Absolute Beginner's Guide to Anatomical and Functional MRI of the Brain

Functional MRI Analysis

Robert W Cox, PhD
National Institute of Mental Health, Bethesda MD USA
robertcox@mail.nih.gov

♦ **Data Issues: Artifacts and Quality**

In all scientific fields, the primary data are subject to contamination which can severely bias the conclusions. In FMRI, the primary data are time series of 3D images, which can be grossly or subtly damaged by scanner and subject artifacts. Scanner artifacts should be minimized by carrying out daily quality checks with FMRI-like phantom scans, to make sure the hardware is functioning at its best and consistently [1]. One common scanner artifact is extra noise from faulty components in receive coils, which can take various forms (e.g., increase in variance, spatially coherent fluctuations, sudden steps in the data). The most common subject-induced artifact is head motion, which can only be corrected to a limited extent; movements that are correlated with task performance will induce signal changes that mimic the timing of neurally-induced signal changes, and this effect can overwhelm the "true" activation map with false positives. It is important that the consumer of FMRI data (e.g., *you*) be aware of what good data looks like so that s/he can promptly detect when problems arise.

♦ **Pre-processing FMRI Time Series Data**

Datasets usually undergo (at least) the following steps prior to analysis for neural activity [2]:

(a) Visual and automated checks for bad data—possibly including editing out of large spikes, and/or marking some time points to be censored out of the data analysis.

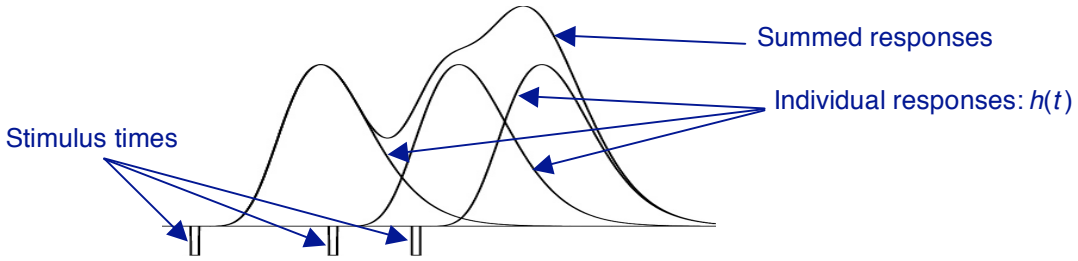
(b) Time shifting (interpolation) of the data so that each volume is at a fixed acquisition time—rather than having the slice acquisition times spread out over 1 *TR*.

(c) Registration of the time series of volumes amongst themselves—to reduce the effects of subject head motion—and to the T_1-weighted anatomical reference volume—possibly including also registration to a standard coordinate space.

(d) Smoothing of the images in space—to reduce noise, reduce the effective number of independent statistical tests that must be made, and to increase overlap with other subjects' results. (The default amount of smoothing applied depends on the software package.)

• **Task-Based FMRI: Pattern Matching in Time** [3]

Linear Convolution Model: The most commonly used analysis methods for task-based FMRI data are essentially pattern matching in time *in each voxel separately* (AKA "massively univariate analysis"), optionally followed by "blob" building in space (*infra*). The pattern that is being searched for in the data comes from combining the known stimulus/task times with a model of the hemodynamic response. The usual model is the linear summation of copies of a time-delayed and blurred *hemodynamic response function* $h(t)$ (HRF), one copy shifted to start at each stimulus/task time:



For example, if there are N_s stimuli that occur at times τ_1, τ_2, …, and the MRI-measurable response at t to an individual stimulus that occurs at time 0 is symbolized by $h(t)$, and the baseline signal is modeled as a constant plus a linear drift in time, then we get the model below

$$\underbrace{Z(t)}_{\substack{\text{voxel data}\\ \text{at time } t}} = \underbrace{\beta_0 + \beta_1 \cdot t}_{\text{baseline model}} + \underbrace{\sum_{s=1}^{N_s} h(t-\tau_s)}_{\text{FMRI signal model}} + \underbrace{\varepsilon(t)}_{\text{noise}}$$

The goal is to find out something about $h(t)$; for example, its amplitude, or its shape. Implicit in the model is the assumption that the responses to repeated stimuli in the same class are the same. This assumption (or something like it) is necessary to be able to analyze repeated stimuli, which are necessary to get decent statistics about $h(t)$—a single activation cycle ("up and down") is not enough to give a decent activation map, so repeated stimuli are needed. If there is more than one class of stimulus, then each class q would get its own hemodynamic response function $h^{(q)}(t)$.

Fixed Shape HRF: In these models, we assume $h(t)=\alpha \cdot r(t)$, where α is unknown and $r(t)$ is some reference function we choose; Mark Cohen's gamma-variate function $r(t)=t^b e^{-t/c}$ for $t>0$ is a popular shape (e.g., b=8.6 and c=0.547; the time delay to the peak is $b \cdot c$ and the FWHM of the peak is approximately $2.4 \cdot c \cdot \sqrt{b}$ for $b>1$). These models have the advantage of having fairly few parameters per voxel: one $\alpha^{(q)}$ for each stimulus class q, plus the baseline parameters (β_0 and β_1 in the models above). The b and c parameters are fixed in these types of models, and are often assumed to be the same for all subjects. A refinement is to separately estimate b and c for each individual using a simple motor or visual FMRI paradigm, prior to the more complicated experiment that you are undoubtedly contemplating.

The BOLD response to a brief stimulus (e.g., a 100 ms flash of light) typically lasts about 10-12 seconds, comprising a 2 s delay, 3-5 s rise and a 4-5 s fall. For long values of TR (3 s or more), using a fixed shape HRF makes a great deal of sense: there isn't enough temporal resolution to try to capture the shape.

Variable Shape HRF: In these models, more parameters are added to the unknown function $h(t)$ in order to let its shape vary. There are two principal motivations: first, to fit the data $Z(t)$ better in each voxel so that the statistical significance of activation is properly assessed; and second, to allow statistical inference on the shape of $h(t)$ itself (e.g., is the activation amplitude stronger over the 4–8 s post-stimulus range or over 8–12 s post-stimulus?).

For example, the widely-used standard SPM variable shape HRF model has the form $h(t)=\alpha_1 \cdot r_a(t)+\alpha_2 \cdot r_a'(t)+\alpha_3 \cdot r_b(t)$, where $r_a(t)$ is of Cohen's form, $r_a'(t)$ (the temporal derivative) allows for small time changes, and $r_b(t)$ is present to allow for small changes in the FWHM of the response. (If there is more than one stimulus class, each class requires a separate set of three α parameters.)

More complicated models (e.g., polynomial, spline, or trigonometric series) allow for more shape flexibility, and have been used to allow for potentially interesting inter-regional HRF shape differences. It is important to note, however, it can become impossible to find activation (i.e., state confidently that $h(t)$ is nonzero) when there are too many parameters for the data, since a very high-dimensional model will fit a pure noise voxel almost as well as it fits a signal+noise case. Similar problems arise when the number of stimulus classes is increased; again, the number of parameters increases (a few α's per q), and it is easy to go too far.

It is best to start with a simple analysis of FMRI time series data, see if the results make sense, then progress to the use of more complicated models to extract more information. In this way, the data analyst can get a feel for how many parameters can be estimated from the datasets. One common conceptual mistake is to group the stimuli into too many classes, so that there are relatively few (under 15–20, say, in an event-related design) responses per class. FMRI datasets are not good enough to reliably assess differential activation between tasks when there are so few samples per task.

Inverse Models: Instead of solving for $h(t)$ in each voxel, one can assume a fixed $h(t)$ and then solve for the stimulus time series $s(t)$ that best fits the data in each voxel. This approach has the potential for finding neural activation patterns for complex continuous stimuli such as video or audio presentations. Such inverse models have not been widely applied, partly because they involve a large number of parameters per voxel (for fitting a general function $s(t)$ to the data $Z(t)$).

Statistical Inference from Linear Models: Under the assumption that the noise is Gaussian and has a known (or estimated) temporal correlation structure, the statistics of linear models are fairly straightforward. The unknown parameters are estimated using a least-squares fitting criterion (e.g., minimize $E = \sum_t \left| Z(t) - \text{model}(t) \right|^2$). The magnitude of E is used to estimate the variance of the noise. From these estimates, the significance of linear combinations of the model parameters can be calculated directly using F- or t-statistics (F-tests for multiple combinations of parameters, t-tests for single combinations). For example, in the fixed shape model, where the only parameter of activation interest is α, the test gives the likelihood p that $\alpha=0$ given the data (and the model assumptions). If this p-value is sufficiently small, we declare this voxel to be "active" and colorize it somehow (usually, the color is based on the amplitude α, but is sometimes instead based on the F- or t-statistic). If we had two stimulus classes and so estimated $\alpha^{(1)}$ and $\alpha^{(2)}$ as the response magnitudes for each type of stimulus, we could determine the p-value for the null hypothesis $\alpha^{(1)}-\alpha^{(2)}=0$; we would presumably colorize voxels in which this p was small (indicating that $\alpha^{(1)}\neq\alpha^{(2)}$) *and* the p-values for $\alpha^{(1)}=0$ and/or $\alpha^{(2)}=0$ were also small—these would be locations where the brain responded to at least one of the types of stimuli *and* responded differently to the two different stimulus types. And so forth—this kind of test is sometimes called a *conjunction analysis*. The limits of this type of inference are your imagination. And the quality of the data.

Nonlinear Models? There is nothing wrong with using nonlinear models for FMRI time series; for example, one could directly solve for the (b,c) parameters in Cohen's $r(t) = t^b e^{-t/c}$ model, in each voxel. The practical drawback to nonlinear models is the difficulty of solving the fitting equations for the parameters. Linear models have the strong advantage that the least-squares criterion leads to linear equations for the unknown parameters; efficient algorithms for solving such equations have been well-established since the 1960s. The same cannot be said for solving nonlinear optimization problems. Nevertheless, nonlinear models

have some attractive features, such as providing the ability to impose constraints on the shape of the expected response. Nonlinear regression has been used, for example, to fit the response to pharmaceutical FMRI challenges (e.g., injection of cocaine).

♦ **Spatial Models of Activation (e.g., "blobs")**

The most common form of FMRI data analysis is voxel-wise: each voxel time series $Z(t)$ is analyzed separately from all others. The attraction is that the full spatial resolution of the echo-planar images is kept. However, we probably wouldn't accept a brain activation map that consisted solely of randomly scattered "on" voxels with no clear spatial structure, no matter how strong the statistics were; instead, we'd go back to the data and try to figure out what went wrong. But if we aren't going to accept an arbitrary spatial map, then we can increase our statistical power by only looking for spatial activation patterns that are "reasonable". There are three commonly-used methods.

Smoothing: One of the simplest ways to produce "reasonable-looking" activation maps is to smooth the FMRI data spatially prior to the temporal analysis (or maybe after analysis). If a 10–15 mm FWHM Gaussian blur is used for this smoothing, for example, then FMRI results can be made to look much like PET results. The drawback to such simple smoothing is obvious: why bother to acquire high-resolution images if the first thing one does is to throw that resolution away? "Smart" smoothing is a variation that only does blurring within the brain volume, or within the gray matter (e.g., as detected from a T1-weighted volume with ≈1 mm resolution). This technique uses the high resolution of FMRI cleverly.

Clusters: A second way to produce "reasonable-looking" activation maps is via a dual-thresholding technique. After a voxel-wise time series analysis, voxels in which $h(t)$ is significantly different from 0 are selected; this first significance threshold is taken to be low, so that a fair number of false positives can be expected. The second thresholding step is to only accept contiguous clusters of voxels that passed the first step; the threshold here is the minimum allowable cluster size. This technique allows for the detection of relatively small amplitude activations, provided these activations cover a large region. Generally, the significance of this dual-threshold technique can't be calculated exactly; instead, it must be approximated using an asymptotic formula (valid for large amounts of smoothing) [4], or by simulation of noise-only images and the detection process (to see how often clusters of a certain size could arise from signal-free data) [5].

Regions of Interest (ROIs): A third method is to pre-select voxels that are to be averaged together; the selection is usually based on some anatomical criterion (e.g., the left hippo-campus). This technique has the advantage that specifically targeted anatomical hypotheses can be addressed precisely, and that the regions can be tailored to each subject's anatomy. It has the disadvantage that intra-ROI differences can be lost (e.g., anterior vs. posterior hippocampus, if the ROI averages over the entire structure). The rise of various automated techniques for parcellating the brain (e.g., FreeSurfer and Automated Anatomical Labeling) make this technique much less laborious than in the past, where manual tracing of the ROIs was required.

Statistical Inference: One major point of using spatial models is that they reduce the multiple comparisons problem. In the case of ROI analyses, it is often the case that only 10-20 ROIs are used; the problem of dealing with 50,000-200,000 comparisons has been anatomically abstracted away. In the case of smoothing, the effective number of comparisons is reduced in proportion to the amount of image blurring carried out.

♦ **Pattern Hunting Analyses**

A very different type of analysis eschews looking for fixed (or parameterized) temporal patterns and instead tries to break an FMRI 3D+time dataset up into a set of component patterns that together explain most of the measurements. The tool for this is dimensional factorization:

$$\underbrace{Z(\mathbf{x},t)}_{\substack{\text{data}\\ \text{at voxel } \mathbf{x}\\ \text{and time } t}} \approx \sum_k u_k(\mathbf{x}) v_k(t)$$

where the $u_k(\mathbf{x})$ are the patterns representing the spatial distribution of amplitude for each temporal component $v_k(t)$. One attraction of this type of analysis is that it does not require the data to fit a pre-conceived model, so that if something un-allowed for in the standard analysis happens, it might be detected (e.g., the amplitude of the response drifts down over a large region with task repetition). Of course, such freedom makes it all too easy to over-interpret the results.

Different methods are used to find the components. If "≈" is interpreted as "least squares", then the factorization above is called Principal Component Analysis (PCA), and is essentially a singular value decomposition of the data. A more elaborate technique is to search for maximum "independence" (statistical unpredictability) among the spatial components $u_k(\mathbf{x})$: this leads to the various flavors of Independent Component Analysis (ICA) [6].

♦ **Network Analyses—The Connectome!**

An important theme in recent years has been the analysis of FMRI data to reveal networks of coordinated brain activity. In task-based FMRI, the idea is to look for coherent fluctuations in the BOLD signal that ride on top of the average activation; in other words, when the per-stimulus activation magnitude is correlated between brain regions. Roughly speaking, these forms of analysis fit some mathematical/statistical model of correlated signal fluctuations to ROI-averaged time series; ROIs are generally used here since the per-voxel level magnitude of these coherent fluctuations is small (these are just a fraction of the BOLD signal, most of which is in the average). There is a wide and confusing variety of methods used for such connectivity analyses, including

- Simple correlation
- Context (e.g., task) dependent correlation—also called Psycho-Physiological Interaction (PPI)
- Dynamic Causal Modeling (DCM)
- Granger Causality (GC)
- Structural Equation Modeling (SEM)
- Structural Vector Autoregression (SVAR)—a technique which combines the features of GC and SEM

The more advanced methods (DCM, GC, SVAR) attempt to infer directionality (causality) in the connections. Some of these methods can also be applied to resting-state FMRI (*infra*).

♦ **Noise Statistics**

The noise in FMRI time series data is complicated, both spatially and temporally. In time, nearby data values have correlated noise values; these correlations are primarily caused by

physiological effects. Respiration and cardiac effects can be large, depending on the brain region, the pulse sequence, and the scanner itself. It is possible to externally monitor breathing and the heartbeat and attempt to filter (regress) these physiological components out of the time series—this usually helps improve the data a little, but the effects are not dramatic.

Allowing for temporal noise correlation in the data analysis and statistics can be important, particularly when carrying out single-subject analyses (e.g., pre-surgical planning), and when using the estimated noise statistics at the group analysis level (*infra*). In the context of least-squares regression, the adjustment for noise correlation is called "pre-whitening", and requires an estimate of the temporal correlation structure of the noise. Different techniques are used to generate these estimates, and it doesn't seem to matter too much which method is adopted.

Physiologically-generated noise will usually be spatially correlated as well. The usual technique for smoothing FMRI data just adds smoothness to the existing data in an uncontrolled way. If the amount of smoothness added is large ($\geq$ 7 mm, say), then this addition probably overwhelms any intrinsic smoothness in the data. However, it is possible to add smoothness in a controlled way, in small steps, and testing the results at each stage to determine how much to add in the next step in each region; in this way, the data can be blurred so that the noise arrives at the desired level of smoothness across the brain.

♦ **Analyses of Resting-State FMRI—More Connectome!**

Neural activity takes energy, leading to the BOLD effect (through somewhat-unknown physiological processes). Even undirected "random" neural activity gives rise to measurable MRI signal changes. However, with no imposed (or observed) task, there is no external reference to use to find these signals, which otherwise look very much like temporally correlated noise. (Aside: *pace* Bob Turner, this form of research should really be called "flat on your back in a very loud metal tube FMRI".)

The goal of resting state FMRI analyses is to find regions of the brain where the spontaneous fluctuations are similar—that is, to find regions that are involved in some network of coherent activity. The usual measurement of "similar" is the correlation coefficient; as a result, the "raw data" is the 6D correlation function $\rho(\mathbf{x},\mathbf{x}')$ for all voxel pairs $(\mathbf{x},\mathbf{x}')$. Since there are typically about 10^5 voxels in a 3D brain image, there are about 10^{10} entries in $\rho(\mathbf{x},\mathbf{x}')$, which is a much bigger pile of numbers than most people want to deal with at once. So various methods are used to summarize the overall correlation function: various component analyses, seed-based approaches using only a small collection of $\mathbf{x}$ locations, collapsing to a 3D map by averaging (in some way) over $\mathbf{x}'$, et cetera.

Spatial correlation in the non-neural noise is a problem for resting state FMRI analysis, since spatial correlation is the "signal" for this type of data. ICA methods can be used to filter out some of the correlated noise, but not all: for example, respiration artifacts can mimic the "default mode" network, so that the spatial "footprint" of the neural network and the artifact are not independent. Spatial correlation multi-channel RF coil images can also be a problem.

♦ **Group Analyses**

Most FMRI studies aim to make some inference about a group of subjects (representing a larger population), or about differences between groups (e.g., "normals" and patients). In principle, all the time series data from all subjects could be analyzed together to get the final group activation (and activation difference) maps, but this is seldom done. Practical difficulties in dealing with perhaps 50 subjects time 1 Gbyte of data per subject make such a

computation unwieldy. Instead, the nearly universal method of group analysis is to carry out the time series analysis on each subjects' data individually, then taking the statistical results (the un-thresholded "activation" maps) from this first-level to a second-level statistical analysis. A simple and common example would be to carry out a *t*-test on the activation magnitudes from each subject, in each voxel.

Most second-level tests are "mixed effects" analyses, "mixed" here meaning that some parameters to be estimated measure randomness in the first-level results (among subjects, usually) and some parameters to be estimated are assumed to be constants (e.g., the mean difference in activation magnitude between controls and patients). At the group level, the simpler forms of analysis (e.g., *t*-tests, ANOVA, ANCOVA) simply take the estimated activation amplitudes from the first level and treat these as the "data" for the second level; these techniques (mostly) amount to carrying out linear regression on the first-level amplitudes. A more complicated form of analysis also takes the estimated *variances* of the first-level amplitudes (in each subject, in each voxel) and uses these in the second-level estimation process. This technique, called "meta analysis" in the statistics literature, allows the uncertainty in the first level results to be carried up to the second level, which in turn will result in down-weighting data that is less reliable. Mixed effects meta analysis has been reported to produce more reliable results from studies that have weak power (e.g., relatively few subjects, or weak activations).

Spatial clustering is almost always done only at the final analysis step—in a group study, the input to the second-level analysis is almost always the un-thresholded maps from the individual subjects. (Sometimes the group analysis is only done on ROI-averaged data from the first-level, but it is more common to do whole-brain group analyses and then later select ROIs or clusters.)

Independent component analysis has been applied to group analysis, of course, including methods that analyze (reduced) time series data from all subjects into a 3-way component decomposition: space×time×subject.

♦ Software Tools

There is a large number of free software packages for FMRI data analysis. In general, the website **http://www.nitrc.org/** provides a repository for many such packages. A few of the widely used packages are:

- SPM = a comprehensive package for FMRI (and more) data analysis; the most widely used such software; open-source Matlab: **http://www.fil.ion.ucl.ac.uk/spm/**
- FSL = another such package; open source C++: **http://www.fmrib.ox.ac.uk/fsl/**
- AFNI = another FMRI package: open source C: **http://afni.nimh.nih.gov/afni**
- FreeSurfer = a specialized package for cortical surface reconstruction and brain parcellation; distributed (freely) in binary formats: **http://surfer.nmr.mgh.harvard.edu/**
- BrainVoyager = a commercial (non-free) package for FMRI data analysis; this is included here because it is widely used: **http://www.brainvoyager.com/**

The choice of which comprehensive package to use might seem to be difficult—they are built around somewhat different assumptions and have very different interfaces. For the "Absolute Beginner", it is probably best to go with whatever the people around you use—that way you can get help locally by pestering your colleagues, rather than have to rely on e-mail lists and message boards. If you are isolated on an desert island (with an MRI scanner and Internet access), then of course I recommend AFNI: after all, I wrote it!

♦ **Final Injunctions**

There are many ways to analyze FMRI datasets.

- Even the linear shift-invariant model described above has many variations, each of which would give different results—only slightly different, we would hope. Fortunately, the BOLD effect is generally robust enough that most straightforward activation analyses give similar results in most cases.

FMRI-based investigators need to be aware of the different techniques, their underlying assumptions about the FMRI signal and noise, their strengths and limitations, and their applicability to any given experimental situation.

- One might hope that FMRI data analysis would be a "one size fits all" situation, but things aren't so pleasant or simple. Just as the experiment design must be tailored to explore your underlying hypotheses, so the data analysis must be tailored to answer the questions that you pose from the data that you have. And there is no "best" analysis method for any given collection of data and hypotheses/questions to test; instead, there are "reasonable" analysis method***s*** to consider, and there is no practical way to decide which one is "right" or "best". If this bothers you, then try more than one method on your data and see if the results differ much.
- People argue about the underlying assumptions that are explicit or implicit in the various methods that are outlined above, debating both the validity of the assumptions and their importance. To judge among these methods and arguments requires (at least) a conceptual appreciation of the techniques.

In other words: understand what you are doing. Or you will do something stupid.

- FMRI analysis is a complex and powerful tool, and "with great power comes great responsibility." In this case, the responsibility is to understand.

♦ **More Fun Things to Read!**

1. Report on a multicenter fMRI quality assurance protocol. L Friedman and GH Glover, *Journal of Magnetic Resonance Imaging* **23**: 827-839 (2006) **http://onlinelibrary.wiley.com/doi/10.1002/jmri.20583/full**
2. Evaluating fMRI preprocessing pipelines. SC Strother, *Engineering in Medicine and Biology Magazine* **25**: 27-41 (2006). **http://freylab.uoregon.edu/MR%20notes/strother_ieee_2006_preprocessing.pdf**
3. *Human Brain Mapping* **27**: issue 5 (2006). Special issue containing discussion of various analysis packages as applied to the same set of data (FIAC = Functional Imaging Analysis Contest). **http://onlinelibrary.wiley.com/doi/10.1002/hbm.v27:5/issuetoc**
4. Developments in Random Field Theory. KJ Worsley, Chapter 15 in *Human Brain Function* **http://ftp.fil.ion.ucl.ac.uk/spm/doc/books/hbf2**
5. Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold. SD Forman, JD Cohen, M Fitzgerald, WF Eddy, MA Mintun, DC Noll, *Magnetic Resonance in Medicine* **33**: 636-647 (1995). **http://onlinelibrary.wiley.com/doi/10.1002/mrm.1910330508/abstract**
6. ICA of Functional MRI Data: An Overview. VD Calhoun, T Adali, LK Hansen, J Larsen, and JJ Pekar. **http://icatb.sourceforge.net/gift/publications/2003_ica_overview.pdf**