Chapter 7. Functional Image Generation Techniques: Single Subject Data

7.1. Introduction

Most of an fMRI researcher's time is spent on data analysis. There is a large amount of data. There is a large number of steps needed to process the data. There is a large number of different techniques for the production of functional activation maps, no one of which is clearly superior—several different analysis methods might be applied to one data set to make sure that the results aren't sensitive to a particular procedure.

Capsule Summary

The basic unit of an fMRI data set is an image time series (2D or 3D), combined with knowledge of the stimuli applied to the subject during the imaging run. At least 2 different stimuli must be used, since fMRI does not measure absolute neural activity. Only changes in the MRI signal can be detected, so to create maps of neural activation, stimuli that will evoke different neural responses must be used. This time series data itself does not comprise a functional image. The techniques for deriving functional activation maps from such data fall into two broad categories: pattern matching methods, and pattern hunting methods. Pattern matching methods presuppose some model for the MRI-measured response to neural activation. The chosen model is fit to the data using some form of regression analysis. The resulting activation map is the set of image voxels that fit the response model with some level of significance. Pattern hunting methods look for spatio-temporal components that explain significant parts of the data, without imposing à priori constraints on the measurable response. Before any analysis is carried out on the image time series, it is advisable to preprocess the data to minimize the effects of subject head motion and other sources of signal contamination.

Data: Signals (Good) and Noise (Bad)

In any detection and characterization problem, it is important to understand both the signal for which one is looking and the noise that is interfering with identification of the signal in the data. In this context, "noise" refers to any component of the data that is not related to the phenomenon of interest. For example, in fMRI, the motion of the brainstem with each heartbeat can be considered to be noise, since it strongly interferes with the detection of neural activation in that region. In a study that concentrated on quantifying brainstem movement, this "noise" would be the signal, and data changes due to neural activity would be the "noise" instead.

Data analysis in fMRI is handicapped by the lack of widely accepted models for either the signal or the noise. As discussed in Chapter 3, the detailed temporal features of the BOLD signal are still matters of controversy. The extent to which these are constant in any given voxel, and the extent to which they vary between voxels and between subjects for non-neural reasons (e.g., venule size distribution, blood vessel pliability, partial volume effects) are active areas of research. The noise in fMRI is principally caused by non-neural physiological effects such as the cardiac and respiratory cycles. Understanding the temporal and spatial distribution of the noise, and its impact on the various signal detection techniques, is also an area of investigation.

In most disciplines, signal detection is based on models of the signal and noise. When these models are completely lacking, ad hoc methods must be used. In fMRI, these models are not entirely lacking, but are still matters of controversy. The result is the proliferation of fMRI data analysis methods, each of which is designed—perhaps implicitly—around some signal+noise models. Each analysis method is also often designed with a particular type of stimulus paradigm in mind. The upshot is that more than one analysis technique can be applied to any given data set, but it requires some understanding of the underlying assumptions and goals to choose among the analysis methods. In many cases, more than one method is applied to a data set in order to see if the activation map is very sensitive to the choice of analysis tool. Fortunately, well-designed and well-executed fMRI experiments produce fairly robust data that give similar results from a variety of analytical methods.

7.2. Comparison to PET Activation Experiments

^{15}O-labeled H_2O PET data are similar to fMRI data in that the signal contrast in both methods is due to changes in blood flow caused by changes in neuronal activity (Chapter 3). PET images have lower spatial and temporal resolution than even coarse fMRI images, which affects the data analysis in several ways. The principal source of noise in PET data is the random nature of radioactive counts, coupled with the limited dose of radiation that is compatible with subject safety. This noise variance is approximately uniform across the brain. To lower the noise level, PET images are usually spatially smoothed in the reconstruction software (10–20 mm of smoothing is commonly reported). As a result, the image noise is highly correlated between voxels, which implies that the statistical tests for activation are not even approximately independent between voxels. Coupled with the fact that the number of PET images per subject is usually small (10–20), this means that effective statistical tests must take into account the spatial correlation in the noise. Techniques that have been developed to do this include the application of principal components methods [Friston, Strother] and of the theory of correlated Gaussian random fields [Worsley].

The noise in fMRI is quite different from the noise in PET. Its variance is quite nonuniform and depends strongly on the tissue composition of the voxel—for example, large blood vessels increase the variance due to the stronger effects of blood pulsations with the cardiac cycle. However, the noise is not strongly correlated between voxels. Since fMRI data are gathered much more rapidly than PET data, there are temporal correlations in the fMRI time series that are not present in the PET images. These facts mean that fMRI data should be processed somewhat differently than PET data. The lack of spatial correlation in the noise means that the effective spatial resolution of fMRI-derived activation maps can be much higher than PET-derived maps. The strong spatial nonuniformity of the noise variance means that activation detection is harder in some brain regions than in others. The temporal correlations in the noise means that it is hard to estimate accurately the statistical significance of fMRI-derived activation maps.

Both types of experiments manipulate the stimuli presented to the subject and expect to see a corresponding change in the image voxel values. It is possible to apply PET data analysis methods to fMRI data, and this is sometimes done; however, the differences between PET and fMRI signal/noise properties has led to the further development of neuroimage analysis methods. Characteristics that make fMRI data easier to analyze than PET data include:

- the much larger number of time samples available for each stimulus condition and each subject;
- the smaller correlation between noise values in spatially separated voxels;
- echo-planar imaging methods can provide whole brain images that repeat faster than the hemodynamic response time.

Characteristics that make the interpretation of fMRI data more difficult than PET data include:

- higher spatial resolution means that small subject movements or small changes in the image will have proportionally larger effects on the signals measured from individual voxels (e.g., with spatial resolution of 10 mm, a 1 mm movement is only 10% of a voxel, but with spatial resolution of 2 mm, the same movement is 50% of a voxel);

- the BOLD effect is a complex interaction of NMR physics, physiology, and microscopic anatomy, which means that comparisons of the magnitudes of signal changes between voxels—intra- or inter-subject—are difficult to justify (not that this stops people from doing just that); this problem is one motivation for using *arterial spin labeling* methods for fMRI;
- temporal correlations in the time series data—the fact that the noise is not "*white*"—make it hard to estimate how many degrees of freedom are actually present in a voxel time series; the temporal correlations are caused by *physiological noise*, which is quite spatially variable, further compounding the difficulty.

7.3. Practical Issues of Computation and Data Management

A typical echo-planar fMRI imaging run acquires 10 images per second for 5 minutes, yielding 3000 images. If the images are reconstructed onto a 64×64 matrix with each voxel stored in 2 bytes, then this image data takes up 24 Mbytes of disk space—in essence, this is a single fMRI datum. Ten such imaging runs per scanning session is a typical number, so that about ¼ Gbyte of data is generated in an hour. (128×128 images are becoming more common, which take up four times as much space: 1 Gbyte per hour.) This is the basic experimental session on a single subject. A typical small study may have 20 scanning sessions. It is important to develop a systematic plan for archiving the raw data, for documenting its acquisition, and for keeping track of the processing steps applied to it. Absent such a plan, investigators quickly come to feel that they are "drowning in data." Significant amounts of disk storage, tapes or other backup media, memory, and processing power are needed to process fMRI data. Fortunately, these things have become relatively inexpensive.

fMRI Software Packages

Several software packages for the statistical analysis and display of neuroimaging data sets are available. Some are commercial systems and some are distributed freely via the Internet. These packages were recently compared [Gold 1998] (as usual, software reviews are mostly out-of-date by the time they appear in print). The major analysis packages are:

• *AFNI*	from the Medical College of Wisconsin.	[freeware]
• *FIASCO*	from Carnegie Mellon University.	[freeware]
• *MED-X*	from Sensor Systems, Inc.	[commercial]
• *SPM*	from the Wellcome Neurological Institute.	[freeware]
• *Stimulate*	from the University of Minnesota.	[freeware]

Some software that performs specialized functions useful for fMRI include:

• *AIR*	from UCLA (image registration)	[freeware]
• ????	from Harvard (cortical flattening)	[freeware]
• ????	from Washington University (cortical flattening)	[freeware]

Most of these packages are designed to run under Unix (some including Linux), since only recently did personal computers acquire the power needed for processing fMRI data. It is also possible to use general purpose image manipulation software such as *Analyze* (Mayo Clinic) for some parts of the data processing. The usual statistical packages (e.g., *SPSS*) are not adequate for processing such large amounts of data, but can be very useful for analysis of a few selected data components (e.g., data averaged over regions of interest—ROIs).

Our personal preference is for the *AFNI* package, but that is largely because it originates from the authors' institution, is in daily use there, and is written by one of authors (RWC). *AFNI* is described briefly in Appendix D. *SPM* is very widely used: it was one of the first available software tools for neuroimage data analysis, and it has accumulated a large number of analysis options over the years.

The field of fMRI data analysis is changing rapidly, and each software system only incorporates a portion of all the published techniques. For this reason, most sites that take fMRI seriously need to employ at least one scientific computer programmer to provide custom analysis options, even if one of the packages above can do the majority of the needed work.

Choosing an Analysis Method

The number of fMRI data analysis methods used in the literature is large and ever increasing. In part, this state of affairs is due to the large number of different experiment types possible with fMRI and human subjects. One analysis paradigm cannot possibly fit the myriad classes of data that neuroscientists can generate. However, many proposed analysis techniques overlap in their applicability. Large scale systematic studies and comparisons of fMRI data analysis methods do not yet exist. Instead, each author of a new methods points out the shortcomings of the old techniques that his approach overcomes, and adduces a data set or two to back his claims. The only practical approach that an fMRI research can take is to choose a couple of data analysis methods that are applicable to his data, use them both in parallel, and see if the results differ significantly.

7.4. Preprocessing the Image Time Series

fMRI data sets are large and are sensitive to many effects besides the neural activation as measured through the BOLD response. Some of the possible artifacts were described in Chapters 2 and 3. The purpose of the preprocessing steps is to detect and/or minimize these effects, so as to improve the detectability of neurally induced image intensity changes.

Image Reconstruction

Image reconstruction is the first step in data processing. Normally, this function is provided by the scanner manufacturer (the software tools above do *not* include reconstruction), but there may be some options that affect the image quality for fMRI purposes. One example is the correction of image distortions that arise from large scale magnetic field inhomogeneities. A special MRI pulse sequence can be used to gather a map of the magnetic field, and then this information can be used during the reconstruction of the fMRI time series. Another example is the final reconstruction matrix. The raw data can be reconstructed onto a finer grid than the actual spatial resolution of the data (e.g., reconstruction of 64×64 data onto a 128×128 grid). This usually improves the visual appearance of the images, but does not add actual resolution. Instead, it will introduce spatial correlation between the noise in each voxel. If this is done, the selection of the statistical threshold for the activation detection should be altered appropriately (cf. §7.6).

At most sites, you will have to live with whatever image reconstruction the scanner manufacturer provides. It is worth looking into the available options, especially with the guidance of an MRI physicist.

Looking at the Images

Simply scanning through the image time series is a good idea, although it can be time consuming. MRI scanner hardware is very complex, and the malfunctioning of any component can significantly compromise image quality in ways that may not be detected by automatic algorithms. For example, in one scanning session at the Medical College of Wisconsin, the gradient power supply subsystem lost part of its capacity halfway through an imaging run. The weaker currents to one gradient coil produced weaker magnetic field gradients, which resulting in the images being distorted by about 40% in the anterior-posterior direction—that is, the subject's head suddenly appeared to stretch out about 4 inches! The data from this scanning session had to be discarded.

It is impossible to anticipate all such possibilities in computer software, so it is important that the images be reviewed by a human. Viewing the images in "cine mode" (like a movie, scanning through time) is relatively quick and will often highlight obvious problems, including subject head movement.

Discarding the First Few Images

If TR is less than about 5·T1, then the first few images in the fMRI time series will be significantly different and brighter than the later images. This is due to the time it takes the magnetization to reach an equilibrium state (to become *saturated*) when it is being subjected to a train of periodic RF excitations. These bright images should not be used in the functional activation data analysis, since the signal changes caused by neural activation are much smaller than the brightness changes during this "warmup" period. For this reason, it is usual to start an fMRI scanning run with a "rest" period rather than an active task/stimulus, since no useful activation results can be derived from this segment of the data.

In practice, it is easy to see how many images to discard by looking at a graph of the time series data from a few representative voxels. About 10·T1/TR images is the usual number; for example, T1≈0.8 s at 1.5 T, so with TR=4 s, 2 or 3 images will show this effect.

Image Time Series Registration

One of the biggest practical problems in fMRI is subject head motion during an imaging run. The signal changes from even tiny motions (under 1 mm) can be larger than the typical 2% BOLD response at 1.5 Tesla. If two voxels differ in intrinsic MRI signal by 10% (a typical value inside the brain), then a movement of 20% of a voxel dimension will cause a 2% change in voxel signal intensity. For 3 mm voxels, this means a 600 μm motion can interfere with the detection of neurally-related changes in voxel signals. At the edge of the brain, neighboring voxel intensities often differ by 70%, meaning that a 100 μm movement could cause a 2% signal change.

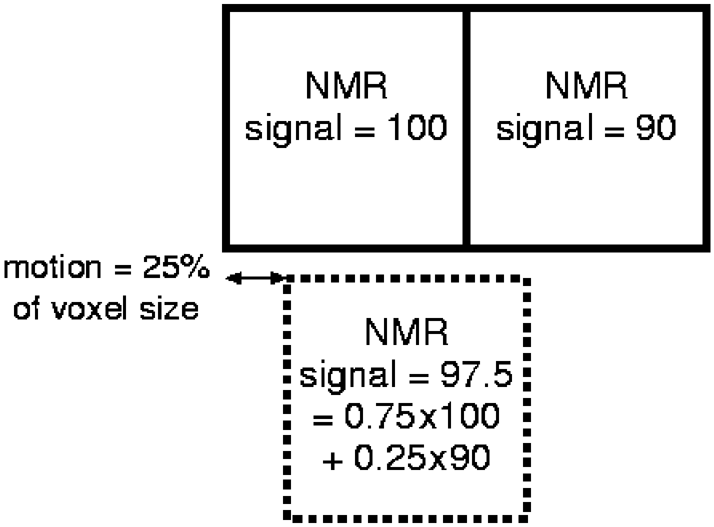

Figure 7.1. Motion of 25% of a voxel edge size causes a 2.5% change in voxel intensity between neighboring voxels that have intrinsic NMR signal intensities that differ by 10%. The squares above represent the original voxels. The dashed square below represents the left voxel after the subject moves slightly to the right (the voxel grid remains fixed and the subject moves "beneath" it). This signal change might be larger than the BOLD effect signal change, particularly in prefrontal and association cortex regions. As discussed in the text, movement-induced signal changes might be confused with BOLD effect signal changes, and/or might mask the presence of the BOLD effect.

The amount of movement is quite variable between individuals. Healthy subjects that have been scanned before are usually more comfortable in the magnet and will move less. Children and patients generally make poorer subjects from this point of view. Head restraints are commonly used, but most subjects do not like them. Also, the skull and brain are not rigidly attached to the scalp, so that even very tight exterior restraints do not rule out movements of the brain.

If the subject's movements occur at the same time as the stimulus, then the signal changes consequent to movement will have the same timing as the signal changes consequent to neural activation. Such stimulus-correlated motion can be caused by sudden changes in visual illumination or by changes in responses required by the subject (e.g., changing from rest to hand motions). When this type of motion occurs, a large number of false "activations" are detected [Hajnal]. Since the signal changes due to movement are largest near the edges of the brain, if a functional activation map shows a "halo" of activations along large portions of the perimeter of the cortex, you can be sure that motion is the culprit.

Contrariwise, if the subject movements occur randomly, the signal changes consequent to motion will be like an extra noise source, and will make it harder to detect true activations. Unless the motions are large (more than 1 voxel size) and frequent, strong activity will be detectable; however, regions of smaller signal change will be masked by the movement noise. This effect is most troublesome when comparing activation maps between groups of subjects. It is important to determine if one group had more movements than the other. For example, if a group of patients had more movements than a group of controls, it might be difficult to determine if smaller volumes of activations seen in the patients was due to movement noise or due to real neurological differences [Weinberg?].

The most commonly used technique for reducing the effects of subject head movement is image registration. Each target image (2D slice or 3D volume) needed to be registered is compared to a common "fiducial" ("base") image, and the displacement—rotation and translation—from the base image is estimated. This estimation is done by minimizing a measure of mismatch between the fiducial image and the target image as the displacements are varied. (The base image should *not* be taken from the early images before the brightness reaches a steady state, since comparisons with later images would be inaccurate—early and later images don't match.) Each target image is then resampled from its original matrix to a new matrix that is aligned with the base image. The various registration methods described in the image processing and fMRI literature differ in the details of how the images are compared to the base image (how the mismatch between two images is defined) and how the resampling to the new matrix is computed.

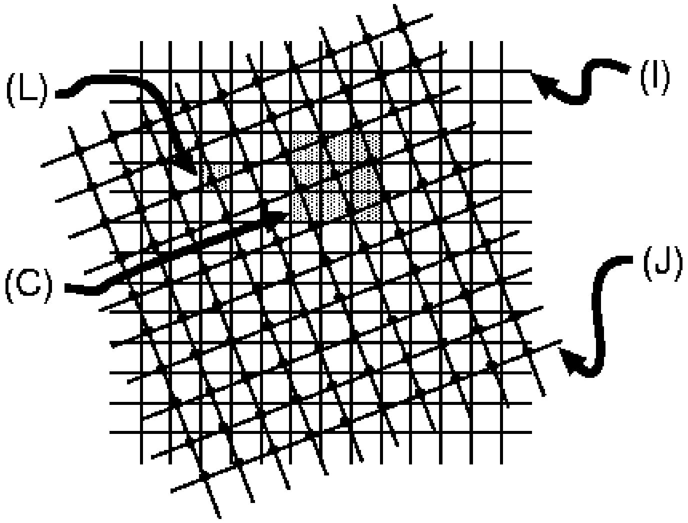

Figure 7.2. The need for resampling when an image must be aligned to a new spatial matrix (grid) is illustrated. Grid (I) is the original grid on which the image is defined; numerical values are associated with each horizontal-vertical grid intersection. Grid (J) is the rotated grid onto which the image must be realigned. Numerical values are needed at each slanted grid intersection (marked with heavy dots). Resampling from the (I) grid to the (J) grid requires interpolating between the values given on the (I) grid. At output point (L), the value is interpolated from its 4 nearest neighbors—this is bilinear interpolation. At output point (C), the value is interpolated from 16 neighbors using bicubic interpolation. Not shown is sinc interpolation, where each output value depends on each input value. For most purposes, sinc interpolation is preferred; however, it is more complicated to program and sinc interpolation software is slower than other resampling methods. (In general, images are 3D and resampling must be done between 3 dimensional grids.)

Image registration does not completely mitigate the effects of subject head movement. It is impossible to estimate the parameters of the motion completely accurately. Another problem is that EPI images are slightly distorted by nonuniformities in the static magnetic field. MRI is analogous to taking a photograph through wavy glass. If the subject moves behind the glass, the image could be moved back to the original position. However, the result would not be identical to a photograph taken when the subject was at the original position, since the distortions induced by the wavy glass would depend on the details of the glass shape between the camera and subject, and these would be different between the two images. Similarly, when the subject moves, the image will be distorted in a different way in the new position. These changes in distortion may not be visible as a warping of the image, but they will cause changes in the voxel values that image registration will not fix.

Image registration can be done in 2D (slice by slice) or in 3D (volumetric). In 2D, rigid movements are described by 3 parameters: *x*-translation, *y*-translation, and rotation. In 3D, rigid movements are described by 6 parameters: translation along each coordinate axis, and rotation about each axis. Generally, 3D registration has been preferred by the fMRI community, since subjects can move their heads in any direction. However, the most common movement while lying supine is a nodding motion. If the 2D slices are gathered in the sagittal plane, then most of the movement will be in-slice, and 2D registration may be sufficient.

One difficulty with 3D registration is that it is based on an obviously false model of the imaging process: 3D images are not gathered in a snapshot, but are assembled over several seconds. It is artificial to assume that there is no motion during a TR interval, and then all the motion occurs before the 3D slice package acquisition starts over. When movements are slow—such as when a subject's head is slowly sinking into a pillow—this model is adequate. For sudden head jerks, this model is clearly inadequate; see Fig. 7.3. For this reason, one strategy is to acquire the fMRI time series in the sagittal plane, where the motions are most likely to occur, and then use 2D registration on each slice separately. At present, there is no automated method to determine whether 2D or 3D registration is preferable for a given data set.

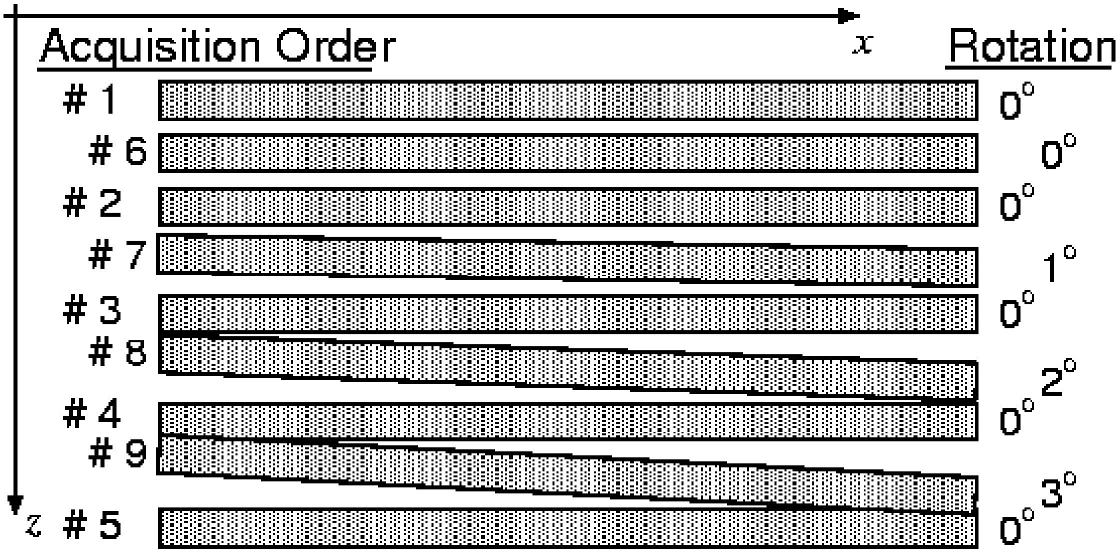

Figure 7.3. The effects of out-of-slice rotation partway through a single volume acquisition. There are 9 slices, gathered in the usual interleaved order. At the time of slices #7-9, the subject rotates his head out of plane (e.g., if the slices are axial, the subject nods his head). The slices are viewed edge-on.

One method for dealing with the difficulty illustrated in Fig. 7.3 is to perform a slice-into-volume registration; that is, to estimate separately for each 2D slice where in the whole 3D volume it came from [Kim]. It is not yet clear if this relatively new technique produces significantly superior results to volumetric registration.

If the out-of-slice movement is as large as depicted in Fig. 7.3, then an NMR effect comes into play to further alter the image intensity. Note that slice #9, as acquired, overlaps with slice #4. This means that the overlapping tissue will have been excited by RF twice within one TR period, unlike most locations. When it is slice #4's turn to be excited again, in the next 3D acquisition, the overlap region will have had less time to recover from the RF excitation. This reduced recovery interval will diminish the NMR signal from that region. Image registration alone does not fix this type of problem; at best, registration will merely straighten the slices out, but cannot correct for the scanner having measured the "wrong" values. In general, this nonlinear effect would be hard to correct for; however, it is generally fairly small, and a linear approximation can be used if it desired to allow for this problem [Friston].

Applying some image registration technique to the data time series is a good idea, and probably necessary if you want to get your results published. At the least, you can use the movement parameter estimates to get a quantitative estimate of the size of the movements. If the largest estimated movements are less than 10% of a voxel dimension, the registration process may not improve the data. It is also useful to view the pre- and post-registration images in cine mode to see if the motion correction has been useful, or if something has gone wrong (e.g., a bad parameter estimate might end up adding motion to the data).

Registering to Anatomical Reference Images

A high-resolution (e.g., 1×1×1 mm^3) 3D image with good gray-white matter contrast makes a very good anatomical reference image. These images are usually acquired with a T1-weighted pulse sequence (e.g., SPGR or MP-RAGE). It is almost universal to overlay functional activation maps on such images, since they provide structural information that the BOLD-sensitized T2*- or T2-weighted images cannot.

The fMRI time series is almost always gathered at a coarser spatial resolution than the anatomical reference data, and the two data sets may be gathered in different orientations as well (e.g., axial EPI and sagittal SPGR). It is important that the software used for fMRI analysis be capable of dealing with this type of mismatch. The most common method determining the proper overlap between two different MRI data sets is to use the scanner coordinate system, which is defined by the gradient coil. Each slice acquisition is specified in terms of these coordinates. If these coordinates are recorded, it is then "merely" a software problem to display two 3D data sets so that the same spatial coordinates are mapped to the same screen coordinates.

In some cases, there might be significant subject movement between the acquisition of the anatomical reference and the fMRI time series. If nothing else, over a period of an hour or so, subjects tend to slide around a little bit (a few mm) on the scanner bed as they try to keep comfortable. One simple method to deal with this problem is to use a fiducial image (for time series registration) that is gathered very close to the time that the anatomical image is acquired. All the fMRI time series from a scanning session will be registered to the same fiducial. This technique is generally adequate, unless the subject moves significantly in the time between the anatomical image acquisition and the closest fMRI time series acquisition. This time should be minimized in order to reduce the likelihood of such motion.

Besides subject motion, subject-induced distortions in the magnetic field can cause the fMRI time series to be misaligned with the anatomical reference. This effect is due to the fact the EPI methods are more sensitive to magnetic field imperfections than the commonly used anatomical imaging techniques. The primary effect on echo-planar images is to shift the images in the phase encoding direction. This problem is larger at higher magnetic field strengths; for example, displacements of up to 3 mm might be seen at 3 Tesla.

Methods for registering un-alike (inter-modality) images have been developed. They mainly differ from intra-modality methods in the way in which they compare images. Intra-modality methods can subtract two images and try to find the motion parameters that make this difference small. Inter-modality registration methods typically try to match edges or other features in the images, rather than directly compare voxel values [ref].

The distortion in echo-planar images in the phase encoding direction is usually not just a simple shift relative to the anatomical image. The amount of displacement depends on the magnetic field change caused by local tissue susceptibility, and that will depend on location in the subject. The effect is that echo-planar images can be warped relative to the underlying anatomy. The best way to solve this problem is to acquire a magnetic *field map* using a special MRI pulse sequence, and use this extra data to undistort the images during reconstruction. Unfortunately, this image acquisition and reconstruction option is not universally available.

Trend Removal

It is very common to see voxel time series that have slow drifts in signal intensity over the time of a scanning run. The most likely cause of this effect is very slow subject movement, but it has also been seen in fMRI applied to anesthetized rats whose skulls are literally screwed into the gradient/RF coil assembly—a situation where movement is essentially impossible. Other effects that can cause slow changes in the NMR signal include (a) slight drifts in the scanner components if they heat up during the imaging run; (b) leakage of weak external RF signals into the scanner room; (c) changes in subject respiration, heartbeat, or posture during the imaging; and (d) accumulation of deoxyhemoglobin downstream from an active site, so that that the signal drifts downward during a set of stimuli, and then clears again when the stimuli cease for minute or so [Frahm].

It is generally necessary to remove these slow drifts when processing the voxel time series. Conceptually, this is a preprocessing step, since it is separate from the activation analysis; however, in practice, trend removal and activation analyses are often carried out simultaneously. The most common technique for trend removal is to fit a straight line ($f(t)=a+b\cdot t$) to each voxel time series separately, using linear least squares, and to remove the constant and linear components. The fluctuations that are left will be analyzed to determine if they correspond to neurally induced signal changes.

The reason for using a straight line is mostly that fMRI time series tend to be short, so that any slow curved trend (up for a while, then down) tends not to be seen fully. For time series longer than several minutes, it is probably a good idea to detrend with a function that can allow for more complex trends (e.g., a quadratic polynomial in t) [Cox].

Filtering out Physiological Fluctuations

After head movement, the biggest sources of noise in fMRI time series are usually the cardiac and respiratory cycles. The heartbeat causes changes in the NMR signal by moving blood through slices at varying rates; blood flowing into a slice has not experienced the same RF excitation history as the stationary in-slice tissue, and so will have a different level of longitudinal magnetization available for conversion into transverse magnetization. An additional effect is the vertical movement of the brainstem with each heartbeat: about 500 μm. Respiration also causes brainstem movement. The motion of the diaphragm muscle and other tissue in the chest causes the magnetic field in the brain to change slightly during the breathing cycle, which in turn causes the NMR signal to vary. (A very large effect can be seen when the subject has any metallic object in or on the chest; for example, the motion of a steel underwire brassiere can produce quite noticeable signal changes in the subject's brain due to the rhythmic changes in the magnetic field.)

It is possible to filter much of this physiological noise out of the fMRI time series data. The simplest approach is to measure or estimate the frequencies of the subject's heartbeat and breathing. All data occurring at these frequencies can be removed from each voxel's time series [Biswal]. The main difficulty with this method is the phenomenon of ***aliasing***, which makes it impossible to distinguish frequencies in sampled data that differ by multiples of $(\Delta t)^{-1}$, where Δt is the data sampling interval. For example, if Δt=4 s (typical for fMRI), then heartbeat-related signals in the range 1.0–1.1 Hz will be indistinguishable from signals in the range 0.0–0.1 Hz. This latter low frequency range is right where the fMRI-measurable activations will generally occur (e.g., with a stimulus cycle of 10 s "on" and 10 s "off", the fMRI signals will be expected at 1/20=0.05 Hz). One way to avoid this problem would be to measure each subject's heart rate and adjust the scanning TR to avoid aliasing the heart rate to anywhere near the stimulus rate. A strong objection to this procedure is that each subject's data would be gathered using a somewhat different methodology, which would make inter-subject comparisons questionable.

A slightly more elaborate method for filtering out physiological noise requires measuring the subject's cardiac and respiratory cycles throughout the imaging experiment. The image time series data can be fit to these extra time series, and components in the image data that are correlated with the measured physiological cycles are subtracted out [Hu, Le]. This extra data can be gathered using external devices or can be inferred from some types of MRI data. Using this technique requires gathering an fMRI time series with the subject in the resting state only, in order to calibrate the fit between the physiological time series and the image time series in each voxel.

Physiological noise filtering methods have not yet become commonplace in the analysis of fMRI data. The simpler method described above is not adequate for relatively slow whole brain imaging, although it works well for single- or few-slice experiments. The second method requires the extra physiological data and more intricate signal processing software.

7.5. Functional Image Generation from Time Series Data: Temporal Pattern Matching Methods

This class of methods for detecting "active" voxels in an image time series is based on positing a model for the NMR signals that would be expected from a neurally active region. The data time series are fit to the model, and those voxels where the fit is statistically significant are declared active. This description allows enough room for many different methods to be developed: various models are plausible, various fitting techniques can be used, and various methods for estimating significance are available.

Most of the methods discussed in this section are applied to each voxel time series in the brain separately; the discussion below will deal with the analysis of a single voxel. Methods for assessing the overall significance of the results, and of spatially clustering "active" voxels together are discussed in §7.7.

Application of Standard Statistical Tools

The earliest—and still widely used—method for detecting stimulus-correlated signal changes is the application of a Student *t*-test. This technique is applicable to experiments in which the stimulus alternates between two conditions—A and B, say. All the values corresponding to A are pooled into one set of numbers, and all the values corresponding to B into another. The difference between the means of the A set and the B set is calculated, as is the *t*-statistic for determining the significance of this difference in means. Voxels with a *t*-statistic above some threshold are declared "active".

If TR is significantly less than the hemodynamic response time of 5–6 s, then there will be some time points that occur during the transition between active and inactive. These data points should not be used in the *t*-statistic analysis, since they don't fit the implicit model response—a voxel value is either "on" or "off", and there is no in-between state. This difficulty can be resolved with a more realistic signal model—the straightforward generalization of the *t*-statistic method is the correlation method, described later.

Other standard statistical tools can be applied to voxel time series analysis. If the "on-or-off" model is plausible for the experiment, nonparametric analogs to the *t*-test can be used (e.g., the Kolmogorov-Smirnov [Weisskoff] and Mann-Whitney tests). These have the advantage that estimates of their significance are robust against non-Gaussianity in the noise. They have the disadvantage that it is difficult to estimate their significance in the present of temporally correlated noise. They also have the disadvantage that they do not calculate a statistical parameter that is plausibly related to neurophysiological activity level—see "What to Display", below. Instead, these methods directly determine only the significance of the difference between the A samples and the B samples.

Time Shifting the Analysis

The hemodynamic response to neural stimulation is relatively slow: the response doesn't plateau until 5–6 s after the stimulus begins. fMRI signal analysis needs to allow for this, especially when TR is less than or equal to this time scale. For "on-or-off" analyses such as the *t*-statistic, it is usually sufficient to group the "on" times starting 5–6 s after the stimulus starts and stopping 5–6 s after the stimulus ends.

For more complex analyses—especially with a TR that resolves the hemodynamic response—it is best to explicitly include the signal delay into the model and its analysis, rather than simply time shift the analysis by some fixed amount. One reason for this is that most 3D imaging is actually done with interleaved 2D slices, as shown in Fig. 7.3. If the slices are spread out uniformly across the TR interval (as is usually done), then two adjacent slices will be gathered about ½·TR apart. This means that the time series from these slices will be shifted different amounts from the stimulus onset and cessation times. Any time series model that includes the hemodynamic time delay, rise time, and fall time, must also properly deal with this inter-slice delay.

One method to deal with this issue is to interpolate the slice data to some standard time point, and then pretend that it was all gathered simultaneously. The only justification for this type of analysis is that it makes the computer software easier to write—one less thing to worry about. Otherwise, this idea goes against the credo of statistical analysis, which is to model the acquisition as it actually occurs and process the data accordingly, rather than force the data into a more convenient form. In addition, resampling the time series data to a standard time point will not completely solve the problem of inter-voxel dispersion in the onset of the hemodynamic response. Voxels may differ in the speed with which the BOLD signal accumulates because of "plumbing" differences: in the arterial supply and in the distribution of venous vessel sizes.

What to Display; or, What is a Brain Activation Map?

In the *t*-statistic method, two parameters are computed for each voxel: the difference in means between the two conditions, and the *t*-statistic corresponding to that difference. Further values can be derived, such a nominal *p*-value based on the assumption of white Gaussian noise. Images formed from any of these values can be displayed and further processed (e.g., averaged between subjects).

In our view, it is not appropriate to display images of statistical thresholds such as the *t*-statistic or a nominal *p*-value and use these as the primary brain activation maps derived from fMRI data. (The situation is a little different in PET, where the noise level is more spatially uniform.) The reason is that two components of the signal go into calculating the any statistical threshold: an estimate of the amount the signal changes with activation, and an estimate of the noise magnitude. The noise level is highly variable across the brain in fMRI data. This means that a voxel with a small BOLD signal change but that has a small amount of noise will have the same *t*-statistic as one with a large BOLD signal change and a larger noise level. It is inappropriate to conclude that these voxels are equivalent—there is no reason to believe that the noise level has a great deal to do with neural activity, whereas there are many reasons to believe that the BOLD signal changes are strongly coupled to changes in neural activity. We believe that primary brain activation maps should be made from statistics that are plausibly related to neurophysiological parameters. In the case of the *t*-statistic method, this would be the difference in the means, since this is an estimate of the BOLD signal change, which is the closest thing we have in most fMRI data to a neurophysiological parameter. (ASL data is probably closer; unfortunately, it also has a higher noise level).

By "primary" activation maps, we mean the derived images that are intended to serve as the most important representation of the experimental results—the images that will be published and the images that will be compared with other data. For many purposes, it is useful to look at maps of *t*-statistics, *p*-values, etc., but these are not appropriate statistics to use in comparisons with other results.

The Correlation Method

This is probably the most widely used method for fMRI activation analysis at this time. It also applies to experiments in which the stimulus alternates between two conditions. The basic idea is to choose a reference ("ideal") time series that has the shape of the expected fMRI response, and then compute the correlation of each voxel time series to the ideal [Bandettini]. Voxels with a large correlation coefficient are deemed to be active. To be explicit, some equations are necessary. Define $x(t)$=data time series for one voxel, and *r*(*t*)=reference time series; let $\bar{x}$=the mean of *x*(*t*) and $\bar{r}$=the mean of *r*(*t*). The estimated correlation coefficient between these two time series is

$$\rho = \frac{\sum_t [x(t)-\bar{x}][r(t)-\bar{r}]}{\left[\sum_t \left(x(t)-\bar{x}\right)^2\right]^{1/2}\left[\sum_t \left(r(t)-\bar{r}\right)^2\right]^{1/2}} .$$

This value will range between –1 and 1. If the noise is *x*(*t*) is assumed to be Gaussian and uncorrelated, then ρ can be converted to an equivalent *t*-statistic via the formula $t = \rho \big/ \sqrt{\left(1-\rho^2\right)/n}$, where *n* is the number of degrees of freedom in the estimate of ρ. (*n*=*N*–*m*, where *m* is the number of nuisance parameters detrended from *x*(*t*); *m*=1 if only the mean is removed, as in the formula above; *m*=2 if a linear trend is also removed, as discussed earlier.) In the case where *r*(*t*) comprises only zeros or ones, this *t*-statistic is exactly the same as would be computed using the direct *t*-statistic method to test the difference of means between the times where *r*(*t*)=1 and where *r*(*t*)=0. The new feature of the correlation method relative to the *t*-statistic method is that the reference waveform can take into account the rise and fall times in the hemodynamic response.

It is necessary to allow for different hemodynamic delays in different voxels. In part, this is due to the multi-slice nature of the imaging process, as mentioned earlier. There is also some scatter in the intrinsic hemodynamic delays between voxels. The reasons for this are not entirely clear. One hypothesis is that voxels with longer delays contain more large venules, which are farther downstream from the activation site and so take longer to fill with oxyhemoglobin. Whatever the case may be, the correlation method is often applied with several reference waveforms *r*(*t*–δ) for a range of delays δ—spaced 1 s apart, say—and for each voxel the value of δ is chosen that maximizes ρ. (This effectively reduces *n* by 1.)

DeYoe et al. have used time shifted correlation in a clever way to map the retinotopic organization of visual cortex [DeYoe]. The stimulus consisted of a flashing checkerboard pattern over a portion of the visual field. The stimulus was presented continuously, but its location in the visual field was changed every 20? s. Correlation was performed with a large number of delay times. Voxels with maximum correlation (as a function of delay) larger than the threshold were included in the brain map. Each active voxel was marked as being mapped to the location in the visual field that corresponded to the stimulus at the voxel's time of maximum correlation. This continuous stimulation protocol and analysis method is an efficient use of scanner time. The retinotopic organization of visual cortex is being used to provide "stimulus off" periods for some parts of the brain while "stimulus on" periods are being provided to other regions.

The correlation method is equivalent to linear least squares regression to the following model equation for the voxel data time series:

$$x(t) = \alpha \cdot r(t) + a + \text{white noise}$$

where α and *a* are the unknown parameters being fit by the least squares method. The solution for α is

$$\alpha = \frac{\sum_t \left[x(t)-\bar{x}\right]\left[r(t)-\bar{r}\right]}{\sum_t \left[r(t)-\bar{r}\right]^2} = \frac{\left[\sum_t \left(x(t)-\bar{x}\right)^2\right]^{1/2}}{\left[\sum_t \left(r(t)-\bar{r}\right)^2\right]^{1/2}} \cdot \rho ;$$

α is the best fit to the amplitude of *r*(*t*) in *x*(*t*). Even though the use of a zero-one *r*(*t*) in the correlation method is equivalent to the *t*-statistic method, the linear regression form of the correlation method has one practical advantage over the *t*-statistic method: it is very easy to include extra regressors to remove unwanted trends in the data at the same time the correlation is being computed (e.g., add $b \cdot t$ to remove a linear drift). In the t-statistic method, detrending must be done as a separate preprocessing step.

Since α does not depend on the noise level in $x(t)$, unlike ρ, α is a reasonable candidate to use as a measure of the BOLD effect. One drawback to this use of α is that its size depends on the scale chosen for $r(t)$: if $r(t)$ were doubled, then α would be halved. The size of α also depends on the scale chosen for $x(t)$: if $x(t)$ were also doubled, then α would again be halved. In the analysis of any one imaging run, this scale factor is unimportant, but it can be confusing when comparing results among multiple scanning sessions. Even if the identical $r(t)$ is used, the scaling factor for $x(t)$ will probably not be the same. This factor is set by the scanner operator (when he sets the RF receiver amplifier level), but is unlikely to be the exactly the same each time. Some clinical scanners set this value automatically during pre-scan operations, and in this case the change in MRI scale factors may not even be known to the data analyst.

A better alternative is to calculate the percent signal change from the baseline. This is most easily done by choosing the reference $r(t)$ to range between 0 and 1; then the estimate of the baseline signal is a, and the percent change is $100\% \cdot \alpha / a$ (this formula must be modified if detrending is included in the regression: the denominator must include the mean baseline of the extra detrending functions). If the density and size distribution of blood vessels is approximately the same across gray matter, then the percent signal change in each voxel will be approximately proportional to the change in the oxyhemoglobin concentration in the voxel.

Choosing the Reference Function

The correlation method does not specify how the data analyst is to choose $r(t)$. In a block trial experiment, where the stimulus durations are several times longer than TR and when $\text{TR} \geq 4$ s—so that the hemodynamic rise and fall times are not well sampled—then a zero-one $r(t)$ is a reasonable choice, with $r(t)$=0 during the "off" periods and $r(t)$=1 during the "on" periods; the transitions between 0 and 1 should be delayed about 5 s from the stimulus onset and cessation. If $\text{TR} < 4$ s, then some intermediate values should be put in to smooth the transitions between 0 and 1.

Early work with fMRI data sometimes used smoothed and/or averaged data from "obviously active" voxels as the reference waveform [Bandettini]. The subjective nature of this approach has led to its falling out of favor. In principal, though, this method could be made objective by using a two-pass approach: first, find voxels that are active using a 0-1 $r(t)$ and a stringent statistical threshold; second, average these voxel time series, perhaps smooth them temporally, and then use the resulting time series as the new $r(t)$ for the final analysis. This technique is really a melding of the pattern matching methods (pass 1) and the pattern hunting methods (pass 2). As far as we know, it has not been frequently used.

A widely used method for generating $r(t)$ is based on the theory of shift-invariant linear systems. In this model, a brief input (the stimulus) to a system (a brain voxel) produces the same output response (the BOLD signal) no matter what the condition of the system or its past history—this is the shift-invariance feature. When multiple stimuli are present, the output response is taken to be the sum of the responses from the individual stimuli—this is the linearity feature. In equations, suppose that a brief stimulus at t=0 produces the response $h(t)$. (This _*impulse response function*_ must be zero for $t < 0$, otherwise the system would respond before the stimulus.) Then if stimuli are actually applied at times t=a and t=b, the model response is $h(t-a)+ h(t-b)$. The subtraction of the stimulus times a and b in $h()$ shifts the individual responses to start at those times. If stimuli are applied continually over an interval $a \leq t \leq b$, then the analog to the sum of responses is an integral: $\int_a^b h(t-q)\,dq$.

The technique for generating the *r*(*t*) for data analysis requires two inputs: the impulse response *h*(*t*) and the stimulus timing function *s*(*t*), which is equal to 0 at times *t* when no stimulus is applied (or when the control stimulus is being applied) and is equal to 1 when the active stimulus is applied. The response function can now be written in integral form (for continuous time) or summation for (for discrete time):

$$r(t)=\int_0^T h(t-q)\cdot s(q)\,dq=\int_t^T h(t-q)\cdot s(q)\,dq \text{ , or}$$

$$r(t)=\sum_{q=0}^{T} h(t-q)\cdot s(q)=\sum_{q=t}^{T} h(t-q)\cdot s(q)\,;$$

here, *T* is the duration of the experiment; in the summation form, the time step is taken to be 1 for convenience. (Equations of this form are called *convolutions*; one says that *r*(*t*) is *h*(*t*) convolved with *s*(*t*).)

The choice of one function *r*(*t*) has now been pushed back to the choice of another function *h*(*t*). This may not seem like progress, but it is, since *h*(*t*) can be chosen without much regard for TR or stimulus timing; instead, *h*(*t*) is to be thought of as a universal function that mimics the response of brain+scanner system. A popular function is $h(t)=t^r\,e^{-t/c}$ for $t>0$ and *h*(*t*)=0 for $t<0$, with parameters *r*=8.6 and *c*=0.51 [Cohen]. This gamma variate function approximately reproduces the features of the hemodynamic response to brief stimuli. The time for this *h*(*t*) to rise to its peak value is $r\cdot c$ after *t*=0; the full width of the response at half the maximum value (*FWHM*) is approximately $2.4\cdot r^{1/2}\cdot c$.

The neural networks in the brain obviously comprise a nonlinear system, and the hemodynamic response measured by fMRI is probably not strictly linear either. Nevertheless, the linear system model described above has proven very useful for modeling fMRI time series data [Glover, Buckner]. There is some evidence of nonlinearity in the response to brief stimuli, in that the amplitude of the BOLD signal does not always add up as the model suggests [Vasquez]. There is also some evidence that previous stimuli may somewhat reduce the amplitude of the response to stimuli that occur within a few seconds [Friston(Volterra)]. It is unknown at present whether these two effects are neural or hemodynamic in origin (or both). Both effects suggest caution when using linear system models in an experiment that mixes short and long stimuli, or has stimuli with overlapping responses mixed with well-separated stimuli. However, practice suggests that the linear systems approach to generating *r*(*t*) works well. Unless TR is very short (1 s or less), the exact choice of *h*(*t*) doesn't seem to matter much, since the shape of the hemodynamic rise and fall will not be sampled very accurately.

Multiple Linear Regression

It is a natural generalization of the correlation method to use multiple reference waveforms to fit the voxel time series. In an experiment with three different stimulus conditions (rest, A, and B), the analysis would include separate response functions for the two active conditions:

$$x(t)=\alpha_A\cdot r_A(t)+\alpha_B\cdot r_B(t)+a+b\cdot t+\text{noise}\,;$$

each reference $r_{A,B}(t)$ could be generated using the convolution method described above. For this model, four parameters are fit for each voxel: the two amplitudes α_A and α_B, the baseline *a*, and the baseline drift rate *b*. Standard linear least squares analysis algorithms can be used to compute the best fit parameters. At least two questions can be asked of the amplitudes: (a) is either α_A or α_B nonzero? (b) is α_A different from α_B (e.g., is $\alpha_A-\alpha_B$ nonzero)? If the noise is assumed to be white and Gaussian, the statistical significance of the answers to these questions can be calculated using *F*- and *t*-tests. Voxels where $\alpha_A-\alpha_B>0$ was significant could be interpreted as regions where processing for stimulus A was more intense. This type of analysis has been used to contrast the response to visual stimuli consisting of human face (A) vs. nonface object (B) images [Haxby]. In primary visual cortex, there was little difference in the α_A and α_B amplitudes, but significant differences were detected in higher visual areas.

Multiple linear regression has many other potential applications in the analysis of fMRI time series data. Uncertainty and/or inter-voxel differences in the shape $h(t)$ of the hemodynamic response function could be modeled by using two different functions $h_1(t)$ and $h_2(t)$ to generate two stimulus response functions $r_1(t)$ and $r_2(t)$, using the convolution method, and then fitting the model

$$x(t) = \alpha_1 \cdot r_1(t) + \alpha_2 \cdot r_2(t) + a + b \cdot t + \text{noise}$$

to each voxel. In this case, the separate parameters α_1 and α_2 may or may not be important in themselves; the only question of interest might be "is either α_1 or α_2 nonzero?" The purpose of the model is to allow the response function model to adapt (a little) to each voxel. For example, time shifts can be approximately allowed for by taking $h_1(t)$ as the expected hemodynamic response function (e.g., a gamma variate function), and taking $h_2(t)$ to be its first derivative $h_1'(t)$. Since $h_1(t+u) \approx h_1(t) + u \cdot h_1'(t)$, this is a linear method for dealing with unknown delays in the hemodynamic response [Friston].

One must be wary of using models with too many parameters. As the number of fit parameters goes up, the number of data points needed to keep the same level of signal detectability and statistical significance also increases. This point is particularly cogent for fMRI, since it is usually necessary to keep the significance level per voxel quite stringent (see §7.7). Table 7.1 shows that the number of samples in time needed to have a given significance level per voxel increases with the number of model fit parameters

		Number of Fit Parameters in the Hemodynamic Response				
		2	**3**	**4**	**5**	**6**
BOLD Signal Change	**1.0 %**	50	57	64	70	75
	0.5 %	180	206	229	251	271

Table 7.1. Entries in the body show the number of samples (N) in a time series needed to reject noise-only data at the level p=10^{-4} per voxel, assuming noise standard deviation of 1% of baseline. The BOLD signal change that is being searched for is assumed to be 1.0% or 0.5% of baseline, and is assumed to be "on" for half the samples. The number of fit parameters in the baseline is assumed to be two. (These assumptions set the statistical threshold for detectability of the signal; the p-value of that threshold in the noise-only case was then determined for each N; the smallest N that has $p < 10^{-4}$ is the value in the table.) For the simple correlation method as usually used, there are 2 fit parameters in the hemodynamic response (amplitude and time delay).

Multiple linear regression can also be used to remove more nuisance effects than a simple linear drift in the time series. For long time series, more complex drifts can be allowed for [Cox]. The approximate method for dealing with inter-slice movement artifacts mentioned earlier is based on linear regression [Friston].

More complex linear signal models can be analyzed with this tool. The validity of the simple convolution model for the BOLD response has been tested using a second order convolution method known as a Volterra series [Friston]. Multiple linear regression is also a key tool in the analysis of event-related fMRI data, as will be discussed later.

Nonlinear Regression (Shape Analysis)

A further generalization of the regression methods is to allow the temporal pattern to depend nonlinearly on the parameters. In this model, it is not only the amplitude of response that is fit to the data in each voxel. The shape of response is also allowed to vary, within some range given by model equations. The model is allowed to be nonlinear (non-additive) in the parameters, unlike the case in multiple linear regression. One example is to define

$$h(t) = t^r e^{-t/c} \qquad \text{[impulse response model]}$$

$$r(t) = \sum_{q=t}^{T} h(t-q) \cdot s(q) \qquad \text{[responses summed over stimuli]}$$

$$x(t) = \alpha \cdot r(t) + a + b \cdot t + \text{noise} \qquad \text{[voxel time series model]}$$

In each voxel, the parameters *r*, *c*, α, *a*, and *b* are estimated from the voxel time series data (typically using a least squares criterion). Parameters *r* and *c* determine the shape of the impulse response, parameter α determines the amplitude of the response, and parameters *a* and *b* determine the baseline and temporal trend. With models of this type, the shape of the response being searched for in each voxel is not completely specified, but is also not completely free. (The model is called nonlinear because the output *x*(*t*) does not depend linearly on some of the parameters; in this example, *r* and *c* appear nonlinearly.)

One reason that nonlinear regression has not been widely used in fMRI analyses is that nonlinear fitting routines are much slower than linear fitting routines. It is also much more difficult to be sure that the "best" (truly *least* squares) fit has been found—it is possible that a false minimum will be found: one that is better than any other set of parameters nearby, but is not as good as some more distant possibility. In linear regression, the algorithms are much better established, executed quickly, and are guaranteed to find the global best fit parameters. Linear system models have worked well for many classes of fMRI experimental paradigms. Nonlinear regression will be useful in situations where the shape of the response function should not be fixed—but can be partially specified—and where the response function should be allowed to vary spatially. These conditions are most likely to be met when the image TR is short (2 s or less), so that the hemodynamic rise and fall times are well-sampled, yielding enough data to make it possible to estimate the response function.

Statistical significance in nonlinear regression is hard to analyze in closed form. If the number of samples in a time series is large compared to the number of parameters, as is usually the case in fMRI, then an *F*-statistic approximation can be used. Approximate confidence intervals can also be assigned to each parameter using *t*-statistics.

A generalization of the voxel-wise regression method outlined above would fix some unknown parameters to be the same across all "active" voxels, or to allow some parameters to vary only regionally—to be the same in one activated cluster, but be different in another. As far as we know, this type of local-global nonlinear regression analysis has not yet been used in fMRI. Defining the models and carrying out the regressions both present several practical difficulties. Nevertheless, the concept has some attraction, since it would provide a very flexible hybrid between the pattern matching and pattern hunting analysis techniques.

Analysis of Single-Event fMRI Data

Some experiments do not lend themselves to rapid repetition. Administration of a pharmaceutical agent is one example; altering the subject's mood by a video presentation is another. These types of stimuli take minutes (at least) to resolve. In such an experiment, there will only be one stimulus event per imaging run. The time course of neural activity is not well specified, although its starting point usually will be. This is the type of situation for which nonlinear regression is called. For pharmaceutical agent fMRI, one signal model that has been used is the difference of two exponentials [Stein??]:

$$r(t)=\begin{cases}0 & \text{for } t<s \\ e^{-c\cdot(t-s)}-e^{-d(t-s)} & \text{for } t>s\end{cases} \qquad \text{[response function]}$$

$$x(t)=\alpha\cdot r(t)+a+b\cdot t+\text{noise} \qquad \text{[voxel signal model]};$$

here, the rate constants *c* and *d* represent the wash-in and wash-out of the agent, and are solved for in each voxel. The start time *s* can be fixed to be the injection time, or can be allowed to vary. Another signal model that has been used to look for signals that generally increase and then decrease is the beta variate:

$$r(t)=\begin{cases}0 & \text{for } t<s \;\text{ or }\; t>f \\ (t-s)^p(f-t)^q & \text{for } s<t<f\end{cases} \qquad \text{[response function]}$$

$$x(t)=\alpha\cdot r(t)+a+b\cdot t+\text{noise} \qquad \text{[voxel signal model]};$$

here, *r*(*t*) is a function that rises up from 0 at the start time *t*=*s* and descends back down to 0 at the final time *t*=*f*. The exact shape depends on the parameters *p* and *q* (which are taken to be positive). As before, the linear parameters (α, *a*, *b*) determine the amplitude of the signal change, the baseline, and the baseline drift, respectively. This model has been applied to subject mood changes induced by video presentations [Montreal HBM abstract].

Analysis of Event-Related fMRI Data

This type of data is characterized by a sequence of responses to individual stimulus events. The stimuli are separated in time so that each response may be separately analyzed. One strength of this experimental approach is that responses may be examined based on post hoc considerations (e.g., reaction time, correct or incorrect response).

Most event-related analyses reported in the fMRI literature have used multiple linear regression as the basic tool [Buckner]. A hemodynamic impulse response function *h*(*t*) is supposed. Responses are divided into categories {1, 2, …, C}. For each category *j*, a 0-1 time series $s_j(t)$ is created; $s_j(t)=1$ if a stimulus of category *j* occurs at time *t*, otherwise $s_j(t)=0$. The voxel signal model is then

$$r_j(t)=\sum_{q=t}^{T}h(t-q)\cdot s_j(q) \qquad \text{[normalized response to stimuli of category } j\text{]}$$

$$x(t)=\sum_{j=1}^{C}\alpha_j\cdot r_j(t)+a+b\cdot t+\text{noise} \qquad \text{[total voxel response]}.$$

In this model, each stimulus category *j* is taken to have its own response amplitude α_j (in each voxel). These are calculated using linear least squares. Tests of various hypotheses (e.g., is $\alpha_1>\alpha_2$?) can be formed using *F*- and *t*-statistics.

It is not strictly necessary that the stimuli be separated far enough in time so that the response to one stimulus does not overlap with its neighbors. It is possible to have overlapping responses (the model above includes this case), as long as there are *some* intervals between stimuli long enough to resolve the rise and fall in *h*(*t*). This recognition allows the use of random inter-stimulus intervals, which has obvious psychological advantages. It also lets the average density of stimuli be higher, which improves the utilization of scanner and subject time. (When generating random stimulus paradigms, it is advisable to analyze each candidate for parameter estimation efficiency and to select the experimental protocol from the best set of timings [Dale].)

The stimulus response model $r_j(t)$ can be made more complex in ways that were discussed earlier. Multiple additive hemodynamic response functions $h_i(t)$ can be used. It is also possible to allow *h*(*t*) to depend nonlinearly on several parameters. Both methods amount to trying to infer the shape of the response function from the data. This class of techniques is sometimes called “deconvolution”, since the goal is to undo the convolution in the signal model to find out something about the underlying process. Deconvolution methods can be linear or nonlinear.

The analysis of event-related fMRI experiments must be designed with the experimental paradigm in mind. An elegant example is the mental rotation experiments of Richter et al., in which the subjects were presented with images of two complex 3D objects and had to decide if the images were views of the identical objects from two different points of view, or if the images were views of mirror-image objects [Richter]. As the angle between the points of view increases, the amount of time to carry out the task lengthens. The signal model must allow for this variation in neural activity following varying stimuli:

$$h(t;p) = \begin{cases} 0 & \text{for } t < 0 \text{ or } t > p \\ 1 & \text{for } 0 < t < p \end{cases} \qquad \text{[response of duration } p\text{]}$$

$$x(t) = \sum_q \alpha_q \cdot h(t - s_q; p_q) + a + b \cdot t + \text{noise} \quad \text{[voxel signal model]},$$

where s_q is the start time of the q^{th} stimulus. The response durations p_q are fit to the data; this is a nonlinear model, since $h(t;p)$ is not linear in p. Richter found that the fitted p_q matched the subject response times well in parietal regions previously associated with mental rotation. This is a creative example of designing the experiment and data analysis procedures together.

7.6. Functional Image Generation from Time Series Data: Pattern Hunting Methods

This class of methods forgoes positing a model for each voxel time series; instead, these techniques search for time series that can account for large numbers of voxels. A common thread is dimensional factorization of the 4D (space+time) data, one form of which is

$$s(x,y,z,t) \approx u_1(x,y,z) \cdot v_1(t) + u_2(x,y,z) \cdot v_2(t) + \cdots;$$

here, $s()$ is the 4D data as a function of voxel location (x,y,z) and time (t). The spatial components $u_i(x,y,z)$ are the “activation patterns” corresponding to the temporal components $v_i(t)$. This factoring is useful because the amount of information contained in a spatio-temporal pattern $u_i(x,y,z) \cdot v_i(t)$ is much smaller than in the original data $s(x,y,z,t)$. For example, if there are 10,000 voxel in the brain and 100 points in time, there are 1,000,000 numbers stored in $s(x,y,z,t)$, but there are only 10,100 numbers in $u_i(x,y,z) \cdot v_i(t)$.

In our view, component decomposition methods in fMRI should be viewed as tools for data exploration and hypothesis generation, rather than the final analytical technique [Somorjai]. One reason for this conclusion is that the investigator must supply some post-analysis judgment to determine which components are neurologically meaningful and which are not. There are many other sources of systematic changes in the fMRI time series beside neural activation, and these will all be picked out by the pattern hunting approaches. The “meaningful” components of activation may be relatively small compared to these other components, since the activated tissue volume is often small. Judgment must be used in picking out the component(s) that correspond to neural activation; this is normally done by finding the $v_i(t)$ that most closely resembles the stimulus timing and/or the expected hemodynamic response. It is also likely that some detected components will not be easily ascribable to any single cause; for example, if stimulus correlated movements occur, the neural activation component(s) and the motion effects component(s) will likely overlap, and will not break cleanly into two separate components as one would hope.

Despite these admonitions, looking for un-preconceived patterns can be useful. It is possible to detect trends in activation magnitude that would escape a simple regression analysis; for example, if the subject's alertness lags during an imaging run, the earlier activations may be larger than later ones. A regression analysis that takes a fixed amplitude (α) for all activations will miss this effect. Certainly it is possible to devise a pattern matching test to look for this trend, but a pattern hunting method might be able to bring it to the investigator's attention.

Most pattern hunting methods that have appeared in the fMRI literature will process all the brain voxels. This can be a drawback, since most of the signal will not be neurally relevant and will inevitably contaminate the results. A hybrid approach would be to use some loose pattern matching criterion to attempt to eliminate "noise-only" voxels from a subsequent pattern hunting analysis. Another way to restrict the pattern hunting would be to put constraints on the type of spatial patterns $u_i(x,y,z)$ and the temporal patterns $v_i(t)$ that are acceptable. To our knowledge, these ideas have not yet been tried on fMRI time series data.

Principal Components

Classical principal components analysis (PCA) is a direct implementation of dimensional factorization. The approximation sign "≈" is taken in the least squares sense; that is, $u_i(x,y,z)$ and $v_i(t)$ are simultaneously determined by minimizing

$$\sum_{x,y,z,t} \left[s(x,y,z,t) - u_1(x,y,z)\cdot v_1(t) + u_2(x,y,z)\cdot v_2(t) + \cdots u_L(x,y,z)\cdot v_L(t) \right]^2 ,$$

where L is the number of components desired; each voxel time series has its mean removed before this analysis. The least squares minimization leads to a matrix eigenvalue problem, which is why the $u_i(x,y,z)$ are sometimes called the "eigenimages". The components are ordered by the amount of total variance they explain in $s(x,y,z,t)$, the first component accounting for the largest amount, etc. It is likely that the first few components will correspond to physiological noise sources (e.g., heartbeat, respiration, movement).

Independent Components

Independent components analysis (ICA) uses the same dimensional factorization as principal components, but

Time Series Clustering

The expansion of $s(x,y,z,t)$ into a sum of spatio-temporal components explicitly assumes that the time series in each voxel is a sum of responses. An alternative approach is to assume that each voxel time series is best explained by exactly one of a set of "master" time series. The goal of time series clustering methods is to simultaneously find the set of master time series and assign one to each voxel. The data model is

$$s(x,y,z,t) = \alpha(x,y,z) v_{j(x,y,z)}(t) ,$$

where $\alpha(x,y,z)$ is the amplitude in the voxel at location (x,y,z), $\{v_1(t), v_2(t), \ldots, v_L(t)\}$ is the set of master time series, L is the number of master time series, and $j(x,y,z)$ is the assignment of master time series to voxels (j is an integer from 1 to L, inclusive); all of these values are to be determined from $s(x,y,z,t)$ by the data analysis software.

This type of data analysis does not lend itself to elegant closed-form mathematical solutions, unlike principal components and linear regression. "Fuzzy clustering" is the most widely used algorithm in the fMRI community for carrying out this kind of decomposition [Scarth, Somorjai]. It is an example of a feature space clustering method, where the features are voxel time series. These methods generally start with a candidate set of master time series. For each voxel, $j(x,y,z)$ is taken as the index of the $v_j(t)$ that best matches the data time series (e.g., in the least squares sense). After this assignment, all the data time series with the same value of j are averaged; this averaged time series replaces $v_j(t)$ in the set of master time series. The process is repeated until the assignments $j(x,y,z)$ and the master time series $v_j(t)$ stabilize. (The "fuzzy" nature of the particular algorithm comes from the way that j is assigned to each data time series, and in the way that the time series belonging to j are averaged.)

7.7. Selection of Statistical Thresholds

The central statistical problem in neuroimaging is the "curse of multiple comparisons"—the fact that although there is an gigantic amount of data, the statistical questions being posed are even more enormous [refs-JCBFM]. When a stimulus is presented to the subject, the question is not "is there brain activation somewhere?" (the answer is almost surely "yes"), but is something like "what is the spatial pattern of the brain activation, and is it different from this other activation pattern?" If there are 10,000 voxels in the brain, 10,000 decisions must be made. If the decision threshold is set so that there is a 5% chance of a false positive ("activation" being declared from a noise time series), then there will be about 500 "active" voxels detected even when imaging a comatose cantaloupe. This is usually unacceptable, so something must be done to reduce the false positive rate.

The Role of Statistical Assumptions

Statistical thresholds are always calculated based on some random model for the noise (and sometimes a random model for the signal, as well). Even nonparametric methods, which claim to be "distribution free", usually assume that each individual measurement is corrupted by a noise value that is independent of all other measurements but is identically distributed. When weighing the merits of a proposed statistical method for fMRI data analysis, it is important to understand what assumptions the authors have made—unfortunately, these are often not spelled out explicitly.

The most common assumption is that the noise is zero-mean Gaussian, that each noise value is independent (both between spatial locations and between time samples—spatially and temporally *white*), and that within each voxel the noise variance is constant in time. These assumptions underlie the classical statistical methods (e.g., t-tests and ANOVA), and they make the statistics of fMRI data analysis methods relatively simple. Since a major portion of the noise is physiological in origin, these assumptions are demonstrably false. Nonetheless, they are very useful, since they allow the derivation of mathematical formulas for statistical significance. Once these assumptions are cast aside, developing good statistical methods and results is much harder.

Adjusting the Degrees of Freedom

Denote the noise in some voxel at time t by $\zeta(t)$. If there are N samples in time, and the noise samples are assumed to be zero-mean independent Gaussians each with variance σ^2, then $R = \sigma^{-2} \cdot \sum_t \varsigma(t)^2$ is distributed like a χ^2 variable with N degrees of freedom. This sum is the least squares objective function in any regression analysis (linear or nonlinear) when the model parameters are at their "true" values. The statistics of R are the basis for F- and t-tests. However, when the noise is correlated in time, the significance levels (p-values) calculated for F- and t-tests using the assumption of N degrees of freedom will be erroneous (almost always too significant: p will be calculated as smaller than it ought to be). For correlated noise, the statistics of R are much more complicated, but a reasonable approximation is to take the statistics of R to be those of a χ^2 variable with a smaller number of degrees of freedom [ref]. *more to come*

Bonferroni Correction

This is the most straightforward method for setting a threshold for the voxel-wise t-statistic or correlation coefficient, and it is the most widely used. Bonferroni's inequality states that if N statistical tests are conducted, each of which has a probability p of producing a false positive, then the probability that at least one false positive occurs in all of the tests is less than or equal to $N \cdot p$. (This result does not depend on assuming that the noise in distinct voxels is independent.) For example, if there are 10,000 brain voxels, and one wants an probability of 0.1 that there is a single false positive voxel in the activation map, one would set the per voxel p-value of the test at $0.1/10{,}000=10^{-5}$. For example, with 60 degrees of freedom, this would correspond to a t-statistic threshold of 4.825, or a correlation coefficient threshold of 0.529. If the noise standard deviation is about 2% of the baseline signal (SNR=50), this means that such an imaging run could detect BOLD signal changes of about 1.2% ($=2\% \cdot 4.825 \big/ \sqrt{60}$). At 1.5 Tesla, this is a reasonable value to expect in primary sensory and motor cortex regions, but it is too large for many higher cognitive areas, where signal changes of perhaps 0.5% are more usual.

Such a stringent criterion can be relaxed in several ways, the more complex of which will be explored later in this section. The simplest method is to lower the threshold and simply accept a few false positive voxels. If the noise samples are independent in distinct voxels, then one would expect approximately $N \cdot p$ false positives. For many purposes, as long as this is much smaller than the number of true positives, the brain activation map won't be noticeably affected. Continuing the example above, raising the per voxel p-value to 10^{-4} lowers the t-threshold to 4.169; $p=10^{-3}$ lowers the t-threshold to 3.373—at this level, 0.9% signal changes could be detected, but 10 false positives would be expected. To detect 0.5% signal changes, the t-threshold would need to be 1.936, corresponding to $p=0.06$—at this level, 600 false positives would be expected.

Spatial Clustering

Voxel-wise detection methods are very good at detecting large signal changes. A major object to these techniques is the potential for rejecting a large region of activation in which each voxel has only a small signal change. This is the price that is paid for it being possible to detect arbitrary spatial patterns of activation.

If there are 10,000 voxels in the brain, then there are $2^{10{,}000} \approx 10^{3{,}000}$ distinct binary (on/off) spatial activation patterns. A basic statistical problem is to distinguish between these patterns using the time series data. Of course, most of these candidate activation maps are absurd (e.g., a voxel checkerboard) and would be rejected out of hand. The idea behind spatial clustering methods is to reject some of these patterns directly, and so improve the detectability of the plausible subset of potential brain maps.

As developed thus far, spatial clustering has been implemented as a dual thresholding method [Forman]. In the first step, a voxel-wise method is used to create an image containing a test statistic (e.g., the correlation coefficient) at each location. A threshold T is selected; all voxels with the test statistic above T pass on to the second step. In the second step, voxels above T are grouped together based on spatial contiguity; for example, clusters of nearest neighbors are formed. The second threshold is cluster size: voxels in clusters below the size threshold are rejected. The two thresholds—per voxel T and cluster size—are selected together to control the false positive rate (see Fig. 7.?). In this way, voxels with small signal changes and so relatively small test statistics still have a chance to be detected, provided that they are clustered together with other active voxels.

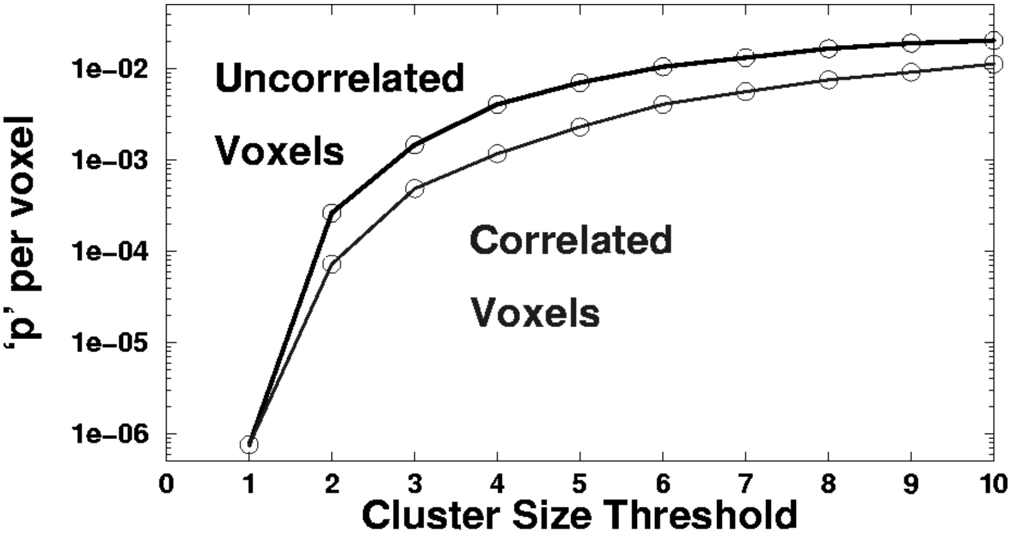

Figure 7.?. Given 7,000 brain voxels in the image time series, this graph shows the *p*-value per voxel needed in the first step of the spatial clustering method in order to have an overall *p*-value of 0.05 after the second step (probability of false positive in the entire image). Clusters here are defined as sets of contiguous voxels sharing a face—being nearest neighbors—with at least on other voxel in the cluster. Even a cluster threshold of 3 voxels allows the per voxel *p*-value to be greatly relaxed. (Calculations were performed with the *AlphaSim* program in the *AFNI* package.)

A minor difficulty with the dual thresholding strategy outlined above is that it is hard to calculate the threshold *T*, given the desired overall false positive rate and the cluster size. The simplest way around this problem is to use random sampling methods to approximate the threshold; simulate a lot of noise-only data, compute the test statistic, and for each level of *T*, carry out the clustering and determine the number of false positives. This can be time consuming (several hours of simulations), but need only be done once for each category of imaging experiment and voxel-wise test statistic.

A theoretical difficulty with dual thresholding is that small clusters with large activations will be rejected in favor of large clusters with weak activations—this is the opposite tradeoff from the original voxel-wise thresholding strategy. If an imaging experiment is alternating two very similar stimuli to look for subtle differences in activation maps, this tradeoff might not be acceptable. Intermediate cluster detection strategies have been developed that allow some of the benefits of both extremes [ref]. These strategies require the investigator to set a budget for the false positive rate allocated to small strong clusters and to weak large clusters.

Correlated Random Fields

If the noise values are correlated between voxels (i.e., the noise is not spatially white), the Bonferroni correction is conservative—the statement that $p_{\text{voxel}}=p_{\text{brain}}/N_{\text{voxel}}$ produces overly small values of p_{voxel}. In the case of PET imaging, the voxel values are so correlated and the number of data samples so small that this objection is insuperable. Instead, methods have been developed to calculate the probability that a functional image computed from spatially correlated random noise will have a peak that is larger than any given threshold value [Worsley]. These methods are produce more reasonable thresholds than the Bonferroni correction when the noise is strongly correlated over a distance of at least 5 voxel dimensions. When the noise is less correlated, these methods are actually worse than the Bonferroni method. PET data is highly correlated by construction, since the high noise level requires strong spatial smoothing. This smoothing produces noise with a simple spatial structure to the correlation function, which is needed in the calculation of the thresholds appropriate for correlated random fields.

fMRI data is not spatially white, but it is not strongly spatially correlated, and its spatial covariance structure is complex since the long-range correlations arise mostly from physiological noise sources. For these two reasons, the theory of correlated random fields is not very useful for setting the thresholds for fMRI data. The only exception is if the fMRI time series data is artificially smoothed prior to signal detection analysis. This practice has been advocated [ref], but in our opinion this is misguided—it essentially makes fMRI data look like PET data, and sacrifices the much greater spatial resolution possible with MR imaging methods. The fact that makes fMRI data tolerable to analyze using the Bonferroni correction is that it is routine to gather many more data samples per subject than with PET (hundreds vs. tens)—this increases the statistical power of the voxel-wise tests enough that the Bonferroni correction is acceptable.

Resting State Data

One might not wish to trust the false positive rates calculated using some large set of statistical assumptions about the data—the necessary *p*-values are far out on the distributional curves and the usual invocation of the law of large numbers and asymptotic Gaussianity to justify the use of classical statistics may not be valid. A way around this problem is to gather a image time series of resting state data and apply an fMRI detection analysis method to it. By observing the number of positives (all false) as a function of the threshold setting, a reasonable value for the threshold can be selected.

The assumption behind this plausible technique is that the state of the subject does not change markedly between the resting state imaging run and a normal fMRI imaging run. This is clearly not entirely correct—there is likely to be more subject movement during a normal fMRI imaging run, and the subject's heart and respiratory rates may change as well. As we said earlier, it is not possible to calculate statistical significance without assumptions. If an investigator is willing to gather a resting state image time series, it can serve as a useful check on the assumptions underlying the statistical methods used to calculate the threshold. One set of assumptions (mathematical) is being traded for a completely different set (physiological and behavioral).

Randomization

Another practical technique for assessing the actual false positive rate compared to the theoretically predicted rate is to randomize the voxel values in time, then carry out the fMRI detection analysis [Bullmore]. The idea is to use actual data, which will include the many nonideal effects of spatial correlations and non-Gaussian noise, to determine the effectiveness of a processing strategy. The purpose of scrambling the data in time is to destroy the actual activation signal, so that whatever supra-threshold activations are detected are false. (In linear regression methods, scrambling the reference vectors in time is equivalent to scrambling the data in time, and much simpler.)

Temporal randomization will destroy any temporal correlations that are present in the fMRI noise. This will have the effect of falsely increasing the effective number of degrees of freedom in the time series, which will make the observed false positive rate be overly optimistic. If the correlations are relatively short range in time (as they usually are in fMRI), this objection can be overcome by randomizing short blocks of adjacent time points.

Randomization may seem similar to the idea of using resting state data, but it is somewhat different. Resting state analysis can only give an overall (whole brain) false positive rate, since one only has a single time series to use. Randomization generates many synthetic time series, and so it is possible to estimate a per voxel false positive rate. This can be important, since different brain regions have different noise characteristics. Randomization analysis takes a lot of computer time, since many different randomizations and analyses must be executed.