulties that result from not actually knowing which voxels are active and which are not, and from having to use simulated data which do not correctly mimic the correlation structure in the true data. We have already seen that they address the second issue through the use of real data as the basis of their simulation technique – resting data represents the null and task data represents the activation state. For the first issue, the ROC curve requires estimates of the two conditional probabilities, the probability of declaring active when the voxel is truly active and the probability of declaring active when the voxel is truly inactive. Now, if the data at hand are resting, or null, data, then in principle there is no task or stimulus to which the brain is reacting, and all voxels are presumed to be "inactive" in that sense. Hence it is possible to estimate the probability of falsely declaring a voxel to be active, for different values of the test statistic.

Estimating the probability of truly declaring a voxel to be active is still difficult, hence they suggest instead the use of task data to obtain the fraction of all detections for a given level of the test statistic. The set of detections from the task data is made up of some truly active voxels and some truly inactive voxels, but again we don't know which belong to which group. However, for a given value of threshold it is possible to plot the proportion of "active positives" against the proportion of "resting positives" (the detected voxels in each of the data sets) to get a modified ROC curve; the proportions can be estimated even if the activation status of the individual voxels is not known. In practice, some additional work is required to get an estimate of the proportion of active positives; details are given in Nandy and Cordes (2003b). Nandy and Cordes show that this modified curve is related to the usual ROC curve; indeed, the former can be transformed to "reconstruct" a conventional ROC curve. The area under the modified curve is always smaller than the area under the conventional curve, since the former is a lower bound for the latter, but this isn't important since it is the relative area for different analysis methods that is compared.

Another framework for comparing analysis pathways is NPAIRS, which stands for *nonparametric prediction, activation, influence, and reproducibility resampling* (Strother et al., 2002); see also Section 3.4. Our earlier discussion of this approach focused on using it to assess preprocessing strategies, but it is more general than that, as the name implies. The authors put forward two criteria for evaluating the quality or validity of neuroimaging results: accurate prediction of output on a test set of data, based on an independent training set; reliable reproduction of the parametric map of the test set. The use of training and test sets places NPAIRS in the family of resampling procedures. Strother and colleagues contrast the use of real data and resampling to ROC analysis, which tends to be based on simulated data. However, as we have discussed just above, more recent work on ROC curves advocates the use of real data as well.

Acknowledging that many resampling possibilities are available, ranging from all varieties of k-fold cross-validation, through jackknife and bootstrap, NPAIRS utilizes a "split-half" paradigm, with subjects as the basic unit. In other words, the training set and testing set are the same size, both consisting of half of the subjects in a study (hence multiple subjects are required to implement the comparison, in contrast to the ROC approaches, which are applicable on single subjects as well). All splits can potentially be considered, since most fMRI studies do not have large numbers of subjects. For each split, one applies a processing stream to the training and testing halves individually. Reproducibility of the statistical parametric maps, one of the criteria for evaluating competing methods, is then summarized by a similarity measure, such as the ordinary correlation coefficient between the voxel values. NPAIRS uses a combination of graphical and analytical techniques, such as the reproducibility histogram, to assess and compare performance of different analysis streams, from preprocessing options through the choice of statistical model.

The influence of individual subjects on the output can also be investigated within NPAIRS.

Kjems et al. (2002) incorporate learning curves as part of the NPAIRS evaluation scheme; these are plots of the testing error as a function of the size of the training set and are again obtained from a cross-validation approach. Here, however, since interest focuses specifically on changes in performance that are linked to the size of the training set, there is obviously no longer a restriction to use split halves. Rather, for different training set sizes, multiple splits are generated; errors are then averaged over fixed sample sizes to yield the desired curves. As in the general NPAIRS approach, the learning curve component introduces a range of interesting graphical techniques for comparing the performance of different analysis streams and the effects of particular components of those paths.

Taken as a whole, this admittedly small body of work on NPAIRS appears to offer a good complement to the more traditional ROC curve analysis already familiar to statisticians. However, it remains to be tested on a wide range of experimental settings and statistical models. The scientists who introduced NPAIRS in these two papers recognize the need to make widely generalizable and applicable statements regarding the efficacy of any aspect of the analysis stream; indeed, they demonstrate their methods on a range of cognitive tasks, a first step. The challenge is to take this – or any other similar framework, including ROC curves – and implement it on a broad enough spectrum of studies that general conclusions can be reached. To my knowledge, this has not yet been done, and it is not surprising, as such an enterprise is beyond the capability of any one laboratory or group of researchers.

11.5 Summary

A recurring thread throughout this book has been that the statistical challenges inherent in the analysis of fMRI data are many and varied. The richness of the data and the complexity of the scientific questions place fMRI squarely in the center of some of the most pressing issues facing modern applied statistics: modeling – both frequentist and Bayesian – of large data sets with complicated structure, assessing significance in multiple testing situations, visualization, and the use of computationally intensive methods. It is not surprising that the range of statistical procedures that have been applied to fMRI data is equally wide and varied. The past 15 years have seen a growth of the popularity of fMRI as an imaging technique, with many universities acquiring research-dedicated MR scanners, and a concomitant flourishing of collaborations between statisticians and neuroimaging scientists. These being early days in the application, we have seen primarily the use of statistics to attempt solutions to the scientific problems. Less common so far has been the inspiration of new statistical methodology, although recent works particularly in the area of spatiotemporal and Bayesian modeling, have approached this.

Ideally, we will find ourselves in the situation where the flow of ideas will be from statistics to neuroimaging and back again. Continued collaborations and the involvement of more statisticians in this exciting and important field will no doubt lead to that outcome.

12

Case Study: Eye Motion Data

In the previous chapters we have encountered a plethora of statistical questions and proposed solutions. Without a doubt it can be bewildering for newcomers to the field to maneuver through all the techniques that have been put forth in the literature for analyzing fMRI data. The purpose of this chapter is to take some of the most basic analyses and demonstrate them on a sample data set. The data appear courtesy of Dr. Rebecca McNamee, University of Pittsburgh.

12.1 Description of the Data

We will look at data from one subject performing simple eye motion (saccade) tasks, in a block design experiment. The two alternating tasks are *antisaccade* and *prosaccade*. In the antisaccade task, a light appears on a screen in the peripheral vision of the subject, and the subject needs to direct his vision to the oppposite orientation. If, for instance, the light appears in the upper right corner of the screen, for the antisaccade task the subject would need to look in the lower left. The antisaccade task requires inhibition of the natural tendency to look where the light appears. For the prosaccade task, again a light appears on the screen, but now the subject is asked to look in the correct location. Saccade tasks are often used in the study of schizophrenia, brain lesions, and other syndromes that may be characterized by loss of inhibition or by deficits in eye movement control.

The data set consists of 30 slices, each of size 64×64, taken over 156 time points. Image volumes were collected every 2.5 seconds. Denoting the antisaccade task by AS and the prosaccade task by PS, the stimulus stream was $\{AS, 1; PS, 1; PS, 12; AS, 12; PS, 12; AS, 12; \ldots; AS, 12; PS, 10\}$. See Figure 12.1 for a graphical description of the stimulus trail.

Preprocessing steps performed on this data set included: removal of spatial outliers, motion correction, outlier correction in image space, removal of linear trends and drift, and Gaussian smoothing with a radius of two voxels.

N.A. Lazar, *The Statistical Analysis of Functional MRI Data*,
DOI: 10.1007/978-0-387-78191-4_12,

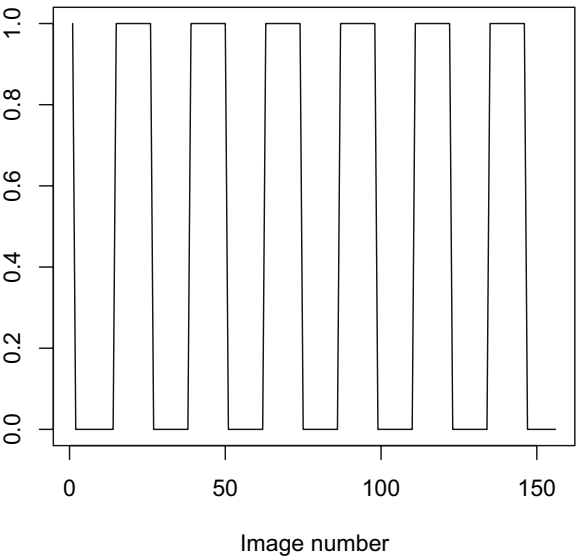


Fig. 12.1. Stimulus trail for block design experiment that generated the data analyzed in this chapter. The main paradigm alternates 6 times between 12 presentations of the prosaccade task and 12 presentations of the antisaccade task.

12.2 Data Analysis

We start with some simple exploratory analysis , based on calculating the correlation coefficient of each of the voxel time series with the stimulus presentation trail. Figure 12.2 shows the histogram of the correlations for all $64 \times 64 \times 30 = 122,880$ voxels in the data set. The average correlation is 0.01, with a standard deviation of 0.093. The minimum correlation over all voxels is -0.51 (indicating behavior that is opposite in trend to that of the stimulus presentation; in this case, higher levels of activation in response to the prosaccade task than to the antisaccade task) and the maximum is 0.45 (higher levels of activation in response to the antisaccade task than to the prosaccade task). Hence none of the time series are very strongly correlated with the experimental paradigm in this individual. As can be seen in the figure, the distribution of the correlation coefficients looks roughly normal.

To explore the correlation behavior more closely, we now consider two slices of possible interest. According to the researcher who provided the data, the fourth slice is expected to show activation; therefore we might expect the correlations in this slice to be higher than those in other slices. The twelfth slice has the highest average correlation of all the slices: 0.048. By contrast, the average correlation in the fourth slice is only 0.021. Both slices have standard deviation of 0.09. The minimum correlation in the fourth slice is -0.29 and

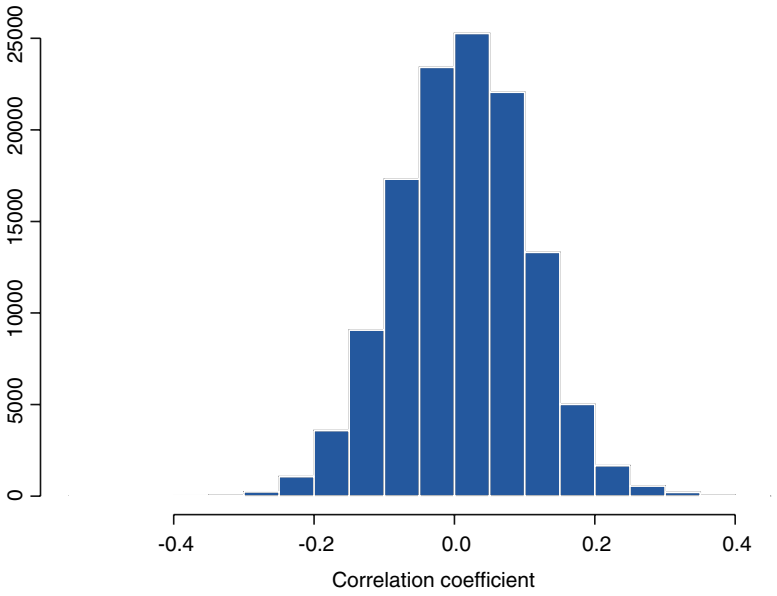


Fig. 12.2. Distribution of correlation coefficients between voxel time courses and stimulus time course, for all 122,880 voxels in the data set. The distribution looks approximately normal, with mean near zero and standard deviation of 0.093

in the twelfth slice, -0.23; the maximum values in the two slices are 0.37 and 0.36, respectively. In terms of the typical summary statistics, the two slices are quite similar.

Histograms of the $64 \times 64 = 4096$ correlations for the two slices are shown in Figure 12.3. These are also rather similar, however the distribution for the twelfth slice is more skewed.

Figure 12.4 displays images of the correlation coefficients in each of the two slices, with contours superimposed. For clarity, only the two highest contours representing the strongest correlations, are plotted. As can be seen in the figure, in the fourth slice the dark shading, which corresponds to higher positive correlation, is quite widespread in the central part of the image; this is the brain, occupying only a relatively small part of the 64×64 grid. The high correlations are furthermore concentrated more or less on the edges of the brain, that is, in the gray matter. This is even more striking in the twelfth slice. Note that since the twelfth slice is further from the top of the head the proportion of this grid covered by brain is greater. Again, the darkest shadings are quite localized in contiguous patches.

Figure 12.5 shows the minimum, mean, and maximum correlation in each slice. The mean is more or less stable around zero as we descend through the brain from the first to the thirtieth, but there is an apparent slight downward

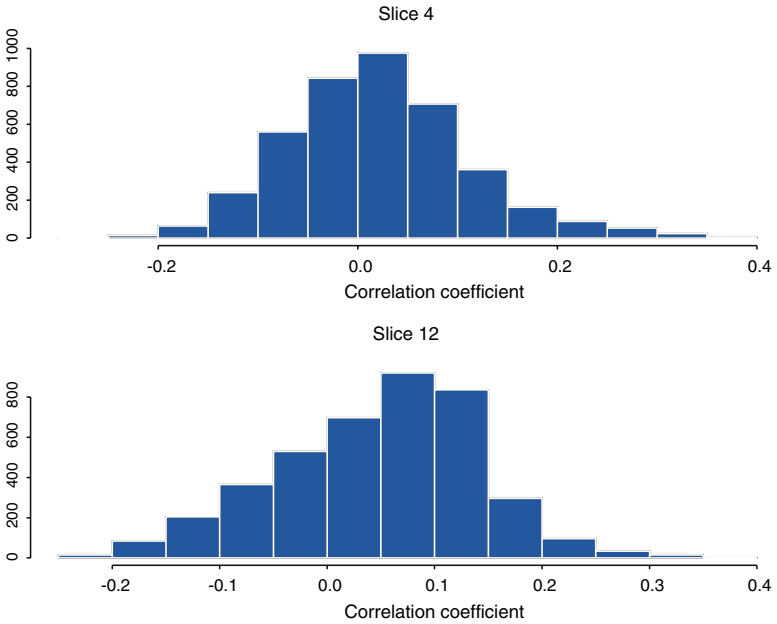
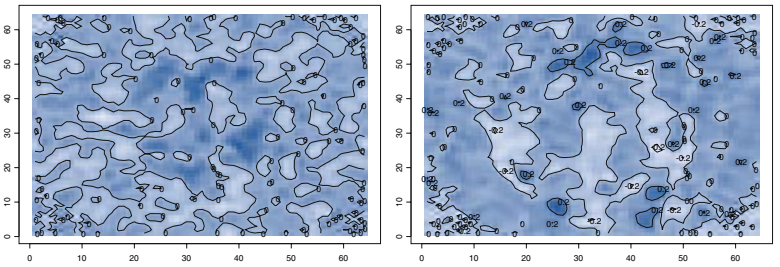


Fig. 12.3. Distribution of the correlation coefficient for two slices of the data set.



(a) Slice 4 (b) Slice 12

Fig. 12.4. Contour plots of the correlation coefficients in the fourth (left) and twelfth (right) slices. Darker shading indicates higher positive correlation.

trend in both the minimum and the maximum from approximately the twentieth slice on. This indicates that, overall, the correlation with the stimulus presentation trail is somewhat weaker in the deeper slices.

Another simple analysis is a two sample t test, as described in Section 5.2. In this study, there are 83 presentations of the prosaccade task and 73 of the antisaccade, so the design is not perfectly balanced. This doesn't matter for the t test, as the sample sizes are close enough. The distributions of the test statistic in the fourth and twelfth slices are shown in Figure 12.6. Interestingly,

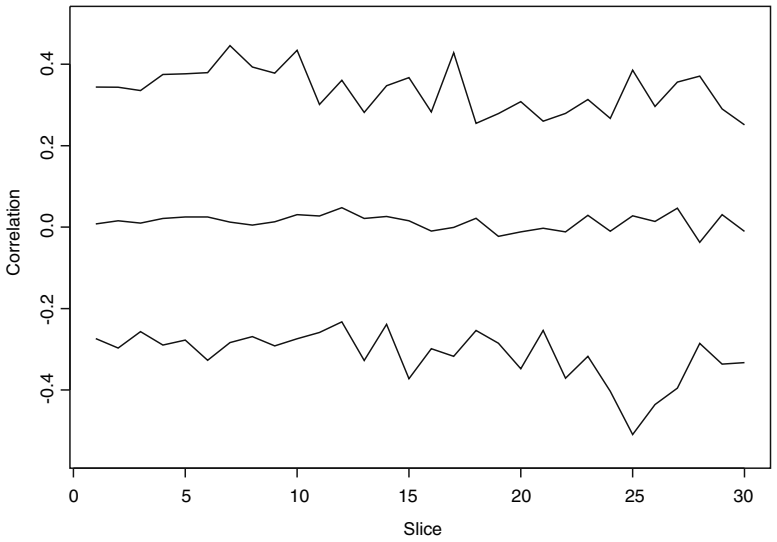
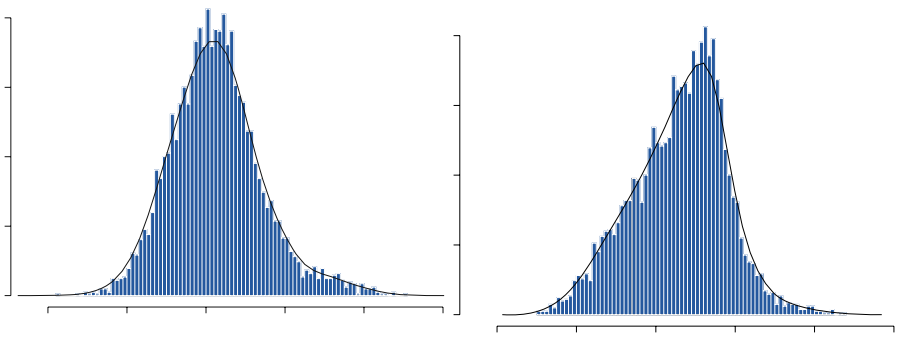


Fig. 12.5. Minimum, mean, and maximum correlation in each of the 30 slices.

while the histogram for the fourth slice looks very normal in shape at first glance, the histogram for the twelfth slice is obviously skewed, possessing a long right tail.



(a) Slice 4 (b) Slice 12

Fig. 12.6. Distributions of the t statistics in the fourth (left) and twelfth (right) slices. The distribution in the twelfth slice is clearly skewed; the distribution in the fourth slice is also skewed, with a small mode in the far right tail.

In fact the distribution in the fourth slice is also skewed, with a smallish mode in the far right tail. This demonstrates the subtlety of the effects that we are searching for, and the difficulty of the inferential process. We show this in another way in Table 12.1 for a few selected slices. The table shows the number of voxels in each of the slices with t values greater than 4 and the number with t values less than -4. In all cases, even the twelfth slice, there are voxels with very high t values (that is, stronger activation during antisaccade conditions than during prosaccade conditions) and none with very low.

Slice	$t > 4$	$t < -4$	$t > 1.655$	$t > 4.35$	$t > 5.14$
4	20	0	434	5	0
5	33	0	553	20	0
12	11	0	667	6	0

Table 12.1. Number of voxels attaining particular thresholds for three slices in the data set. The first two columns give the numbers of voxels with t values above 4 and below -4, respectively. The lack of balance in these numbers is one indication of the slight skewness in the direction of the effect of interest and demonstrates the subtlety of the response. The third column corresponds to thresholding without any correction for multiplicity, at level $\alpha = 0.05$. The fourth and fifth columns show Bonferroni-adjusted thresholds based on a single slice and the whole brain, respectively.

We can also use these slices to explore some of the thresholding methods discussed in Chapter 10. This is also presented in Table 12.1. With $\alpha = 0.05$ and no correction for multiplicity, the t threshold on 154 degrees of freedom is 1.655 (compared to 1.645 with the z score), resulting in hundreds of significant voxels in every slice; this is clearly too many since most of those are outside the brain. An adjustment for multiple testing is needed to avoid many false discoveries.

With the same $\alpha = 0.05$, and applying a Bonferroni correction on each slice separately, the adjusted level is $0.05/4096 = 0.0000122$; this corresponds to an equivalent t score of 4.35 for t with 154 degrees of freedom (or 4.22 if we use the z score instead). With this adjustment, the number of detected voxels in each slice is dramatically reduced, from several hundred to at most tens. Of course, the appropriate correction in this case should be over all slices, for a t cutoff of 5.14 or a z of 4.93, in which case no significant voxels remain, belying the robustness and well-understood behavior of the antisaccade task.

Slicewise false discovery rate procedures applied to the same three slices with $q = 0.05$ yield equivalent t thresholds of 3.12 (fourth slice; 90 significant voxels), 3.02 (fifth slice; 125 significant voxels), and 3.33 (twelfth slice; 46 significant voxels), respectively.

Figure 12.7 shows the results of three thresholding approaches (no correction, slicewise Bonferroni, and slicewise false discovery rate control) for the

fourth slice; as reference, the first panel in the figure is a plot of the slice at an arbitrary point in time. Ordinarily, a high-resolution anatomical image is taken at the same time as the functional data are acquired and this is used as background for presenting assessments of significant voxels. We don't have such an image for this data set, and so the low-resolution functional is given instead in order to point out the general shape and location of the brain in this slice.

Even when no correction for multiple testing is performed, the majority of the voxels declared active are in the brain. The problem is that the detected activation is so widespread as to be meaningless; the effect is more localized than indicated by these results. Furthermore, many "active" voxels are detected outside of the brain altogther. At the other extreme, the slice-wise Bonferroni correction removes almost all traces of activation, although the detected voxels are in an appropriate location for the tasks in question. Control of the false discovery rate yields a good middle ground, as is expected. The significantly active voxels are localized in regions that are relevant to the task, organized contiguously, and confined to the brain.

In and of itself, this analysis stream (voxel-by-voxel t tests followed by a correction for multiple testing) is sufficient, if not sophisticated. The more elaborate analyses derived and advocated by various researchers, as described in the chapters devoted to statistical methodology, aim to refine these results in different ways. But it is evident that one could be satisfied with the simpler analysis, and indeed one often is. Additionally, in terms of understanding the outcome of an experiment, it is not clear on the face of it how much one gains from applying more complicated models. Obviously, if the goal is to delve, for example, into questions of connectivity, the analysis presented here is wholly inadequate. For basic detection of regions of activation, on the other hand, the "variations on a theme" should be taken as just that.

The remainder of this chapter will explore briefly other analyses that could be carried out on these data, demonstrating some of the approaches explained in previous chapters. For ease of calculation and comparison, the fourth slice will be used for all analysis.

First, we consider clustering the voxel time courses, as described in Section 6.3.1; this is a simple spatiotemporal approach. As discussed there, some researchers advocate crudely thresholding the activation map prior to clustering in order to reduce the number of voxels and ease the computational burden. In fact, at least for a single slice of data, clustering on the unthresholded data is not that much more time consuming than clustering on a reduced data set. Results from both approaches are therefore included here.

Without prescreening we look at $k = 3, 4, \ldots, 10$ clusters, in Figures 12.8 and 12.9. It is interesting to note that with $k = 3$, one cluster is used for voxels outside the head, as can be seen by the clear shape of the brain in the plot. As more clusters are added, this feature is retained and the additional clusters partition the brain area more finely. However, starting with $k = 6$ clusters, some of the air voxels are joined with the cluster that defines the

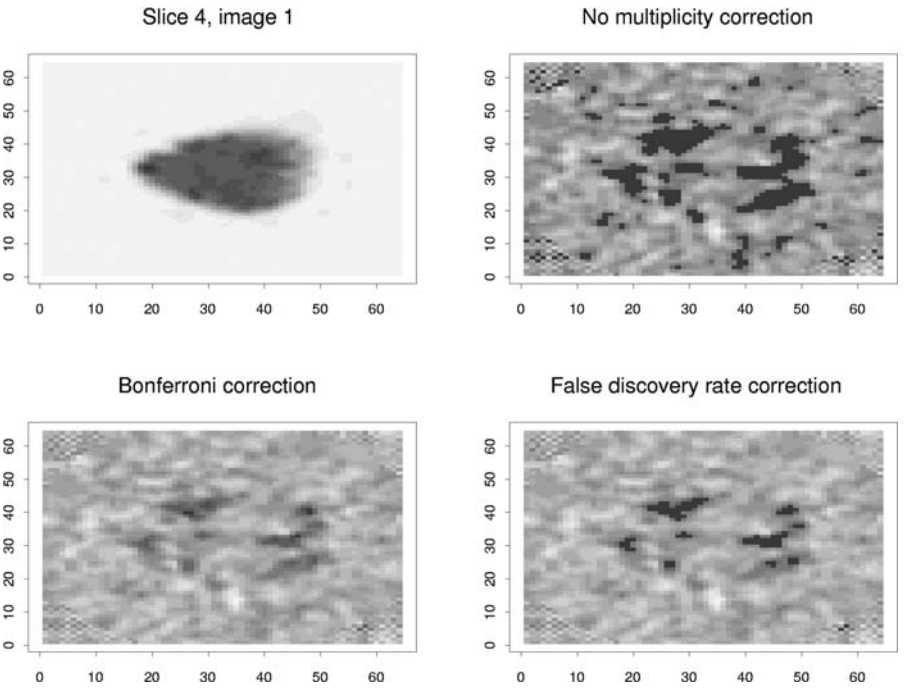


Fig. 12.7. Fourth slice of the data set. For reference, the slice at an arbitrary point in time is shown in the top left panel. The other panels show the significant voxels when there is no correction for multiple testing (top right), when the slicewise Bonferroni correction is applied (bottom left), and when the slicewise false discovery rate procedure is applied (bottom right). The results of thresholding are qualitatively similar in all three cases, but differ meaningfully in the details.

outer edge of the brain. This pattern, too, is consistent from $k = 6$ to $k = 10$ clusters. Furthermore, it is not apparent that adding clusters actually succeeds in isolating the voxels in which activity is taking place, if we compare to Figure 12.7. So, while the initial result for $k = 3$ is promising, in that brain voxels and nonbrain voxels belong to disjoint clusters, this method does not do a very good job at detecting activation.

As a crude prescreening method, we take only those voxels for which the calculated t statistic is greater than 2. This leaves 287 voxels. These are shown in the top left panel of Figure 12.10. The shape of the brain is only roughly apparent, and note also that some of the retained voxels are clearly outside the brain. Now the main utility of the clustering algorithm is to isolate those air voxels into a single cluster, and this it accomplishes. Note that some brain voxels are included in the cluster as well; this is not a problem, since the rough threshold should include voxels that are not truly in the regions of activation.

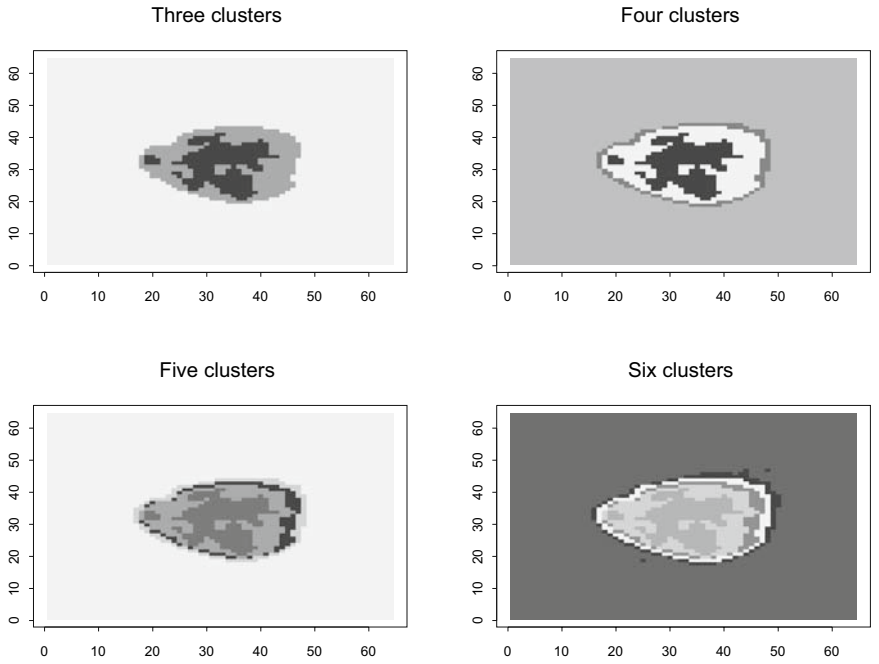


Fig. 12.8. Clustering all the voxels in the fourth slice with hierarchical clustering; distance is between the voxel time courses. Results of $k = 3, 4, 5, 6$ clusters.

Here, too, we see the phenomenon that successive clusters seem to be "peeling away" different layers or rings within the brain slice. Simply clustering the voxel time courses is apparently not sufficient to separate regions of activation from other voxels with similar behavior, even with some prescreening.

Basic principal components analysis is an example of a multivariate approach. Here, the 287 prescreened voxels are used in the analysis, which for purposes of demonstration builds components that are linear combinations of voxels as opposed to time (recall that the roles of voxels and time are interchangeable; the components based on one are a transformation of those based on the other). A scree plot of the first ten components is in Figure 12.11. The first component explains 60% of the variance; the first ten together explain 86%.

A representation of the first six principal components is given in Figure 12.12. To obtain these figures, only voxels with loadings greater than 0.08 (in absolute value) on the component are plotted. The first component only has voxels with high positive loadings, whereas the other five have voxels with both negative and positive loadings. The first component is made up of voxels around the edge of the brain image; depending on where exactly these are located relative to the actual brain space, this could represent either head

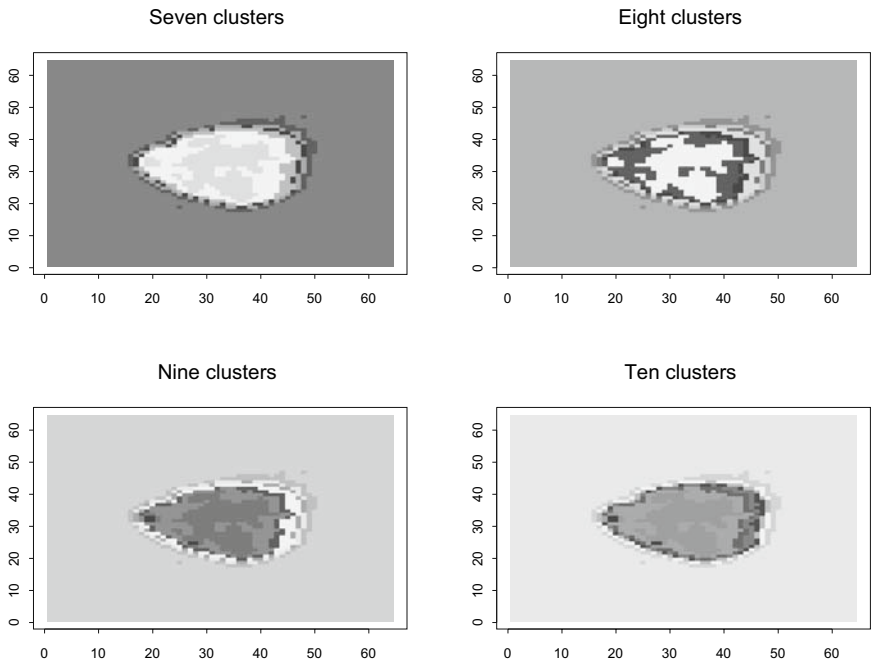


Fig. 12.9. Clustering all the voxels in the fourth slice with hierarchical clustering; distance is between the voxel time courses. Results of $k = 7, 8, 9, 10$ clusters.

motion or gray matter. The second component contrasts voxels on the edge of the brain with voxels in the middle; the middle voxels correspond roughly to the task-related areas. The third component is almost a complement of the second. The fourth component contrasts the front of the brain and the back. The fifth and sixth components aren't easily interpreted, although the sixth is close to a contrast between the left and right sides of the brain.

Figure 12.13 shows the distribution of the number of principal components (out of the first six) to which each of the 287 prescreened voxels belongs. Only 155 have high loadings, according to our definition, on any of the first six principal components; of these, 65 load on a single component, 47 on two, 24 on three, 12 on four, 6 on five, and 1 voxel loads on all six. As can be seen in the figure there is a clustering in this distribution, with voxels that load on more than one component (a darker appearance means loading on more components) appearing laterally in the frontal eye field (the dark patch near the front of the brain, towards the bottom of the plot) and in the occipital cortex (the darkish patch near the back of the brain image on the far right of the plot).

Another way to look at the results from these analyses is to consider the average time course of the different sets of voxels, for example those retained

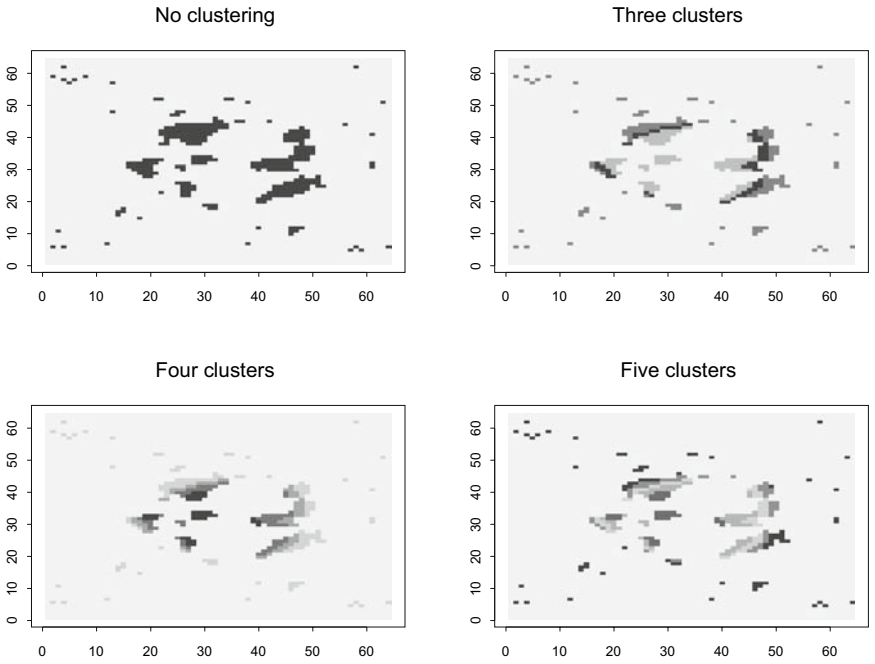


Fig. 12.10. Clustering on 287 prescreened voxels, $k = 3, 4, 5$.

by the Bonferroni correction, or those that load on the first principal component. Insofar as task-related voxels are being distinguished by a particular method, the average time course should reflect the boxcar shape of the block design. Some sample average time courses are given in Figures 12.14, 12.15, and 12.16.

Figure 12.14 shows the signal averaged (at each time point) over all the voxels retained by the slicewise Bonferroni correction and by the slicewise false discovery rate procedure. The two series exhibit largely similar behavior, very roughly paralleling the design of the experiment.

In Figure 12.15 we see the average time courses for the case $k = 3$ clusters based on the 287 prescreened voxels. These series differ widely in their average level, although closer inspection of the time courses individually reveals that the general pattern of each is the same. However, this does indicate that the clusters are picking up something other than just the overall trend of relatedness to the stimulus trail. Since the clustering is based on distance between the voxel time courses, it makes sense that voxels with similar levels of activation (and not just similar patterns) will cluster together. It is interesting to note that of the ten most significant voxels (that is, those with the highest value of the t statistic calculated earlier), seven belong to the first cluster, two belong to the second, and one to the third. While the first cluster thus

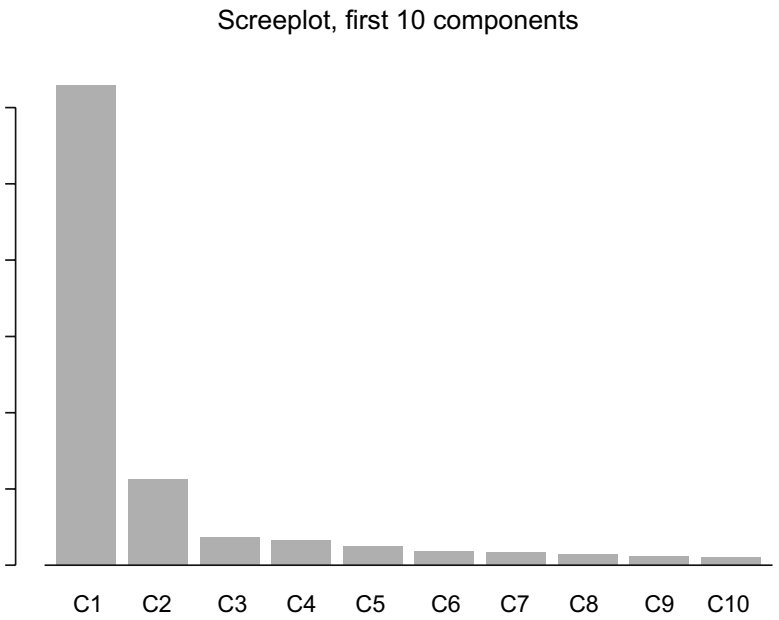


Fig. 12.11. Screeplot of the first ten principal components, fourth slice of the data, prescreened voxels only.

is most highly associated with activation, the suspicion raised earlier that the clusters are largely based on location receives some additional validation from this analysis.

Finally, Figure 12.16 shows the average time courses for the voxels that load highly (loading greater than 0.08 in absolute value) on the first four principal components. The picture is similar to what we have already seen. Again, there is not a simple correspondence between the derived principal components and overall levels of activation as detected by the t statistic. For example, of the ten most significant voxels, four do not load strongly on any of the first six principal components, three load on only one, two load on two, and one loads on four components. Only two of the ten load on the first component, and only two load on the second; these components are not solely isolating extremely active voxels. The similarity of the average time courses, both in pattern and in overall level of signal, is further proof of this. In addition, the seven voxels that load on at least five of the first six principal components (one loads on all six, the rest load on five) are not among the most active; they have t values between 3.2 and 3.6.

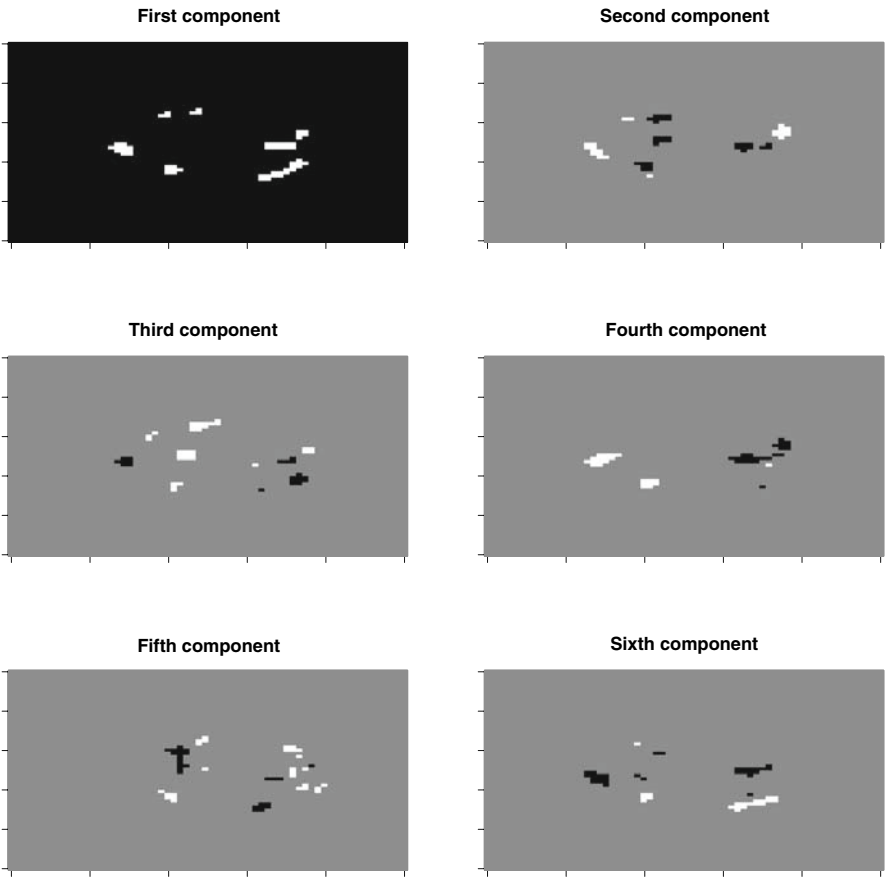


Fig. 12.12. Voxels with loadings greater than 0.08 in absolute value, for each of the first six principal components.

12.3 Summary

In this chapter we have shown just a small sampling of the possible analysis; we have looked primarily at a single slice and have simplified by either performing univariate analyses or prescreening prior to analyses that are more multivariate in nature. We have not fit any of the complex models that are available in software packages such as SPM, or from individual researchers. Even with the purposely restricted analysis, however, we have been able to uncover many interesting relationships among the voxels, we have located areas of apparent scientific interest, and we have set ourselves up to perform deeper and more intricate analysis of the data at hand.

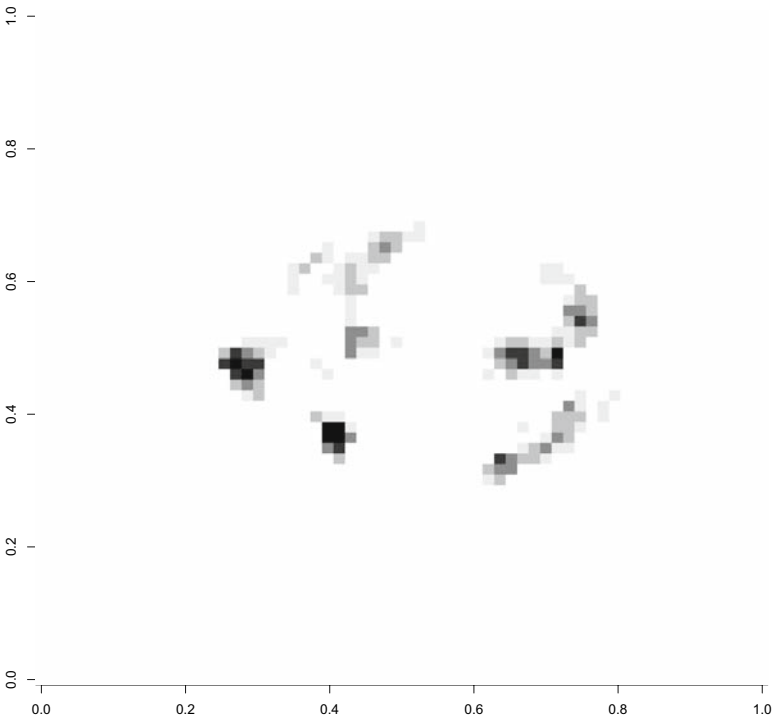


Fig. 12.13. Distribution of the number of principal components, out of the first six, on which each of the voxels in the analysis has high (positive or negative) loading. Darker colors correspond to higher numbers of components. Almost half – 132 out of 287 voxels – do not load highly on any of the first six principal components. Of the rest, 65 load on only one component and 47 on two. The voxels that load on four or more of the principal components, the dark patches in the image, seem to correspond roughly to task-relevant areas.

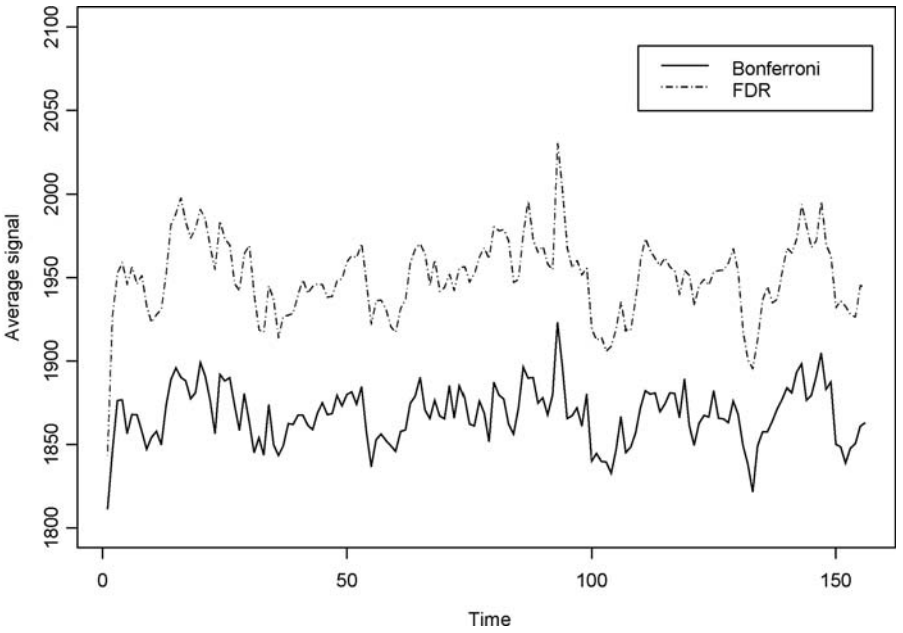


Fig. 12.14. Average time course for voxels retained by slicewise Bonferroni correction and by slicewise control of the false discovery rate. The patterns exhibited by the two sets of voxels are similar, following the design of the experiment.

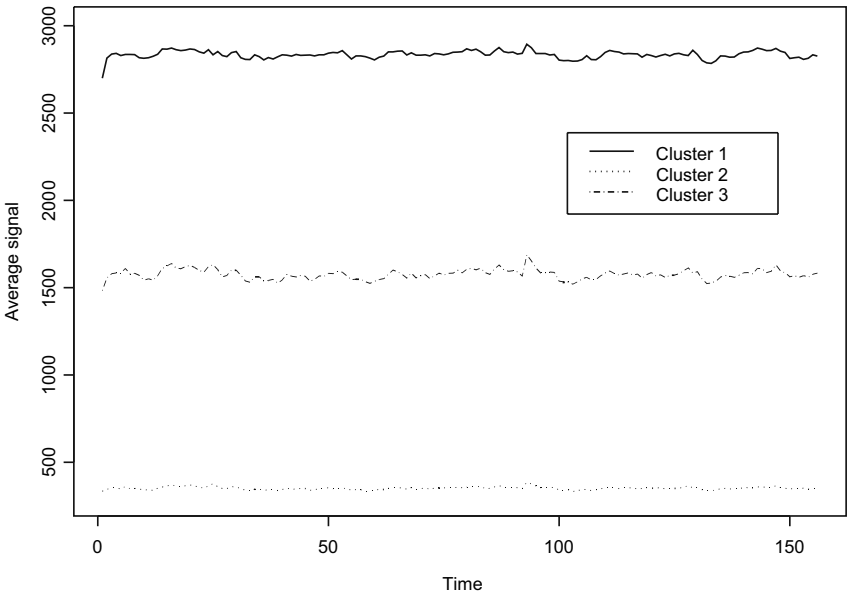


Fig. 12.15. Average time courses for the case $k = 3$, hierarchical clustering on the 287 prescreened voxels.

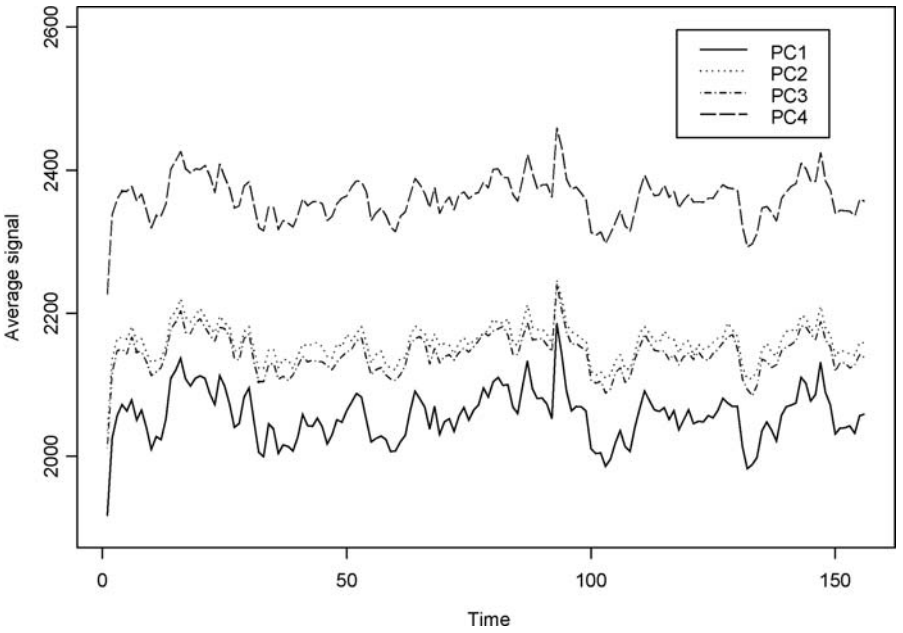


Fig. 12.16. Average time courses for the voxels in the first four principal components, among the prescreened voxels.

A

Survey of Major fMRI Software Packages

The focus of this appendix is fMRI software. Although there is both commercial software and freely downloadable software available for the analysis of fMRI data, I concentrate here on the latter, since most analyses published in the literature are performed in one of two noncommercial packages: AFNI and SPM. Thus there is no discussion in this appendix of popular commercial packages such as Brain Voyager and Analyze. Many groups also develop their own procedures using programming languages such as MATLAB. Software is constantly being updated and upgraded; readers are advised that the content of this appendix is current to the time of writing, and they should check the relevant links for the newest versions. Homepages for the various software packages give detailed explanation on downloading and use, hence I don't provide this information here. SPM and AFNI both have active user lists (email fora for discussion); SPM also has several "wikis" that users new and old can consult and contribute to. The electronic resources for these packages in particular are extensive and are an important source of knowledge and information for the communities of users.

Also worthy of note here is the Internet Analysis Tools Registry whose address is `http://www.cma.mgh.harvard.edu/iatr`. It contains a listing of image analysis tools available to the community. As of January 2008, 186 tools, both commercial and freeware, are registered at the website. The page for each tool includes a brief description, contact information (email addresses of developers, and a link to the tool's website), software requirements, whether or not it is open source, whether or not it is freeware, and lists of technical and applied references. Researchers can also review and evaluate the tools in the registry.

A.1 Analysis of Functional NeuroImages: AFNI

The package AFNI (homepage `http://afni.nimh.nih.gov`) was developed by Robert Cox starting in 1994, originally as a program for translating images

into Talairach coordinates. Today it is one of the most widely used packages for the analysis of fMRI data, providing a full range of tools for statistical modeling and inference, and the visual display of results. AFNI is a collection of programs written in C, and runs in Unix environments (including Linux and Mac OS X).

Among the capabilities of AFNI are transformation of brain images to Talairach coordinates; display of axial, coronal, or sagittal views of the data; display of voxel time series; linear models, including the simple correlation method for computing activation maps (see Chapter 5, Sections 5.2 and 5.4 for details); thresholding using contiguity thresholds or FDR control (see Chapter 10, Sections 10.1 and 10.4).

For information on how to join the AFNI mailing lists and message board, see the websites `http://afni.nimh.nih.gov/afni/community/lists` and `http://afni.nimh.nih.gov/afni/community/board`, respectively.

A.2 Statistical Parametric Mapping: SPM

The package SPM (homepage `http://www.fil.ion.ucl.ac.uk/spm`) is a suite of MATLAB programs for the analysis of brain imaging data in general, including imaging modalities beyond fMRI. It was originally developed in 1991 by Karl Friston to analyze images collected using positron emission tomography (PET). SPM can be run in both Unix (including Linux) and Windows environments.

Among the capabilities of SPM are realignment of image sequences; automated spatial normalization; segmentation of images; spatial smoothing; data analysis via a general linear model approach (maximum likelihood and Bayes estimation); display of statistical maps; display of posterior probability maps; analysis of functional connectivity.

SPM is perhaps the leading software package for the analysis of fMRI data in terms of popularity, and as such has played a prominent role in shaping how practitioners think about the statistical aspects of their data. The general linear model approach – and in particular the random effect model, the canonical HRF model, the ways in which SPM presents the output of an analysis, have all become standards in the literature. This is useful, on the one hand, since it provides a uniform frame of reference for researchers from different laboratories; however, this uniformity can pose problems for neuroimagers who use other software packages, or indeed who develop their own programs. I have seen referee reports, for example, in which it was asked why the data weren't analyzed using SPM, as if this were the only available option. As a field, we need, I think, to be wary of such trends. Fortunately, most of the people involved in software development seem to believe this, too.

For information on how to join the SPM discussion list, see the website `http://www.fil.ion.ucl.ac.uk/smp/support`. The wiki-books on SPM can

be found at the sites http://en.wikibooks.org/wiki/SPM and htpp://en.wikipedia.org/wiki/Statistical_parametric_mapping.

A.3 Other Packages

Although SPM and AFNI are widely used, there are other prominent packages also available. Some of these are described here. The list is by no means exhaustive, however.

Functional Imaging Analysis Software, Computational Olio (FIASCO; homepage http://www.stat.cmu.edu/~fiasco/) was developed at Carnegie Mellon University's Statistics Department, primarily by Bill Eddy. FIASCO is a collection of shell scripts and executables written in C and Python. It performs preprocessing (detrending, motion correction, and so forth), fits linear models to the data, thresholds and displays images. Users can also write their own procedures to customize their analyses.

Automated Image Registration (AIR; homepage http://bishopw.loni.ucla.edu/AIR5/index.html) was developed by Roger Woods to perform automated registration of two- and three-dimensional images, both within and across subjects. Registration across imaging modalities is also supported. AIR is written in C and runs in Unix, Windows, and Apple environments.

FMRIB Software Library (FSL; homepage http://www.fmrib.ox.ac.uk/fsl) is written mainly by the members of the Analysis Group, FMRIB, at Oxford University. FSL is a library of tools for image analysis and statistical processing of fMRI data, among other modalities. It runs in Apple, PC (Linux and Windows), and Unix environments. Among the capabilities of FSL for functional imaging are: general linear model analysis; Bayesian analysis; model-free analysis via Independent Component Analysis; spatial mixture modeling; thresholding using the permutation test, Gaussian random field, and false discovery rate approaches; interactive display of three- and four-dimensional images; registration and segmentation of images. FSL has an email list for users; archives and information on joining this list can be found at http://www.jiscmail.ac.uk/lists/fsl.html.

VoxBo (homepage http://www.voxbo.org) is a suite of C/C++ programs that runs in a Linux environment, including OS X for Mac and Cygwin for Windows. VoxBo performs standard preprocessing (motion correction, normalization, smoothing); data analysis via the general linear model for block and event-related designs; and graphical presentation at the voxel level (voxel time series, for example). The analysis focus is on the univariate general linear model; other types of analysis are not supported in VoxBo. A characteristic of VoxBo is its *scheduling mechanisms*, which allow for easy batch processing of fMRI data sets.

Like SPM, VoxBo has a wiki, found at http://voxbo.org/wiki/index.php/Main_Page. There are also several mailing lists for this package; for infor-

mation on the lists and how to join them, see `http://www.voxbo.org/lists.html`.

fMRIstat (homepage `http://www.math.mcgill.ca/keith/fmristat`) was developed by Keith Worsley of McGill University. It is a MATLAB-based collection of tools and can be run in Windows and Linux environments. fMRIstat features a variety of linear model analyses, analysis of the hemodynamic response function, thresholding via random field theory or false discovery rate control, and an advanced suite of visualization modules.

A.4 Comparison of Imaging Software Packages

Gold et al. (1998) report a descriptive comparison of many of the packages (both freeware and commercial) available in the late 1990s for the analysis of fMRI data. The comparison considers operating system; availability of source code; completeness of documentation (including ease of learning and the inclusion of a graphical user interface – GUI); necessary preprocessing steps; inclusion of image realignment routines; capability to input images of different dimension; types of statistical analysis; image display features; inclusion of spatial transformations; and corrections for multiple testing.

As might be expected, Gold et al. (1998) find that each package has advantages and drawbacks. The choice of software depends, to a large extent, on the requirements of the particular laboratory or group. Hence it is not possible to conclusively recommend one package over the others. AFNI and SPM, for example, have extensive GUIs, which make them easy to use. On the other hand, SPM relies heavily on MATLAB, a potential barrier for users who would therefore be required to obtain the latter in order to run the former. FIASCO doesn't have a GUI at all; rather, routines are invoked on command-line operations, in a hierarchical structure (scripts call scripts); while some may see this as a drawback, it does in fact allow users a great deal of flexibility in customizing analysis.

The following table summarizes some of the features of the packages described in this appendix. In practice, many people find it helpful to use different packages for different parts of their analyses in order to build on the strengths of each. The software developers themselves generally take an ecumenical "mix and match" approach; useful analyses from one package are also often quickly adopted by others.

Feature	AFNI	SPM	FIASCO	FSL	VoxBo	fMRIstat
Computing environment	Unix	Linux Windows	Unix	Linux Windows	Linux	Linux Windows
Source code language	C	MATLAB	C/Python	Unknown	C/C++	MATLAB
GUI	Yes	Yes	No	Yes	Yes	No
Email list	Yes	Yes	No	Yes	Yes	No
Other user resources	No	Wiki	No	No	Wiki	No
Statistical analysis	LM+	LM+	LM+	LM+	LM	LM+
Multiple testing	CT/FDR	RF/FDR	FDR	CT/FDR	CT+	RF/FDR

Table A.1. Summary of fMRI software analysis packages. Readers are encouraged to visit the homepages of the packages, listed in the text for a more in-depth discussion of each one. Across from Statistical analysis, an entry of "LM" means that the linear model is the sole method of analysis; an entry of "LM+" means that other modes of analysis are also available. Across from Multiple testing, "CT" refers to contiguity cluster thresholding, "FDR" refers to false discovery rate control, and "RF" refers to random field theory.

B

Glossary of fMRI Terms

Aliasing Distortion of the measured signal. If the sampling rate is too low, the highest frequencies in the signal will be expressed as lower frequencies and the reconstructed image will be contaminated with artifacts: the high frequencies are aliased as low frequencies.

Anterior Toward the front of the brain (see *rostral*).

Axial slice A view of the brain, as if looking from the top of the head.

B_0 field The static magnetic field generated by the MRI scanner.

B_1 field The magnetic field caused by the introduction of energy into the scanner system.

Bandwidth The range of frequencies over which the data are sampled. Determines the thickness of the slices.

Bloch equation The equation that describes how the net magnetization of a tissue changes over time. The Bloch equation is the sum of three terms: a precession term, involving the gyromagnetic ratio and the field strength; a T_1 term; and a T_2 term.

Blood Oxygenation Level Dependent (BOLD) response The changes in oxygen levels of the blood as a result of neuronal activity in reaction to a presented stimulus. Under stimulation, blood flow increases to the involved regions of the brain. However, the extra oxygen that travels along with the blood is not recruited to the working cells, causing changes in the relative levels of oxyhemoglobin and deoxyhemoglobin. These changes are measured as the BOLD response. Stereotypically, this response manifests itself in: (i) an initial dip below baseline levels; (ii) an increase to peak levels; (iii) decay following the cessation of the stimulus, possibly to a level below the baseline, before recovering baseline. The peak BOLD response is small, on the order of 2-3% signal change at 1.5T, occurring approximately 6 seconds after stimulus presentation. Most functional MR imaging measures BOLD contrast.

Caudal Toward the back of the brain (see *posterior*).

Cerebrospinal fluid (CSF) A clear liquid that surrounds the brain and spinal cord.

Coronal slice A view of the brain, as if looking from the front of the head.

Crosstalk The phenomenon by which radiofrequency pulses that are applied to adjacent slice positions excite the same region of tissue, resulting in contamination of the measured signal. Crosstalk is often avoided by leaving a gap between slices and by not acquiring the slices in sequential order.

Deoxyhemoglobin Hemoglobin (the protein in red blood cells) without oxygen ("blue blood"). Deoxyhemoglobin is paramagnetic.

Diamagnetism Having a weak repulsion from a magnetic field. Most tissues in the body are diamagnetic. When diamagnetic substances are placed in an external magnetic field, the effective strength of the field is reduced.

Dorsal Toward the top of the brain (see *superior*).

Echo-planar imaging (EPI) A data acquisition sequence that traverses k-space in a boustrephedonic fashion, alternating the reading of rows from left to right and from right to left. The fast switching of gradient direction allows for rapid image acquisition.

Excitation The application of energy at the resonant frequencies of tissue inside of the static magnetic field.

Ferromagnetism Having a strong attraction to a magnetic field. When ferromagnetic substances are placed in an external magnetic field, they become permanently magnetized. Iron, cobalt, and nickel are ferromagnetic.

Field of view (FOV) The dimensions of the acquired image of the tissue.

Field inhomogeneity Nonuniformity of the static magnetic field, caused by imperfections in the magnet and interference from external sources.

Flip angle The change in precession angle from aligning with the z-axis (direction of the magnetic field) to aligning with the transverse (x, y)-plane following application of a pulse. A 90^o flip angle flips the net magnetization into the (x, y)-plane; lesser flip angles tilt the net magnetization in the direction of that plane.

Fourier transform The mathematical operation that is used to reconstruct data acquired from the MR scanner into an image recognizable as a brain. If $\rho(x, y)$ is the density of hydrogen nuclei at location (x, y), and $I(t)$ is the signal that is stored by the scanner at time t, then it can be shown that the measured signal is the Fourier transform of the values ρ, which are ultimately of interest. The discrete (sampled on an $n \times n$ grid) two dimensional Fourier transform of a function $f(x, y)$ is given by

$$g(w, v) = \frac{1}{n^2} \sum_{x=0}^{n-1} \sum_{y=0}^{n-1} f(x, y) e^{-i2\pi(wx/n+vy/n)}.$$

Free induction decay (FID) The shape of the signal that is picked up and measured by the receiver coil, which is a slowly decaying sinusoidal wave.

Frequency encoding gradient The gradient applied in the x direction of a slice, this determines location along the x-axis. Also called the *readout gradient*.

Gap The space left between consecutive slices of data. Used to limit crosstalk.

Ghost An artifact that manifests itself in a shadow image superimposed on the main reconstructed image. Ghosts are the result of improper alignment of the gradients.

Gradient A magnetic field that varies spatially, usually linearly. The gradients, when applied in each of the x, y, and z directions, induce temporary inhomogeneities in the static field, so that each locus in the tissue being imaged has a unique resonant frequency. Hence protons at each voxel, determined by the gradients, are differentially excited.

Gradient coil The physical mechanism by which the gradients are applied. Coils are electrical devices made up of loops of wire.

Gray matter The tissue in the brain in which neural activity takes place. Areas in gray matter consist of nerve cell bodies with no myelin (fatty) covering.

Gyromagnetic ratio The ratio of the field strength (the magnetic moment) to the frequency, or angular momentum, of the nucleus. Denoted γ. Each type of nucleus has its own value of the gyromagnetic ratio. For ^{1}H (hydrogen) $\gamma = 42.58$ MHz/T. γ depends only on the charge and the mass of the atomic nucleus, hence it is constant for a given type.

Hemodynamic response The changes that occur in the measured MR signal due to brain activity. The hemodynamic response is a result of the decrease in levels of deoxygenated blood around active regions. It is characterized by a short lag following stimulation, after which a rise in signal is observed. This reaches a peak some 6-7 seconds after stimulus presentation (with some variability); if no further stimulation is present, the signal will slowly decay, often dipping below the starting levels before returning eventually to baseline.

Image space The space in which the data are viewed. Most statistical analysis takes place in image space.

Inferior Toward the bottom of the brain (see *ventral*).

k-space The space in which the data are acquired. Some data preprocessing, and some statistical analysis, may be performed in k-space.

Larmor equation This equation provides the rate at which a proton precesses around the external (static) magnetic field: $\omega = \gamma \mathbf{B}_0$, where γ is the gyromagnetic ratio and $\mathbf{B}_0$ is the strength of the field (typically 1.5 or 3T for imaging experiments on humans).

Lateral Away from the middle of the brain; toward the edge of the brain.

Longitudinal magnetization The magnetic field in the direction of the main, external, field, denoted as the z-axis. Longitudinal magnetization is caused by the fact that when tissue is placed in the strong external field,

more protons align parallel to the field than anti-parallel to it, resulting in a net magnetization in the z direction.

Longitudinal relaxation time The time required for the magnetization in the longitudinal direction to return to its starting (maximal) value. Following the application of a pulse that tips the field into the transverse plane, the longitudinal magnetization is zero. When the pulse is turned off, the magnetization parallel to the main field begins to recover, finally reachieving the initial value. Also known as *spin-lattice relaxation* or T_1 *recovery*.

Medial Toward the middle of the brain.

Net magnetization The total magnetization of the system at a given time. This is influenced by the external field, as well as any gradient pulses that have been applied.

Neurological convention The display of brain images so that the right side of the image corresponds to the right side of the brain, and the left side of the image corresponds to the left side of the brain.

Nuclear magnetic moment The (small) magnetic field generated by an individual nucleus with nonzero spin.

Oblique slice A view of the brain that is not from any of the three main directional planes (axial, coronal, or sagittal). Oblique slices are obtained by linearly combining the slice selection, phase encoding, and frequency encoding gradients.

Oxyhemoglobin Hemoglobin (the protein in red blood cells) that is loaded with oxygen. Oxyhemoglobin is bright red in color.

Paramagnetism Having an attraction to a magnetic field. When a paramagnetic substance is placed in an external magnetic field, it becomes magnetized; it becomes demagnetized once the field is turned off or it is removed from the field. Paramagnetic substances increase the effective strength of the magnetic field in which they are placed.

Phantom An object, usually filled with a liquid or gel of known properties, used for testing the MR system.

Phase coherence The state of the protons after the application of a radiofrequency pulse, whereby they all precess around the central axis of the magnet in unison.

Phase encoding gradient The gradient applied in the y direction of a slice, this determines location along the y-axis.

Posterior Toward the back of the brain (see *caudal*).

Precession The motion of an object with a central axis, such as a spinning top or a proton, around that axis. When there is no external magnetic field, a proton will rotate around its central axis, thereby generating a small magnetic field. When there is an external field, a proton will rotate around its central axis, but will also precess around the center of the external field. The rate of precession of the proton around the external magnetic field is determined by the *Larmor equation*.

Proton One of the elemental particles, along with neutrons and electrons, that make up the atom. Different atoms have different numbers of protons, and this determines their properties. The nuclei of hydrogen atoms have a single proton; since hydrogen is the most common element in the human body, fMRI usually images hydrogen, and the properties of single protons are of prime importance in understanding how the imaging technique works.

Pulse sequence A description of the various energies injected into the MR system over the course of a study, which allows the scanner to create the required images.

Radiofrequency (RF) coil Coil that transmits or receives RF pulses. RF coils may be transmitters only, receivers only, or both.

Radiofrequency (RF) pulse An insertion of energy into the system that generates a weak magnetic field and "flips" the net magnetization vector into the transverse plane.

Radiological convention The display of brain images so that the right side of the image corresponds to the left side of the brain, and the left side of the image corresponds to the right side of the brain.

Resonance The result when the frequency of the RF pulse matches the rate of precession of the protons in a region of tissue. Resonance adds energy to the system; without it, the flip of the magnetization vector into the (x, y)-plane could not occur.

Rostral Toward the front of the brain (see *anterior*).

Sagittal slice A view of the brain, as if looking from the side of the head.

Shimming coil Part of the MR coil system that corrects for field inhomogeneities in the external magnetic field.

Signal to noise ratio (SNR) The size of the signal in the data relative to the external variability.

Slice A single cross-section of an imaging volume, that is, a plane of data in the axial, coronal, or sagittal direction (see *axial slice*, *coronal slice*, *sagittal slice*).

Slice selection The process of exciting the protons within a chosen slice of the imaging volume by the application of gradients and electromagnetic pulses.

Spin A nucleus with the nuclear magnetic resonance (NMR) property, namely, one possessing both a magnetic moment and angular momentum. Only these nuclei can be used in magnetic resonance imaging. Nuclei with even-valued atomic masses cannot be spins. The most commonly used spin for functional MRI is hydrogen, as it is the most prevalent of the nuclei having the NMR property.

Spin-lattice relaxation time The time required for the magnetization in the longitudinal direction to return to its starting (maximal) value. Following the application of a pulse that tips the field into the transverse plane, the longitudinal magnetization is zero. When the pulse is turned off, the magnetization parallel to the main field begins to recover, finally

reachieving the initial value. Also known as *longitudinal relaxation* or T_1 *recovery*.

Spin-spin relaxation time The time required for the magnetization in the (x, y)-plane to decay to zero after the radiofrequency pulse is turned off. When the RF pulse is turned off, the protons, which had been precessing in phase, begin to lose phase coherence, causing the loss of magnetization in the transverse plane. Also called *transverse relaxation* or T_2 *decay*.

Spiral imaging A data acquisition sequence that traverses k-space in a spiraling pattern. Images maybe be acquired in a spiral-in fashion, which starts at the edge of k-space and ends in the center; or in a spiral-out fashion, which starts at the center of k-space and ends at the edge.

Static magnetic field The strong external magnetic field inside the MR scanner. The strength of this field is a constant, measured in Tesla (T), and is usually 1.5 or 3T for research on humans.

Superior Toward the top of the brain (see *dorsal*).

Susceptibility artifact A distortion in the acquired image due to field inhomogeneities where sinuses (air) and tissue neighbor each other.

T_1 **recovery** The time required for the magnetization in the longitudinal direction to return to its starting (maximal) value. Following the application of a pulse that tips the field into the transverse plane, the longitudinal magnetization is zero. When the pulse is turned off, the magnetization parallel to the main field begins to recover, finally reachieving the initial value. Also known as *longitudinal relaxation* or *spin-lattice relaxation*.

T_2 **decay** The time required for the magnetization in the (x, y)-plane to decay to zero after the radiofrequency pulse is turned off. When the RF pulse is turned off, the protons, which had been precessing in phase, begin to lose phase coherence, causing the loss of magnetization in the transverse plane. Also called *transverse relaxation* or *spin-spin relaxation*.

T_2^* **decay** The time until magnetization in the (x, y)-plane decays to zero, due to both loss of phase coherence and local inhomogeneities in the magnetic field. T_2^* is shorter than T_2, as it incorporates two sources of decay, and is the basis for BOLD fMRI imaging.

Tesla The unit of measurement of a magnetic field. The magnets used for fMRI research typically are 1.5 or 3 Tesla (T) strong.

Time to echo (TE) The interval between application of the radiofrequency pulse and acquisition of the signal. Measured in milliseconds.

Time to repetition (TR) The interval between successive applications of radiofrequency pulses. Measured in seconds.

Tissue contrast (T_1, T_2, and T_2^* weighting) The use of the different values of T_1, T_2, and T_2^* in different types of tissue to enhance certain features of the image. T_2 weighted images are sensitive to regions filled with fluid, such as tumors and other pathologies. T_2^* weighted images are sensitive to the amount of deoxyhemoglobin in the tissue, making them particularly useful for BOLD functional MR. T_1 weighting is not a significant factor in fMRI.

Transverse magnetization The magnetization in the (x, y)-plane, following a radiofrequency pulse that flips the field. After a 90^{o} pulse, all magnetization is in the transverse plane.

Transverse relaxation time The time required for the magnetization in the (x, y)-plane to decay to zero after the radiofrequency pulse is turned off. When the RF pulse is turned off, the protons, which had been precessing in phase, begin to lose phase coherence, causing the loss of magnetization in the transverse plane. Also called T_2 *decay*.

Ventral Toward the bottom of the brain (see *inferior*).

Voxel Three-dimensional volume element. The basic unit of fMRI data, a voxel comprises millions of neurons.

White matter The part of the brain responsible for transmitting information between areas of gray matter. Consists of nerve cells covered by a fatty myelin sheath.

References

Abbott, D. F., Opdam, H. I., Briellman, R. S., and Jackson, G. D. (2005). "Brief breath holding may confound functional magnetic resonance imaging studies." *Human Brain Mapping*, **24**, 284–290.

Adler, R. J. (1981). *The Geometry of Random Fields*. New York: John Wiley & Sons.

Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). "The variability of human, BOLD hemodynamic responses." *Neuroimage*, **8**, 360–369.

Andersen, A. H., Gash, D. M., and Avison, M. J. (1999). "Principal component analysis of the dynamic response measured by fMRI: A generalized linear systems framework." *Magnetic Resonance Imaging*, **17**, 795–815.

Bagarinao, E., Matsuo, K., Nakai, T., and Sato, S. (2003). "Estimation of general linear model coefficients for real-time application." *NeuroImage*, **19**, 422–429.

Balslev, D., Nielsen, F. A., Frutiger, S. A., Sidtis, J. J., Christiansen, T. B., Svarer, C., Strother, S. C., Rottenberg, D. A., Hansen, L. K., Paulson, O. B., and Law, I. (2002). "Cluster analysis of activity-time series in motor learning." *Human Brain Mapping*, **15**, 135–145.

Bandettini, P. A. and Cox, R. W. (2000). "Event-related fMRI contrast when using constant interstimulus interval: Theory and experiment." *Magnetic Resonance in Medicine*, **43**, 540–548.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C., and Hyde, J. S. (1993). "Processing strategies for time-course data sets in functional MRI of the human brain." *Magnetic Resonance in Medicine*, **30**, 161–173.

Baumgartner, R., Scarth, G., Teichtmeister, C., Somorjai, R., and Moser, E. (1997). "Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: Reproducibility." *Journal of Magnetic Resonance Imaging*, **7**, 1094–1101.

Baumgartner, R., Somorjai, R., Summers, R., and Richter, W. (2001). "Ranking fMRI time courses by minimum spanning trees: Assessing coactivation in fMRI." *NeuroImage*, **13**, 734–742.

Baumgartner, R., Windischberger, C., and Moser, E. (1998). "Quantification in functional magnetic resonance imaging: Fuzzy clustering vs. correlation analysis." *Magnetic Resonance Imaging*, **16**, 115–125.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). "General multilevel linear modeling for group analysis in FMRI." *NeuroImage*, **20**, 1052–1063.

Beckmann, C. F. and Smith, S. M. (2005). "Tensorial extensions of independent component analysis for multisubject FMRI analysis." *NeuroImage*, **25**, 294–311.

Benjamini, Y. and 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.

— (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics." *Journal of Educational and Behavioral Statistics*, **25**, 60–83.

Berk, R. H. and Cohen, A. (1979). "Asymptotically optimal methods of combining tests." *Journal of the American Statistical Association*, **74**, 812–814.

Bernardo, J. and Smith, A. F. M. (2000). *Bayesian Theory*. New York: John Wiley & Sons.

Besag, J. (1986). "On the statistical analysis of dirty pictures." *Journal of the Royal Statistical Society, Series B*, **48**, 259–302.

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics (with discussion)." *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Birn, R. M., Cox, R. W., and Bandettini, P. A. (2002). "Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing." *NeuroImage*, **15**, 252–264.

Biswal, B. B. and Ulmer, J. L. (1999). "Blind source separation of multiple signal sources of fMRI data sets using independent component analysis." *Journal of Computer Assisted Tomography*, **23**, 265–271.

Bowman, F. D. and Patel, R. (2004). "Identifying spatial relationships in neural processing using a multiple classification approach." *NeuroImage*, **23**, 260–268.

Brammer, M. J. (2001). "Head motion and its correction." In *Functional MRI: An Introduction to Methods*, eds. P. Jezzard, P. M. Matthews, and S. M. Smith. Oxford: Oxford University Press.

Breakspear, M., Brammer, M. J., Bullmore, E. T., Das, P., and Williams, L. M. (2004). "Spatiotemporal wavelet resampling for functional neuroimaging data." *Human Brain Mapping*, **23**, 1–25.

Brown, M. A. and Semelka, R. C. (1995). *MRI: Basic Principles and Applications*. New York: Wiley-Liss.

Buckner, R. L. (1998). "Event-related fMRI and the hemodynamic response." *Human Brain Mapping*, **6**, 373–377.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., and Rosen, B. R. (1996). "Detection of cortical activation during averaged single trials of a cognitive task using functional

magnetic resonance imaging." *Proceedings of the National Academy of Sciences (USA)*, **93**, 14878–14883.

Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996a). "Statistical methods of estimation and inference for functional MR image analysis." *Magnetic Resonance in Medicine*, **35**, 261–277.

Bullmore, E., Fadili, J., Breakspear, M., Salvador, R., Suckling, J., and Brammer, M. (2003). "Wavelets and statistical analysis of functional magnetic resonance images of the human brain." *Statistical Methods in Medical Research*, **12**, 375–399.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., and Brammer, M. (2001). "Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains." *Human Brain Mapping*, **12**, 61–78.

Bullmore, E. T., Rabe-Hesketh, S., Morris, R. G., Williams, S. C. R., Gregory, L., Gray, J. A., and Brammer, M. J. (1996b). "Functional magnetic resonance image analysis of a large-scale neurocognitive network." *NeuroImage*, **4**, 16–33.

Buonocore, M. H. and Zhu, D. C. (2001). "Image-based ghost correction for interleaved EPI." *Magnetic Resonance in Medicine*, **45**, 96–108.

Buračas, G. T. and Boynton, G. M. (2002). "Efficient design of event-related fMRI experiments using m-sequences." *NeuroImage*, **16**, 801–813.

Burock, M. A. and Dale, A. M. (2000). "Estimation and detection of event-related fMRI signals with temporally correlated noise: A statistically efficient and unbiased approach." *Human Brain Mapping*, **11**, 249–269.

Calhoun, V. D., Adali, T., Hansen, L. K., Larsen, J., and Pekar, J. J. (2003a). "ICA of functional MRI data: An overview." *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, 281–288.

Calhoun, V., Adali, T., Kraut, M., and Pearlson, G. (2000). "A weighted least-squares algorithm for estimation and visualization of relative latencies in event-related functional MRI." *Magnetic Resonance in Medicine*, **44**, 947–954.

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001a). "A method for making group inferences from functional MRI data using independent component analysis." *Human Brain Mapping*, **14**, 140–151.

— (2001b). "Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms." *Human Brain Mapping*, **13**, 43–53.

Calhoun, V. D., Adali, T., Pearlson, G. D., van Zijl, P. C. M., and Pekar, J. J. (2002). "Independent component analysis of fMRI data in the complex domain." *Magnetic Resonance in Medicine*, **48**, 180–192.

Calhoun, V. D., Adali, T., Pekar, J. J., and Pearlson, G. D. (2003b). "Latency (in)sensitive ICA: Group independent component analysis of fMRI data in the temporal frequency domain." *NeuroImage*, **20**, 1661–1669.

Calhoun, V. D., Stevens, M. C., Pearlson, G. D., and Kiehl, K. A. (2004). "fMRI analysis with the general linear model: Removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms." *NeuroImage*, **22**, 252–257.

Caparelli, E. C., Tomasi, D., Arnold, S., Chang, L., and Ernst, T. (2003). "k-space based summary motion detection for functional magnetic resonance imaging." *NeuroImage*, **20**, 1411–1418.

Carew, J. D., Wahba, G., Xie, X., Nordheim, E. V., and Meyerand, M. E. (2003). "Optimal spline smoothing of fMRI time series by generalized cross-validation." *NeuroImage*, **18**, 950–961.

Carlson, N. R. (1981). *Physiology of Behavior*. 2nd ed. Boston: Allyn and Bacon, Inc.

Casey, B. J., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., Rosen, B. R., Truwitt, C. L., and Turski, P. A. (1998). "Reproducibility of fMRI results across four institutions using a spatial working memory task." *NeuroImage*, **8**, 249–261.

Chau, W. and McIntosh, A. R. (2005). "The Talairach coordinate of a point in the MNI space: How to interpret it." *NeuroImage*, **25**, 408–416.

Chen, F. (2004). "A two-stage method for approximate spatial inferences by combining site-specific analyses." Ph.D. thesis, Department of Statistics, Carnegie Mellon University.

Chen, N.-K., Egorova, S., Guttmann, C. R. G., and Panych, L. P. (2003a). "Functional MRI with variable echo time acquisition." *NeuroImage*, **20**, 2062–2070.

Chen, C.-C., Tyler, C. W., and Baseler, H. A. (2003b). "Statistical properties of BOLD magnetic resonance activity in the human brain." *NeuroImage*, **20**, 1096–1109.

Chuang, K.-H. and Chen, J.-H. (2001). "IMPACT: Image-based physiological artifacts estimation and correction technique for functional MRI." *Magnetic Resonance in Medicine*, **46**, 344–353.

Chui, C. (1992). *An Introduction to Wavelets*. London: Academic Press.

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

Clark, V. P. (2002). "Orthogonal polynomial regression for the detection of response variability in event-related fMRI." *NeuroImage*, **17**, 344–363.

Cox, R. W. and Jesmanowicz, A. (1999). "Real-time 3D image registration for functional MRI." *Magnetic Resonance in Medicine*, **42**, 1014–1018.

Cressie, N. and Whitford, H. J. (1986). "How to use the two sample t-test." *Biometrical Journal*, **28**, 131–148.

Crivello, F., Schormann, T., Tzourio-Mazoyer, N., Roland, P. E., Zilles, K., and Mazoyer, B. M. (2002). "Comparison of spatial normalization procedures and their impact on functional maps." *Human Brain Mappping*, **16**, 228–250.

da Rocha Amaral, S., Rabbani, S. R., and Caticha, N. (2004). "Multigrid priors for a Bayesian approach to fMRI." *NeuroImage*, **23**, 654–662.

Dale, A. M. (1999). "Optimal experimental design for event-related fMRI." *Human Brain Mapping*, **8**, 109–114.

Dale, A. M. and Buckner, R. L. (1997). "Selective averaging of rapidly presented individual trials using fMRI." *Human Brain Mapping*, **5**, 329–340.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Desco, M., Hernandez, J. A., Santos, A., and Brammer, M. (2001). "Multiresolution analysis in fMRI: Sensitivity and specificity in the detection of brain activation." *Human Brain Mapping*, **14**, 16–27.

Donoho, D. and Johnstone, I. (1994). "Ideal spatial adaptation by wavelet shrinkage." *Biometrika*, **81**, 425–455.

Eddy, W. F., Fitzgerald, M., and Noll, D. C. (1996). "Improved image registration by using Fourier interpolation." *Magnetic Resonance in Medicine*, **36**, 923–931.

Eddy, W. F. and Young, T. K. (2000). "Optimizing MR resampling." In *Handbook of Medical Image Processing and Analysis*, ed. I. Bankman. San Diego: Academic Press.

Edelstein, W. A., Glover, G. H., Hardy, C. J., and Redington, R. W. (1986). "The intrinsic signal-to-noise ratio in NMR imaging." *Magnetic Resonance in Medicine*, **3**, 604–618.

Edward, V., Windischberger, C., Cunnington, R., Erdler, M., Lanzenberger, R., Mayer, D., Endl, W., and Beisteiner, R. (2000). "Quantification of fMRI artifact reduction by a novel plaster cast head holder." *Human Brain Mapping*, **11**, 207–213.

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

Esposito, F., Scarabino, T., Hyvärinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., and Di Salle, F. (2005). "Independent component analysis of fMRI group studies by self-organizing clustering." *NeuroImage*, **25**, 193–205.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., and Peters, T. M. (1993). "3D statistical neuroanatomical models from 305 MRI volumes." *Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference*, **3**, 1813–1817.

Fadili, M. J. and Bullmore, E. T. (2002). "Wavelet-generalized least squares: A new BLU estimator of linear regression models with $1/f$ errors." *NeuroImage*, **15**, 217–232.

— (2004). "A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps." *NeuroImage*, **23**, 1112–1128.

Fadili, M. J., Ruan, S., Bloyet, D., and Mazoyer, B. (2000). "A multistep unsupervised fuzzy clustering analysis of fMRI time series." *Human Brain Mapping*, **10**, 160–178.

Filzmoser, P., Baumgartner, R., and Moser, E. (1999). "A hierarchical clustering method for analyzing functional MR images." *Magnetic Resonance Imaging*, **17**, 817–826.

Fisher, R. A. (1950). *Statistical Methods for Research Workers*. 11th ed. London: Oliver and Boyd.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). "Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold." *Magnetic Resonance in Medicine*, **33**, 636–647.

Fransson, P., Merboldt, K.-D., Petersson, K. M., Ingvar, M., and Frahm, J. (2002). "On the effects of spatial filtering – A comparative fMRI study of episodic memory encoding at high and low resolution." *NeuroImage*, **16**, 977–984.

Freire, L. and Mangin, J.-F. (2001). "Motion correction algorithms may create spurious brain activations in the absence of subject motion." *NeuroImage*, **14**, 709–722.

Friman, O., Borga, M., Lundberg, P., and Knutsson, H. (2002a). "Detection of neural activity in fMRI using maximum correlation modeling." *NeuroImage*, **15**, 386–395.

— (2002b). "Exploratory fMRI analysis by autocorrelation maximization." *NeuroImage*, **16**, 454–464.

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., and Knutsson, H. (2001). "Detection of neural activity in functional MRI using canonical correlation analysis." *Magnetic Resonance in Medicine*, **45**, 323–330.

Friman, O. and Westin, C.-F. (2005). "Resampling fMRI time series." *NeuroImage*, **25**, 859–867.

Friston, K. J. (1998). "Imaging neuroscience: Principles or maps?" *Proceedings of the National Academy of Sciences (USA)*, **95**, 796–802.

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

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. (1993). "Functional connectivity: The principal-component analysis of large (PET) data sets." *Journal of Cerebral Blood Flow and Metabolism*, **13**, 5–14.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, B. J., Williams, C. R., Frackowiak, R. S. J., and Turner, R. (1995). "Analysis of fMRI time-series revisited." *NeuroImage*, **2**, 45–53.

Friston, K. J., Holmes, A. P., and Worsley, K. J. (1999a). "How many subjects constitute a study?" *NeuroImage*, **10**, 1–5.

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

Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., and Poline, J.-B. (2000a). "To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis." *NeuroImage*, **12**, 196–208.

Friston, K. J. and Penny, W. (2003). "Posterior probability maps and SPMs." *NeuroImage*, **19**, 1240–1249.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). "Classical and Bayesian inference in neuroimaging: Theory." *NeuroImage*, **16**, 465–483.

Friston, K., Phillips, J., Chawla, D., and Büchel, C. (1999b). "Revealing interactions among brain systems with nonlinear PCA." *Human Brain Mapping*, **8**, 92–97.

— (2000b). "Nonlinear PCA: Characterizing interactions between modes of brain activity." *Philosophical Transactions of the Royal Society of London, Series B*, **355**, 135–146.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. (2005). "Mixed-effects and fMRI studies." *NeuroImage*, **24**, 244–252.

Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N. A., and Dale, A. M. (1999c). "Stochastic designs in event-related fMRI." *NeuroImage*, **10**, 607–619.

Gaillard, W. D., Grandin, C. B., and Xu, B. (2001). "Developmental aspects of pediatric fMRI: Considerations for image acquisition, analysis, and interpretation." *NeuroImage*, **13**, 239–249.

Genovese, C. R. (2000). "A Bayesian time-course model for functional magnetic resonance imaging (with discussion)." *Journal of the American Statistical Association*, **95**, 691–719.

Genovese, C. R., Lazar, N. A., and Nichols, T. E. (2002). "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *NeuroImage*, **15**, 870–878.

Genovese, C. R., Noll, D. C., and Eddy, W. F. (1997). "Estimating test-retest reliability in functional MR imaging. I: Statistical methodology." *Magnetic Resonance in Medicine*, **38**, 497–507.

Gibbons, R. D., Lazar, N. A., Bhaumik, D. K., Sclove, S. L., Chen, H. Y., Thulborn, K. R., Sweeney, J. A., Hur, K., and Patterson, D. (2004). "Estimation and classification of fMRI hemodynamic response patterns." *NeuroImage*, **22**, 804–814.

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

Glover, G. H. and Law, C. S. (2001). "Spiral-in/out BOLD fMRI for increased SNR and reduced susceptibility artifacts." *Magnetic Resonance in Medicine*, **46**, 515–522.

Glover, G. H., Li, T.-Q., and Ress, D. (2000). "Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR." *Magnetic Resonance in Medicine*, **44**, 162–167.

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D. L., Flaum, M., and Andreasen, N. C. (1998). "Functional MRI statistical software packages: A comparative analysis." *Human Brain Mapping*, **6**, 73–84.

Gonçalves, M. S. and Hall, D. A. (2003). "Connectivity analysis with structural equation modeling: An example of the effects of voxel selection." *NeuroImage*, **20**, 1455–1467.

Gonzalez Andino, S. L., Grave de Peralta Menendez, R., Thut, G., Spinelli, L., Blanke, O., Michel, C. M., Seeck, M., and Landis, T. (2000). "Measuring the complexity of time series: An application to neurophysiological signals." *Human Brain Mapping*, **11**, 46–57.

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

— (2001). "Bayesian spatiotemporal inference in functional magnetic resonance imaging." *Biometrics*, **57**, 554–562.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A., and Hansen, L. K. (1999). "On clustering fMRI time series." *NeuroImage*, **9**, 298–310.

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

Hajnal, J. V., Bydder, G. M., and Young, I. R. (1995). "FMRI: Does correlation imply activation?" *NMR in Biomedicine*, **8**, 97–100.

Hajnal, J. V., Myers, R., Oatridge, A., Schwieso, J. E., Young, I. R., and Bydder, G. M. (1994). "Artifacts due to stimulus correlated motion in functional imaging of the brain." *Magnetic Resonance in Medicine*, **31**, 283–291.

Hampson, M., Peterson, B. S., Skudlarski, P., Gatenby, J. C., and Gore, J. C. (2002). "Detection of functional connectivity using temporal correlations in MR images." *Human Brain Mapping*, **15**, 247–262.

Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses." *NeuroImage*, **21**, 1639–1651.

Hansen, L. K. and Larsen, J. (1996). "Unsupervised learning and generalization." *Proceedings of the IEEE International Conference on Neural Networks*, **1**, 25–30.

Hansen, L. K., Larsen, J., Nielsen, F. A., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., and Paulson, O. B. (1999). "Generalizable patterns in neuroimaging: How many principal components?" *NeuroImage*, **9**, 534–544.

Harms, M. P. and Melcher, J. R. (2003). "Detection and quantification of a wide range of fMRI temporal responses using a physiologically-motivated basis set." *Human Brain Mapping*, **20**, 168–183.

Harrison, L., Penny, W. D., and Friston, K. (2003). "Multivariate autoregressive modeling of fMRI time series." *NeuroImage*, **19**, 1477–1491.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hartvig, N. V. (2002). "A stochastic geometry model for functional magnetic resonance images." *Scandinavian Journal of Statistics*, **29**, 333–353.

Hartvig, N. V. and Jensen, J. L. (2000). "Spatial mixture modeling of fMRI data." *Human Brain Mapping*, **11**, 233–248.

Hashemi, R. H., Bradley, W. G. J., and Lisanti, C. J. (2004). *MRI: The Basics*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins.

Hayasaka, S. and Nichols, T. E. (2003). "Validating cluster size inference: Random field and permutation methods." *NeuroImage*, **20**, 2343–2356.

— (2004). "Combining voxel intensity and cluster extent with permutation test framework." *NeuroImage*, **23**, 54–63.

Hedges, L. V. (1992). "Meta-analysis." *Journal of Educational Statistics*, **17**, 279–296.

Henson, R. N. A., Price, C. J., Rugg, M. D., Turner, R., and Friston, K. J. (2002). "Detecting latency differences in event-related BOLD responses: Application to words versus non-words and initial versus repeated face presentations." *NeuroImage*, **15**, 83–97.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). "Validating the independent components of neuroimaging time series via clustering and visualization." *NeuroImage*, **22**, 1214–1222.

Hinshaw, W. S. and Lent, A. H. (1983). "An introduction to NMR imaging: From the Bloch equation to the imaging equation." *Proceedings of the IEEE*, **71**, 338–350.

Holmes, A. P., Blair, R. C., Watson, J. D. G., and Ford, I. (1996). "Nonparametric analysis of statistical images from functional mapping experiments." *Journal of Cerebral Blood Flow and Metabolism*, **16**, 7–22.

Horwitz, B. (2003). "The elusive concept of brain connectivity." *NeuroImage*, **19**, 466–470.

Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, **24**, 417–441, 498–520.

— (1936). "Relations between two sets of variates." *Biometrika*, **28**, 321–377.

Hu, X., Le, T. H., Parrish, T., and Erhard, P. (1995). "Retrospective estimation and correction of physiological fluctuation in functional MRI." *Magnetic Resonance in Medicine*, **34**, 201–212.

Hu, D., Yan, L., Liu, Y., Zhou, Z., Friston, K. J., Tan, C., and Wu, D. (2005). "Unified SPM-ICA for fMRI analysis." *NeuroImage*, **25**, 746–755.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates, Inc.

Hyvärinen, A. and Oja, E. (2000). "Independent component analysis: Algorithms and applications." *Neural Networks*, **13**, 411–430.

Jernigan, T. L., Gamst, A. C., Fennema-Notestine, C., and Ostergaard, A. L. (2003). "More "mapping" in brain mapping: Statistical comparison of effects." *Human Brain Mapping*, **19**, 90–95.

Jessen, F., Manka, C., Scheef, L., Granath, D.-O., Schild, H. H., and Heun, R. (2002). "Novelty detection and repetition suppression in a passive picture viewing task: A possible approach for the evaluation of neuropsychiatric disorders." *Human Brain Mapping*, **17**, 230-236.

Jezzard, P. and Clare, S. (2001). "Principles of nuclear magnetic resonance and MRI." In *Functional MRI: An Introduction to Methods*, eds. P. Jezzard, P. M. Matthews, and S. M. Smith. Oxford: Oxford University Press.

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd ed. Berlin: Springer.

Jones, R. A., Brookes, J. A., and Moonen, C. T. W. (2001). "Ultra-fast fMRI." In *Functional MRI: An Introduction to Methods*, eds. P. Jezzard, P. M. Matthews, and S. M. Smith. Oxford: Oxford University Press.

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

Kershaw, J., Ardekani, B. A., and Kanno, I. (1999). "Application of Bayesian inference to fMRI data analysis." *IEEE Transactions on Medical Imaging*, **18**, 1138–1153.

Kershaw, J., Kashikura, K., Zhang, X., Abe, S., and Kanno, I. (2001). "Bayesian technique for investigating linearity in event-related BOLD fMRI." *Magnetic Resonance in Medicine*, **45**, 1081–1094.

Kherif, F., Poline, J.-B., Flandin, G., Benali, H., Simon, O., Dehaene, S., and Worsley, K. J. (2002). "Multivariate model specification for fMRI data." *NeuroImage*, **16**, 1068–1083.

Kherif, F., Poline, J.-B., Mériaux, S., Benali, H., Flandin, G., and Brett, M. (2003). "Group analysis in functional neuroimaging: Selecting subjects using similarity measures." *NeuroImage*, **20**, 2197–2208.

Kiebel, S. and Friston, K. J. (2002). "Anatomically informed basis functions in multisubject studies." *Human Brain Mapping*, **16**, 36–46.

Kiebel, S. J., Goebel, R., and Friston, K. J. (2000). "Anatomically informed basis functions." *NeuroImage*, **11**, 656–667.

Kiehl, K. A. and Liddle, P. F. (2003). "Reproducibility of the hemodynamic response to auditory oddball stimuli: A six-week test-retest study." *Human Brain Mapping*, **18**, 42–52.

Kjems, U., Hansen, L. K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., and Strother, S. C. (2002). "The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves." *NeuroImage*, **15**, 772–786.

Kochunov, P. V., Lancaster, J. L., and Fox, P. T. (1999). "Accurate high-speed spatial normalization using an octree method." *NeuroImage*, **10**, 724–737.

Kochunov, P., Lancaster, J., Thompson, P., Boyer, A., Hardies, J., and Fox, P. (2000). "Evaluation of octree regional spatial normalization method for regional anatomical matching." *Human Brain Mapping*, **11**, 193–206.

Konishi, S., Donaldson, D. I., and Buckner, R. L. (2001). "Transient activation during block transition." *NeuroImage*, **13**, 364–374.

Kruggel, F. and von Cramon, D. Y. (1999). "Modeling the hemodynamic response in single-trial functional MRI experiments." *Magnetic Resonance in Medicine*, **42**, 787–797.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner,

R., Cheng, H. M., Brady, T. J., and Rosen, B. R. (1992). "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation." *Proceedings of the National Academy of Sciences (USA)*, **89**, 5675–5679.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., and Strother, S. (2003). "The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics." *NeuroImage*, **18**, 10–27.

Lahaye, P.-J., Poline, J.-B., Flandin, G., Dodel, S., and Garnero, L. (2003). "Functional connectivity: Studying nonlinear, delayed interactions between BOLD signals." *NeuroImage*, **20**, 962–974.

Laird, A. R., Rogers, B. P., and Meyerand, M. E. (2004). "Comparison of Fourier and wavelet resampling methods." *Magnetic Resonance in Medicine*, **51**, 418–422.

Lancaster, H. O. (1961). "The combination of probabilities: An application of orthonormal functions." *Australian Journal of Statistics*, **3**, 20–33.

Lange, N. (2003). "What can modern statistics offer imaging neuroscience?" *Statistical Methods in Medical Research*, **12**, 447–469.

Lange, N. and Zeger, S. L. (1997). "Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion." *Applied Statistics*, **46**, 1–29.

Lazar, N. A., Eddy, W. F., Genovese, C. R., and Welling, J. (2001). "Statistical issues in fMRI for brain imaging." *International Statistical Review*, **69**, 105–127.

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

Levin, D. N. and Uftring, S. J. (2001). "Detecting brain activation in fMRI data without prior knowledge of mental event timing." *NeuroImage*, **13**, 153–160.

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

Littell, R. C. and Folks, J. L. (1973). "Asymptotic optimality of Fisher's method of combining independent tests II." *Journal of the American Statistical Association*, **68**, 193–194.

Liu, T. T. (2004). "Efficiency, power, and entropy in event-related fMRI with multiple trial types – Part II: Design of experiments." *NeuroImage*, **21**, 401–413.

Liu, T. T. and Frank, L. R. (2004). "Efficiency, power, and entropy in event-related fMRI with multiple trial types – Part I: Theory." *NeuroImage*, **21**, 387–400.

Liu, T. T., Frank, L. R., Wong, E. C., and Buxton, R. B. (2001). "Detection power, estimation efficiency, and predictability in event-related fMRI." *NeuroImage*, **13**, 759–773.

Liu, J. Z., Zhang, L., Brown, R. W., and Yue, G. H. (2004). "Reproducibility of fMRI at 1.5 T in a strictly controlled motor task." *Magnetic Resonance in Medicine*, **52**, 751–760.

Locascio, J. J., Jennings, P. J., Moore, C. I., and Corkin, S. (1997). "Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging." *Human Brain Mapping*, **5**, 168–193.

Logan, B. R. and Rowe, D. B. (2004). "An evaluation of thresholding techniques in fMRI analysis." *NeuroImage*, **22**, 95–108.

Long, C., Brown, E. N., Manoach, D., and Solo, V. (2004). "Spatiotemporal wavelet analysis for functional MRI." *NeuroImage*, **23**, 500–516.

Lu, Y., Jiang, T., and Zang, Y. (2003). "Region growing method for the analysis of functional MRI data." *NeuroImage*, **20**, 455–465.

Luo, W.-L. and Nichols, T. E. (2003). "Diagnosis and exploration of massively univariate neuroimaging models." *NeuroImage*, **19**, 1014–1032.

Maccotta, L., Zaks, J. M., and Buckner, R. L. (2001). "Rapid self-paced event-related functional MRI: Feasibility and implications of stimulus- versus response-locked timing." *NeuroImage*, **14**, 1105–1121.

Magnotta, V. A., Bockholt, H. J., Johnson, H. J., Christensen, G. E., and Andreasen, N. C. (2003). "Subcortical, cerebellar, and magnetic resonance based consistent brain image registration." *NeuroImage*, **19**, 233–245.

Maitra, R., Roys, S. R., and Gullapalli, R. P. (2002). "Test-retest reliability estimation of functional MRI data." *Magnetic Resonance in Medicine*, **48**, 62–70.

Mansfield, P. (1977). "Multi-planar image formation using NMR spin echoes." *Journal of Physical Chemistry*, **10**, L55–58.

Marchini, J. and Presanis, A. (2004). "Comparing methods of analyzing fMRI statistical parametric maps." *NeuroImage*, **22**, 1203–1213.

Marchini, J. L. and Ripley, B. D. (2000). "A new statistical approach to detecting significant activation in functional MRI." *NeuroImage*, **12**, 366–380.

Marchini, J. L. and Smith, S. M. (2003). "On bias in the estimation of autocorrelations for fMRI voxel time-series analysis." *NeuroImage*, **18**, 83–90.

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J.-B. (2003). "Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information." *Human Brain Mapping*, **19**, 1–17.

Marx, E., Deutschländer, A., Stephan, T., Dieterich, M., Wiesmann, M., and Brandt, T. (2004). "Eyes open and eyes closed as rest conditions: Impact on brain activation patterns." *NeuroImage*, **21**, 1818–1824.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). "A probabilistic atlas of the human brain: Theory and rationale for its development." *NeuroImage*, **2**, 89–101.

McGonigle, D. J., Howseman, A. M., Athwal, B. S., Friston, K. J., Frackowiak, R. S. J., and Holmes, A. P. (2000). "Variability in fMRI: An examination of intersession differences." *NeuroImage*, **11**, 708–734.

McIntosh, A. R., Chau, W. K., and Protzner, A. B. (2004). "Spatiotemporal analysis of event-related fMRI data using partial least squares." *NeuroImage*, **23**, 764–775.

McKeown, M. J. (2000). "Detection of consistently task-related activations in fMRI data with hybrid independent component analysis." *NeuroImage*, **11**, 24–35.

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

McNamee, R. L. and Eddy, W. F. (2001). "Visual analysis of variance: A tool for quantitative assessment of fMRI data processing and analysis." *Magnetic Resonance in Imaging*, **46**, 1202–1208.

— (2005). "A simple method for removing cardiac and respiratory effects from fMRI data." Unpublished manuscript.

McNamee, R. L. and Lazar, N. A. (2004). "Assessing the sensitivity of fMRI group maps." *NeuroImage*, **22**, 920–931.

Mechelli, A., Penny, W. D., Price, C. J., Gitelman, D. R., and Friston, K. J. (2002). "Effective connectivity and intersubject variability: Using a multisubject network to test differences and commonalities." *NeuroImage*, **17**, 1459–1469.

Miller, R. G. (1986). *Beyond ANOVA, Basics of Applied Statistics*. New York: John Wiley & Sons.

Moelker, A. and Pattynama, P. M. T. (2003). "Acoustic noise concerns in functional magnetic resonance imaging." *Human Brain Mapping*, **20**, 123–141.

Morgan, V. L., Pickens, D. R., Hartmann, S. L., and Price, R. R. (2001). "Comparison of functional MRI image realignment tools using a computer-generated phantom." *Magnetic Resonance in Medicine*, **46**, 510–514.

Morrison, D. F. (1978). *Multivariate Statistical Methods*. 2nd ed. Singapore: McGraw-Hill Book Company.

Müller, K., Lohmann, G., Bosch, V., and von Cramon, D. Y. (2001). "On multivariate spectral analysis of fMRI time series." *NeuroImage*, **14**, 347–356.

Nandy, R. and Cordes, D. (2003a). "Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data." *Magnetic Resonance in Medicine*, **50**, 354–365.

Nandy, R. R. and Cordes, D. (2003b). "Novel ROC-type method for testing the efficiency of multivariate statistical methods in fMRI." *Magnetic Resonance in Medicine*, **49**, 1152–1162.

Nandy, R. and Cordes, D. (2004a). "Improving the spatial specificity of canonical correlation analyis in fMRI." *Magnetic Resonance in Medicine*, **52**, 947–952.

Nandy, R. R. and Cordes, D. (2004b). "New approaches to receiver operating characteristic methods in functional magnetic resonance imaging with real data using repeated trials." *Magnetic Resonance in Medicine*, **52**, 1424–1431.

Neumann, J. and Lohmann, G. (2003). "Bayesian second-level analysis of functional magnetic resonance images." *NeuroImage*, **20**, 1346–1355.

Neumann, J., Lohmann, G., Zysset, S., and von Cramon, D. Y. (2003). "Within-subject variability of BOLD response dynamics." *NeuroImage*, **19**, 784–796.

Newman, S. D., Twieg, D. B., and Carpenter, P. A. (2001). "Baseline conditions and subtractive logic in neuroimaging." *Human Brain Mapping*, **14**, 228–235.

Ngan, S.-C. and Hu, X. (1999). "Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity." *Magnetic Resonance in Medicine*, **41**, 939–946.

Nichols, T. and Hayasaka, S. (2003). "Controlling the familywise error rate in functional neuroimaging: A comparative review." *Statistical Methods in Medical Research*, **12**, 419–446.

Nichols, T. E. and Holmes, A. P. (2001). "Nonparametric permutation tests for functional neuroimaging: A primer with examples." *Human Brain Mapping*, **15**, 1–25.

Noll, D. C., Genovese, C. R., Nystrom, L. E., Vazquez, A. L., Forman, S. D., Eddy, W. F., and Cohen, J. D. (1997). "Estimating test-retest reliability in functional MR imaging. II: Application to motor and cognitive activation studies." *Magnetic Resonance in Medicine*, **38**, 508–517.

Oakes, T. R., Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., and Davidson, R. J. (2005). "Comparison of fMRI motion correction software tools." *NeuroImage*, **28**, 529–543.

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990). "Brain magnetic resonance imaging with contrast dependent on blood oxygenation." *Proceedings of the National Academy of Sciences (USA)*, **87**, 9868–9872.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and 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.

Ogden, R. T. and Parzen, E. (1996). "Change-point approach to data analytic wavelet thresholding." *Statistics and Computing*, **6**, 93–99.

Park, C., Lazar, N. A., Ahn, J., and Sornborger, A. (2007). "Do different parts of the brain have the same dependence structure? A multiscale analysis of the temporal and spatial characteristics of resting fMRI data." Unpublished manuscript.

Pauling, L. and Coryell, C. (1936). "The magnetic properties and structure of hemoglobin, oxyhemoglobin, and carbon monoxyhemoglobin." *Proceedings of the National Academy of Sciences (USA)*, **22**, 210–216.

Pavlicová, M., Cressie, N., and Santner, T. J. (2006). "Testing for activation in data from FMRI experiments." *Journal of Data Science*, **4**, 275–289.

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

Penny, W., Kiebel, S., and Friston, K. (2003). "Variational Bayesian inference for fMRI time series." *NeuroImage*, **19**, 727–741.

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

Pounds, S. and Cheng, C. (2004). "Improving false discovery rate estimation." *Bioinformatics*, **20**, 1737–1745.

Pounds, S. and Morris, S. W. (2003). "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values." *Bioinformatics*, **19**, 1236–1242.

Preston, A. R., Thomason, M. E., Ochsner, K. N., Cooper, J. C., and Glover, G. H. (2004). "Comparison of spiral-in/out and spiral-out BOLD fMRI at 1.5 and 3 T." *NeuroImage*, **21**, 291–301.

Purdon, P. L., Solo, V., Weisskoff, R. M., and Brown, E. N. (2001). "Locally regularized spatiotemporal modeling and model comparison for functional MRI." *NeuroImage*, **14**, 912–923.

Purdon, P. L. and Weisskoff, R. M. (1998). "Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI." *Human Brain Mapping*, **6**, 239–249.

Quenouille, M. (1949). "Approximate tests of correlation in time-series." *Journal of the Royal Statistical Society, Series B*, **11**, 68–84.

Rajapske, J. C., Kruggel, F., Maisog, J. M., and von Cramon, D. Y. (1998). "Modeling hemodynamic response for analysis of functional MRI time-series." *Human Brain Mapping*, **6**, 283–300.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Berlin: Springer.

Razavi, M., Grabowski, T. J., Vispoel, W. P., Monahan, P., Mehta, S., Eaton, B., and Bolinger, L. (2003). "Model assessment and model building in fMRI." *Human Brain Mapping*, **20**, 227–238.

Robert, P. and Escoufier, Y. (1976). "A unifying tool for linear multivariate statistical methods: The RV-coefficient." *Applied Statistics*, **25**, 257–265.

Roy, A., Bhaumik, D. K., Gibbons, R. D., Lazar, N. A., Sweeney, J. A., Aryal, S., Kapur, K., and Patterson, D. (2005). "Estimation and classification of BOLD responses over multiple trials." Unpublished manuscript.

Ruttimann, U. E., Unser, M., Rawlings, R. R., Rio, D., Ramsey, N. F., Mattay, V. S., Hommer, D. W., Frank, J. A., and Weinberger, D. R. (1998).

"Statistical analysis of functional MRI data in the wavelet domain." *IEEE Transactions on Medical Imaging*, **17**, 142–154.

Saad, Z. S., DeYoe, E. A., and Ropella, K. M. (2003a). "Estimation of fMRI response delays." *NeuroImage*, **18**, 494–504.

Saad, Z. S., Ropella, K. M., Cox, R. W., and DeYoe, E. A. (2001). "Analysis and use of fMRI response delays." *Human Brain Mapping*, **13**, 74–93.

Saad, Z. S., Ropella, K. M., DeYoe, E. A., and Bandettini, P. A. (2003b). "The spatial extent of the BOLD response." *NeuroImage*, **19**, 132–144.

Salli, E., Korvenoja, A., Visa, A., Katila, T., and Aronen, H. J. (2001). "Reproducibility of fMRI: Effect of the use of contextual information." *NeuroImage*, **13**, 459–471.

Savoy, R. (2001). "The scanner as a psychophysical laboratory." In *Functional MRI: An Introduction to Methods*, eds. P. Jezzard, P. M. Matthews, and S. M. Smith. Oxford: Oxford University Press.

Scarff, C. J., Dort, J. C., Eggermont, J. J., and Goodyear, B. G. (2004). "The effect of MR scanner noise on auditory cortex activity using fMRI." *Human Brain Mapping*, **22**, 341–349.

Schmithorst, V. J. and Holland, S. K. (2004). "Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data." *Journal of Magnetic Resonance Imaging*, **19**, 365–368.

Seto, E., Sela, G., McIlroy, W. E., Black, S. E., Staines, W. R., Bronskill, M. J., McIntosh, A. R., and Graham, S. J. (2001). "Quantifying head motion associated with motor tasks used in fMRI." *NeuroImage*, **14**, 284–297.

Shaw, M. E., Strother, S. C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., and Egan, G. (2003). "Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics." *NeuroImage*, **19**, 988–1001.

Skudlarski, P., Constable, R. T., and Gore, J. C. (1999). "ROC analysis of statistical methods used in functional MRI: Individual subjects." *NeuroImage*, **9**, 311–329.

Smith, S. M. (2001). "Preparing fMRI data for statistical analysis." In *Functional MRI: An Introduction to Methods*, eds. P. Jezzard, P. M. Matthews, and S. M. Smith. Oxford: Oxford University Press.

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. M., and McGonigle, D. J. (2005). "Variability in fMRI: A re-examination of inter-session differences." *Human Brain Mapping*, **24**, 248–257.

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian variable selection with application to functional magnetic resonance imaging." *Journal of the American Statistical Association*, **102**, 417–431.

Smith, M., Pütz, B., Auer, D., and Fahrmeir, L. (2003). "Assessing brain activity through spatial Bayesian variable selection." *NeuroImage*, **20**, 802–815.

Stanberry, L., Nandy, R., and Cordes, D. (2003). "Cluster analysis of fMRI data using dendrogram sharpening." *Human Brain Mapping*, **20**, 201–219.

Stephan, T., Marx, E., Brückmann, H., Brandt, T., and Dieterich, M. (2002). "Lid closure mimics head movement in fMRI." *NeuroImage*, **16**, 1156–1158.

Stone, J. V., Porrill, J., Porter, N. R., and Wilkinson, I. D. (2002). "Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions." *NeuroImage*, **15**, 407–421.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). "The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework." *NeuroImage*, **15**, 747–771.

Sun, F. T., Miller, L. M., and D'Esposito, M. (2004). "Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data." *NeuroImage*, **21**, 647–658.

Svensén, M., Kruggel, F., and Benali, H. (2002). "ICA of fMRI group study data." *NeuroImage*, **16**, 551–563.

Talairach, J. and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. New York: Thieme.

Tanabe, J., Miller, D., Tregellas, J., Freedman, R., and Meyer, F. G. (2002). "Comparison of detrending methods for optimal fMRI preprocessing." *NeuroImage*, **15**, 902–907.

Thirion, B. and Faugeras, O. (2003). "Dynamical components analysis of fMRI data through kernel PCA." *NeuroImage*, **20**, 34–49.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., and Poline, J.-B. (2007). "Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses." *NeuroImage*, **35**, 105–120.

Thulborn, K. R. (1999). "Visual feedback to stabilize head position for fMRI." *Magnetic Resonance in Medicine*, **41**, 1039–1043.

Thulborn, K. R., Waterton, J. C., Matthews, P. M., and Radda, G. K. (1982). "Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field." *Biochimica et Biophysica Acta*, **714**, 265–270.

Tippett, L. H. C. (1931). *The Method of Statistics*. 1st ed. London: Williams and Norgate.

Toga, A. W., ed. (1998). *Brain Warping*. San Diego: Academic Press.

Tukey, J. (1958). "Bias and confidence in not quite large samples." *Annals of Mathematical Statistics*, **29**, 614.

Turkheimer, F., Pettigrew, K., Sokoloff, L., Smith, C. B., and Schmidt, K. (2000). "Selection of an adaptive test statistic for use with multiple comparison analyses of neuroimaging data." *NeuroImage*, **12**, 219–229.

Turkheimer, F. E., Smith, C. B., and Schmidt, K. (2001). "Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data." *NeuroImage*, **13**, 920–930.

van de Ven, V. G., Formisano, E., Prvulovic, D., Roeder, C. H., and Linden, D. E. J. (2004). "Functional connectivity as revealed by spatial independent

component analysis of fMRI measurements during rest." *Human Brain Mapping*, **22**, 165–178.

Van De Ville, D., Blu, T., and Unser, M. (2004). "Integrated wavelet processing and spatial statistical testing of fMRI data." *NeuroImage*, **23**, 1472–1485.

Veltman, D. J., Mechelli, A., Friston, K. J., and Price, C. J. (2002). "The importance of distributed sampling in blocked functional magnetic resonance imaging designs." *NeuroImage*, **17**, 1203–1206.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.

Visscher, K. M., Miezin, F. M., Kelly, J. E., Buckner, R. L., Donaldson, D. I., McAvoy, M. P., Bhalodia, V. M., and Petersen, S. E. (2003). "Mixed block/event-related designs separate transient and sustained activity in fMRI." *NeuroImage*, **19**, 1694–1708.

Viviani, R., Grön, G., and Spitzer, M. (2005). "Functional principal component analysis of fMRI data." *Human Brain Mapping*, **24**, 109–129.

Wager, T. D. and Nichols, T. E. (2003). "Optimization of experimental design in fMRI: A general framework using a genetic algorithm." *NeuroImage*, **18**, 293–309.

Ward, H. A., Riederer, S. J., Grimm, R. C., Ehman, R. L., Felmlee, J. P., and Jack, C. R. (2000). "Prospective multiaxial motion correction for fMRI." *Magnetic Resonance in Medicine*, **43**, 459–469.

Welch, B. L. (1937). "The significance of the difference between two means when the population variances are unequal." *Biometrika*, **29**, 350–362.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M., and Andreasen, N. C. (2001). "Anatomic and functional variability: The effects of filter size in group fMRI data analysis." *NeuroImage*, **13**, 577–588.

Wicker, B. and Fonlupt, P. (2003). "Generalized least-squares method applied to fMRI time series with empirically determined correlation matrix." *NeuroImage*, **18**, 588–594.

Wilkinson, B. (1951). "A statistical consideration in psychological research." *Psychological Bulletin*, **48**, 156–158.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004a). "Multilevel linear modelling for FMRI group analysis using Bayesian inference." *NeuroImage*, **21**, 1732–1747.

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

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004c). "Fully Bayesian spatio-temporal modeling of FMRI data." *IEEE Transactions on Medical Imaging*, **23**, 213–231.

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). "Temporal autocorrelation in univariate linear modeling of FMRI data." *NeuroImage*, **14**.

Worsley, K. J. (1994). "Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields." *Advances in Applied Probability*, **26**, 13–42.

— (1996). "The geometry of random images." *Chance*, **9**, 27–40.

— (2000). "Comment on "A Bayesian time-course model for functional magnetic resonance imaging data"." *Journal of the American Statistical Association*, **95**, 711–716.

— (2003). "Detecting activation in fMRI data." *Statistical Methods in Medical Research*, **12**, 401–418.

Worsley, K. J. and Friston, K. J. (1995). "Analysis of fMRI time-series revisited– Again." *NeuroImage*, **5**, 173–181.

— (2000). "A test for a conjunction." *Statistics and Probability Letters*, **47**, 135–140.

Zhuang, J., LaConte, S., Peltier, S., Zhang, K., and Hu, X. (2005). "Connectivity exploration with structural equation modeling: An fMRI study of bimanual motor coordination." *NeuroImage*, **25**, 462–470.

Index

Morphometrics with R

Julien Claude

This richly illustrated book describes the use of interactive and dynamic graphics as part of multidimensional data analysis. Chapters include clustering, Quantifying shape and size variation is essential in evolutionary biology and in many other disciplines. Since the "morphometric revolution of the 90s," an increasing number of publications in applied and theoretical morphometrics emerged in the new discipline of statistical shape analysis. It explains how to use R for morphometrics and provides a series of examples of codes and displays covering approaches ranging from traditional morphometrics to modern statistical shape analysis such as the analysis of landmark data, Thin Plate Splines, and Fourier analysis of outlines.

2007, Approx. 330 pp Softcover ISBN 978-0-387-77789-4

Statistical Learning from a Regression Perspective

Richard A. Berk

This book considers statistical learning applications when interest centers on the conditional distribution of the response variable, given a set of predictors, and when it is important to characterize how the predictors are related to the response. Real applications are emphasized, especially those with practical implications. The material is written for graduate students in the social and life sciences and for researchers who want to apply statistical learning procedures to scientific and policy problems. Intuitive explanations and visual representations are prominent. All of the analyses included are done in R.

2008, Approx. 370 pp. Hardcover ISBN 978-0-387-77500-5

Applied Spatial Data Analysis with R

Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio

This book is divided into two basic parts, the first presenting R packages, functions, classes and methods for handling spatial data. This part is of interest to users who need to access and visualise spatial data. The second part showcases more specialised kinds of spatial data analysis, including spatial point pattern analysis, interpolation and geostatistics, areal data analysis and disease mapping. This book will be of interest to researchers who intend to use R to handle, visualise, and analyse spatial data. It will also be of interest to spatial data analysts who do not use R, but who are interested in practical aspects of implementing software for spatial data analysis.

2008, Approx. 410 pp. Softcover ISBN 978-0-387-78170-9