# Statistical Inference in Functional Magnetic Resonance Imaging

Christopher R. Genovese

*Carnegie Mellon University*

## Under Review

See also Department of Statistics Technical Report 674 which contains extensive additional material. It can be downloaded from `http://www.stat.cmu.edu/~genovese/papers/fmri/mrmodel.ps`.

**Running Head:** Statistical Inference in fMRI

**Keywords:** Magnetic Resonance, Functional Neuroimaging, Spatio-temporal inference

**Acknowledgments:** This work is supported by NSF Grants DMS 9505007 and the Center for the Neural Basis of Cognition. The author would like to thank Pat Carpenter, Marcel Just, Doug Noll, and John Sweeney for providing the data used in this paper and for helpful discussions. The author would also like to thank Rob Kass, Bill Eddy, and Marsha Lovett for helpful discussions and comments.

**Correspondence:**

Christopher R. Genovese
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

| | |
|---|---|
| Phone: | (412) 268-7836 |
| Fax: | (412) 268-7828 |
| E-mail: | `genovese@stat.cmu.edu` |
| Home Page: | `http://www.stat.cmu.edu/~genovese/` |

# ABSTRACT

Functional Magnetic Resonance Imaging (fMRI) is a new technique for studying the workings of the active human brain. During an fMRI experiment, a sequence of Magnetic Resonanc images is acquired while a subject performs specific behavioral tasks. Changes in the measured signal can be used to identify and characterize the brain activity resulting from task performance and thus help to understand how higher cognition emerges from the brain's architecture.

The data obtained from an fMRI experiment are a realization of a complex spatio-temporal process with many sources of variation, both biological and technological. The noise is complicated, and the task-related signal changes are small in amplitude. Here, we describe a new and detailed statistical model for fMRI data and present inferential methods that enable investigators to directly target their scientific questions of interest, many of which are inaccessible to current methods. Our model allows for the complexity of the noise process, flexibly parameterizes the task-related signal changes, and allows for non-linearity and non-additivity in the system response.

## 1. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a rapidly developing tool that enables cognitive psychologists and neuroscientists to study the human brain *in action.* During an fMRI experiment, a subject performs a carefully choreographed sequence of behavioral tasks while Magnetic Resonance (MR) images of the subject's brain are acquired at regular intervals. The tasks are designed to exercise specific motor, sensory, or cognitive processes, and the measured MR signal contains information about the nature and location of the neural activity that results when those processes are engaged. Psychologists hope to use MR data to build and test theoretical models of human cognition, but to do this, they must solve a quintessentially statistical problem. Our goal in this paper is to put forward a new standard of inference for fMRI data. We develop a model and inferential framework with three key objectives that are not met by current methods:

- Account for as many important features of the data as possible,
- Address directly the scientific questions of interest with rigorous statistical procedures,
- Allow for natural extension to encompass new results on the underlying physical processes.

An fMRI experimental design specifies the tasks the subject is to perform along with their timing and duration. Each distinct task the subject performs defines one experimental condition. A task condition is usually replicated several times during the experiment, and each distinct block of time during which a given condition applies is called a task *epoch* within that condition. Figure 1a illustrates the design of a very simple experiment with two task conditions: maintaining visual fixation on a marked point in the center of a projection screen and steadily tapping index finger against thumb while maintaining this visual fixation. Every epoch in this design involves repeatedly performing the specified task for 24 seconds or 8 image acquisitions. The goal of this experiment is to study the pattern of activity produced by the simple motor function of finger tapping. The fixation-only condition serves as a control: brain activity in response to finger tapping with fixation but not to fixation alone is attributed to the neural processing evoked by tapping. To enable this "subtractive" logic, the two conditions are designed to differ only by the involvement of the motor function under study. Much more complex designs for answering more intricate questions can be based on the same considerations.

The principal data from an fMRI experiment are a sequence of three-dimensional images of the subject's brain. Each image consists of measurements of the MR signal over a grid of small, regular volume elements, called *voxels*. The MR signal for a voxel in an image is related to the density of a particular nuclear species, usually hydrogen, averaged over the voxel and a small time interval. Taken together, the sequence of MR images yields a time series of measurements for each voxel. By varying the acquisition scheme, we can select the voxel size, with voxels typically ranging from 3 to 50 cubic mm and with corresponding images containing from 400,000 to 100,000 voxels. There is, however, a trade-off between spatial and temporal resolution: images with large voxels can be acquired more quickly than images with small voxels. The time to acquire an entire image typically ranges from less than 1 second to over 8 seconds, depending on the acquisition scheme.

An MR image reveals the anatomical structure of the brain, but in *functional* MRI, we are not interested in the images *per se* but rather in small, systematic changes in the measured MR signal over time. These changes are caused by a localized blood-flow, or hemodynamic, response in the brain to concentrated neural activity during performance of the experimental tasks. This hemodynamic response is detectable in the images because of the different magnetic properties of oxygenated and de-oxygenated blood [58], which gives rise to the Blood Oxygenation Level Dependent (BOLD) effect [49, 6, 44]. Figure 2 shows the time series of measurements from two voxels in the finger-tapping study. For the first voxel, there is a systematic rise in the signal during tapping but not fixation; no such task-related signal changes are evident for the second voxel.

The statistical problem in fMRI is to identify and characterize the task-related signal changes in a way that helps scientists test their models and predictions about how the brain works. Current methods for analyzing fMRI data attack what we call the *localization problem*: How can data from an fMRI experiment be used to identify the parts of the brain that activate during performance of a given brain function? Localization is viewed in fMRI as a problem of classification, where each voxel is classified as *active* or *inactive* with respect to a comparison between two conditions. The assumption is that in an active voxel there is greater neural activity in response to one condition than to the other; hence, that voxel is believed to play a role in any brain functions required by the former condition but not the latter. Thus, in a well-designed fMRI experiment in which the conditions being compared differ only with respect to the process under study, it is possible to localize very specific brain functions.

Most current methods analyze the fMRI data one voxel time series at a time, treating all voxels as independent. Although other classification procedures have been considered [62, 51], the majority of methods use a statistical hypothesis test to identify active voxels. A wide variety of testing procedures has been employed in fMRI. The simplest and still most widespread is the two sample t-test in which the average MR signal values from images associated with any two experimental conditions are compared. Figure 3 shows one slice of a mean image for the tapping versus fixation experiment with the voxels classified as active by a t-test marked in white. The active voxels are those for which the mean MR signal during the tapping epochs is “significantly” larger than the mean signal during the fixation epochs. Other frequently used tests include the Kolmogorov-Smirnov test [4, 46], and the split sample t-test [24], which classifies a voxel as active only if separate t-tests for the first and second halves of the data both indicate significance. Two generalizations of the t-test attempt to account for features that the t-test treats simplistically. The first is to test for a non-zero correlation coefficient between the voxel time-series and a fixed reference curve designed to approximate the shape of the hemodynamic signal perturbation [5]. Whereas the t-test implicitly uses an ideal square wave for the hemodynamic signal change, this test allows a more realistic shape for this curve. The second is to use an Analysis of Variance (ANOVA) model to block on time and thus account for uncontrolled changes over the course of the experiment [12, 13]. More recent methods attempt to get the best of both of these by using a general linear model to capture temporal variation, possibly after spatial and temporal smoothing [26, 66]. For experimental designs that alternate periodically between two conditions, spectral methods (e.g., F-tests based on periodogram ordinates [9]) make it possible to test for large power in any frequency band [63, 25], particularly the epoch frequency and its harmonics.

It is tempting to begin an assault on the statistical problems of fMRI by trying to improve the testing procedures used for classification, but there are two limitations of the current paradigm that need to be addressed: the complexity of the data and the richness of the scientific questions. The data obtained from an fMRI experiment are a realization of a very complex spatio-temporal process with many sources of variation, both biological and technological. First, the noise in the data is complicated: there are nonlinear signal drifts, non-homogenous (and often non-local) correlations across space and time, heavy tails, and a variety of phenomena caused by subject movement, physiological effects, and instrument artifacts. The noise distribution also depends on the image acquisition scheme. Second, the task-related signal changes are generated by a hemodynamic response that varies over the brain and has both non-linear and non-additive features. The problem is further complicated by the highly irregular tissue boundaries in the brain and the presence of confounding factors such as large blood vessels.

While the various hypothesis tests and classification procedures often give reasonable results and can reliably detect large signal changes, the simplistic assumptions underlying these methods introduce non-trivial inefficiency into the inferences. An interesting illustration of what can go wrong lies in the choice of thresholds for such hypothesis tests. The theoretical values, even after correction for multiple comparisons, are generally too small, leaving no clear system for choosing the thresholds [32]. This suggests that the tests are not accounting for all relevant sources of variation. Another limitation of the classification approach is that, as new information comes to light regarding the processes generating the data, the classification procedures need to be continually replaced.

These problems have evoked two responses in the fMRI literature: (1) the search for new procedures for classification and (2) the development of pre-processing algorithms to "correct" the data for artifacts prior to analysis, e.g., linear detrending to correct for signal drifts. Neither approach addresses the basic problems just mentioned, and both leave many unaccounted sources of variation. While it is true that the results obtained thus far are often reasonable and that a good deal has been learned, reasonable is not enough to support the strong claims for which the classifications are being presented as evidence. More recent efforts have attempted to deal with some of these problems. Forman et al. [23] adjust voxelwise tests by accepting as active only those voxels that lie in a cluster of voxels of at least a specified size with test statistic above threshold. Worsley [67, 65] uses the distributional properties of the extremes of random fields to select a global threshold for voxelwise tests. Holmes *et al.* [37] propose randomization tests to protect against violations of the assumptions while maintaining a nearly specified type I error for testing the omnibus hypothesis. Lange and Zeger [45] develop a model in the spectral domain for the voxel time series. Their model allows variation in the hemodynamic response and their inferences are based on tests with parameter estimates under the model. See also [31, 20] for further discussion.

Despite these improvements, none of the current methods overcomes the second and most important limitation of the localization paradigm: only a severely restricted range of questions can be directly addressed by localization methods. The classification procedures currently used in fMRI are based on localization and as such target only a single question: where did the activity occur? Localization is an important step; however, fMRI is being applied to a diverse set of scientific questions beyond localization, questions about more complex relationships in the pattern of responses. For example, the primary interest of cognitive psychologists centers on testing and refining their cognitive theories, a goal for which localization is of only indirect interest. As these scientists gain experience with fMRI and as the technology develops, the experimental designs are becoming sophisticated enough to tackle new classes of questions like the following:

- Graded Responses. Many theoretical predictions revolve around how the system response changes as the input is varied along a single dimension, such as task difficulty. The data contains information about response magnitude, but estimating the specific form of the relationship between input and response requires that the response be quantified along a continuous scale.
- Dissociations. Cognitive theories often make a distinction between different types of processing (e.g., spatial and symbolic) to the point of predicting a separate neural implementation. One piece of evidence in support of such a theory is called a dissociation, a situation in which two types of processing yield qualitatively distinct responses. Functional MRI can provide evidence for dissociations, not only in the location of responses but also through the temporal pattern of responses and through the effect of input manipulations on the nature of the response.
- Functional Connectivity. A brain system consists of a distributed hierarchy of specialized components that combine to produce some complex function. In many cases, scientists have identified the specific function of a system's components but would like to know how these components interact. To characterize the connectivity of such a system, it is necessary to learn about the flow of information among the components, as specified by the relative magnitude and timing of activation responses.

In each of these cases, a rich set of features describing the response is essential to answering the question, but classification methods do not make these features accessible. As a consequence, investigators are left to make *ad hoc* interpretations of the classification results in order to find support for their predictions. A generic example that illustrates the scope of the problem is the common practice of comparing conditions through counts of active voxels (relative to a control) derived from the classifications. These counts estimate the extent of the set of active voxels for each target function, but as taken from the classifications, there is no useful measure of uncertainty associated with these estimates. Hence, the comparison is made with no baseline against the likelihood of chance fluctuations. If we try to construct a hypothesis test for the difference in the counts, we encounter a highly composite null hypothesis, and the natural test statistics (which are not functions of the classification results) have intractable null distributions in general. To move the science forward, the statistical methodology must advance to bridge the gap between the information in the data and the questions scientists want answered.

We take a new approach to inference from fMRI data, moving away from testing and classification towards estimation under a model for the data, and towards inferences that are directly relevant to the scientific questions of interest. Our inferences are based on a detailed, nonlinear hierarchical model for the data that represent the components of variation so that the model parameters are meaningful with respect to the underlying processes generating the data. The result is an inferential framework that offers greater sensitivity and wider inferential scope than current methods. At the data level, our model decomposes the measured signal $Y_v(t)$ at location $v$ and time $t$ into several interpretable pieces

$$Y_v(t) = \mu_v + d_v(t) + a_v(t\,;\ \mu_v, \boldsymbol{\gamma}_v, \boldsymbol{\theta}_v) + \epsilon_v(t), \tag{1}$$

where each term captures a distinct component of variation in the process. Constraints on these terms are based on the available substantive information and play a critical role in identifying the model components. The model parameters are related across deeper levels of the hierarchy.

We describe the basic structure of our modeling framework in section 2 and show how we incorporporate the available prior information into the model in section 3. While the model improves the accuracy of inferences made from the data, what is more important is the range of questions that can be addressed directly under the model. In section 4, we illustrate some of the possibilities in light of a specific data example. In section 5, we describe the computational techniques that we use to fit the model and implement our inferential procedures. Finally, in section 6, we discuss some more general extensions to the model and some directions for future research.

## 2. Modeling fMRI Data

Let $Y(t)$ be the observed MR signal at time $t$ from a specific voxel, where $t = 0, \Delta, \ldots, T\Delta$ for sampling interval $\Delta$. Our voxelwise model decomposes this time series into four distinct components

$$Y(t) = \mu + d(t) + a(t\,;\ \mu, \boldsymbol{\gamma}, \boldsymbol{\theta}) + \sigma\,\epsilon(t), \tag{2}$$

where $\mu$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, and the function $d()$ are model parameters and $\epsilon$ is a parameterized noise process with mean 0 and variance 1. This equation defines the likelihood for the model; the specification of deeper levels in the hierarchy is given in Section 3. The four additive components in equation (2) will be called the baseline signal, drift profile, activation profile, and noise, respectively. Below, we clarify the role and parameterization of each component.

### 2.1. Baseline Signal

The real-valued parameter $\mu$ in equation (2) represents the magnitude of the baseline signal at the given voxel, defined as the mean signal over time in the absence of activation and noise. The baseline signal can vary by an order of magnitude across the imaged volume. While some of this variation reflects differences in nuclear density across the brain tissue, much of it arises from other sources, including differential position and orientation with respect to the receiver electronics, inhomogeneities in the receiver electronics, and local magnetic anomalies in the tissue. Nonetheless, the baseline $\mu$ is usually well-determined from the data.

### 2.2. Drift Profile

The measured MR signal at a voxel tends to drift over the course of an fMRI experiment, and the magnitude of these changes often far exceeds both the ambient noise level and the amplitude of the task-related signal change. The drift profile $d(t)$ in equation (2) represents signal drift as a function of time. We treat the function itself as the model parameter here, within a potentially complicated space of smooth functions. Our empirical study of many fMRI data sets suggests that the drift profile has the following basic properties. It tends to be smooth but can undergo occasional rapid, localized changes. The drift profile also exhibits a diverse range of shapes, often highly nonlinear and heterogeneous over time. See Figure 4. Much of the drift variation seems to be of biological origin since the largest changes and most interesting features of the drift only arise when imaging living tissue. Moreover, the drift has interesting spatial structure: its shape and magnitude can vary greatly across voxels, and even neighboring voxels can display completely different behavior.

It is this diversity in shape that poses the main challenge to modeling the drift. We want a flexible parameterization that allows a range of functional forms consistent with the data, but we also must discourage spurious structure, particularly structure that might be confounded with the task-related signal changes. We parameterize the drift profile as a function on $[0, 1]$ and rescale it onto the time interval of the experiment. We also constrain $d(t)$ to be orthogonal to constants (and thus the baseline) with respect to the empirical inner product, $\sum_{t=0}^{T} d(t) = 0$. This maintains the conceptual separation between the baseline and drift. Although a natural starting point would be to model $d(t)$ with low degree polynomials, we have found that these generally lead to a poor fit to the data because the drift frequently changes its character over the course of the experiment. Instead we take $d$ as belonging to a space of splines of some degree $D$.

A spline of degree $D$ is determined by the number and position of its knots and coefficients in an associated basis of functions [17]. Let $0 \leq K \leq K_{\max}$ denote the number of knots and $\boldsymbol{\kappa}$ denote the vector of knot locations, where $0 < \kappa_1 < \cdots < \kappa_K < 1$. We consider two strategies for constructing the splines that offer different ways to balance the conflicting goals of diversity and parsimony: (i) use a small number (e.g., 1-4) of adaptively placed knots and (ii) use a large number (e.g, one per time point) of fixed knots with regularization to eliminate spurious structure. In the former case, both the number of knots and the knot positions are parameters in the model, and $d$ thus varies across a union of standard spline spaces. By keeping the number of knots small relative to the number of data, spurious structure is discouraged and the drift profile does not become confounded with the activation profile. The latter case corresponds to a modified smoothing spline [61], where the curve is overparameterized but the regularization (prior) enforces smoothness. The knots for a smoothing spline are usually placed at every data point, but sufficient flexibility is often gained with 1/2 to 1/4 as many depending on the length of the time series. These strategies for modeling the drift are both quite effective and require very different algorithms for computation as described in Section 5. The adaptive strategy is computationally more costly but does a somewhat better job capturing sharp changes in the profile without excessive wiggliness.

Given $K$ and $\boldsymbol{\kappa}$, the splines with those knots form a vector space, and a drift profile $d(t)$ uniquely determines a set of coefficients $\boldsymbol{\delta}$ in any basis for this space. Note that the choice of basis is arbitrary, for the true parameter here is the drift profile $d(t)$ itself. If we change the basis, the coefficients change but the profile remains the same. One possible choice is the power basis, defined by the set of functions $1, t, \ldots, t^D, (t - \kappa_1)_+^D, \ldots, (t - \kappa_K)_+^D$. The power basis is neither numerically well-conditioned nor structurally convenient, so we reparameterize in one of two ways. The first is to use a basis generated by orthonormalizing the power basis; we call these orthogonal splines. This basis makes changing the knots and updating the basis computationally efficient and is thus our choice in the few knot case. Alternatively, B-splines [17] are more efficient when the number of knots is large, because they provide a stable and localized representation.

As an example, Figure 4 displays fits under our model to two adjacent voxel time series. These data show very different drift profiles and highlight the need for both complex and simple drift structure under the same model. In Figure 4a, when the drift is complicated, the model fit includes a larger number of knots and places them to account for the principle features of the curve. On the other hand, in Figure 4b where the drift shows little structure, the model fit eliminates the knots completely and thereby achieves a reasonable fit without introducing too much complexity. The need to vary structure along this continuum guides our treatment of drift in the model.

### 2.3. Activation Profile

The hemodynamic response to neural activity manifests itself as a perturbation in the MR signal. Figure 5 illustrates the basic shape of this signal change as a function of time for a response to a single period of activity. The measured signal begins at baseline and remains there for some time (on the order of 1/2 to 3 seconds) after the beginning of task performance. It is currently unknown whether this delay represents a genuine system lag or an undetectable signal change. As local blood vessels dilate, the MR signal then begins to rise over 3 to 8 seconds, as the balance of oxygenated to deoxygenated blood shifts within the voxel. If task performance continues, the signal levels off, where the height of the plateau is usually in the range of 1-3% of the baseline signal and rarely more than 5%. The plateau height is associated with the intensity of the hemodynamic response, and by inference the degree of neural activity, within the voxel. While task performance is maintained and for some time after it ends, the signal holds at the plateau, and then begins a slow decay back to baseline. This decay usually takes longer than the corresponding rise [18], on the order of 10 seconds. The signal will sometimes dip below the baseline for an extended period before returning [16], which has two important implications: (i) a large signal dip during one task epoch distorts the signal in adjacent epoch, complicating analysis; and (ii) the dip may itself serve as a sensitive discriminator for evaluating local activity [57]. While the dip can be informative, it is often handled in current analyses by dropping data at the beginning of each task epoch. See below and [68, 1, 10] for more systematic approaches to incorporporating the dip.

Our parameterization of the response curve is intended to capture the effects of the underlying biological mechanisms while maintaining flexibility and computational efficiency. The shape of the response to an isolated epoch of task performance is determined by the vector $\boldsymbol{\theta}$ of shape parameters. This shape vector can vary from voxel to voxel, allowing the model to capture changes in the underlying response across the tissue. There are eight possible components in $\boldsymbol{\theta}$, defined as follows: (A) lag between task beginning and the signal rise (lag-on), (B) time for signal attack to plateau (attack), (C) acceleration of the attack (rise) (D) lag between task end and the signal decay (lag-off), (E) time of first return to baseline (decay), (F) acceleration of the decay (fall) (G) relative height of the dip to the plateau (dip), and (H) skewness of the dip (skew). The correspondence between these parameters and the shape of the curve is illustrated in Figure 5. Here, the rise and fall parameters determine the sharpness of the attack and decay; the dip and skew parameters determine the shape and duration of the dip. In a given fit, not all of the shape parameters need be varied; a basic configuration allows only four parameters: lag-on, attack, lag-off, decay, with no dip and with rise and fall fixed.

For a specific $\boldsymbol{\theta}$, the modeled response to an isolated epoch of task performance is proportional to what we call a bell function $b(t - t_0;\, \boldsymbol{\theta})$, where $t_0$ is the known time at which the task epoch begins. The basic bell function takes the form $b(t;\, \boldsymbol{\theta}) = b_{\text{attack}}(t;\, \boldsymbol{\theta}) \times b_{\text{decay}}(t;\, \boldsymbol{\theta})$, where $b_{\text{attack}}$ and $b_{\text{decay}}$ both have at least two continuous derivatives, $b_{\text{attack}}$ rises monotonically from 0 to 1 over the interval $[0, 1]$, and $b_{\text{decay}}$ falls from 1 to 0 over the interval $[0, 1]$. The $b_{\text{attack}}$ function is shifted by the lag-on parameter and is scaled and modulated by the attack and rise parameters. The $b_{\text{decay}}$ function is shifted by the known length of epoch plus the lag-off parameter and is scaled and modulated by the decay and fall parameters. We currently use piecewise polynomials for these functions; other choices (e.g., exponential, sinusoidal) are reasonable but somewhat less convenient. As the shape parameters change, the bell function can take on a variety of forms. For example, when the stimulus length is shorter than the time for the curve to reach plateau, the corresponding bell is shortened and rounded, which is consistent with empirical observations of responses to short stimuli. To include the dip, we can extend the parameterization of $b_{\text{decay}}$. A convenient way to do this is to define the bell as $b = b_{\text{attack}} \times (b_{\text{decay}} - b_{\text{dip}})$ where $b_{\text{dip}}$ is a basic bell whose shape is determined by the dip components of $\boldsymbol{\theta}$.

The amplitude of the response to an isolated epoch of task performance is determined by the *responsiveness* parameters. If the experimental design includes $C$ conditions, then $\gamma_c \geq 0$ for $c = 1, \ldots, C$ denotes the amplitude of the task-related signal change for the $c^{\text{th}}$ condition as a proportion of the baseline $\mu$. These parameters measure the degree to which the given voxel activates in response to the stimulus or task associated with condition $c$. The modeled response to an isolated epoch associated with condition $c$ consequently takes the form $\mu\,\gamma_c\, b(t - t_0;\, \boldsymbol{\theta})$. As stated, the model treats every epoch within a particular condition identically. This is not a bad assumption since the responses are generally consistent, but there are small variations in the response amplitudes across epochs within a condition. We can include a layer in the hierarchy to capture these variations, although we do not do so for the results in this paper.

The activation profile, $a(t)$ in equation (2), combines the responses from every task epoch in the experiment. This combination can be additive or non-additive, the latter suggested by recent evidence [60] that responses closely spaced in time combine sub-additively. In the additive regime, $a(t) = \mu \sum_k \gamma_{c_k}\, b(t - t_k;\, \boldsymbol{\theta})$, where the sum is over epochs, $c_k$ is the condition associated with epoch $k$, and $t_k$ is the start time of epoch $k$.

Our model makes several assumptions about properties of the hemodynamic response that relate to open empirical questions. First, parameterizing the responsiveness as a proportional change is natural in this context because the absolute magnitude of the signal change appears to depend on the magnitude of the baseline. More empirical study is needed to systematically validate this aspect of the model. Second, the model decouples response amplitude and shape. While this is conceptually simple and computationally efficient, there is some recent evidence for a relationship between the two features [60]. For instance, large responses may exhibit a broader plateau and later decay. As these relationships are clarified by further research, we will adjust the model to incorporate this dependence, although this will add a non-trivial computational cost. Finally, by parameterizing attack and decay as times, our model implies that these changes will be steeper for larger responses. This assumption is consistent with a biological model in which the change in blood volume increases with the intensity of the response yielding a more rapid change in the signal, and it also fits the data well. Nonetheless, this assumption needs to be further tested.

### 2.4. Noise Distribution

The noise in fMRI data is not simple. Important features of the noise distribution include subject movement, signal drifts, spatial and temporal correlations, outliers, physiological effects, instrument artifacts, and changes in noise variance with signal magnitude. All of these sources of variations affect the data in different ways and make it more difficult to isolate the hemodynamic response. The noise distribution is also sensitive to the specific scheme used to acquire the images. The results in this paper use a simple white-noise model, which serves as a useful initial approximation.

Extension to account for many of these noise features is straightforward.

Perhaps the most serious source of variation in the data is subject movement, which blurs the mapping between voxels in the image and anatomical locations in the brain. Movements as small as 2mm can appear as sudden, drastic signal changes in some voxels, and several large movements can render a data set unusable. Numerical techniques to align a sequence of images after acquisition (see in particular [22] but also [64]) are demonstrably quite effective at adjusting for rigid movements within the slice plane, although full three-dimensional alignment still needs development. Ideally, we would incorporate movement into the model, but since this currently presents a major computational obstacle, we carry out our analysis after movement correction.

Another important source of variation is caused by the subject's physiological cycles: respiration, heartbeat, and peristalsis. These introduce temporal variations that may be confounded with the activation response to the experimental tasks. The primary effect seems to be a non-rigid motion of the brain resulting from changes in blood pressure. Measurements of the physiological cycles can be recorded accurately at high sampling rate during fMRI experiments, making it possible to account for these fluctutations. See [38] for one effective method.

## 3. Prior Information

The use of prior information is a critical aspect of inference, particularly in complex or high-dimensional problems, since it enforces substantive constraints and restricts the parameter space to a reasonable form. Note that this is not the exclusive domain of Bayesian statistics [33, 55]. In fMRI, there is considerable information available regarding the various processes generating the data. Each component of our model has itself been the object of research in the MR literature, and our experience working with these data has yielded further insights. In this section, we describe the available prior information for each model component and illustrate how we use it.

We take a Bayesian approach here and express the prior information as prior distributions within a hierarchical model. Philosophy aside, we believe this approach is both natural and advantageous in this problem for several reasons. First, a hierarchical model makes it possible to include variation in the structure of the model while still accounting for the uncertainty in that structure. Second, it is straightforward to estimate any function on the parameter space with an accurate assessment of its uncertainty, even when the model allows for several qualitatively different types of structure (e.g., discrete components or disjoint submodels). This broadens the effective scope of inference under the model. Third, the approach offers a mechanism for feedback so that we can refine the specification of the model as we analyze more data and learn more about the underlying processes. Fourth, because posterior quantities are conditioned on the observed data, they are not vulnerable to spatial selection biases. Finally, the hierarchical structure of the model facilitates extensions to refine our handling of spatial structure, across-epoch variations, and previously unrecognized sources of variation. Note that a reasonable non-Bayesian interpretation takes our method as using a likelihood penalized by soft constraints on the parameters to regularize the fit.

### 3.1. Baseline Signal

Uncertainty about the baseline signal $\mu$ arises primarily from five sources: variation in the spin density across the tissue, signal fluctuation as a function of voxel position, magnetic anomalies in the tissue besides activation (cf., T2*), signal leakage from surrounding voxels (called "partial voluming"), and differential coil sensitivity across the imaged volume. Each image is also scaled by a known but arbitrary factor determined by gains in the amplifiers and pre-amplifiers, by corrections during reconstruction, and by various acquisition decisions (e.g., voxel volume). With some effort, these effects can be mapped out to derive a fairly accurate prior estimate of the baselines. However, because $\mu$ is usually well-determined from the data, inferences about $\mu$ are not very sensitive to the choice of prior. For simplicity, we use a scaled $t_1$ distribution centered on a fixed value $\mu_0$, which provides a conservative assessment of our prior uncertainty. The value of $\mu_0$ can be set separately for each voxel with scout images obtained prior to the experiment, but by default we set $\mu_0$ to a typical large signal intensity in the brain, e.g., $\mu_0 = 2000$.

### 3.2. Drift Profile

The prior for $d$ must give weight to the observed properties—general smoothness with several potential change points, some of which may be sharp—while discouraging spurious structure (e.g., oscillatory behavior) that may be confounded with activation. Let $\mathcal{S}(D, K)$ denote the orthogonal complement to the constant function in the space of splines of degree $D$ with $K$ knots on $[0, 1]$. Let $\mathcal{S}(D, K, \boldsymbol{\kappa})$ denote the subset of this space whose knots are fixed at positions $0 < \kappa_1 < \cdots < \kappa_K < 1$, and let $\mathcal{S}(D)$ denote the union of the $\mathcal{S}(D, K)$ for $K = 0, \ldots, K_{\text{max}}$. We denote the prior for the function-valued parameter $d$ by $\pi_{\text{drift}}$. If the number or position of knots is fixed, we condition on $K$ and/or $\boldsymbol{\kappa}$. The degree $D$ is fixed throughout, typically at 3.

The two strategies for modeling $d$ described in Section 2.2 are (i) placing a few knots adaptively, and (ii) fixing the knots but using a large number. Under the first strategy, we restrict $K_{\text{max}}$ to a small number and put a rapidly decreasing prior, such as a truncated Poisson, on $K$ over the range $\{0, \ldots, K_{\text{max}}\}$. The prior $\pi_{\text{drift}}(\boldsymbol{\kappa} \mid K)$ is derived from a Dirichlet$(\alpha_1, \ldots, \alpha_{K+1})$ on the simplex of knot seperations $\kappa_i - \kappa_{i-1}$ where $\kappa_0 \equiv 0$ and $\kappa_{K+1} \equiv 1$ and by default all $\alpha_i = 2$. Under the second strategy, $K$ is fixed to a value on the order of the number of time points $T$ and the knots $\boldsymbol{\kappa}$ are taken as regularly spaced. In both cases, given the knots, $\pi_{\text{drift}}(d \mid K, \boldsymbol{\kappa})$ is defined by

$$\pi_{\text{drift}}(d) \propto e^{-\frac{1}{2\lambda}Q(d)}, \tag{3}$$

where

$$Q(d) = a_n \int_0^1 d^2(t)\,dt + a_c \int_0^1 |d''|^2(t)\,dt, \tag{4}$$

and where $a_n, a_c, \lambda > 0$ are fixed constants. The constants $a_n$ and $a_c$ determine the relative penalty ascribed to norm and curvature of the profile, and $\lambda$ mediates the overall level of smoothness given this weighting, with smaller $\lambda$ indicating a smoother profile. The standard smoothing spline does not include the norm term, but for fMRI data, we have a good idea of the range of magnitudes exhibited by the drift.

The form of the prior in equation (3) does not depend on the basis we use to represent the drift profile. Given $K$ and $\boldsymbol{\kappa}$, there corresponds to any profile $d \in \mathcal{S}$ a unique vector of coefficients $\boldsymbol{\delta}$ with respect to a basis for $\mathcal{S}(D, K, \boldsymbol{\kappa})$, and the quadratic form $Q$ in $d$ induces a quadratic form in $\boldsymbol{\delta}$ whose kernel is a symmetric, non-negative definite matrix. We take $\boldsymbol{\delta}$ to have the corresponding normal distribution, which has mean 0. If $a_n = 0$, this distribution allows complete uncertainty in the linear part of the drift. For the many knots strategy, there is a big computational advantage to using the B-spline basis because the quadratic forms above are expressed in terms of banded matrices. Since a B-splines basis forms a partition of unity over the corresponding interval, the redundancy caused by our constraint $\sum d(t) = 0$ requires that the coefficients themselves sum to zero, which is easily enforced.

We select the smoothing parameter in one of two ways. The first is to fix $\lambda$ to yield a specified effective degrees of freedom. We define the degrees of freedom as the trace of the effective smoothing matrix [36] because this form is the most efficient to compute. One distinction between using the smoothing spline as a general smoother and as a component in a model like this is that the relative size of $\lambda$ and $\sigma^2$ determines the degree of smoothing. This requires an extra iterative step in the optimization. The second method is to allow $\lambda$ to vary as a model hyper-parameter with a distribution that is weighted towards zero. We use a fixed Exponential($\lambda_0$) distribution where $\lambda_0$ is chosen so this prior has a specified effective degrees of freedom as its mean. This formulation causes $\pi_{\text{drift}}$ to put more mass on profiles with less structure, discouraging spurious features.

### 3.3. Responsiveness

The responsiveness parameters $\boldsymbol{\gamma}$ describe the magnitude of the task-related signal change in each experimental condition as a proportion of baseline. These changes are small relative to the noise level, with responsiveness rarely exceeding 5% for current imaging configurations. Even though the magnitude of signal changes varies across tasks, brain regions, and subjects, the observed distributions of (estimated) signal changes from previous studies yield useful information for constraining $\boldsymbol{\gamma}$. We use such results to clarify the shape of the upper part of the response distribution. For instance, to constrain the upper range of responsiveness values, we can use a robust performance standard—the human visual system. The primary visual area, called V1, tends to exhibit the strongest BOLD response of any area of cortex yet studied with fMRI. The lower range of responsiveness values is more difficult to specify empirically since small responses will often go undetected.

We choose our prior $\pi_{\text{resp}}$ for $\boldsymbol{\gamma}$ to match the available information. We base our default priors on data from a particular suite of studies, but more specialized information (e.g., about a specific task or imaging configuration) can be incorporated with ease. Because the BOLD mechanism leads to a positive overall signal change $\pi_{\text{resp}}$ has support on $\{\boldsymbol{\gamma} \geq 0\}$. (There is an open question about whether "de-activation" in the sense of a negative BOLD response can occur; if so, the positivity constraint can be lifted without loss of generality when it is warranted.) Our specification for $\pi_{\text{resp}}$ puts non-zero mass at 0 for each condition, corresponding to no response whatsoever. All conditions are taken as independent and identically distributed given the subset of conditions with non-zero responsiveness. Specifically,

$$\pi_{\text{resp}}(\boldsymbol{\gamma}) = \sum_{\boldsymbol{j} \subset \{1,\ldots,C\}} \eta_{\boldsymbol{j}} \prod_{k \in \boldsymbol{j}} f(\gamma_k) \prod_{l \notin \boldsymbol{j}} \text{point-mass}_0(\gamma_l) \tag{5}$$

where the $\eta$'s are non-negative constants that sum to 1 and $f$ is a continuous density on $(0, \infty)$. The density $f$ decays towards both 0 and $\infty$, with its range and upper tail behavior calibrated to the available prior information. In our experiments with the form of $f$, results show sensitivity primarily to upper tail behavior, and a suitable Gamma density (e.g., with parameters 2 and 50) allows a reasonable fit to our prior constraints.

The atoms in the priors for $\gamma$ are equivalent to including a collection of sub-models in which various responsiveness parameters are constrained to be 0. This is a critical part of the model because it accounts for the substantial uncertainty concerning whether or not there is a response and prevents overfitting. To maintain identifiability, we exclude the sub-model in which every $\gamma_c > 0$. We usually choose $\boldsymbol{\eta}$ so that $\eta_\emptyset$ is large and the components of $\boldsymbol{\gamma}$ are independent. A convenient alternative is to put mass only on the null and the saturated models. A value of $1 - \eta_\emptyset$ in the range 1/1000 to 1/100 appears to be reasonable.

### 3.4. Shape

Although the mechanism behind the hemodynamic response is not yet fully understood, a body of empirical work aimed at understanding how the response manifests itself in the MR signal provides constraints that we use to construct the prior $\pi_{\text{shape}}$ for the shape parameters $\boldsymbol{\theta}$. We currently take the shape parameters to be non-negative, although allowing the two lag parameters to take negative values may be useful for broadening the range of shapes fit by the bell functions. We also define $\pi_{\text{shape}}$ to make the components independent of each other and $\boldsymbol{\gamma}$. The lag-on and lag-off parameters are likely to be similar, on the order of 1/2 second. Attack seems to be generally shorter than decay, the former ranging from 3–8 seconds and the latter from 5-15 seconds. The rise and fall parameters describe the shape of the attack and decay; as we have parameterized these, they lie between -1 and 1 inherently. The height of the undershoot below baseline is parameterized as a proportion of the plateau height; 1/3 seems to be a representative value, although more study will clarify this further. We have little information to constrain dip skew but it is a naturally bounded parameter. We define $\pi_{\text{shape}}$ by giving rise, fall, and skew uniform distributions over their natural range and the other shape parameters suitably calibrated Gamma distributions. In practice, the specific values of the hyperparameters defining $\pi_{\text{shape}}$ depend on the image acquisition scheme and experimental design, since these can affect the response characteristics.

### 3.5. Noise Parameters

Many of the statistical challenges surrounding the analysis of fMRI data arise from the complexity of the noise distribution. We have studied the noise with data from a large number of studies, for a variety of acquisition methods, and imaging different types of objects, from air to "phantoms" to human subjects. In this paper, we use a simple white noise model to capture the basic fluctuations, but there is much room for extension to more complicated spatio-temporal distributions. We put a Gamma prior on the noise precision $1/\sigma^2$ (e.g., parameters 1.6 and 200 by default for echo-planar images on a 1.5T scanner), where the mean is selected to match the measured overall signal to noise ratio for the scanner and the variance chosen to make the distribution reasonably diffuse.

## 4. Making Inferences from fMRI Data

Our model for fMRI data provides a flexible inferential framework that is consistent with current research on the processes generating the data. However, a good model is only the first step. We also need a way to use the model that makes it possible to address the full range of relevant scientific questions, to test the predictions of competing theories, and to generalize inferences to a broader population. Our approach here is to relate the questions of interest to particular functions on the model's parameter space and to derive inferences under the model through the posterior distributions of these functions.

### 4.1. An Example Experiment

Why are some sentences more difficult to understand than others? One answer is that more difficult tasks require that a greater amount of information be maintained in memory during task performance. For instance, when parsing complex sentences, any nested clauses, modifiers, or unresolved ambiguities must be held in memory until they can be assigned a role in the meaning of the sentence. The brain has a set of general mechanisms, known as *working memory* [3], for maintaining such information during processing. Working memory plays a vital role in essentially every cognitive task. The amount of available working memory dictates how much information can be maintained, what associations can be made, and to how many aspects of the environment attention can be given. A useful way to think about working memory is as a cognitive resource available to the system; it can be allocated in many ways at any level of a computation. The goal of research on working memory is to understand how the brain allocates and uses this resource.

Cognitive psychologists have developed theories of working memory and its role in cognition. These theories provide an abstract representation of the processes that underlie task performance and make specific, quantitative predictions, *e.g.,* what processes will be used and when, what distribution of responses will result, and how long subjects will take to complete processing. Current cognitive theories explain human performance on difficult tasks by attributing to each individual a limited supply of the working memory resource [12, 42]. Working memory utilization during a particular cognitive task is considered a good measure of how "hard" that task is [42]. As working memory is engaged in a sequence of increasingly difficult tasks, the individual must work harder, and when resource limitations restrict further allocation, performance degrades. Some current cognitive theories [11, 42, 43] describe working memory by a hierarchy of resource pools that are specialized to particular types of computations. Other theories posit a single, general pool [2].

Questions regarding the mechanism of working memory abound: Is the resource limitation view a valid one? How does working memory utilization increase with the difficulty of the task? How are the resource pools arranged? How can two distinct pools be distinguished from a single, more general pool? How can the interactions among different resource pools be determined? These are critical questions for understanding working memory and evaluating current cognitive theories.

Here, we consider an fMRI experiment [43] designed to study how working memory utilization changes with task difficulty. The hope is that these changes are quantifiable through fMRI at a finer level of detail than is possible with behavioral data alone. During the experiment, the subject reads a sequence of visually-presented sentences and responds to a question about each sentence by pushing an appropriate button. The experiment manipulates the type and difficulty of the sentences presented to the subject. This experiment and the data are part of a study described in [43]. We use the data from a single subject (study #1078, provided by Drs. Carpenter and Just) for our analyses here, and our results for this subject support the original conclusions drawn by the investigators. Our focus here is not on the results themselves; rather, we wish to illustrate how our model can be used to address directly a range of scientific questions beyond localization.

The experimental design specifies six task conditions arranged in thirty-eight task epochs, as illustrated in Figure 1b. The first condition is simple rest, which serves as a buffer between every pair of tasks and as a baseline control. The second condition requires that the subject maintain visual fixation on a marked point in the center of his or her visual field. This is the primary control condition. The third condition is a trivial version of the task with no semantic content—reading strings of consonants. This task involves all the same stages of processing as do meaningful sentences (e.g., visual encoding, eye movements, button pushing to answer questions) except for the high-level functions underlying comprehension. Hence, differences between the trivial condition and the other sentence conditions can serve to isolate the processes under study. The remaining three conditions involve reading and comprehending increasingly difficult sentences. The sentences at different levels of difficulty are distinguished by different syntactic and semantic structures that increase the cognitive load required to understand them. Each task epoch involves processing several sentences of the given type in succession, and the multiple epochs for the non-control conditions are distributed across time in four balanced blocks. We will label the conditions as Rest (R), Fixation (F), Trivial (Tr), Task Level 1 ($\mathrm{T}_1$), Task Level 2 ($\mathrm{T}_2$), and Task Level 3 ($\mathrm{T}_3$).

One goal for this experiment is to clarify the relationship between intensity of response and sentence difficulty. Two of the specific questions underlying the study are as follows:

1. Do responses to the three task levels increase monotonically?
2. What is the functional relationship between difficulty and response and how does this relationship vary spatially?

Note that changes in the intensity of response can manifest themselves in the data in two ways: through changes in the responsiveness within voxels and changes in the extent of the region showing a significant response. Which of these measures is most relevant is a scientific issue, but statistically, we can use both types of effects (or some combination of the two) to quantify response changes.

These data are analyzed using the BRAIN software package (see section 5) to fit the model. For the Gamma prior on the responsiveness parameters, we use $(1, 50)$ as the hyperparameters for all conditions. The model is fit with the four-parameter shape configuration described above with hyperparameters $(2, 4.3)$, $(4, 1)$, $(2, 4.3)$, and $(4, 0.43)$ on the Gamma priors for lag-on, attack, lag-off, and decay respectively. The Gamma prior on the noise precision is given hyperparameters $(1.6, 200)$. Relative weightings for the drift penalties are set to $a_n = 0.01$ and $a_c = 1.0$, and $\lambda_0$ is set to target 5 effective degrees of freedom. The fixed knot strategy for drift is used for both the maximum posterior estimation and MCMC sampling phases of the analysis, but the MCMC results do not vary greatly when adaptive knots are used.

### 4.2. Estimation

The simplest method for making inferences under our model is to estimate the model parameters and their associated uncertainties. Maximum posterior estimates and their standard errors provide a convenient alternative to a testing approach and can be used to compute analogues of the statistical classification maps that are currently standard in fMRI. Almost all of the results obtained this way can be improved through full posterior inference (see section 4.3 below), but the maximization approach is of interest as a fast and effective approximation. Even if full posterior inferences are desired instead, the estimates provide a good starting point for MCMC simulations, and the approximate covariance matrix is useful for tuning Metropolis steps [48, 59].

The output of posterior maximization is an approximation to the joint posterior distribution of the parameters. Using this approximation, we can derive estimates and their standard errors, but we can also compute a variety of interesting probabilities that relate the parameters to the questions of interest. As discussed previously, a critical feature of the posterior is that it is supported on a disjoint union of sub-models, for example, corresponding to each subset of responsiveness parameters that are constrained to be 0. The posterior derived from maximization is based on the Normal approximation to the posterior within each sub-model and an approximation to the posterior probabilities of the different sub-models. (See section 5.) Below, we consider a variety of analyses that address questions posed in the sentence-comprehension experiment.

The greatest interest in fMRI analyses typically centers on the task-related signal changes described by the activation profile. One useful way to characterize these changes is through constrasts among the responsiveness parameters $\boldsymbol{\gamma}$ of the form $\boldsymbol{\alpha} \cdot \boldsymbol{\gamma} = \sum_c \alpha_c \gamma_c$ for real constants with $\sum_c \alpha_c = 0$. The posterior distribution of $\boldsymbol{\alpha} \cdot \boldsymbol{\gamma}$ is a mixture of the posteriors conditional on each sub-model. While the posterior of the contrasts will be most accurately computed using the sampling procedure discussed below, a convenient approximation, which is typically good although somewhat conservative, is to take the posterior of $\boldsymbol{\alpha} \cdot \boldsymbol{\gamma}$ as a mixture of (degenerate) multivariate Normals suitably truncated to the non-negative orthant. The degeneracies arise from the components of $\boldsymbol{\gamma}$ that are constrained to 0 in each sub-model. In the sentence-comprehension experiment, an example of a simple contrast is what we call a domination probability: the posterior probability that the response to one condition is greater than to another. For example, Figure 6b shows a map of domination probabilities $\mathrm{P}\{\, \gamma_{T_3} > \gamma_{Tr} \mid \boldsymbol{Y} \,\}$. The corresponding slice of the mean image is given in Figure 6a for reference. The domination probability map highlights the voxels where reading the most difficult sentence type evokes a greater response than reading consonant strings. The great number of voxels with probabilities near zero reflects that the majority of voxels have most of the posterior mass on the null model. The $T_3$ versus Tr comparison is expected to be the most extreme and thus bounds the region where full monotonicity is achieved. As opposed to more complicated contrasts, the domination probabilities can be computed directly from the posterior approximation without a further mixture approximation. We can also examine directly the estimated signal changes for each voxel. Other contrasts can be similarly computed, such as the interaction $\gamma_{\mathrm{T}_3} - 2\gamma_{\mathrm{T}_2} + \gamma_{\mathrm{T}_1}$ which evaluates the changes between adjacent difficulty levels.

For comparison, Figure 6c shows a "traditional" t-map for the same slice. This map displays the two-sample t-statistics at each voxel for the $T_3$ versus Tr comparison after linear detrending of the data. The map is thresholded at ±4 for classification, an arbitrary but often used value. Once the negatives (white) are excluded from the t-map, the two maps identify similar clusters of voxels where $T_3$ dominates Tr, but there is a scattering of voxels at which the two disagree. We examined individually all such time-courses, and the results qualitatively support the results of our model fit and are consistent with our findings over a variety of studies. There are several reasons to expect greater sensitivity from our model than from the simple test in general. The model captures variation in the response structure and in the heterogeneous signal drifts. Moreover, signal changes ascribed to activation by our model must be consistent with the basic form of the profile.

### 4.3. Posterior Inferences

The maximum posterior estimates and asymptotic uncertainties provide a good approximation to the posterior that can be used to address a variety of questions, including but not limited to localization. By using Markov Chain Monte Carlo (MCMC) simulation techiques [59, 54, 8, 27], we can refine the computed posterior and can account for more complex features in the model. The result of an MCMC simulation is a sample from the full posterior distribution of the parameters, including variation across sub-models. Any functional of this distribution can be easily estimated from a sufficiently large sample, offering great flexibility for tuning the analysis to the scientific questions of interest. This applies to quantities within a voxel, such as the responsiveness contrasts described above and various features of the response shape, but it also applies to comparisons of these quantities across voxels. The key point here is that our approach makes the question accessible quantitatively and provides a measure of uncertainty with respect to the question.

*Assessing Monotonicity: Responsiveness.* With respect to the responsiveness parameters, the monotonicity question centers on whether $\gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}}$ for a given voxel. This would imply that the size of the hemodynamic perturbation increases with task difficulty whenever semantic processing is required. The posterior monotonicity probabilities

$$\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\} \tag{6}$$

quantify the support in the data for monotonicity in each voxel. Figure 6d shows a map of these probabilities computed from the data. The picture reveals that the structure in the main cluster (lower right), which has been a persistent feature through most of the images displayed here, is consistent with the monotonicity hypothesis. On the other hand, the small cluster at the middle right shows little indication of monotonicity. There are several possible reasons why monotonicity is not apparent for the latter cluster, but we leave the interpretation of this result to scientific argument. An interesting variant on the monotonicity probabilities is to compute $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > 0, \gamma_{\mathrm{Tr}} = 0 \mid \boldsymbol{Y} \,\}$ which in principle can distinguish areas recruited specifically for semantic processing.

Although it is possible to construct a classification procedure for monotonicity, it is neither as natural nor as effective. A hypothesis test for classifying monotonicity could be constructed by combining the results of one-sided, pairwise tests between successive conditions. However, since equality does not preclude monotonicity (e.g., $\gamma_{\mathrm{T}_3} > \gamma_{\mathrm{T}_2} = \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}}$ is still considered monotonic), the null hypothesis in this case is composite. The error rates of the combined test are also difficult to compute, providing an unsatisfying assessment of uncertainty.

*Assessing Monotonicity: Extent.* To study the extent of activity, we must consider many voxels simultaneously. A common approach to this question in fMRI is to compare the counts of voxels that are classified active in each condition. This approach yields an estimated difference with no measure of uncertainty for evaluating the difference; an alternative under our model solves this problem. For each voxel $v$, let $N_{iv}$ be the indicator of the event $\{\gamma_{\mathrm{T_i,v}} > \gamma_{\mathrm{Tr,v}}\}$, for $i = 1, 2, 3$, and let $N_i = \sum_v N_{iv}$ for each $i$. We use the posterior distribution of $(N_1, N_2, N_3)$ to address the monotonicity question. In general, the $(N_{1v}, N_{2v}, N_{3v})$ all have different distributions, but using the Fast Fourier Transform (FFT), the convolution can be computed efficiently. The joint probability mass function of each vector $(N_{1v}, N_{2v}, N_{3v})$ is supported on the lattice $\{0, 1\}^3$ and can be extended by zeros to a larger lattice containing $\{0, 1, \ldots, V\}^3$, where $V$ is the total number of voxels being considered. If we assume the voxels to be independent, the distribution of $(N_1, N_2, N_3)$ can be obtained by multiplying the individual Fourier transforms over the larger lattice and inverting the transform. In practice, we can take $V$ to be much smaller than the total number of voxels because, for the vast majority of voxels, there is only negligible mass away from 0 for any of the $N_{iv}$'s. $V$ will also be small if we are focusing on local changes. This makes the computation feasible since the lattice scales as $V^3$. With these data, only 147 voxels have posterior probability bigger than 0.001 away from $(0, 0, 0)$; we thus use a lattice of edge length 256 to compute the Fourier transform. An alternative strategy is to simulate draws from the distribution of $(N_1, N_2, N_3)$ by generating and adding $(N_{1v}, N_{2v}, N_{3v})$'s. This is computationally efficient for both small and large $V$ but does add some uncertainty to the estimated probability. In our example, $\mathrm{P}\{\, N_3 \geq N_2 \geq N_1 \mid \boldsymbol{Y} \,\} \approx 0.67$, which appears consistent with monotonicity in extent. As mentioned earlier, there is no good way to address this question by combining voxelwise classifications.

There are several variations of this idea: it is possible to look for strict ordering of the sets of active voxels across conditions and to study the changes in extent local to a given cluster of activity. Note that in making such posterior inferences, we can restrict our attention to apparently responsive voxels (e.g., as measured by a domination probability above some threshold) without selection bias. In other words, if $F(\boldsymbol{Y})$ is a functional of the joint posterior over the parameter space, then $P\left(A \mid F(\boldsymbol{Y}), \boldsymbol{Y}\right) = P\left(A \mid \boldsymbol{Y}\right)$ for any subset $A$ of the parameter space, since $F(\boldsymbol{Y})$ is trivially $\boldsymbol{Y}$-measurable.

*Assessing Monotonicity: Integrated Response.* An alternative to looking for changes in extent or magnitude individually is to combine the two measures. One useful way to do this is to integrate the responsiveness over a specified region of interest, preferably a region defined *a priori* from the mean functional or structural images with anatomical expertise. Although the precise location and shape of these regions may vary across individuals, the regions can sometimes be taken as comparable from a functional point of view, providing a tool for making comparisons across subjects.

To illustrate how our method facilitates region-of-interest analysis, we examine the posterior of the functionals $\Gamma_c(R) = \int_R \gamma_c(v)\, dv$ for conditions $c = \mathrm{Tr}, \mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3$ and for a fixed set of voxels $R$. Figure 7 shows $\mathrm{P}\{\, \Gamma_c(R) > u \,\}$ as a function of $u > 0$ for each of these conditions, where $R$ is an arbitrary region of 21 contiguous voxels in the language area (surrounding the most notable cluster of activity in the other maps). These curves can computed directly from the posterior sample; either the empirical distribution or a normal kernel density estimate provides an approximation that is easy to use. We used the latter here. The curves in the figure are not all 1 at $u = 0$ because there is some posterior probability of zero responsiveness in each condition. The curves shown in the figure strongly suggest the desired monotonicity relationship.

*Characterizing the Functional Form.* Given monotonicity, the next step is to investigate the specific form of the relationship between response and task difficulty and to compare it with the predictions of cognitive theories. These theories make specific predictions about the functional form of this relationship for each individual. For example, consider the following simplified predictions of a resource-based theory: an individual with a large resource supply should only show an increase in response intensity for the most difficult tasks, whereas an individual with a moderate supply should show an increased response for somewhat easier tasks as well.

There are several approaches to characterizing this functional form; here we examine a graphical technique that makes it easy to identify clusters where the response-difficulty relationship is similar. Specifically, we derive the joint posterior distribution of successive responsiveness differences, $(\gamma_{\mathrm{T}_2} - \gamma_{\mathrm{T}_1}, \gamma_{\mathrm{T}_3} - \gamma_{\mathrm{T}_2})$. When monotonicity holds, most of the mass will be concentrated in the positive quadrant. How the mass is distributed in this quadrant indicates the support for each the four possible monotonic forms, that is, the two segments of the curve being flat-flat, flat-up, up-flat, or up-up. Each "pixel" in Figure 8 shows this joint distribution for a given voxel. Only the positive quadrant is shown in each case, and voxels with less than 0.001 of the mass in that quadrant are left blank. Two voxels in the Figure are marked with arrows. These distributions exemplify a difference in response shape. The marked voxel on the right tends to show a large change in responsiveness between $\mathrm{T}_1$ and $\mathrm{T}_2$ and a smaller but non-trivial change between $\mathrm{T}_2$ and $\mathrm{T}_3$. The marked voxel on the left, on the other hand, shows most of its responsiveness change between $\mathrm{T}_1$ and $\mathrm{T}_2$; the distribution of $\gamma_{\mathrm{T}_3} - \gamma_{\mathrm{T}_2}$ is concentrated near zero. Note that this analysis treats the differences among the conditions as ordinally related; more specific information about the differences in task difficulty would be required to fit parametric forms to the responsiveness curves.

## 5. Computational Techniques

Fitting our model to data and implementing the inferential procedures described above raises a number of computational challenges. Each of the many time series in an fMRI data set must be processed automatically despite great variation in structure among the voxels. The large number of voxels requires efficient algorithms and the diversity among voxels requires robust methods that can adapt to structural differences. In this section, we describe our computational techniques for model fitting and inference. These are implemented in the publicly available software package Bayesian Response Analysis and Inference for Neuroimaging (BRAIN) [30].

### 5.1. Initial Data Processing

The raw data produced by an MR scanner are collected in the Fourier domain and the images reconstructed. The raw data are subject to several sources of bias and mis-calibration and must be specially processed to yield good quality images. While it would be ideal to integrate these sources of variation into the model, it is not yet computationally feasible. Instead, we pre-process the data with the FIASCO (Functional Image Analysis Software, Computational *Olio*) software package [21] and use the resulting spatio-temporal data as input to our procedures.

### 5.2. Posterior Maximization

Maximum posterior estimates are computed via direct numerical optimization with the log unnormalized posterior as objective function. The procedure can accept arbitrarily defined prior distributions which are then interpolated to compute derivatives, but it markedly improves performance to use smooth priors and analytical derivatives. Numerical difference approximations, when necessary, are computed by Richardson Extrapolation [15] to improve accuracy. We take care near the boundaries of the parameter space to acquire a well-defined value when differencing is used. We use the BFGS version of the variable metric optimization algorithm [52] while enforcing bounds on the parameters through an active set method. Upon arrival at the maximizer, the algorithm is restarted after a perturbation to the parameters to validate and refine the result. The standard errors of the estimates are derived from the inverse observed Fisher information at the mode which is obtained from the computed Hessian.

When a number of sub-models are to be included in the analysis (e.g., various responsiveness parameters are allowed to take the value 0), the posterior (and likelihood) are maximized for each such model at every voxel. We estimate the posterior probabilities of the submodels with a form of the Laplace approximation [19] or with the Schwarz criterion [53]. When parameters are pushed to their bounds and thus a smaller sub-model, we give a conservative estimate of the probability of the larger model, but we are still assessing more detailed corrections to the Schwarz criterion as in [50]. All inferential results are averaged over the different sub-models using the computed posterior probabilities. The parameters maintain a consistent interpretation across all the models.

### 5.3. MCMC Sampling

To obtain a more accurate representation of the posterior, we turn to Markov Chain Monte Carlo (MCMC) sampling. For each voxel, we obtain a sample from the posterior and use this to derive inferences. Our sampling strategy is a mix of Metropolis and Gibbs steps with a fixed scan order across the components. Sampling occurs in three stages: an optional pre-scan for adjusting the initial Metropolis jumping distributions, a period of burn-in where no output is recorded, and final sampling. The maximum posterior estimates are used as the starting point for each chain.

The pre-scan stage automates the initialization of the Metropolis candidate distributions because with many thousands of voxels, it is not convenient to monitor and tune the individual chains by hand. The approximate covariance matrix from the posterior maximization is dilated by a fixed factor and then decomposed by a Cholesky factorization [34] to obtain the variances of the conditional distributions of each parameter conditional on the previous parameters. These variances are used to derive the initial jumping widths. The pre-scan phase consists of a brief sampling run during which the rejection rate and average length of moves is recorded in blocks of samples for all of the parameters that require a Metropolis chain. After each block, the jumping widths are adjusted by interpolating the recorded measures over previous blocks to either bring the rejection rate closer to a fixed target (e.g., 50%) or to maximize the average move length over sub-blocks. Both of these are heuristic criteria, but they tend to yield good jumping widths.

The burn-in phase is a long sampling run during which the chain is allowed to equilibrate towards its stationary distribution. No output is recorded during this phase. The length of the burn-in is configurable, but by default, we burn-in for 5000 samples in each parameter.

The sampling phase begins at the end of burn-in and continues for a specified number of samples. Since there are so many chains running over the data set, efficiency is critical in practice, and we generally run the chain as long as we can tolerate for the analysis. Our default is 10,000 samples per component after burn-in for standard image sizes, but for very high-resolution images this must be reduced to get a reasonable run time. When feasible, we run the chains longer and sub-sample to reduce correlations in the recorded sequence. We have several ways to speed up the computations, including parallelizing the computation, eliminating uninteresting voxels outside the head, and ordering the computation based on the preliminary estimates.

One area that needs development here is convergence diagnosis since multiple chains and graphical monitoring are inconvenient in practice. We currently use only rudimentary measures of chain performance during analysis, but we are working to improve this. As part of a "quality control" effort, we studied the performance of our sampling scheme on a collection of voxel time series from several experiments. Graphical diagnostics, correlations among parameters, and various standard convergence diagnostics [14] based on parallel chains with different starting points suggest that the chains are mixing quite well, and equilibrating sufficiently and also that the Normal approximation is reasonable in most cases. However, more systematic study is needed.

For the default configuration, our sampling algorithm is arranged as follows. The baseline parameter $\mu$ mediates a number of the other parameters and so has a complicated complete conditional even with the default priors. We use a symmetric random walk Metropolis chain for $\mu$, but as part of the move, we adjust the responsiveness parameters by the ratio $\mu'/\mu$ (candidate to current) so that the activation profile does not change. The complete conditional for the drift and responsiveness parameters can be sampled directly. We first sample $\boldsymbol{\gamma}$ and then the drift profile conditional on $\boldsymbol{\gamma}$ because the non-negativity constraint on the responsiveness complicates its distribution. The conditional distribution for $\boldsymbol{\gamma}$ given everything but the drift is a multi-variate Normal truncated to the positive orthant. To sample from this distribution, we draw the components of $\boldsymbol{\gamma}$ one at a time from successive univariate conditional distributions. The Cholesky factorizations of the covariance matrix and its inverse allow us to derive the mean and variance of these conditional distributions iteratively. We then draw from a univariate truncated Normal using the inverse distribution function method when the mean is large enough to ensure precision in computing the Normal distribution function and a rejection method (based on an Exponential approximation to the Normal tail) otherwise. The drift profile can then be drawn as a whole from its complete conditional. The shape parameters capture most of the nonlinearity in the model. We choose from among two different types of Metropolis moves for these parameters: (i) a log Normal random walk in the parameters individually, and (ii) coupled jumps in related pairs (lag-on and attack, lag-off and decay, etc.). As an example of the latter, we use one move type that keeps lag-on + attack constant while varying their relative size and another that changes the sum while keeping the relative size the same. These diverse moves provide an automatic reparameterization voxel to voxel that reduces the correlation among the parameters and improves mixing of the shape. The smoothing hyper-parameter for the drift is sampled using a log Normal random walk and poses no complication. Finally, the noise precision is drawn from its complete conditional which depends on the residuals and the degree of drift smoothing. We take a great deal of effort to make all these computations as efficient as possible and employ a number of coding tricks to reduce overhead.

### 5.3.1. Model Jumping

For varying the structure of the model in discrete ways, we use the reversible jump framework developed in [35]. This allows a single Markov chain to travel among distinct model spaces while maintaining detailed balance; it consequently becomes feasible to work with posteriors that have support in all of these spaces. The resulting inferences can be expressed within a particular model or by averaging across models. We use this model jumping technology to allow for varied structure in the responsivenss parameters and the drift profile (the latter only when using adaptive knots). Since these components maintain their interpretation in every sub-model, we average over the models to account for uncertainty in the structure.

Because the prior for $\boldsymbol{\gamma}$ has mass at 0, the Markov chain must move over disjoint sub-models defined by the non-zero responsiveness parameters. At each sampling iteration, $\boldsymbol{\gamma}$ is updated by the Gibbs' step as described above, and then with some probability a model jumping move is attempted. There are two types of moves, made with equal probability, inclusion of a zero component or the removal of a non-zero component. If a model jump is to be made, we select at random a candidate parameter of the appropriate type (zero or nonzero). Our basic moves involve both the responsiveness parameter and the baseline because inclusion or removal of a condition impacts which measurements provide information about the baseline signal. Without this adjustment, there would be substantial lack of fit and few such moves would be accepted. Let $\ell_1$ and $\ell_2$ denote the lengths in the design corresponding to the zeroed conditions and the candidate condition, respectively. The simplest move takes $(\mu, \gamma)$ to $((\ell_1 + \ell_2(1+\gamma))/(\ell_1 + \ell_2)\mu, 0)$ for removal and $(\mu, 0)$ to $((\ell_1 + \ell_2)/(\ell_1 + \ell_2(1 + z))\mu, z)$, where $z$ is a random responsiveness candidate that is independent of $\mu$. This move follows the template given in [35] and satisfies detailed balance. A generalization of this which improves mixing is to perturb $\mu$ by an independent random amount (i.e., $(\mu, \gamma)$ to $((\ell_1 + \ell_2(1 + \gamma))/(\ell_1 + \ell_2)\mu + w, 0)$ for a Normal $w$ with small variance and similarly in the other direction).

Our approach is similar for the drift when adaptive knots are used. Because the drift profile is the parameter, the posterior is invariant under mappings that leave the profile unchanged. The orthogonalized splines work well for the adaptive handling of drift because (i) updates to the knots are computationally efficient [34] and (ii) orthogonality of the basis functions allows the components to be treated independently. Both the number $K$ and positions $\boldsymbol{\kappa}$ of the knots are allowed to vary, although we enforce an upper bound on the number of knots. At every sampling iteration, we take a Gibbs' step as described above to change the structure of the drift profile. Then with some probability, we take a model jumping move, one of three types with equal probability: changing the position of a knot, adding a knot, and removing a knot. When $K = 0$, then no knots can be removed, and when $K = K_{\max}$, no knots can be added, but this does not impact detailed balance. In the adaptive case, we typically will have only a few knots (e.g., 2–5), but it facilitates mixing among the models to allow $K_{\max}$ to be larger than this [35]. When adding, removing, or moving a knot, the effected knot position is selected at random, and the basis is then reformatted to make it easier to update that component. Moving a knot involves randomly perturbing the selected knot within the bounds of its neighbor; the dimension of the model is fixed but this is a jump between different subspaces. The simplest way to add or remove a knot is to change a single coefficient, setting it to zero when removing or drawing it from a distribution independent of the profile when adding. This attains detailed balance but does not mix very well because only a small perturbation to a single component leads to an acceptable change in the profile. We fix this problem by also updating the other components of the profile as part of the move. The dimension matching requirement of [35] is satisfied, mixing is improved, and detailed balance is maintained.

## 6. Discussion

Our model has several notable advantages over traditional methods for analyzing fMRI data. It attempts to capture the structure of the time course directly, dealing with drift and allowing for changes in the shape of the hemodynamic response. The fit to the data is thus more precise than the implicit forms underlying most classification tests. Among recently developed methods, only the model of [45] has similar advantages. Our model also handles complex experimental designs and accounts for important features of the drift and response, e.g., the undershoot dip. Moreover, it is modular, adaptable and is built on substantive information about the underlying processes.

The inferences we can derive under our model can address questions of localization and thus subsume the traditional classification-based methods. In our opinion, however, the primary contribution of this approach is that it makes accessible to direct analysis a wide range of more general questions as well, including questions about changes in the response across conditions and across voxels and about temporal patterns in the response that distinguish different types of processing. This flexibility allows scientists to directly target the questions they want to address with accurate and relevant measures of uncertainty that enable them to evaluate the analysis.

There are several weaknesses in our approach that remain to be addressed. The first is that fitting the model, particularly via MCMC simulation, is very computationally intensive, requiring on the order of a day to analyze a single subject's data. With careful parallelization and improvements in computing technology, however, we expect this problem to become steadily less severe over time. Second, the noise model used for the results in this paper is still rather simple, and we will extend it to deal with the noise complexity, especially physiological variations and spatial dependence. Third, structural independence of the parameters across voxels is likely a simplistic assumption, and we can gain precision by combining across voxels with similar functional properties. Accounting for this spatial structure dynamically is a challenging problem; we discuss our approach below. Finally, inferences under the model depend somewhat on the priors, and it may be informative to systematically evaluate the nature of this dependence. In our experiments, we have found that specific shape of the priors has only a small impact on the results provided the basic range of the parameters is suitably constrained. Our goal has been to include generally accepted information about the basic processe, so the priors we use reflect reasonably uncontroversial constraints. Nonetheless, further model validation remains a priority.

The basic implementation of our model presented in this paper can be extended in several directions. First, we can allow variations in the response amplitude across epochs and within each condition to capture this variation in the data [32]. Second and similarly, we can allow variations in the response shape across conditions within a voxel; this is particularly important for designs in which the conditions occur on vastly different time scales. Third, we can parameterize the model by an ANOVA decomposition of the response amplitudes to accomodate blocked designs.

Finally, we need to account for spatial relationships among the fundamental processes that generate the data. The shape of the hemodynamic response function, the amplitude of the response, the impact of physiological variations, and other such features exhibit complicated dependence across voxels. Modeling these relationships increases the precision of inferences because multiple voxels contribute information about features they have in common. The task is not to segment the image *per se* but to identify regions with consistent physiologic and functional properties. A particular challenge here is that tissue boundaries in the brain are convoluted and piecemeal, so methods based on canonical Markov Random Fields [7, 28, 41, 29] are likely to oversmooth. We are currently working to adapt the method described in [39, 40] that incorporates disjoint region descriptors as parameters at a deeper level in the hierarchy. The simplest approach is, given a set of disjoint dependence regions, to take the relevant parameters (e.g., responsiveness and shape) to be constant on the given regions. More generally, a within-region component of variance will likely be required. This spatial model offers both improvements in precision and a more accurate basis for making inferences about the spatial pattern of responses.

Another issue is that all the methods for fMRI analyses we have thus far described apply to data from a single subject, but to generalize results to a broader population, the results must be compared and combined across subjects. Unfortunately, there is large variability in the physical geometry of the cerebral cortex across individuals. The most common method to combine fMRI data across subjects is to map the subjects' brains onto a common coordinate system, the Talairach atlas [56], and then average in this coordinate system. The Talairach atlas was derived from a detailed study of six human brains, and the mapping for a given subject is computed using only a few gross measurements of that subject's brain. This averaging procedure is far from satisfactory, however, because large inter-subject variations remain (see [47] for a demonstration).

While several more sophisticated methods of mapping across subjects' brains show promise, in some cases, our model can be used to combine results without an explicit anatomical map. This applies when the question of interest is abstracted away from the anatomy and internalizes anatomical differences. Let $G$ be a function on the (total) model parameter space that does not depend on the explicit coordinate system of the image for a given subject. For example, $G$ might be the integrated response over a pre-specified and anatomically defined region of interest or the indicator that two tasks yield distinct temporal patterns of response. Suppose for discussion purposes that $G$ depends only on $\boldsymbol{\gamma}$, and that the $J$ subjects in the experiment contribute data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_J$. If we are willing to assume that these data are drawn i.i.d. from some population distribution, we can combine the posterior distribution of $G$ across subjects to make inferences about that population. For example, the population expectation of $G$ can be estimated by the average of the posterior expectations: $E(G(\boldsymbol{\gamma})) \approx (1/J) \sum_{j=1}^{J} E(G(\boldsymbol{\gamma}) \mid \boldsymbol{Y}_j)$. Variances can be estimated similarly using the standard conditioning identity. Although this may represent an unorthodox use of the posterior distribution, it is an intuitive way to combine information across subjects.

## 7. Conclusions

There is tremendous diversity in the range of questions to which fMRI is being applied. Scientists' choices of how to address these questions with the data are influenced by two conflicting forces: the desire for standardized procedures for statistical analysis and the desire for precise and scientifically relevant inferences. The localization paradigm has been so widely embraced in large part because the corresponding statistical analyses are automatic. But automaticity has a cost: as the questions posed become more sophisticated, the chain of inference between data and conclusions is strained and stretched, and scientists are forced to make interpretive leaps to connect the "where" to the "why". We have proposed a different approach, in which scientists pose a set of questions of interest and tune their inferential procedures to address these specific questions. The advantages are improvements in both the precision and scientific relevance of the inferences; the cost is that more careful thinking is required to relate the statistical and scientific aspects of the problem. We use this inferential approach in the context of a detailed model for fMRI data that we designed to accurately capture the critical sources of variation. The model is modular and extendable and offers improved precision relative to current methods of fMRI analysis.

Beyond fMRI, every aspect of our framework is applicable in some way to more general spatio-temporal problems, from the specification of the model as a sum of nonlinear functions with distinct structure to the design of inferential procedures that target specific scientific questions to the computational techniques for fitting the model with a vast supply of data. When analyzing large and complex data sets, there is a natural tendency for scientists to search for simple and automatic statistical procedures. But as computational resources continue to improve, a more substantive approach like the one described here becomes more and more practical, and the more complex the problem, the greater the potential gain.

# References

[1] G.K. Aguirre, E. Zarahn, and M. D'Esosito. Empirical analyses of BOLD fMRI statistics ii: Spatially smoothed data collected under null-hypothesis and experimental conditions. *NeuroImage*, 5(3):199–212, 1997.

[2] J. R. Anderson, L. M. Reder, and C. Lebiere. Working memory: Activation limits on retrieval. *Cognitive Psychology*, 30:221–256, 1996.

[3] A.D. Baddeley. *Working Memory.* Oxford University Press, New York, 1986.

[4] J. R. Baker, R. M. Weisskoff, C. E. Stern, D. N. Kennedy, A. Jiang, K. K. Kwong, L. B. Kolodny, T. L. Davis, J. L. Boxerman, B. R. Buchbinder, V. J. Weeden, J. W. Belliveau, and B. R. Rosen. Statistical assessment of functional MRI signal change. In *Proceedings of the Society for Magnetic Resonance, Second Annual Meeting*, volume 2, page 626. SMR, 1994.

[5] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. Hyde. Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, 30:161–173, 1993.

[6] J.W. Belliveau, D.N. Kennedy, R.C. McKinstry, B.R. Buchbinder, R.M. Weisskoff, M.S. Cohen, J.M. Vevea, T.J. Brady, and B.R. Rosen. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254:716–719, 1992.

[7] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, 36:192–236, 1974.

[8] J. Besag and P. J. Green. Spatial statistics and bayesian computation. *Journal of the Royal Statistical Society*, 55(1):25–37, 1993.

[9] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer-Verlag, New York, 1991.

[10] R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of perfusion and deoxyhemoglobin changes during brain activation. *NeuroImage*, 5(4):32, 1997.

[11] P.A. Carpenter and M.A. Just. The role of working memory in language comprehension. In D. Klahr and K. Kotovsky, editors, *Complex Information Processing: The Impact of Herbert A. Simon.* Erlbaum, 1989.

[12] J.D. Cohen, S.D. Forman, T.S. Braver, B.J. Casey, D. Servan-Schreiber, and D.C. Noll. Activation of prefrontal cortex in a non-spatial working memory task with functional MRI. *Human Brain Mapping*, 1:293–304, 1994.

[13] J.D. Cohen, D.C. Noll, and W. Schneider. Functional magnetic resonance imaging: Overview and methods for psychological research. *Behavior Research Methods, Instruments, Computers*, 25(2):101–113, 1993.

[14] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. Technical report, Harvard School of Public Health, 1995.

[15] G. Dahlquist and A. Björck. *Numerical Methods.* Prentice Hall, 1974.

[16] T. L. Davis, R. M. Weisskoff, K. K. Kwong, R. Savoy, and B. R. Rosen. Susceptibility contrast undershoot is not matched by inflow contrast undershoot. In *Proceedings of the Society for Magnetic Resonance, Second Annual Meeting*, page 435. SMR, 1994.

[17] C. de Boor. *A Practical Guide to Splines.* Springer-Verlag, 1978.

[18] J. A. Detre, Z. Wang, M. M. Stecker, and R. A. Zimmerman. Vascular transit times in calcarine cortex: Kinetic analysis of r2* changes observed using localized 1h spectroscopy. *Magnetic Resonance in Medicine*, 34:326–330, 1995.

[19] T. J. DiCiccio, R. E. Kass, A. Raftery, and L. Wasserman. Computing Bayes Factors by combining simulation and asymptotic approximations. Technical Report 630, Department of Statistics, Carnegie Mellon University, 1995.

[20] W. F. Eddy. Comment on Lange and Zeger. *Journal of the Royal Statistical Society C*, 46:19–20, 1997.

[21] W. F. Eddy, M. Fitzgerald, C. R. Genovese, A. Mockus, and D.C. Noll. Functional image analysis software - computational olio. In A. Prat, editor, *Proceedings in Computational Statistics*, volume 12 pp. 39-49. Physica-Verlag, Heidelberg, (1996).

[22] W. F. Eddy, M. Fitzgerald, and D. C. Noll. Improved image registration using Fourier interpolation. *Magn. Reson. Med.*, 36:923–931, 1996.

[23] S. Forman, J. C. Cohen, M. Fitzgerald, W.F. Eddy, M.A. Mintun, and D. C. Noll. Improved assessment of significant change in functional magnetic resonance fMRI: Use of a cluster size threshold. *Magn. Reson. Med.*, 33:636–647, (1995).

[24] K. J. Friston, C. D. Frith, and R. S. J. Frackowiak. *Human Brain Mapping*, 1:69–79, 1994.

[25] K. J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1994.

[26] K.J. Friston, A.P. Holmes, J.B. Poline, P.J. Grasby, S.C.R. Williams, R.S.J. Frackowiak, and R. Turner. Analysis of fMRI time series revisited. *NeuroImage*, 2:45–53, 1995.

[27] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.*, 85(410):398–408, 1990.

[28] S. Geman and Geman D. Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intell.*, 6:721–741, 1984.

[29] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the ISI, Bulletin of the ISI*, volume 52, 1987.

[30] C. R. Genovese. Bayesian Response Analysis and Inference for Neuroimaging (BRAIN). Software Package, 1997.

[31] C. R. Genovese. Comment on non-linear fourier time series analysis for human brian mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society C*, 46:23–24, 1997.

[32] C. R. Genovese, D. C. Noll, and W. F. Eddy. Estimating test-retest reliability in fMRI I. *Magn. Res. Med.*, 38:497–507, 1997.

[33] C.R. Genovese, P.B. Stark, and M.J. Thompson. Uncertainties for two dimensional models of solar rotation from helioseismic eigenfrequency splitting. *Astrophysical Journal*, 443:843–854, 1995.

[34] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1989.

[35] P. J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[36] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.

[37] A.P. Holmes, R.C. Blair, J.D.G. Watson, and I. Ford. Non-parametric analysis of statistic images from functional mapping experiments. *J. Cerebral Blood Flow and Metabolism*, under review.

[38] X. Hu, T. H. Le, T. Parrish, and P Erhard. Retrospective estimation and correction of physiological fluctuation in functional mri. *Magnetic Resonance in Medicine*, 34:201–212, 1995.

[39] V. E. Johnson. A model for segmentation and analysis of noisy images. *J. Amer. Stat. Assoc.*, 89:230–241, 1994.

[40] V. E. Johnson, J. Bowsher, R. Jaszczak, and T. Turkington. Analysis and reconstruction of medical images using prior information. In C. Gatsonis, J. S. Hodges, R. E. Kass, and N. D. Singpurwalla, editors, *Case Studies in Bayesian Statistics*, volume 2, pages 149–218. Springer-Verlag, 1995.

[41] V. E. Johnson, W. H. Wong, X. Hu, and C. T. Chen. Aspects of image restoration using gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):412–235, 1991.

[42] M. A. Just and P. A. Carpenter. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99:122–149, 1992.

[43] M.A. Just, P.A. Carpenter, T. A. Keller, W. F. Eddy, and K. R. Thulborn. Brain activation modulated by sentence comprehension. *Science*, 274:114, 1996.

[44] K.K. Kwong, J.W. Belliveau, D.A. Chesler, I.E. Goldberg, R.M. Weisskoff, B.P. Poncelet, D.N. Kennedy, B.E. Hoppel, M.S. Cohen, R. Turner, H. Cheng, T.J. Brady, and B.R. Rosen. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. U.S.A.*, 89:5675, 1992.

[45] N. Lange and S. Zeger. Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society C Applied Statistics*, 46:1–29, 1997.

[46] E. L. Lehmann. *Nonparameterics: Statistical Methods Based on Ranks.* Holden Day, Oakland, CA, (1975).

[47] B. Luna, K.R. Thulborn, M.H. Strojwas, B.J. McCurtain, R.A. Berman, C.R. Genovese, and J.A. Sweeney. Dorsal cortical regions subserving visually-guided saccades in humans: An fMRI study. *Cerebral Cortex*, 8:40–47, 1998.

[48] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[49] S. Ogawa, D.W. Tank, D.W. Menon, J.M. Ellermann, S. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping using MRI. *Proc. Natl. Acad. Sci. U.S.A.*, 89:5951–5955, 1992.

[50] D. Pauler. *The Schwarz Criterion for Mixed Effects Models.* PhD thesis, Carnegie Mellon, 1996.

[51] J. B. Poline and B. Mazoyer. Cluster analysis in individual functiona brain images: Some new techniques to enhance the sensitivity of activation detection methods. *Human Brain Mapping*, 2:103–111, 1994.

[52] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, second edition, 1992.

[53] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.

[54] A. F. M. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *J. Roy. Statist. Soc. B*, 55(1), 1993.

[55] P.B. Stark. Strict bounds and applications. In P.C. Sabatier, editor, *Some Topics on Inverse Problems*, pages 220–230. World Scientific, Singapore, 1988.

[56] J. Talairach and P. Tounoux. *Coplanar Stereotaxic Atlas of the Juman Brain. Three-dimensional Proportional System: An Approach to Cerebral Imaging.* Thieme, 1988.

[57] K. R. Thulborn. Personal communication, march, 1998.

[58] K.R. Thulborn, J.C. Waterton, P.M. Matthews, and G.K. Radda. Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochem. Biophys. Acta*, 714:265–270, 1982.

[59] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1727, 1994.

[60] A. Vazquez. Non-linear temporal aspects of the blood oxygenation response in functional magnetic resonance imaging. Master's thesis, Bioengineering, University of Pittsburgh, 1996.

[61] G. Wahba. *Spline Models for Observational Data.* SIAM, 1990.

[62] J. B. Weaver, A. Y. Saykin, R. B. Burr, H. Riordan, and A. Maerlender. Principal component analysis of functional MRI of memory. In *Proceedings of the Society for Magnetic Resonance, Second Annual Meeting*, page 808. SMR, 1994.

[63] R. M. Weisskoff, J. Baker, J. Belliveau, T. L. Davis, K. K. Kwong, M. S. Cohen, and B. R. Rosen. Power spectrum analysis of functionally-weighted MR data: What's in the noise? In *Proceedings of the Society for Magnetic Resonance in Medicine, Twelfth Annual Meeting*, page 7. MRM, 1993.

[64] R.P. Woods, S.R. Cherry, and J.C Mazziotta. Rapid automated algorithm for aligning and reslicing PET images. *J. Comp. Assist. Tomog.*, 16:620–633, 1992.

[65] K. J. Worsley. Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Stat.*, 23:640–669, 1995.

[66] K. J. Worsley and K.J. Friston. Analysis of fMRI time series revisited – again. *NeuroImage*, 2:173–181, 1995.

[67] K.J. Worsley, A.C. Evans, S. Marrett, and P. Neelin. Detecting changes in random fields and applications to medical images. *JASA*, to appear.

[68] E. Zarahn, G. K. Aguirre, and M. D'Esposito. Empirical analyses of BOLD fMRI statistics i: Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage*, 5(3):179–198, 1997.

# Figure Captions

**Figure 1.** Two fMRI experimental designs, indicating the task being performed at every time throughout the experiment. The horizontal axis shows the corresponding image index, and the heights of the line segments serve only to visually separate the conditions. Panel (a) describes a simple, alternating two-condition design. Panel (b) shows a more complicated design with six conditions (R,F,Tr,T1,T2,T3).

**Figure 2.** Two voxelwise time series from the finger tapping experiment. The vertical lines show the separation between the conditions, and the conditions are labelled on the horizontal axes along with the image index (1–64). The vertical axes give the measured signal values. The time series in the top panel shows an apparently active voxel; note the correspondence between tapping and the pattern of signal change. The time series in the bottom panel shows little evidence of activity.

**Figure 3.** A t-map from the finger tapping experiment for a single slice of the subject's brain overlaid on the corresponding mean image. The white pixels indicate the locations for which a t-statistic comparing the signal in the tapping and rest conditions exceeded 4. This is an "axial" slice, orthogonal to the long axis of the subject's body. The image is shown according to radiological convention, so the right side of the image is the left side of the subject's brain. The bottom of the image is the back of the subject's head.

**Figure 4.** Two voxel time series from adjacent voxels. One (a) shows substantial signal drift and the other (b) shows little. The superimposed curve in both cases shows the fitted value under our model; neither exhibits a strong activation response. The vertical axis in each case is the signal intensity, in arbitrary units, and the horizontal axis is the image index.

**Figure 5.** A typical shape for the hemodynamic response in fMRI data to a single period of task performance. The time during which the task is performed is marked, and the curve shows the pattern of signal change that results. This curve is a polynomial bell function, as described in the text. The labels indicate the role of the various shape parameters in our model. The rise, fall, and skew parameters here affect the shape of the corresponding part of the curve.

**Figure 6.** Various results using the example data for a single slice of the subject's brain. The solid outline copied on each map encloses the brain to facilitate comparison across panels. The gray-scale in each panel refers to a different quantity as described below.

**(a)** The mean over time of the functional images for a single slice of the example data set. The gray-scale for this panel shows the signal intensities in the image. The fuzzy ring of voxels surrounding the brain is the fat outside the subject's skull.

**(b)** Domination probabilities $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\}$. The large number of nearly white voxels results from a posterior mass for the corresponding responsiveness parameters concentrated at zero. The gray-scale for this panel shows the probability values.

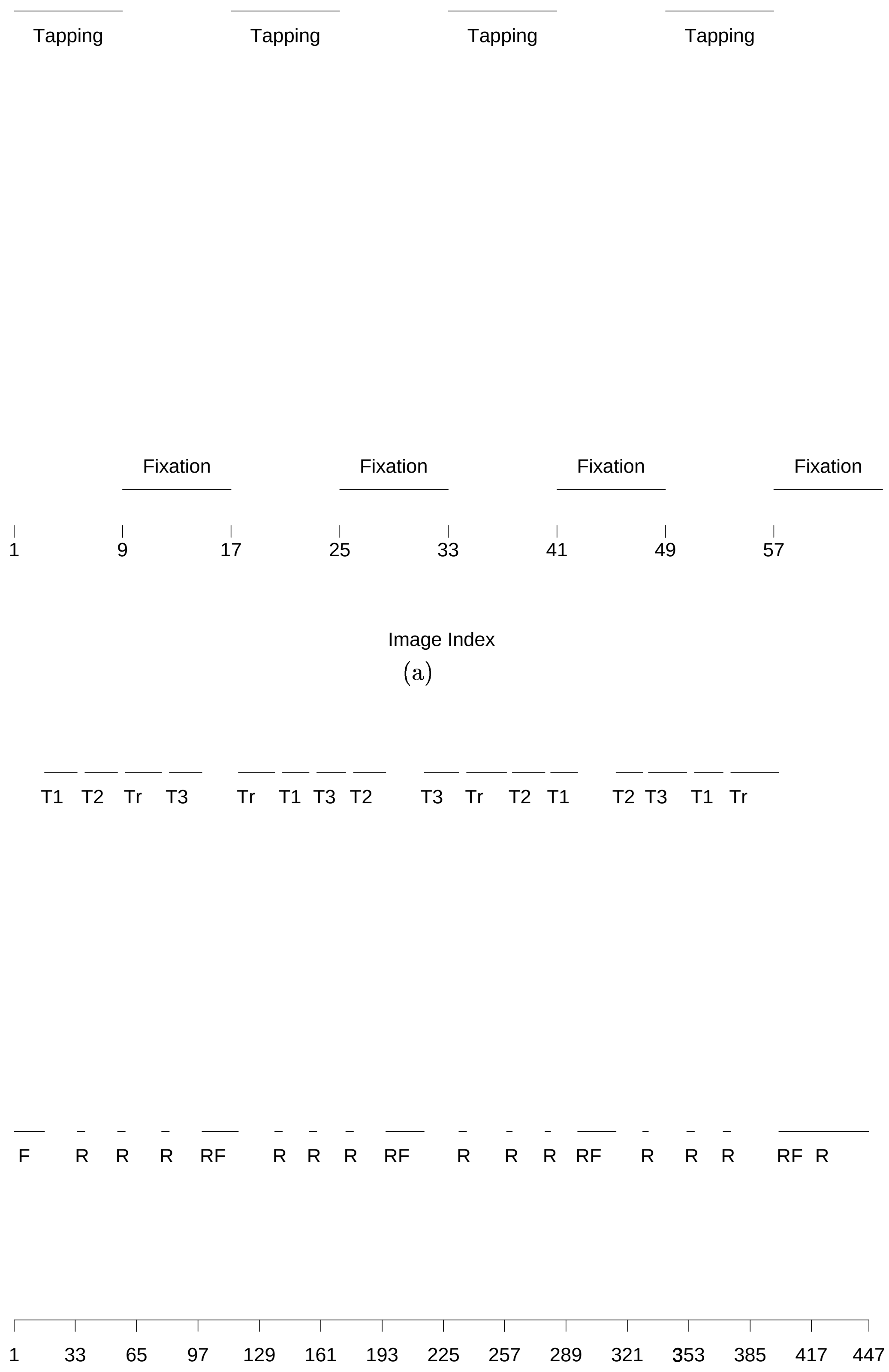
**(c)** A traditional t-map thresholded at the arbitrary but often used value of ±4. The nominal significance levels suggested by theory do not give the expected error rates, most likely because of complexity in the noise distribution that is unaccounted for by the test. The gray-scale for this panel shows the t values.

**(d)** Monotonicity probabilities $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\}$. The gray-scale for this panel shows the probability values.

**Figure 7.** Estimates of the marginal probabilities $\mathrm{P}\{\, \Gamma_c(R) > u\, \}$ as a function of $u$ for four task conditions Tr, $T_1$, $T_2$, and $T_3$. The selected region is a set of 21 contiguous voxels surrounding the main cluster on the lower right in the previous maps.

**Figure 8.** Samples from joint posterior distributions for $(\gamma_{\mathrm{T}_2} - \gamma_{\mathrm{T}_1}, \gamma_{\mathrm{T}_3} - \gamma_{\mathrm{T}_2})$ restricted to the positive quadrant. The figure shows the results for the 21 contiguous voxels surrounding the prominent cluster in the lower right Figure 6d, with one icon per voxel. The axes for each icon range from 0 to 0.05 in each direction. The structure of the distributions gives an indication of the shape of the response changes from $\mathrm{T}_1$ to $\mathrm{T}_2$ to $\mathrm{T}_3$. Two voxels are marked with arrows, one on the left and one on the right. The marked voxel on the right tends to show a more pronounced change between $\mathrm{T}_1$ and $\mathrm{T}_2$ than between $\mathrm{T}_2$ and $\mathrm{T}_3$. The marked voxel on the left tends to show a very small change between $\mathrm{T}_1$ and $\mathrm{T}_2$ but a large change between $\mathrm{T}_2$ and $\mathrm{T}_3$.

Figure 1


Tapping
Tapping
Tapping
Tapping
Fixation
Fixation
Fixation
Fixation
1
9
17
25
33
41
49
57
Image Index
T1
T2
Tr
T3
Tr
T1
T3
T2
T3
Tr
T2
T1
T2
T3
T1
Tr
F
R
R
R
RF
R
R
R
RF
R
R
R
RF
R
R
R
RF
R
1
33
65
97
129
161
193
225
257
289
321
353
385
417
447
Image Index

(a)

(b)

Figure 2

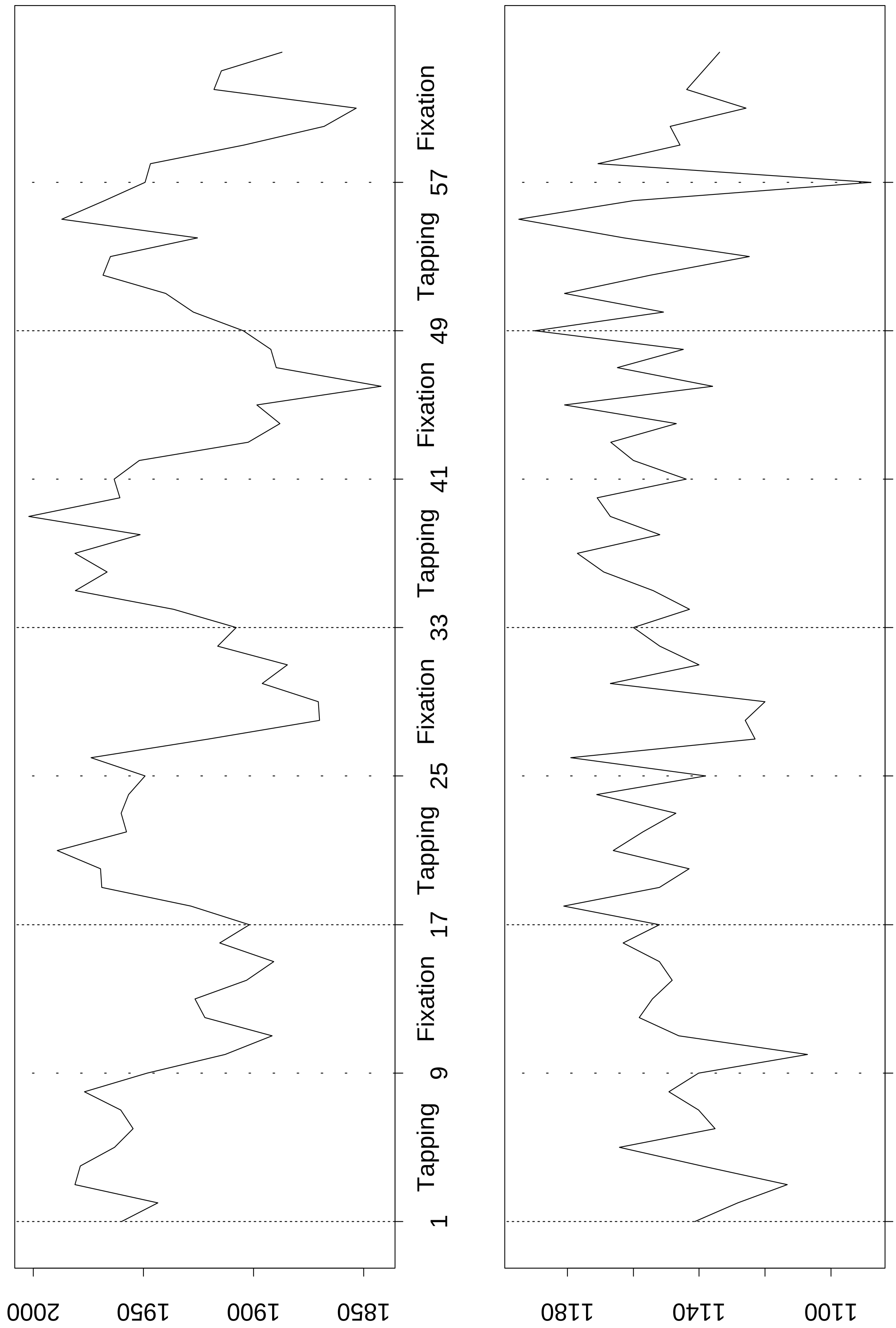

2000
1950
1900
1850
1180
1140
1100
1
9
17
25
33
41
49
57
Tapping
Fixation
Tapping
Fixation
Tapping
Fixation
Tapping
Fixation

Figure 3

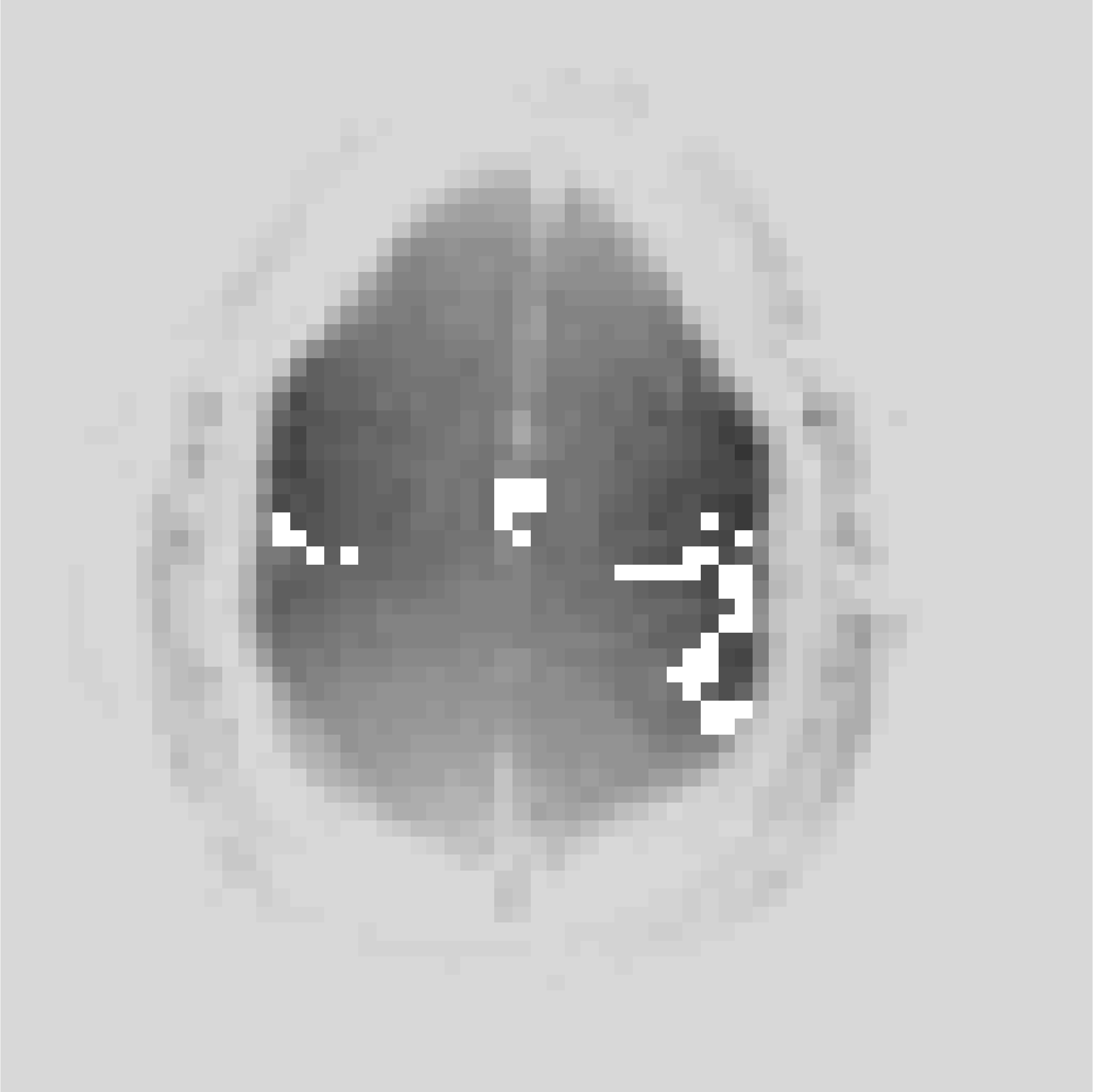
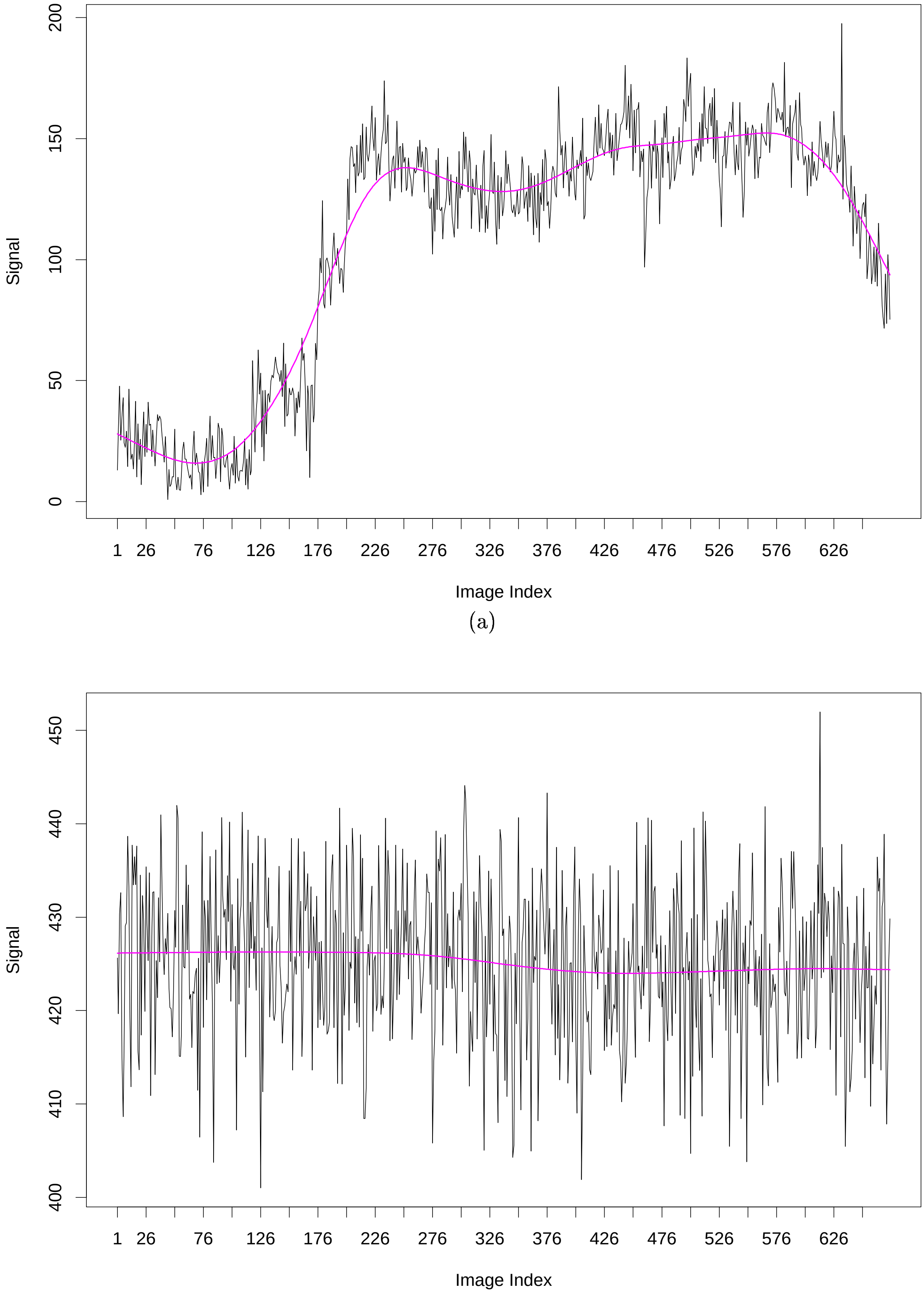


Figure 4


Signal
200
150
100
50
0
1
26
76
126
176
226
276
326
376
426
476
526
576
626
Image Index
Signal
450
440
430
420
410
400
1
26
76
126
176
226
276
326
376
426
476
526
576
626
Image Index

(a)

(b)

Figure 5

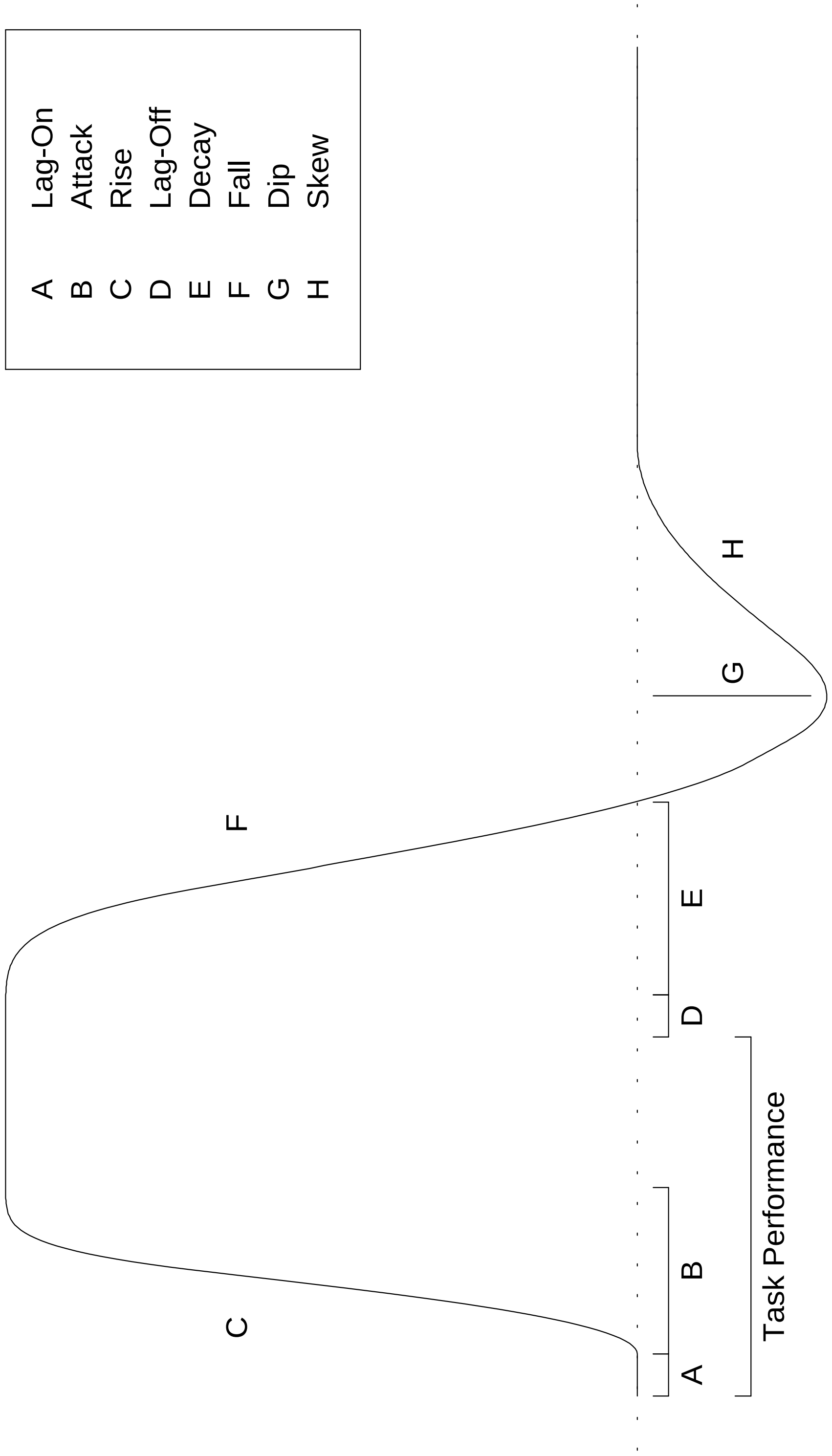
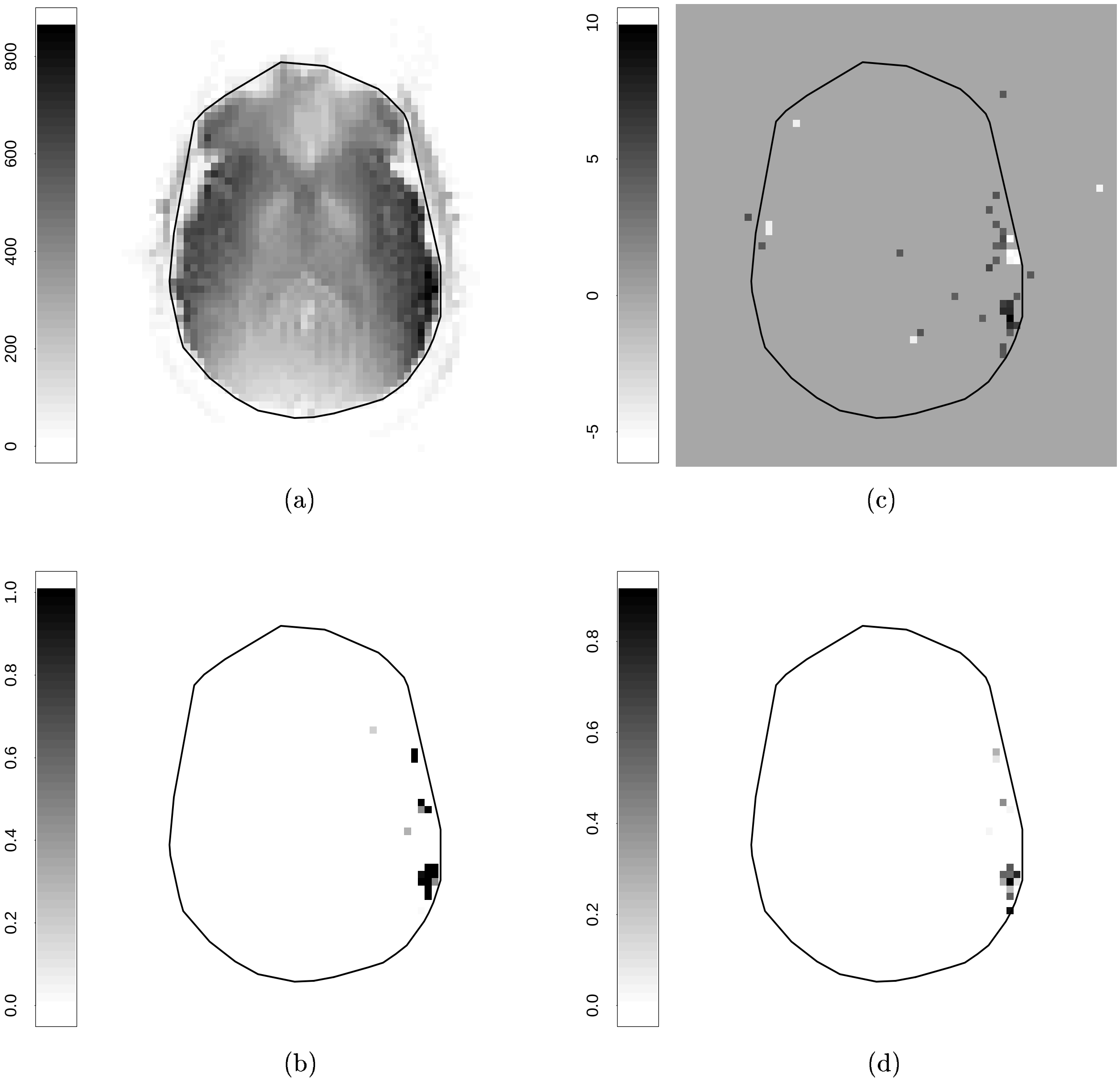

A Lag-On
B Attack
C Rise
D Lag-Off
E Decay
F Fall
G Dip
H Skew
A
B
C
D
E
F
G
H
Task Performance

Figure 6



(a)

(c)

(b)

(d)

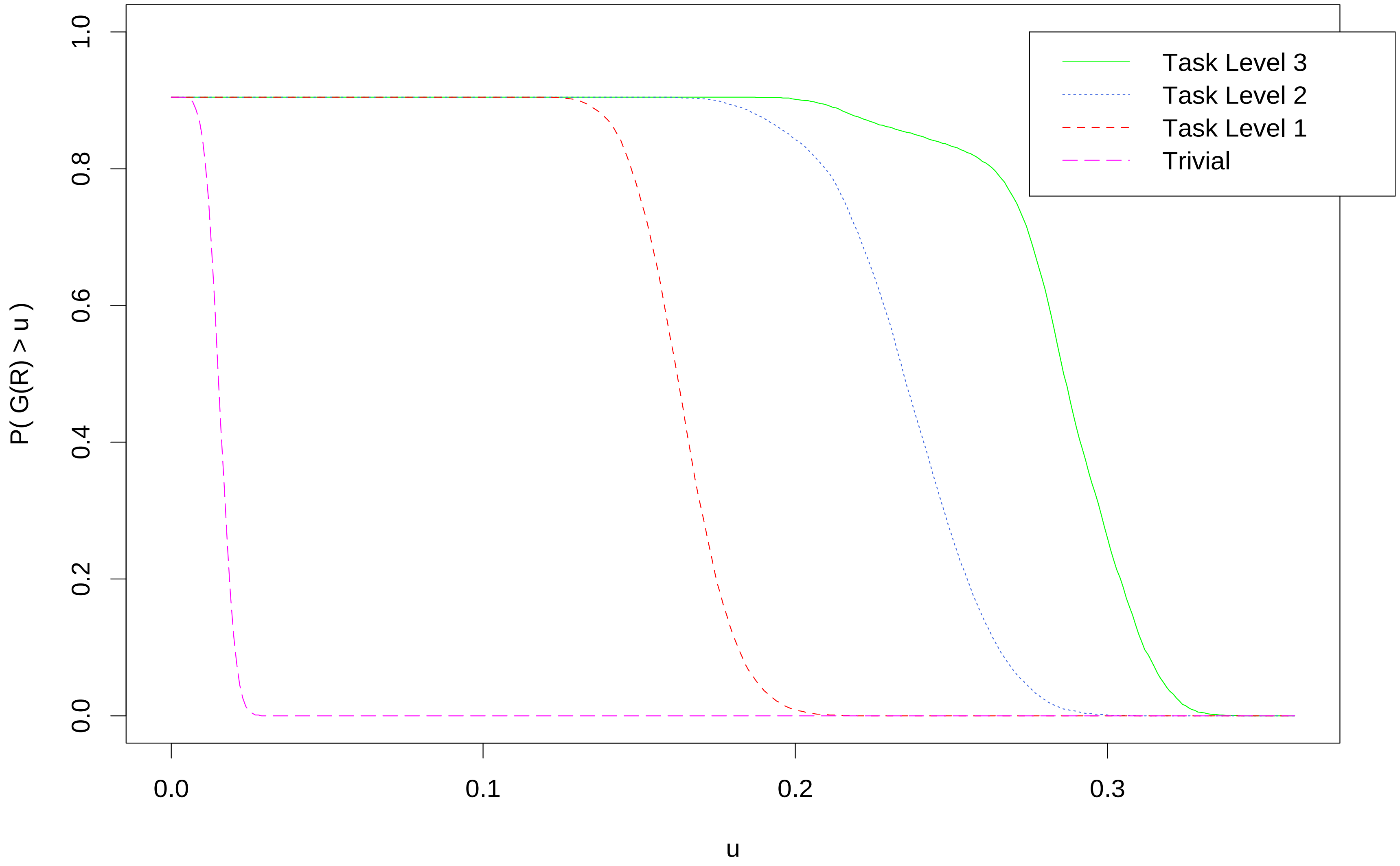

Task Level 3
Task Level 2
Task Level 1
Trivial
P( G(R) > u )
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
u

Figure 7

Figure 8