Statistical Inference in Functional Magnetic Resonance Imaging

Christopher R. Genovese

Carnegie Mellon University

Technical Report

Department of Statistics, Carnegie Mellon University

Running Head: Statistical Inference in fMRI

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

Acknowledgements: 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`

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 sequence of behavioral tasks while Magnetic Resonance (MR) images of the subject's brain are acquired. 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 accomplishing this is a quintessentially statistical problem.

Consider a very simple experiment with two task conditions. While MR images are obtained at regular intervals, the participant alternates between periods of rest (condition 1) and periods of tapping her index finger against her thumb at a steady rate (condition 2), focusing throughout on a marked point on a projection screen. Each MR image is three-dimensional and consists of measurements over a grid of small volume elements. Figure 1 shows the series of measurements taken from two such volume elements with vertical lines dividing the conditions; the pattern of task performance during the experiment is given in Figure 2a. The rest periods here serve as a control for the motor task being studied; brain activity in response to finger tapping but not to rest is attributed to neural processing that is unique to the motor task. The effects of brain activity appear as slight but systematic changes in the signal over time. For example, in Figure 1, small signal changes that coincide with the experimental design are evident in the first time series but not in the second. Current methods for analyzing fMRI data are based on detecting such patterns; each location is classified as active or inactive, usually by a statistical hypothesis test. One common method is to compare the level of the signal between the two conditions for each volume element by a two-sample t-test. The result is a classification map which shows how each volume element is classified; one slice of such a map is shown in Figure 3. The white locations, here overlaid on a mean image, show the volume elements in which the average signal during the tapping condition is "significantly" larger than that during the rest condition. These locations correspond reasonably well to the motor areas on the cortex. This example demonstrates only the simplest application of the technology; fMRI can be used to address more important scientific questions as well. However, there are many statistical issues that must be considered when analyzing such data.

Perhaps the most salient is the problem of multiple comparisons because the classification map requires performing thousands of hypothesis tests; other more serious issues involve the complexity of the noise and the underlying structure of the signal. In this paper, we will argue that a new framework for statistical analysis is needed to face the many challenges, scientific and statistical, that arise in making inferences from fMRI data.

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. The noise in the data is complicated: there are nonlinear trends, 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. Neural activity manifests itself in these data as a perturbation in the measured signal whose magnitude is typically smaller than the noise level. This perturbation is the result of a blood-flow response in the brain with both non-linear and non-additive features. Moreover, the spatial structure of the problem is complicated by highly irregular tissue boundaries and the presence of confounding factors such as large blood vessels.

We take a new approach to inference from fMRI data that can address this inherent complexity. Whereas current methods use classification procedures based on simplistic statistical models, we develop a detailed, nonlinear hierarchical model for the data that represents the structure of the underlying processes as accurately as possible. If $Y_v(t)$ is the measured signal at location v and time t, we take

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

where each term captures a distinct component of variation in the process, as explained in section 4. The model parameters are then related across deeper levels of the hierarchy. Since the model accounts for the detailed structure of the signal and can incorporate a wide range of noise processes, estimates under the model are more sensitive discriminators between artifact and activity than are the hypothesis tests used for classification. The model also provides an accurate assessment of uncertainty that is essential to making well-supported inferences, and it makes new types of questions accessible to direct analysis.

Although fMRI data are collected in images, the essential content here centers on *inference*—inference about spatio-temporal fluctuations in the measured signal and about fundamental motor, sensory, and cognitive processes. Thus one of the key statistical challenges is to develop procedures that enable scientists to make rigorous inferences regarding the questions they want to answer. This requires an appreciation of the breadth and variety of questions being asked. However, the classification procedures currently used in fMRI target only a single question: where did the activity take place? Strong scientific claims must then be based on classification maps, even though many of the scientific questions require more than just the location of activity to be answered. In contrast, our model provides an inferential framework that can be used to directly address a broad class of scientific questions that cannot be addressed by classification alone.

Beyond its implications for clinical and basic neuroscientific research, fMRI holds a great deal that is of interest to statisticians. It is a prototype for a class of statistical problems that is arising with increasing frequency in applications—namely, large data sets realized from a complex process with both spatial and temporal structure. The goal here is not prediction, as in many conventional spatial applications, but the elucidation of subtle structure over space and time amidst many intertwined sources of variation, all of which are subject to strong substantive constraints. The methods we develop to handle these data—from data management to visualization to spatio-temporal model fitting—carry over to diverse applications. Finally, fMRI is a highly inter-disciplinary field, and it offers a great opportunity for statisticians to have a critical impact on neuroscience and diagnostic radiology for many years to come. There is a tremendous supply of statistical problems that need solution and a tremendous demand for statistical guidance and methodology. Indeed, over the past few years, we have worked closely with scientists from many disciplines involved with fMRI, including psychologists, neuroscientists, neurologists, radiologists, physicists, engineers, and statisticians. The whole enterprise rests on the successful sharing of perspectives from all of these disciplines, and it is important here to acknowledge the influence of our many collaborators on our view of the field. Their time and effort spent communicating the scientific background of their field, the issues that are central to their research, and the questions that they are trying to answer is reflected in all of this work.

Our goal in this paper is to put forward a new standard for inference from fMRI data. As a foundation for our approach, in section 2 we describe the basics of fMRI and illustrate some of the scientific issues that need to be addressed if fMRI is to achieve its full potential. In section 3, we introduce what we call the localization paradigm on which current methods are based and discuss the limitations of this strategy. We describe the basic structure of our modeling framework in section 4 and show how we incorporporate the available prior information into the model in section 5. 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 6, we illustrate some of the possibilities in light of a specific data example. A full case-study using our approach is given in [36]. In section 7, we describe the computational techniques that we use to fit the model and implement our inferential procedures. Finally, in section 8, we discuss some more general extensions to the model and some directions for future research.

2. fMRI and Cognitive Neuroscience

2.1. Overview of fMRI

Magnetic Resonance (MR) is a consequence of the characteristic behavior of certain atomic nuclei (e.g., hydrogen) in a magnetic field. In particular, a precise modulation of the ambient magnetic field excites these nuclei so that they precess in phase with each other for a short period of time. The motion in phase of so many nuclei causes a macroscopic oscillation in the magnetic field, which in turn induces a detectable current in any nearby coil of wire. This is the MR signal, and the greater the density of the selected nucleus in the scanning volume, the greater the measured signal. The key to applications of the MR phenomenon (and the source of the word resonance) is the selectivity of this excitation—even though an entire volume is exposed to the magnetic manipulations, only nuclei of a selected species will be excited.

An MR scanner is thus a machine for measuring the density of a chosen nuclear species within a given volume of space. It consists of (i) a large, super-cooled magnet to produce a strong, uniform field throughout the scanning volume, (ii) electronic components to modulate the magnetic field in a controlled way very quickly, and (iii) a receiver coil to convert changes in the field generated by the nuclei to detectable currents. The MR scanner can be used in several ways. By spectral decomposition of the received signal, the chemical composition of a substance can be determined; this is Magnetic Resonance Spectroscopy. By encoding spatial information in the phase and frequency of nuclear precession during scanning, high-resolution images of living tissue can be obtained non-invasively; this is Magnetic Resonance Imaging (MRI).

In MRI, each image shows the density of the chosen nuclear species (usually hydrogen) over the given volume. Although two-dimensional slices are usually presented, MR images are actually three-dimensional. The scanning volume is divided into a grid of small volume elements, or *voxels*, and an average nuclear density is recorded for each voxel in each image. We will call this the measured MR signal for the given voxel. It is worth noting that through the complicated spatial encoding, the MR scanner actually records the Fourier transform of the entire image first, and the MR signal values associated with individual voxels are subsequently reconstructed. There are many image acquisition schemes, each with a different sampling path through Fourier space.

Functional MRI uses a sequence of MR images to glean information about neural activity. This information is not obtained directly from the MR signal but rather from perturbations to the signal caused by a local magnetic anomaly that is a by-product of neural metabolism. Specifically, as the firing frequency for a group of neurons increases, the neurons' metabolism also rises. Through a mechanism that is not fully understood, the metabolic increases cause the local blood vessels to dilate; an inflow of oxygenated blood into the area results. Since oxygenated and de-oxygenated blood have different magnetic properties [68], the MR signal changes slightly as the balance of oxygenated and de-oxygenated blood in a region changes. This Blood Oxygenation Level Dependent (BOLD) effect is detectable in the MR signal and makes it possible to localize neural activity via the MR phenomenon, albeit indirectly.

Functional MRI developed from this observation [54, 6, 49] and is now the premier technique for studying brain function. While other functional neuroimaging modalities, such as Positron Emmission Tomography (PET) [71], were developed earlier, fMRI offers superb sensitivity and an unprecedented combination of good spatial and temporal resolution, without any of the limitations that result from exposing subjects to radioactivity. While there are several other potential mechanisms for functional imaging with MR, the BOLD effect has the most promising signal-to-noise characteristics and is the most commonly used. The BOLD perturbations are nonetheless small relative to the noise level, which is a non-trivial part of the statistical challenge here. Moreover, since the BOLD effect results from a *hemodynamic response* (i.e., changes in the local blood flow), it is also slower and broader than the neural activity for which it is a proxy. There is an inherent trade-off between spatial and temporal resolution in fMRI data: as images are acquired faster, the minimum voxel size increases, and as voxels are made smaller, the maximum acquisition rate decreases. The resolution of a typical fMRI study is on the order of 1-3 s by 30-50 mm^3.

2.2. Neuroscientific Questions

To appreciate the potential of fMRI for clarifying how the brain works, it helps to understand the nature and scope of the questions being asked by neuroscientists. Here, we present an array of representative scientific issues that illustrate the questions neuroscientists face and point to the ways in which fMRI can shed light on the workings of the brain. The analyses presented in section 6 and [36] draw from experiments related to these questions.

2.2.1. Maintaining Information During Processing: Working Memory

Why are some sentences more difficult to understand than others? Why is it harder to multiply two large numbers together than to add them? Why is the video game Pac-Man more challenging than the game Pong? One answer is that more difficult tasks require that a greater amount of information be maintained in memory while the task is being performed. 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. In the multiplication of large numbers, many intermediate products need to be computed and stored until they can all be combined into a final answer. In Pac-Man, the player must attend to the locations of numerous threats and opportunities while planning a path through a maze. The brain has a set of general mechanisms, known as *working memory* [3], for maintaining such information and intermediate results 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 into working memory is to understand how the brain implements working memory—how this resource is allocated and used—in order to elucidate the dynamics of human cognition.

In the past, the only empirical tools available to address this question were comparisons based on behavioral data. Psychologists studied people's performance across tasks that differentially exercised working memory. Commonly used behavioral measures include the time to perform the task, the rate of making errors, and a space-time trace of eye-movements during the task. These types of data help to clarify the nature and degree of working memory's impact on performance.

As an example, consider the "n-back" experiment of [14]: A sequence of letters is presented, one at a time, and the subject is instructed to read each aloud and push a button whenever the letter that appears also appeared exactly n letters previously. Working memory is engaged in this task because the identity and relative position of the most recent n letters must be maintained for potential matching *while* new letters are being processed. Results show that subjects' times and error rates both increase with n, until subjects can no longer perform the task adequately.

To account for this kind of data, psychologists have developed cognitive theories based on computational models. These theories provide an abstract representation of the processes that underlie a given set of tasks and make specific predictions about performance, *e.g.,* what component processes will be used and when, the order and timing of processing in each component, the actions to be initiated, and the distribution of responses that result. For example, current cognitive theories explain the results of "n-back" experiments by attributing to each individual a limited supply of the working memory resource [14, 46]. The utilization of this resource by a cognitive task is considered a good measure of how "hard" that task is [46]. As working memory is engaged in a sequence of increasingly difficult tasks, the individual must work harder, until the limit is reached, when there is a serious degradation in performance. But there is more to the structure of working memory than a single limited resource. Current theories [12] also posit that working memory allocation is hierarchically arranged with different pools of resources specialized to particular types of computations. For instance, experiments in [59, 62, 9] suggest distinct working memory pools for symbolic and spatial information. Similarly, the model of [46, 13] posits several distinct resource pools for sentence comprehension [47] and others for mental rotation of objects [45]. The size and degree of specialization of these pools can vary, and if the capacity of a particular resource pool has been used, it can draw support from another unrelated but untapped pool.

Questions regarding the mechanism of working memory abound: Is the resource limitation view a valid one? How does capacity 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. An obstacle to answering these questions is that the current theories provide a richer set of predictions than standard behavioral data alone can test. However, with the emergence of fMRI, more refined data are available that can provide new and more detailed tests of these theories.

2.2.2. Fine Control of Sensory Input: Eye Movements

It is easy to take for granted the precision and speed with which our eyes can move and focus onto any target in the field of view, but the eye movement system is a complex nework of distributed neural processes that is only beginning to be understood. The eyes move continually during visual processing, usually without explicit control or awareness, to scan the key features of a scene. This is essential to our ability to attend to multiple features of our environment and to extract useful information from complex visual stimuli. Eye movement abnormalities are a common neurobehavioral associate of schizophrenia, neurodevelopmental disorders and many neurologic diseases, so the study of how the brain controls eye movements also promises to yield deeper insights into the brain abnormalities at the root of these conditions.

Eye movements have been studied extensively in both humans and monkeys. Two basic processes of interest are specific to the occulomotor system: (i) saccades, the rapid repositioning of the eye, and (ii) pursuit, the visual tracking of a moving object. In experiments, subjects perform a task that exercises a particular mixture of the component processes that subserve eye movements. A simple example is called the memory guided saccade task [65]. A subject visually fixates on a cross in the center of the visual field while a flash of light appears and disappears in the periphery, at a random time and location. After a delay period, the subject must make a saccade to the location of the flash. This task involves processing of visual stimuli (the cross and flash), motor control of the eye suppress and then make the saccade, spatial working memory to hold the position of the flash "on-line" during the delay, and other processes that regulate what the subject is attending to at any time. In monkey studies, data are obtained from direct electrical recordings of changes in the firing rate of neurons in a well-localized region of interest. The monkey is trained to perform an eye movement task on cue, and an electrode is inserted into the monkey's brain to record activity from single neurons in a specific area. The task is repeated many times with the electrode placed across a range of locations; in this way, a detailed map of the brain areas involved in the task can be constructed. These data provide good spatial and temporal resolution for a single brain region, but because of the difficulty in placing electrodes and maintaining recording, it is infeasible to target and record several separated brain regions simultaneously. Since direct recording is too invasive for use in humans, human eye movements have previously been studied only by monitoring eye position and movement and by studying patients after stroke or invasive surgery.

A general picture has emerged from these data regarding what areas of the brain subserve eye-movement processes. Although there is a strong homology between humans and monkeys, there are also notable differences, and a detailed delineation of the system in the human brain is still needed. To develop and test a complete theory of how eye movements are implemented requires study of the component sub-processes of the human system, of the interactions among these sub-processes, and of their functional connectivity. Specifically, neuroscientists are interested in the sequence and relative timing of each processing step. They would like to determine if pursuit and saccade are implemented separately and to characterize the pathways linking these sub-systems. Addressing these issues requires detailed spatial and temporal information across widely separated regions of the brain. The advent of neuroimaging has opened the door to exciting new advances in this area, not only because of the possibility of studying the human system but because of the capability of studying the entire brain simultaneously.

2.3. The Potential of fMRI

The brain is a computational system built on a hierarchy of simple functions that interact and cascade to produce complex behaviors. The fundamental goal of cognitive neuroscience is to understand how the building blocks of cognition (e.g., working memory, visual processing, attention) emerge from the brain's basic architecture. The potential of fMRI to address this goal–and neuro-scientific questions like those above—lies in its ability to measure the brain's responses at a fine level of spatial and temporal detail *during* cognitive processing. Functional MRI data offer a unique, global view of the system dynamics, a view that bridges the divide between the micro-level information obtained by recording single neurons and the macro-level information obtained by studying behavior.

During an fMRI experiment, the subject is placed in the scanner and performs a carefully designed sequence of tasks while MR brain images are acquired at regular intervals. Through the particular sequence of tasks performed, the experiment manipulates the input to the system in terms of the mixture of basic cognitive processes that is required at each point in time. By studying the relationship between input and response, we learn how the basic processes involved act and interact to produce more complex brain functions. Cognitive theories predict specific aspects of this relationship and can often be tested by an appropriately designed experiment.

Hence, neuroscientific questions translate into questions about the pattern of responses produced by an experiment. Below we describe several questions about the nature of the response that arise again and again in the context of neuroscientific research. In each case, the key challenge is to develop statistical methods with which relevant inferences about the pattern of response can be made. As we will show, current statistical techniques can deal effectively with only the simplest of these; in contrast, the methods we present in this paper can handle all of them.

Localization. The most basic question to ask about a neural process is where in the brain that process is implemented. That question may not always have an easy answer because computation in the brain may be widely distributed. Nonetheless, fMRI can be useful for localizing a process. Since the BOLD effect is a perturbation to the MR signal that is associated with neural activity, the involvement of a particular region of the brain can be gauged by comparing the measured response when the process is engaged to the response when the process is not engaged. This leads to the simplest fMRI experiment: the subject alternates between two tasks, where each is performed repeatedly for a certain length of time before the next task begins. One of these tasks invokes the process under study, and the other task, which serves as a control, does not. (See Figure 2a.) A commonly used control is visual fixation, where the subject holds his or her gaze on a marked location in the center of the visual field. In contrast, a pure rest condition is not necessarily a good control because it does not constrain the subject to a consistent behavior. The logic of localization is that the voxels whose MR signals show a response to the task of interest but not the control are involved in the process under study [22]. There are, however, several assumptions behind this logic that can complicate the interpretation of localization results; see [?].

Graded Responses. Many theoretical predictions revolve around how the system response changes as the input is varied along a single dimension. For example, as sentence difficulty increases, we expect a greater proportion of the available working memory resources to be allocated to comprehending the sentence. A theory that specifies a particular arrangement and dependence among resource pools predicts a specific functional relationship between sentence difficulty and working memory utilization. With fMRI, this relationship can be estimated and the theory tested. An experiment designed to accomplish this requires several different conditions, including at least one control. The conditions correspond to versions of a task that are graded with respect to the degree of involvement they require from the process under study [64]. The subject performs each version of the task repeatedly for a certain length of time, with the conditions in suitably randomized order. (See Figure 2b.)

Dissociations. Cognitive theories often make a fundamental distinction between different types of processing (e.g., spatial and symbolic working memory) to the point of predicting a separate neural implementation. One piece of evidence in support of such a theory would be a dissociation, a situation in which one type of processing is present while the other is absent and vice versa. Historically, neurological patients whose brain damage impaired one process but not another provide evidence of dissociations, although these are infrequent and rarely clean-cut. Functional MRI data can provide other types of dissociations, that can be used to distinguish two cognitive processes in unimpaired subjects. For example, fMRI experiments can identify what we will call location, pattern, and manipulation dissociations. Location dissociations, which are the weakest of the three, occur when two types of processing activate distinct brain regions. This provides some evidence, although it is not definitive, that the processes are implemented in distinct sub-networks. Pattern dissociations occur when two processes lead to temporal responses of different shapes. The memory guided saccade task described above invokes several distinct processes, including an "executive" process that initiate eye movements and other behaviors and a working memory process that maintains the destination of the eventual saccade. The executive process is expected to give rise to short bursts of activity whereas the working memory process should exhibit activity throughout the delay period. This distinction in the temporal pattern of activity can be detected and used to distinguish the two types of processing. Finally, manipulation dissociations occur when the two processes are differentially sensitive to a particular manipulation of the task (e.g., increasing its difficulty). In all of these cases, the experimental design can be tailored to maximize the discriminability of the responses.

Inferences across Subjects. In order to generalize the results of an experiment to a broader population, it is necessary for investigators to relate their results across multiple subjects. If all brains looked alike, it would be a simple matter to compare the results from many fMRI experiments. Unfortunately, this is not the case: while the functional organization of the brain is topologically similar across individuals, the physical geometry of the cerebral cortex can vary substantially from person to person. The approach most commonly used to circumvent this difficulty in functional neuroimaging is to map subjects' brains onto a common coordinate system, the Talairach atlas [66], and then average over subjects. The Talairach atlas was generated 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. Averaging in Talairach coordinates is far from satisfactory, however; it adds a critical source of variation to the inferences (see [52] for a demonstration) and lends an illusory aura of anatomic certainty to the results. In sections 6 and 8, we will present an alternative that can, in some cases, yield inferences across subjects that do not suffer from these difficulties.

3. Current Methods and The Localization Paradigm

3.1. Development of the Paradigm

The goal of the earliest fMRI studies [6, 49, 54] was to demonstrate that the BOLD effect could reliably detect and locate activation. To obtain as large an effect as possible, these studies used fundamental tasks such as viewing visual stimuli (e.g., flashing checkerboard patterns or LED's) or making simple motor operations (e.g., finger tapping) that tend to yield robust responses. The experimental designs were simple, with one task and one control condition. PET studies and neural recordings in animals served as a standard for where the activity should be during performance of these tasks. The success of these early efforts caused tremendous excitement in the neuroimaging community, and in response, there was a wave of research on new acquisition techniques to improve the quality of fMRI data. The introduction of statistical methods to the field paralleled this development. Because of the need to detect and locate activity, the focus was on the *localization problem*: How can data from an fMRI experiment be used to identify which parts of the brain activate during performance of a given task?

Localization is viewed in fMRI as a problem of classification, where each voxel is classified as *active* or *inactive* with respect to the task of interest. The assumption is that in an active voxel a significant proportion of neurons respond to the task whereas in an inactive voxel few if any do. For each voxel, the sequence of MR signals across the images form a time-series, and an active voxel may be distinguished by the presence of the BOLD perturbation at times during which the task is being performed. Hence, a natural approach is to compare the mean signal in the task and control conditions using a statistical hypothesis test, applied voxel by voxel. Although other classification procedures have been considered [74, 56], this testing-based approach is the statistical analysis used by the majority of fMRI researchers.

A variety of testing procedures has been employed in fMRI, and although most of these have a statistical heritage, few if any were introduced to the field by statisticians. The most commonly used is the two sample t-test in which the MR signal values from images associated with any two experimental conditions are compared. Nonparameteric tests, such as the Kolmogorov-Smirnov test [4, 51], also appear frequently in the fMRI literature. A variant of the t-test that is sometimes used, called the split sample t-test [27], classifies a voxel as active only if t-tests carried out separately for data from the first and second halves of the experiment both indicate significance. Another common test statistic is the correlation coefficient of the voxel time-series with a fixed reference curve designed to approximate the shape of the BOLD signal perturbation [5]. Whereas the t-test implicitly uses an ideal square wave as the pattern of signal change, the correlation allows a more realistic shape for this curve. Another generalization of the t-test is based on an Analysis of Variance (ANOVA) model that blocks on time to account for uncontrolled changes over the course of the experiment [14, 15]. Here, F-statistics are used to test if specified contrasts among the conditions are non-zero. Similar but more recent methods use a general linear model to capture temporal variation, possibly after spatial and temporal smoothing [29, 77]. Finally, if the experimental design alternates periodically between task and control, spectral methods (e.g., F-tests based on periodogram ordinates [10]) can be used to test for large power in any frequency band [75, 28].

3.2. Failures of the Paradigm: Model Assumptions

While the various hypothesis tests often give reasonable results and can reliably detect large signal changes, there are several inherent complexities in fMRI that undermine the effectiveness of these procedures. First, the noise in fMRI data is complicated, and the testing procedures are based on simplistic statistical models. For example, signal drift over time is pervasive, and as demonstrated in Figure 4, these often highly nonlinear trends can vary in shape even across neighboring voxels. The properties of the noise also depend on the image acquisition scheme that is used; Figure 5 shows a striking and nonlocal correlation function induced by one such scheme. Such features of the noise can wreak havoc on the classifying tests. Second, the structure of the system's hemodynamic response is far from simple. For example, the temporal shape of the activity-induced perturbation varies across the brain and is very sensitive to the local vasculature. Nor is it an ideal box car. Rather, the signal changes lag the task, both at beginning and end, rise and fall in distinct ways, and can dip below baseline for extended periods. Figure 6 shows a typical signal profile for the hemodynamic response; Third, applying a set of hypothesis tests to characterize a spatio-temporal response gives rise to a number of technical difficulties. For example, an fMRI classification map usually requires performing 10,000 to 100,000 tests for each contrast among conditions. Since the signal changes are small relative to the noise level, common adjustments for multiple comparisons (e.g., Bonferroni or Tukey) often (conservatively) eliminate signs of activity. An equally serious problem is the lack of consensus in the fMRI community on how to choose significance thresholds for the tests. Because of unaccounted features of the noise, theoretical significance levels tend to be inappropriate for fMRI data. While systematic approaches have been proposed [35], thresholds are usually set to arbitrary values, often "by eye". This practice raises the clear danger that the expectations of the investigator will influence the threshold choice, explicitly or implicitly, and thus distort the conclusions. Another issue is that the voxels classified as active are often used for downstream analysis with potentially serious selection biases as a consequence.

These problems have evoked two responses in the fMRI literature: (1) the search for new testing 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 mentioned above, 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. Moreover, interpreting the results of the classifications usually requires a good amount of *ad hoc* reasoning without support from the data, for instance to argue that certain apparently active voxels are truly active while others are spuriously so.

More recent efforts with the involvement of statisticians have attempted to deal with some of these problems. Forman et al. [26] developed an adjustment to voxelwise tests, called the cluster size threshold. Here, a voxel is classified active only if its test statistic is above threshold and if the voxel is contained in a contiguous cluster of at least a specified size whose voxelwise statistics are also above threshold. The idea is to increase the sensitivity of classification by using the fact that real activation tends to be more clustered than the artifactual activation caused by noise. While it is often more robust than voxelwise tests, this cluster-size threshold also depends on simplistic assumptions, tends to be highly conservative, and is very sensitive to the specified minimal cluster size.

Worsley [80, 81, 78] constructs a hypothesis test for detecting activation using distributional properties of the extremes of Gaussian random fields. Under the null model of no differences between the task and control conditions, the method assumes that the difference of the average images is a homogeneous, mean zero, Gaussian random field. It derives the distribution of the Euler characteristic of the level sets of the field (roughly, the number of connected components) under this assumption. The test statistic is the Euler characteristic (with correction for boundary effects) of the observed level set of the random field at a specified threshold. Worsley [79] extends these results to t, F, and χ^2 random fields that arise from commonly used test statistics. While the mathematical results here are very elegant, the method suffers from several practical problems: (i) the parametric assumptions underlying the calculations are generally violated in practice, (ii) the test is extremely sensitive to the tails of the noise distribution, and (iii) the method does not take advantage of the fact that the interesting alternatives have a specific form, namely localized shifts in the mean.

Holmes *et al.* [41] has proposed the use of randomization tests to protect against violations of the assumptions. By considering the maximal voxelwise test statistic in a given region, they test the omnibus alternative of no differences in activation in that region with a nearly specified Type I error. A serious difficulty with this method, especially in an fMRI context, is the computational burden, for the number of relabelings grows quickly with the length of the experiment and number of experimental conditions. Moreover, trends in the signal render the notion of overall significance less useful.

A very different and more comprehensive approach has been put forward by Lange and Zeger [50]. They model the voxel time series in the spectral domain using a regression framework in which the shape of the hemodynamic response function is allowed to vary. As the first published work to explicitly model the voxel time series in fMRI data, this method makes an important contribution to the field. By fitting in the spectral domain, the Lange and Zeger method reduces the impact of autocorrelation and drift on the results. The primary limitations of the method are limitations of scope. It applies most effectively to periodic experimental designs, where two conditions alternate in even blocks; this limits its range of application and its ability to address more complex questions. Furthermore, the model for the shape of the hemodynamic response is quite simple and does not account for several important features of the system. See also [34, 23] for further discussion.

3.3. Failures of the Paradigm: Inferential Relevance

A more critical failure is that the localization paradigm limits the range of scientific questions that can be addressed with fMRI data. This problem persists no matter how refined the classification methods become because they only use part of the information in the data, namely the average signal change across conditions. Localization remains an important step, but the key point is that it is only the first step. It allows an investigator to validate the effectiveness of an experimental manipulation and to check its consistency across studies. While localization can help neurosurgeons decide where to cut, it is often of only indirect interest for cognitive psychologists whose primary interest centers on theories that describe the integrated function of the brain as a computational system. These scientists need to answer questions like those in section 2, which deal with changes in the response and more complex spatial relationships. Classification provides useful information only about a subset of these questions. To move the science forward, the statistical methodology must advance to enable well-supported inferences regarding more general questions.

To better understand the scope of questions that localization can address, consider the problem from a more abstract perspective. Suppose we have data realized from a general spatio-temporal process $\boldsymbol{Y} = (Y_{\boldsymbol{x},t})$ that can be decomposed as $\boldsymbol{Y} = \boldsymbol{H} + \boldsymbol{N}$. Here, $\boldsymbol{H}$ represents the structure of interest (e.g., the hemodynamic response in fMRI) and may itself be stochastic; $\boldsymbol{N}$ represents the noise or nuisance processes. For any scientific question that can be addressed with these data, we can define a function f on the range of $\boldsymbol{H}$ such that the scientific question can be expressed mathematically in terms of $f(\boldsymbol{H})$. Inferences (or predictions if $\boldsymbol{H}$ is stochastic) from the data $\boldsymbol{Y}$ about $f(\boldsymbol{H})$ thus provide evidence to help answer the question; to be useful, any such inference must provide an assessment of its uncertainty. A given statistical method supports inferences about a particular set $\mathcal{Q}$ of such functions. Any number of questions from $\mathcal{Q}$ may be addressed with the data, but making inferences about multiple functions with the same data requires some consideration of dependence and multiple comparisons.

Now, let $\mathcal{Q}_c$ denote the set of questions addressable from fMRI data using classification methods. By the nature of the classification procedures, these are *binary, coordinate-wise, condition contrasts*: Each function $f \in \mathcal{Q}_c$ is binary-valued and indicates whether there is a particular difference in the response among selected conditions at a single location. A test at voxel v provides an estimate of one such f_v, and the error rate and power function of the test give some sense of the uncertainty in this inference. However, these separate inferences cannot be easily combined in practice. Consider for example the question of whether the total area of activity is greater in response to one condition than to another. This is usually addressed in fMRI by counting the number of voxels classified as active (with respect to a control) in each condition and then comparing the counts across conditions. But the counts alone provide no baseline for comparison against the likelihood of chance fluctations (What is a large difference? What is a small difference?) and hence do not provide a complete inference. If we try to construct a hypothesis test for the difference in the counts, we encounter a number of problems. The null hypothesis of equal extent does not constrain which voxels are active, and hence the null is highly composite. Even if we were to take the null as extremely restrictive—that the active sets are identical—the voxels within the region being tested would have different response properties, so the count difference would not have a tractable null distribution. While it is possible to use the error rates and power functions of the classification hypothesis tests to compute various probabilities under different alternatives, there is an overwhelming array of possible alternatives so that these probabilities are not useful in practice. (The overly-simplistic models underlying the tests imply that the error rates are inaccurate in any case.) The classification maps thus serve as estimates of a set-valued parameter, namely, the regions of activity with respect to a particular contrast among conditions. But these estimates are not accompanied by useful uncertainties that are needed to construct more general inferential statements. All of these problems can be circumvented by using the information in the data directly rather than after classification.

What is often of interest instead of $\mathcal{Q}_c$ is a set of *coordinate-free spatio-temporal questions*, i.e., real-valued functions that relate to more detailed changes in the response across space and time. These questions include spatial distinctions (Are the responses to pursuit and saccadic eye movements separated?), characterizations of system behavior (What is the rate of increase in signal change with sentence difficulty?), and distributed comparisons (To what degree is working memory involved in the different areas that subserve eye movements?). The classification approach, while providing generally comforting and often reasonable results, offers no way to quantify answers to these questions. Investigators are left interpreting the classification results to answer more detailed questions, without true statistical support. By moving to a more detailed representation of the underlying processes, we can use more of the information in the data and improve both the accuracy of our statistical inferences and their scientific relevance.

3.4. Other Methodological Issues

There are other limitations to current statistical practice in fMRI. Current approaches cast the sources of variation in the data as "artifacts" to be "corrected" serially through several data processing stages prior to analysis. But treating the different effects separately reduces the inferential efficiency of the methods. Moreover, as new research improves our understanding of the processes that generate fMRI data and reveals the details that need to be accounted for, the classification procedures will not readily adapt. For instance, voxelwise hypothesis tests cannot easily incorporate complex spatial structure in the active set or large-scale spatial dependence in the data. Nor do the classification procedures help diagnose or identify sources of variation in the data.

A first step to improvement is to move away from classification and hypothesis testing based on simplistic models towards estimation of interpretable parameters within a detailed model of the data. A second step is to move beyond the localization paradigm by isolating the scientific questions underlying a study and devising inferential procedures that directly address these questions. This is the framework we propose.

4. Modeling fMRI Data

It is tempting to begin an assault on the statistical problems of fMRI by trying to improve the testing procedures used for classification. However, this strategy is not sufficient to handle the the complexity of the data and the richness of the underlying scientific questions. We began at a lower level, studying the processes that generate the data and characterizing the critical sources of variation. We integrate these findings into a family of detailed models for fMRI data on which can be based more accurate and general inferences.

In particular, we develop an inferential framework that can achieve three key objectives:

- Account for as many important features of the data as possible
- Address directly the scientific questions of interest with rigorous statistical procedures
- Extend naturally to encompass new research about the fundamental physical processes behind the measurements.

This framework is built on a nonlinear, hierarchical, spatio-temporal model. The data from an fMRI experiment arise from the interaction of several component processes, many of which correspond to identifiable biological and technological phenomena. We construct our model to explicitly represent as many of these components as possible so that the model parameters are meaningful with respect to the underlying processes. Many other sources of variation in the data (e.g., physiological noise) are accounted for through an appropriate noise distribution.

In this paper, we present a relatively simple version of this framework in which we focus on the temporal (voxelwise) structure in the process and do not fully account for some features of the noise. A number of challenges nonetheless arise here, and even at this level, our methods lead to improved sensitivity and a broader scientific scope. In section 8, we describe several extensions that account for more complicated noise distributions and for spatial structure. Modeling the spatial structure requires new levels in the hierarchy that relate the temporal responses across the convoluted folds of the tissue; we are exploring a detailed spatial extension of the model for a future paper.

Let $Y(t)$ be the observed MR signal for time $t = 0, \Delta, \ldots, T\Delta$ at a specific voxel, where Δ is the sampling interval. Our basic voxelwise model assumes that this time series takes the form

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

where $\mu, \boldsymbol{\gamma}, \boldsymbol{\theta}$, and the function $d()$ are model parameters and ϵ is a parameterized noise process. The four additive components in equation (2) will be called the baseline signal, drift profile, activation profile, and noise, respectively. This equation shows the likelihood level in the hierarchy; section 5 describes the priors used at the deeper levels. Below, we clarify the role and parameterization of each component in the model.

4.1. Baseline Signal

The MR signal associated with a particular voxel is, to first order, a measurement of the density of a particular nuclear species (e.g., hydrogen) within the voxel, averaged over some small time interval. (See section 3.1.) The measured signal can vary by an order of magnitude across the imaged volume. While some of this variation reflects differences in nuclear density across the highly structured brain tissue, much of it arises from other sources. For example, with certain receiver coils, the signal strength measured from a particular location diminishes as the distance between the source (e.g., "spinning" nucleus) and receiver increases. It also depends on the relative orientation of the receiver (essentially a coil of wire) and the magnetization at the given location. A more important source of variation in the measured signal is caused by local magnetic anomalies in the tissue (e.g., chemical substances with magnetic properties). These anomalies alter the precession of the nuclei, causing a non-uniform loss of phase coherence that distorts the signal. Other distortions result from flaws and inhomogeneities in the receiver electronics that make the receiver differentially sensitive across the coil surface.

The real-valued parameter μ in equation (2) represents the magnitude of the baseline signal at the given voxel, that is the mean signal over time in the absence of activation and noise. The amplitude of the BOLD perturbation and, to some extent, the variance of the noise are both dependent on the baseline signal, although for fMRI μ is a nuisance parameter. Notice that in general $\mu \neq \mathrm{E}[Y(t)]$ since the drift and activation terms can have non-zero values. The baseline μ is typically well-determined from the data; experimental designs that incorporate rest periods at the beginning and end of the experiment (during imaging) make estimates of μ particularly precise.

4.2. Drift Profile

The measured MR signal at a voxel tends to drift over the course of an fMRI experiment. Since the magnitude of these changes often far exceeds both the ambient noise level and the amplitude of the task-related BOLD perturbation, signal drift represents a significant source of variation in fMRI data. The drift likely has a number of causes, but these have not yet been conclusively identified. Changes in instrument calibration, equilibration of the tissue and scanner, subject movement, and physiological deformation of the brain are known to play a role. Much of the variation seems to be of biological origin since the largest changes and most interesting features of the drift arise when imaging living tissue.

Our empirical study of many fMRI data sets suggests that the drift has the following basic properties. While the drift tends to be smooth, it 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. Moreover, the drift has interesting spatial structure; its shape and magnitude can vary greatly across voxels. Even neighboring voxels can display completely different behavior. While there are groups of voxels with similar drift properties, there is a convoluted network of boundaries across which these properties may change suddenly. Another notable although anecdotal effect is that the severity of the drift tends to increase with the field strength of the MR scanner's main magnet, even though the overall signal-to-noise improves for higher fields.

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 activation-induced changes to the signal. 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. We balance the conflicting goals of diversity and parsimony by appropriate regularization over this space; see section 5. Although a natural starting point for modeling $d(t)$ would be to use 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. Since we want d to represent a smooth but temporally heterogeneous function, we need a more flexible model for the drift profile. A more effective choice is to take d from an appropriate spline space of some degree D. (Typically, we take $D = 3$.)

A spline of degree D is determined by the number and position of its knots and coefficients in an associated basis of functions [19]. 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: (i) use a small number (e.g., up to 4) of adaptively placed knots and (ii) use a moderate number (e.g, 6 to 12) 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. With a small number of knots relative to the number of time points and with the potential to reduce the complexity of the function (e.g., when $K = 0$, $d(t)$ is a pure polynomial), spurious structure is discouraged and the drift profile cannot become confounded with the activation profile. At the same time, the adaptive placement of the knots provides a great deal of flexibility in the shapes that can be attained. In the latter case, d varies over a fixed spline space, but the regularization bears much more of the burden of maintaining a reasonable shape for the curve.

We parameterize the drift profile as a function on $[0, 1]$ and rescale it onto the time interval of the experiment. It is also convenient to constrain $d(t)$ to be orthogonal to constants (and thus the baseline) with respect to the empirical inner product, *i.e.*, $\sum_{t=0}^{T} d(t) = 0$. This maintains the conceptual separation between the baseline and drift. 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. A common remedy for the problems of the power basis is to use a B-spline basis [19] because these functions provide a stable and localized representation. B-splines allow greater computational efficiency when the number of knots is large, but in the few knot case (see section 5), we use the basis functions generated by orthonormalizing the power basis with respect to the empirical inner product. In particular, this makes changing the knots and updating the basis more computationally efficient. When we wish to emphasize the specific parameterization, we will denote the drift profile by $d(t; K, \boldsymbol{\kappa}, \boldsymbol{\delta})$.

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 chooses 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 reduces to a essentially pure polynomial and thereby achieves a good fit without introducing too much complexity. The need to vary structure along this continuum guides our treatment of drift in the model.

4.3. Activation Profile

The hemodynamic response to neural activity manifests itself as a perturbation in the MR signal. Figure 6 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 a undetectable signal change. Eventually, through a metabolic process, the increased neural activity causes dilation of the local blood vessels. The MR signal then begins to rise as the balance of oxygenated to deoxygenated blood shifts within the voxel, which takes place over 3 to 8 seconds. 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, the signal holds at the plateau, although in some cases, there may be important fine structure in the signal during this period. When the task ends, the signal remains elevated for a time then begins a slow decay back to baseline. This decay is typically longer than the corresponding rise [20] (on the order of 10 seconds), and more significantly, the signal often dips below the baseline for an extended period before returning [18]. This dip is typically dealt with by dropping data at the beginnings of task epochs. See [82, 1, 11] for more systematic approaches to incorporporating the dip. The importance of the dip is two-fold: (i) a large signal dip during one task epoch distorts the signal in adjacent epoch, complicating analysis; and (ii) the dip may serve as a sensitive discriminator for evaluating local activity. The latter possiblity results from the fact that the BOLD perturbation is a blood-flow response and only a proxy for neural activity. Large blood vessels draining from an active area can lead to apparent activation in distant voxels. Of greater interest is a response from the microvasculature (e.g., capillaries) that serves the active neurons locally. BOLD responses in the microvasculature would provide a much sharper picture of the distribution of activity but may also tend to produce a weaker signal. It has been conjectured that several available image acquisition techniques put different weights on the signal contribution from different sized blood vessels. There has been recent speculation, as yet unproved, that the signal dip appears using all of these techniques, suggesting that the dip reflects changes at the microvasculure level [67]. Further research on the source and meaning of the dip is currently underway.

The activation profile, $a(t)$ in equation (2), is the combined pattern of signal changes over the course of the experiment. The profile at a given voxel depends on the responses in that voxel to the different experimental conditions and to the shape of the response function there; the amplitude of the response also scales with the baseline signal. The shape of the response curve can also vary from voxel to voxel since it depends on the nature of the tissue and the distribution of local vasculature. Suppose that there are C experimental conditions being evaluated in the experiment. We make no assumptions about the order or duration of the time periods associated with these conditions. For each condition $c = 1, \ldots, C$, we define a *responsiveness* parameter, $\gamma_c \geq 0$, that measures the degree to which the given voxel activates in response to the stimulus or task associated with condition c. The responsiveness is expressed as the proportional change; specifically, the amplitude of the perturbation for the c^{th} task condition is taken to be $\mu\gamma_c$, where μ is the baseline signal. Parameterizing the responsiveness as a proportion is natural in this context because the absolute magnitude of the signal change scales with the magnitude of the baseline.

By assigning a single parameter to each condition, the model treats every instance of a particular condition (i.e., every epoch of the task repeated at any point during the experiment) identically. This is not a bad assumption since the responses are generally consistent, but in general, there may be some variations in the level of response across an experiment. This includes random variations from epoch to epoch and systematic changes such as a downward trend caused by practice with the task. Such extensions can be incorporated by adding another layer to the hierarchy; see section 8.

To model the shape of the activation profile, we specify a family of smooth functions that captures the form illustrated by Figure 6. There are a number of ways to parameterize such a family, but a simple and flexible approach is to use what we call *bell* functions. The basic bell function is of the form $b_{\text{attack}}(t) \times b_{\text{decay}}(t)$ where b_{attack} and b_{decay} both have at least two continuous derivatives, b_{attack} rises from 0 to 1 over a finite interval, and b_{decay} falls from 1 to 0 over a finite interval. We currently use polynomial bells, where b_{attack} and b_{decay} are piecewise polynomial functions; other choices (e.g., exponential, sinusoidal) are possible but somewhat less convenient. We can deal with the dip during decay by parameterizing a dip directly into b_{decay}, but it is more computationally convenient to separate the dip into a separate term. Hence, we define the response curve to be of the form $b_{\text{attack}} \times (b_{\text{decay}} - b_{\text{dip}})$, where b_{dip} is a non-negative, compactly supported, bell-shaped function shifted relative to b_{decay}. The specific form of the response curve is determined by a vector of *shape* parameters, denoted $\boldsymbol{\theta}$. There are eight possible components 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 6. 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. 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 rise and fall fixed. Our parameterization of the response curve is intended to capture the effects of the underlying biological mechanisms while maintaining flexibility and computational efficiency. The bell formulation is useful because it decouples different features of the curve. 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. Another advantage is the ease of extending this parameterization as new features of the response come to light.

A critical aspect our model for the response is that the height and shape of the response are decoupled, i.e., the final response curve is the bell function scaled by $\mu\gamma$. The advantage of this is simplicity and a reduced computational load in fitting the model. However, there is some evidence for a relationship between height and shape; we take this up in section 8. Another issue is that our parameterization specifies times for signal attack and decay, which implies that the attack and decay will be steeper for a larger response. This is consistent with a biological model in which the change in blood volume increases with the intensity of the response thus yielding a more rapid change in the signal; it also fits the data well. However, this assumption needs to be tested.

The above description specifies how the model captures an isolated response, but there is more to the story. Specifically, there is evidence [72] that the profiles from two responses closely spaced in time, combine sub-additively. This corresponds to a marginally decreasing response in the presence of substantial vasal dilation. Nonetheless, we do expect signal from two combined responses to be at least as large as either of the two alone. This suggests a sub-additive function that lies somewhere between a pointwise maximum and an additive combination. The model specifies a one-parameter family of operators that combine the profiles from overlapping respones. One choice for this family is a convex combination of addition and a smooth approximation to pointwise maximum. The additive case corresponds to a physiological model in which activation leads to an increased blood flow independently of recent changes; the maximal case corresponds to a physiological model in which further increase in blood-flow is damped when the vessels are already sufficiently dilated. The reality likely lies somewhere between these two extremes but has not yet been established in empirical work.

4.4. 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, frequent 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. Any such movement 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. As an example, Figure 7 shows an MR time-series a voxels in a data set apparently corrupted by large movement. Keeping the subject still (while performing difficult or timed tasks) is thus very important. While restraint systems have been devised, these can be quite uncomfortable and even disturbing for subjects. There is consequently a need for numerical techniques to register, or align, a sequence of images after acquisition (see in particular [25] but also [76]). These methods are demonstrably quite effective at adjusting for rigid movements within the slice plane, although full three-dimensional registration still needs development. Ideally, we could incorporate subject movement directly into the model, but since this currently presents a major computational obstacle, we must carry out statistical analysis after movement correction.

The physiological cycles of the subjects, including respiration, heartbeat, and peristalsis introduce temporal variations that may be confounded with the activation response to the experimental tasks. Although much is still being learned about these fluctuations, the primary effect seems to be a non-rigid motion of the brain resulting from changes in blood pressure. As such, boundaries between tissue and cerebro-spinal fluid (CSF) are most severely affected. The respiratory effects are, perhaps surprisingly, dominant over the cardiac except near the brain stem. We record physiological data—respiratory traces from a flexible band around the chest and cardiac traces by a pulse oximeter at the finger—at a high sampling rate during every experiment, which makes it possible to account for these effects.

Other features of the noise have been studied but remain to be dealt with during analysis. For certain acquisition schemes, spatial correlations can be significant. As Figure 5 shows, these correlations need not be isotropic or local. There has been little research into the impact of this dependence. Temporal autocorrelation in the voxel time series, while present even after drift and physiological fluctuations are accounted for, seems somewhat less severe. Shot-noise and other heavy-tailed transients are quite common, and instabilities in the magnets and electronic components of the scanner can distort the measured signal. Large blood vessels also contaminate the signal of surrounding voxels but are difficult to classify automatically.

5. 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 [37, 63]. In fMRI, there is considerable information 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. The goal of this section is to describe some of the prior information avalable for each model component and to illustrate how we use it.

We take a Bayesian approach here and incorporate the available 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 very structure of the model while still accounting for the uncertainty in that structure. We use this property for instance to allow distinct forms for both the drift and activation components of the model (see below). Second, under this formulation, it is a simple matter to compute an estimate of any functional on the parameter space with an assessment of its uncertainty, thus broadening the effective scope of inference. Third, the approach offers a mechanism for feedback so that, as we analyze more data, we have a direct way to incorporate what has been learned into future studies and thus refine the specification of the model components. Fourth, the spatial variations in the parameters are complicated and nonhomogeneous over the brain, making them difficult to capture with a simple parameterization. An extension of our model to include spatial structure requires adding new levels to the hierarchy that connect the behavior across voxels in specific ways. This provides the flexibility to constrain the allowed spatial variations and gives us more local control over them. 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. Posterior means and maximizers can then be viewed as particular shrinkage estimators, and posterior probabilities can be viewed as a measure of the strength of evidence for specific propositions.

5.1. Baseline Signal

Uncertainty about the baseline signal μ 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 other than 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, the effects of these fluctuations can be mapped out prior to the experiment and a fairly accurate prior estimate of the baselines can be constructed. However, μ is usually very well determined from the data, so inferences about μ are not very sensitive to the choice of prior. We thus use by default a simple symmetric distribution centered on a fixed value μ_0. In this paper, we use a t_1 distribution, which provides a conservative assessment of our prior uncertainty. The value of μ_0 is set separately for each voxel, using scout images obtained prior to the experiment when available. While it may also be reasonable to use a flat (improper) prior on μ, we avoid this.

5.2. Drift Profile

Although the sources of the signal drift have not been fully identified, some persistent features of the drift have become clear, as described in section 4.2. The drift can take a wide range of possible forms, and there is substantial variability in its magnitude across voxels. The prior for d must give weight to particular properties—general smoothness with several potential change points, some of which may be sharp—within a broad class of functions while discouraging spurious structure (e.g., oscillatory behavior) that may confound with activation. We take the support of this prior to be an appropriate space of splines because of the flexiblity these functions offer.

Let $\mathcal{S}(D, K)$ denote the orthogonal complement to the constant (with respect to the empirical inner product) 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 $K = 0, \ldots, K_{\max}$. We denote the prior for the function-valued parameter d by π_{drift}. If the number or position of knots is fixed, we condition on K and/or $\boldsymbol{\kappa}$. The degree D is fixed throughout. We consider two basic strategies for modeling the drift profile on these spaces:

(i) Fix K to be on the order of the number of observations over time, and fix the positions $\boldsymbol{\kappa}$ to a regularly spaced grid. Then define π_{drift} as the corresponding standard Sobolev prior over $\mathcal{S}(D, K, \boldsymbol{\kappa})$, see equation (3);

(ii) Restrict $K_{\max}$ to a small number, put a decreasing prior on K over $\{0, \ldots, K_{\max}\}$, let $\pi_{\text{drift}}(\boldsymbol{\kappa} \mid K)$ be a diffuse prior on the appropriate simplex, and set $\pi_{\text{drift}}(\cdot \mid K, \boldsymbol{\kappa})$ to a standard Sobolev prior over $\mathcal{S}(D, K, \boldsymbol{\kappa})$;

The first strategy corresponds to a modified smoothing spline [73]. 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. The Sobolev prior for d is given by

$$\pi_{\text{drift}}(d) \propto e^{-\frac{1}{2\lambda}\left[a_n \, \|d\|^2_{2,w} + a_c \, |d|^2_{2,2,w}\right]}, \tag{3}$$

where $a_n, a_c, \lambda > 0$ are constants,

$$\|d\|_{2,w} = \left[\int_0^1 d^2(t)\, w(t)\, dt\right]^{1/2} \text{and} |d|_{2,2,w} = \left[\int_0^1 |d''|^2(t)\, w(t)\, dt\right]^{1/2} \tag{4}$$

are the weighted $\mathcal{L}^2$ norm and weighted $\mathcal{L}^2_2$ semi-norm, respectively, with a non-negative, non-zero weighting function w. The constants a_n and a_c determine the relative penalty ascribed to norm and curvature of the profile, and λ mediates the overall level of smoothness given this weighting, with smaller λ 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

This form for the prior 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 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. When K is large, there is a 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 λ to yield a specified effective degrees of freedom. We define the degrees of freedom as the trace of the effective smoothing matrix [40] 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 λ and σ^2 determines the degree of smoothing. This requires an extra iterative step in the optimization. The second method is to allow λ to vary as a model hyper-parameter with a distribution (e.g., Exponential) that is weighted towards zero. This requires a substantial improvement in fit to add complexity to the drift profile. We choose the mean of this prior on λ to achieve approximately a specified effective degrees of freedom for the profile.

The second strategy for modeling drift allows a variable (but limited) number of adaptively placed knots. This is intended to achieve a greater flexibilty for dealing with localized changes in the profile, which are reasonably common. The Sobolev prior underlying the smoothing spline is homogenous and has some difficulty fitting sharp, sudden changes in the way we desire. If the smoothing parameter is set to allow for a sharp change of a particular magnitude, then the fit must trade off capturing that change accurately with allowing greater flexibility in the rest of the profile (over which there is more data contributing to the likelihood). Typically, the fit can then be improved by only roughly approximating the sharp change while allowing wilder fluctuations elsewhere, but the latter are not consistent with our specification of the drift. Indeed, we want $\pi_{\rm drift}$ to place more mass on functions with sparser structure. By limiting the maximum number of knots $K_{\rm max}$ but placing them carefully, the model can fit the change points accurately while achieving a more appropriate level of flexibility for the rest of the curve. This adaptive prior more closely captures the desired behavior. Conditional on K and $\boldsymbol{\kappa}$, we use a Sobolev prior defined as above, although with potentially different constants a_n and b_n for each number of knots. We specify a rapidly decreasing prior for K on $\{0, \ldots, K_{\rm max}\}$ to penalize structure when none is needed. Typically, we would like between 0 and 3 knots, but to improve mixing in posterior simulation (see section 7) we set $K_{\rm max}$ larger (e.g., $5-7$). We specify a prior for $\boldsymbol{\kappa}$ given K through the distribution of the separations between adjacent knots (and the ends of the interval). In particular, if $\kappa_0 \equiv 0$ and $\kappa_{K+1} \equiv 1$, we take the distances between the knots $\kappa_i - \kappa_{i-1}$ for $i = 1, \ldots, K+1$ to be Dirichlet$(\alpha_1, \ldots, \alpha_{K+1})$. For convenience, we usually set all α_i's equal to a common value (by default $\alpha = 2$). Note that when α is a positive integer, the knot positions are distributed as selected uniform order statistics. This version of $\pi_{\rm drift}$ can be effectively implemented for posterior simulation, but it does introduce a hefty computational cost. See section 7 for details.

5.3. Responsiveness

The responsiveness parameters $\boldsymbol{\gamma}$ describe the magnitude of the activation-induced 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%. Even though the magnitude of signal changes vary somewhat across tasks, brain regions, and subjects, the observed distributions from previous studies still yield useful information for constraining $\boldsymbol{\gamma}$. 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, handles the initial cortical processing of visual stimuli in the brain. V1 tends to exhibit the strongest BOLD response of any area of cortex yet studied with fMRI. The cellular and vascular structure here have been studied in detail and the characteristics of visual stimuli that evoke the strongest response have been isolated empirically. Similarly, the study of signal changes across other areas and tasks clarifies the shape of the upper part of the response distribution and shows how the distribution scales with imaging configuration (e.g., magnetic field strength).

The lower range of responsiveness values is more difficult to specify empirically since small responses will often go undetected. To deal with this, we consider two models for the distribution of neural activity in the brain. In the *isolated activation model*, activity is concentrated in several distinct regions, and other areas are not responsive to the given task. In the *distributed activation model*, the activity is distributed over a large portion of cortex, although certain regions may bear a greater part of the load. It is currently an open question as to which of these views is more accurate, but the isolated activation model is implicitly assumed in most discussions regarding fMRI. Under the isolated activation model, we would expect many voxels with zero responsiveness and relatively few with a very small response. Under the distributed activation model, we would expect many voxels with a small to moderate response. We use the isolated activation model by default.

We choose a prior π_{resp} for $\boldsymbol{\gamma}$ to match the available information, under either the isolated or distributed model. 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 (of course, the activation profile can extend below baseline as discussed earlier), π_{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. Under the isolated activation model, π_{resp} puts non-zero mass at 0 (no response) in each condition. Specifically,

$$\pi_{\text{resp}}(\boldsymbol{\gamma}) = \sum_{\boldsymbol{j} \subset \{1,\ldots,C\}} \eta_{\boldsymbol{j}} \prod_{k \in \boldsymbol{j}} f(\gamma_k) \prod_{l \notin \boldsymbol{j}} \delta_0(\gamma_l) \tag{5}$$

where the η's are non-negative constants that sum to 1, δ_0 is a point mass at 0, and f is a continuous density on $(0, \infty)$. The density f decays towards both 0 and ∞, with its range and upper tail behavior calibrated to the available prior information. We have experimented with a variety of forms for f, but our results show sensitivity primarily to the tail behavior. A suitable Gamma density allows a reasonable fit to our prior constraints and is more convenient than some of the more complex forms we have tried. The inclusion of different sub-models in the prior is a critical part of the model since it accounts for the substantial uncertainty concerning whether or not there is a response. 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 saturated models. A value of $1 - \eta_\emptyset$ in the range 1/500 to 1/100 yields, for typical image sizes, about 50 to 100 voxels expected to show a response. Under the distributed activation model, all components of $\boldsymbol{\gamma}$ are positive, although most will be quite small. In this case, we take each component to be independent with a density given as a step function with an exponential upper tail. We use two or three steps with decreasing weights, which correspond to populations of voxels with different degrees of involvement in the task. Note that in both models particular components of $\boldsymbol{\gamma}$ can be fixed to 0 *a priori* if desired, which is most relevant to pure rest conditions.

5.4. 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 π_{shape} for the shape parameters $\boldsymbol{\theta}$. We take the parameters to be non-negative although allowing the lag-off parameter to take some negative values may be useful for broadening the range of shapes fit by the bell functions. We also define π_{shape} to make the components independent of each other and $\boldsymbol{\gamma}$. This is likely optimistic since the shape might depend on response intensity (see section 8), but currently the relationship between the two is not well understood. 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 π_{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 π_{shape} depend on the image acquisition scheme and the task lengths in the experiment, since these can affect the response characteristics.

5.5. 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.

6. 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.

The posterior distribution offers several practical advantages as a basis for deriving such inferences. First, probabilities provide a readily interpretable scale for evaluating results. In contrast, arbitrarily thresholded test statistics make it difficult to evaluate the strength of evidence behind the inference. Second, the distributions of both simple and complex quantities can be computed directly even when the distributions allow for several qualitatively different types of structure (e.g., discrete components or disjoint submodels). Thus, for instance, there is no conceptual difference between computing a distribution related to a single voxel and a distribution describing many voxels. Third, because posterior quantities are conditioned on the observed data, they are not vulnerable to the spatial selection biases that arise when choosing voxels for analysis based on the results of earlier classifications.

To illustrate the broad range of questions that can be addressed with our approach, we describe several possible analyses in light of a specific example. Throughout, we target the analysis to specific scientific questions (translated into statistical terms). Nonetheless, we do not focus here on the results *per se*. Rather, our goal in this paper is to demonstrate the power and flexibility that this approach offers investigators, whatever their scientific questions may be.

6.1. An Example Experiment

We consider an fMRI experiment designed to study how the usage of cognitive resources (e.g., working memory) changes with the difficulty of processing (e.g., comprehending different types of sentences). The hope is that these changes are quantifiable through fMRI at a finer level of detail than is possible with behavioral data alone. This experiment and the data are based on a study presented in [47]. Although we only show data from a single participant, our analyses support the original conclusions of the investigators. The key point here is that our method provides a more rigorous basis for statistical inference and at the same time broadens the range of questions that investigators can *directly* address with fMRI data.

The experimental design specifies six task conditions; see Figure 2b for an illustration. We will discuss these conditions in order of increasing complexity. 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 participant fixate on a marked point in the center of the 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. For each sentence, the participant reads a sentence and then answers a question about that sentence with the push of a button. Each task epoch involves processing several sentences, and each condition consists of several epochs replicated at different times throughout the experiment. The sentences at different levels of difficulty are distinguished by different syntactic and semantic structures that increase the cognitive load required to understand them. We will label the conditions, in order, as Rest (R), Fixation (F), Trivial (Tr), Task Level 1 (T_1), Task Level 2 (T_2), and Task Level 3 (T_3).

Each of these tasks is performed repeatedly for a period of time, and these task epochs occur several times over the course of the experiment. The order of task performance is randomized so that the epochs are balanced in time. It is worth noting that both the tasks being performed here and the experimental design itself are more complex than is now typical in fMRI, which has several implications for the analysis. Since this study has several experimental conditions with more than one useful control, it is possible to examine a variety of contrasts and relationships among the responses. Since non-control conditions may be separated by at most a short period of rest, the responses from different conditions may overlap, which necessitates deconfounding the contribution to the net response at any time. Our model does this since it fits the entire response, but simple averages of the signal within an epoch cannot. Moreover, the sentence processing tasks are complicated enough so that the experimenters do not have control over when the subject completes the processing. Task performance thus need not be aligned in time with image acquisition, and so we need to account for the timing differences in fitting the response curves. The design also includes periods of fixation and rest at the beginning and end of the experiment, which improves the precision of the baseline and drift estimates. As fMRI gets applied to more subtle and sophisticated questions, experimental designs like this one will become increasingly common, and the statistical methods used will need to deal with these complexities, as ours does.

One goal for this experiment is to clarify the relationship between intensity of response and sentence difficulty. Two of the specific questions being asked are

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?

A third question asks how this relationship depends on behavioral covariates (e.g., an individual's working memory capacity); we take this up in section 8 where we discuss inference across subjects. 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. These may be physiologically related; as new neurons within a voxel are recruited by a process, the local intensity of the hemodynamic response may increase, and as more distant neurons are recruited, hemodynamic changes may then extend to neighboring voxels. Of course, new active regions that emerge with increasing task intensity may instead represent distinct cognitive processes that are only required (or detectable) at higher task difficulty. Which of these views applies is a scientific issue, but from a statistical perspective, we can use both types of effects (or some combination of the two) to measure response changes. Both types of changes are commonly used in fMRI for comparisons across conditions, the within-voxel responses to classify active voxels and the changes in extent to compare conditions.

We now turn to analyses of some data from this experiment. We begin with the simplest methods, those closest to what is currently standard in fMRI, then exemplify methods for making more sophisticated inferences.

6.2. Estimation

The simplest method for making inferences under our model is to estimate the model parameters and their associated uncertainties. Maximum posterior estimates can be computed by numerical optimization, and the standard errors of the estimates can be derived through a normal approximation to the posterior at the mode. These estimates and their uncertainties provide a convenient alternative to a testing approach and can be used to compute analogues of the statistical classification maps that are standard in fMRI. Almost all of the results obtained this way can be improved through full posterior inference (see section 6.3 below), but the maximization approach is of interest as a fast and convenient approximation. Even if full posterior inferences are desired instead, the estimates from maximization provide a good starting point for MCMC simulations, and the approximate covariance matrix is useful for tuning Metropolis steps [53, 69].

Differences Between Conditions. The greatest interest typically centers on the responsiveness parameters ($\boldsymbol{\gamma}$) as well as contrasts among them. For simple comparisons, the estimates and standard errors can be used directly to assess the strength of evidence for responsiveness in a particular condition. Since researchers in the fMRI community commonly use images to summarize voxelwise results of these comparison, we construct a simple analogue of these test-statistic maps that can be used to similar effect. Our maps are built from normalized contrasts such as

$$\frac{\widehat{\gamma}_c}{\mathrm{SE}(\widehat{\gamma}_c)} \quad \text{or} \quad \frac{\widehat{\gamma}_c - \widehat{\gamma}_{c'}}{\mathrm{SE}(\widehat{\gamma}_c - \widehat{\gamma}_{c'})}, \tag{6}$$

where c and c' represent different experimental conditions. One advantage of these contrast statistics is that they offer a familiar measure on which to base an analysis and will thus be readily used in the fMRI community. However, the real benefit is the improved sensitivity due to the statistical efficiency of our model. For example, Figure 8b presents one slice of a normalized contrast map for the T_3 versus Tr comparison. Figure 8c shows a corresponding "traditional" t-map for the same slice after linear detrending and thresholding at 4, an arbitrary but often used value. The corresponding slice of the mean image is given in Figure 8a for reference. Comparing these figures, we note that the normalized contrast map is cleaner, with fewer scattered peaks than the t-map. We generally see this improvement.

An important question is whether the crispness of our contrast maps results from improved sensitivity or from a more conservative assessment. To answer this question, we considered individually the time courses for those voxels showing activity in one of the maps but not the other. In every case in which the t-statistic shows no activity but the normalized contrast does, either the t-statistic failed to detect the differences because of an inflated variance estimate resulting from lack of fit to the signal drift or the normalized contrast was able to identify a weak response because it allows structural variations (e.g., lags, dips, attack/decay) that the t-test does not. In the cases in which the t-statistic shows activity but not the normalized contrast, the average signal changes are large relative to the noise because of nuisance variation (e.g., movement, drift, physiological noise) that does not have the shape of an activation response. The distinction between the two statistics, in both directions, is that the constraints embodied in our model more effectively isolate activity from other sources of variation.

Domination Probabilities. An alternative way to evaluate pairwise changes (e.g., T_3 versus Tr) depends on posterior probabilities. We use a normal approximation to the joint posterior of $\boldsymbol{\gamma}$ (with some correction for parameters at the boundary [55]) to evaluate posterior probabilities that the response to one condition is greater than to another. For example, Figure 8d shows a map of domination probabilities $\mathrm{P}\{\,\gamma_{T_3} > \gamma_{Tr} \mid \boldsymbol{Y}\,\}$. While the revealed structure is similar to that displayed by the normalized contrasts, the domination probabilities give a more interpretable measure of the strength of evidence for a particular ordering. More sophisticated applications of this technique are illustrated in the next subsection.

Model Selection and Averaging. A practical difficulty arises in computing the maximum posterior estimates in that discrete components in the model (e.g., different drift bases or responsiveness models) are not easy to deal with during optimization. To make the optimization efficient, it is ideal for the posterior to be smooth on a simple domain. While it is possible to implement a non-smooth multivariate optimization, the result is very slow and inconvenient. Hence, for the maximization phase of the analysis, we deal with the different possible structures as different possible models. For the full model, we fix a drift basis to a reasonable spline space and include all of the responsiveness terms. However, this model no longer reflects uncertainty about whether or not there is a response to individual conditions, and the estimated responsiveness parameters will be more likely to take small but non-trivial values for lack of a better option. Therefore, we define sub-models $M_{j_1 \cdots j_m}$ such that $\gamma_{j_i} = 0$ *a priori* for integers $1 \le j_1 < j_2 < \cdots < j_m < C$. The posterior can be maximized separately over every such sub-model and the results suitably combined. One option is model selection, where a specific model is chosen using a criterion like BIC [2, 58]. Since the number of conditions is moderate (e.g., 2-7) in most fMRI experiments, this is generally a practical strategy. An alternative is to average the results across the different models. This makes sense because the parameters are all interpretable in every sub-model. The posterior becomes a mixture of the restricted posteriors over the individual sub-models, where the mixture weights are the posterior mass attributed to each $M_{j_1 \cdots j_m}$. Using the results of the maximization within each sub-model, we approximate the mixture weights by first approximating Bayes factors [48] using a Laplace approximation [70] (or optionally a localized form of Laplace's method [21]) and then combining the Bayes factors with the prior weights to derive posterior probabilities. The standard errors derived from the mixture distribution are larger than those under the individual models, more accurately reflecting the underlying uncertainties. These standard errors are then incorporated into the maps and probabilities described above. In practice and in the analyses presented here, we consider two models, the full model with all the terms and the null model which contains no activation profile.

Characterizing the Functional Form. Finally, a useful diagnostic for evaluating the nature of the change in response across conditions is to fit a simple weighted regression to the estimated responsiveness parameters. Figure 8e shows the binned slope coefficients from a linear fit to the parameter vector $(\widehat{\gamma}_{T_1}, \widehat{\gamma}_{T_2}, \widehat{\gamma}_{T_3})$. with weights derived from the standard errors of the estimates. Only voxels for which $\mathrm{P}\{\max(\gamma_{T_1}, \gamma_{T_2}, \gamma_{T_3}) > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y}\} > 0.01$ were fit; the rest are left blank in the figure. The slopes are binned according to the t-statistic for the coefficient in order to make the image visually interpretable. The goal of this diagnostic is to identify clusters with similar types of response. The results in the map are somewhat unclear, but the map does suggest two notable clusters, over which the unbinned slopes are similar. A more formal approach to this question is developed below.

6.3. 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 [69, 61, 8, 30], we can compute posterior probabilities more accurately 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 varying sub-models and atomic components, from which any functional of this distribution can be easily estimated. All of the methods described in the previous sub-section can be based on posterior samples without loss of generality and with some gain in flexibility (because there is no restriction to a smooth posterior). For example, we can estimate both the domination probabilities such as $\mathrm{P}\{\gamma_{\mathrm{T}_3} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y}\}$ and the variant $\mathrm{P}\{\gamma_{\mathrm{T}_3} > 0, \gamma_{\mathrm{Tr}} = 0 \mid \boldsymbol{Y}\}$, which uses the atomic component in the responsiveness distribution. If the latter is large, it suggests that the voxel is responsive to the study task but *not* the trivial task and in principle, can distinguish areas recruited specifically for semantic processing. (See section 7 for details on the sampling procedures.) Our goal here is to show how a posterior sample can also be used to address new kinds of questions with fMRI data. Genovese et al. [36] provides further examples from the study of eye movements.

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

$$\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\} \tag{7}$$

quantify the support in the data for monotonicity in each voxel. Figure 8f 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. The key point here is that our approach makes the question accessible quantitatively and provides a measure of uncertainty with respect to the question. Note that several variants of the monotonicity probabilities above that add more stringent conditions can just as easily be computed.

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., $\gamma_{\mathrm{T}_3} > \gamma_{\mathrm{T}_2} = \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}}$ still counts as 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. This is a particular problem because the the BOLD signal changes have a limited dynamic range, possibly leading to a ceiling effect when the response is large. In such a case, the monotonicity probabilities may be reduced somewhat when the uncertainties are large, but the hypothesis test will have a drastically inflated tendency to erroneously classify such voxels as non-monotonic.

Assessing Monotonicity: Extent. To study the extent of activity, we must consider many voxels simultaneously. The standard approach to this question is to count the number of voxels that are classified active in each condition and compare the resulting counts. One problem with this approach, however, is that it provides no measure of uncertainty for the estimated difference in counts. Our approach 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 will use the posterior distribution of (N_1, N_2, N_3) to address the monotonicity question, getting either estimated differences with associated uncertainties or the probability $\mathrm{P}\{\, N_3 \geq N_2 \geq N_1 \mid \boldsymbol{Y} \,\}$ which embodies both. 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 $P\,(A \mid F(\boldsymbol{Y}), \boldsymbol{Y}) = P\,(A \mid \boldsymbol{Y})$ 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; that is, examine the posterior of the functionals $\Gamma_c(R) = \int_R \gamma_c(v)\, dv$ for condition c and for a fixed set of voxels R. A new technique that some fMRI researchers are developing will soon allow them to demarcate *anatomical* regions of interest, subject by subject, by mapping between functional and high-resolution structural images. Although the precise location and shape of these regions may vary across individuals, the regions can be taken as comparable from a functional point of view. Hence, many comparisons of interest will focus on changes in the response within such regions. As we will discuss in section 8, this provides a tool for making inferences across subjects.

To illustrate how our method facilitates region-of-interest analysis, we compute the distribution of $\Gamma_c(R)$ for $c = \mathrm{Tr}, \mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3$ for an arbitrary region of 21 contiguous voxels in the language area (surrounding the most notable cluster of activity in the other maps). Figure 9 shows $\mathrm{P}\{\, \Gamma_c(R) > u \,\}$ as a function of $u > 0$ for each of these conditions. These curves can be computed directly from the posterior sample, or we can use the sample to construct the distribution of of the sum $\Gamma_c(R)$. Either the empirical distribution or a normal kernel density estimate provides an approximation that is easy to work with; we used the latter here. The curves 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 monotonicity. These plots allow scientists to compare the integrated response over the region across conditions or subjects. This technique is explored in more detail in [36].

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 (e.g., the two segments of the curve being flat-flat, flat-up, up-flat, or up-up). Each "pixel" in Figure 10 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 T_1 and T_2 and a smaller but non-trivial change between T_2 and T_3. The marked voxel on the left, on the other hand, shows most of its responsiveness change between T_1 and T_2; the distribution of $\gamma_{T_3} - \gamma_{T_2}$ 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.

Summary. The inferential procedures described in this section show that it is possible to use fMRI data to directly address complex scientific questions, many of which cannot be easily addressed with currently available methods. Our approach allows scientists to design procedures that target their specific questions and quantitatively assess the effectiveness of the data for answering them. The inferences are built on the foundation of an accurate model for the data and account for the full range of uncertainties involved.

7. Computational Techniques

Fitting our model to data and implementing the inferential procedures described above raises a number of computational challenges. The images from fMRI experiments range from 10,000 to 100,000 voxels in the brain itself, and each of the time series 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 the computational techniques that we employ. The model fitting and a wide range of inferential queries are implemented in the publicly available software package Bayesian Response Analysis and Inference for Neuroimaging (BRAIN) [33].

7.1. Initial Data Processing

As described earlier, the raw data produced by an MR scanner is collected in the Fourier domain. The raw data are subject to several sources of bias and mis-calibration and must be specially processed to reconstruct good quality images. While it would be ideal if we could integrate these sources of variation into the model, that is not computationally feasible at this time. Instead, we use the FIASCO (Functional Image Analysis Software, Computational *Olio*) software package [24] to carry out all of these pre-processing steps and take the resulting spatio-temporal data as input to our procedures.

The FIASCO processing stream involves five core steps: mean correction, baseline correction, deghosting, image registration, and reconstruction. Mean correction involves a multiplicative normalization of each image to a common global mean. Baseline correction is an adjustment for high spatial-frequency bias. Deghosting reduces the intensity of aliasing artifacts ("ghosts") in the images. Image registration refers to the alignment of successive images to account for and reduce the impact of subject movement. In reconstruction, the raw data are transformed from the Fourier to the image domain.

7.2. Posterior Maximization

Maximum posterior estimates are computed via direct numerical optimization; the objective function is the log un-normalized posterior. The procedure requires that the priors be at least C^2, so the non-standard priors used during posterior sampling must be approximated by a smooth distribution. While it is possible to use a non-smooth objective function, the algorithms for optimization in this case [57, 60] are *very* inefficient. The procedure can accept arbitrarily defined distributions which are then interpolated to compute derivatives, but it is usually desirable for performance reasons to obtain derivatives analytically. Numerical difference approximations, when necessary, are computed by Richardson Extrapolation [17] to improve accuracy. We use the BFGS version of the variable metric optimization algorithm [57] 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 at the optimum. When available, as with the default priors, we use analytic second derivatives. Otherwise, if analytic first derivatives are available, we use first differences of the gradient, refined with Richardson Extrapolation. When no derivatives are available analytically, we approximate the Hessian components with second differences, again refined with Richardson Extrapolation. We take care near the boundaries of the parameter space to acquire a well-defined value when differencing is used. When the optimal parameter values are sufficiently far from the boundary, we apply central differences, but near a boundary we use forward or backward differences. Although this loses some precision in calculated derivatives, it tends to avoid catastrophic failure of the procedure and give workable results.

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 Bayes Factors with a form of the Laplace approximation [21] or with the Schwarz criterion [58]. When parameters are pushed to their bounds and thus a smaller sub-model, the larger but equivalent model is penalized by the full difference in degrees of freedom. We are still assessing more detailed corrections to the Schwarz criterion as in [55]. The Bayes Factors are used to compute posterior probabilities over the sub-models, and results are averaged over the different models. All the parameters maintain a consistent interpretation across all the models, although the hemodynamic shape parameters play no useful role in the null (no response) sub-model.

7.3. MCMC Sampling

Most of the posterior inferences we would like to make require a more accurate expression of the posterior distribution than is available through a simple normal approximation. By drawing a large sample from the posterior, most functionals of the distribution can be computed quite accurately. The sample is obtained by running a Markov Chain Monte Carlo (MCMC) simulation independently for each voxel. As spatial structure is added to the model, the simulations for different voxels will necessarily become linked.

7.3.1. Basic Sampling Strategy

For the voxelwise model, the sampling strategy is a mix of Metropolis and Gibbs steps, depending on the priors used. We generally use a fixed scan order across the components for convenience, although a random scan is not difficult to implement. Sampling occurs in three stages: an optional pre-scan where the initial Metropolis jumping distributions are adjusted, a period of burn-in where no output is recorded, and the final sampling. The maximum posterior estimates are used as the starting point for the chain; if these were not computed, the parameters are initialized by several iterations of a backfitting algorithm [40] where each component is estimated in succession using the partial residuals with respect the other components.

The pre-scan stage is intended to automate the initialization of the Metropolis candidate distributions. With many thousands of voxels, it is not convenient to monitor and tune the individual chains by hand. Although the BRAIN software produces diagnostics that expose problems with the choice of candidate distributions, it is far preferable (and less time consuming) to start with reasonable values. The approximate covariance matrix from the posterior maximization is dilated by a fixed factor and then decomposed into successive conditional distributions using the Cholesky factorization [38]. 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. Both of these are heuristic criteria, but they tend to lead to good jumping widths with little effort.

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. When possible, we sub-sample to reduce correlations in the sequence, although naturally this adds quite a bit to the running time of the chains. 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. With independent voxels, the algorithm is easily parallelizable, and we can reduce the total number of voxels substantially by masking out those outside the head. The voxels can also be ordered based on the preliminary asymptotic results (e.g., maximum posterior estimates). In this way, all of the interesting structure can be explored in a fraction of the total running time. Our default is 10,000 samples after burn-in for standard image sizes, but for very high-resolution images this is often reduced. 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 examined the performance of the sampling scheme on a sample of voxel time series (from several experiments) with different types of structure. With graphical diagnostics, correlations among parameters, and various standard convergence diagnostics [16] based on parallel chains with different starting points, we evaluated the effectiveness of sampling with the configurations we use in practice. These results suggested that the chains are mixing quite well and also that the normal approximation is quite reasonable in most cases.

Our sampling strategy varies among the model components; here we describe the strategy for each component assuming the default priors and an additive combination of overlapping responses. There are variations on these strategies when the defaults are not used, but this gives a general picture of our sampling algorithm. The baseline parameter 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 μ, but as part of the move, we adjust the responsiveness parameters by the ratio of the candidate to the current baseline so that the activation profile does not change as a result. 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 two seperate Metropolis steps, one of which keeps lag-on + attack constant while varying their relative size and the other which 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 hyperparameter 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 low-level coding tricks to substantially reduce the overhead.

7.3.2. Model Jumping

For varying the structure of the model in discrete ways, we use the reversible jump framework developed in [39]. 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.

As described in section 5, we allow the responsiveness parameters to take the value 0 (no response) with non-zero prior probability. This is equivalent to including a family of sub-models over which every subset of the experimental conditions is constrained to yield no response. 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, inclusion of a zero component or the removal of a non-zero component. The probabilities of the different move types are constrained if detailed balance is to be maintained for transitions across the spaces; it is usually simplest to keep the probabilities of moving in each direction equal. If a model jump is to be made, we select a candidate condition (i.e., one of the responsiveness paramters) of the appropriate type (zero or nonzero) at random for inclusion or removal. Our basic moves involve both the responsiveness parameter and the baseline. The baseline is adjusted simultaneously as part of the move since the 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 ℓ_1 and ℓ_2 denote the lengths in the design corresponding to the zeroed conditions not including the candidate and the candidate condition. The simplest move takes (μ, γ) 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 μ. This move follows the template given in [39] and hence satisfies detailed balance. A generalization of this is to perturb μ by an independent random amount (i.e., (μ, γ) 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). Although it adds variation to the transition, it appears to increase the rate of flow across models as long as the variance of w is not too large.

This satisfies Green's dimension matching condition but with a higher order since we are generating perturbations in both variables.

Our approach is similar for the drift when adaptive knots are used. Here, we take advantage of the viewpoint that the drift profile is the parameter, that is, 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 [38] 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: changing the position of a knot, adding a knot, and removing a knot. Again, the probabilities of these move types are constrained; we typically take the probabilities to be equal for every possible move. When there $K = 0$, 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 wish to have only a few knots (e.g., 2–5), but it facilitates mixing among the models to allow $K_{\max}$ to be larger. Specifically, to move the knots across a large portion of the domain usually involves moving through a low likelihood region which is unlikely to occur; it is thus easier to add knots in a new location and remove unneeded knots elsewhere in order to substantially change the structure of the profile. See [39] for a further discussion of this point. When adding, removing, or moving a knot, the effected knot position is selected at random. 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 [39] is satisfied, mixing is improved, and detailed balance is maintained.

8. Discussion

8.1. Assessment

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 the more recently developed methods, only the model of [50] has similar advantages. Our model also handles complex experimental designs and accounts for important features of the drift and response, such as 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 subsume the traditional classifications. We can address questions of localization but also a wide range of more general questions, including those about changes and spatial relationships in the response. Our methods allow scientists to directly target the questions they want to address. All of our inferences are accompanied by a measure of uncertainty in contrast to the results of classifications. Moreover, many spatial selection biases are avoided by using posterior probabilities, and our approach offers a direct treatment for multiple comparisons in terms of the probability calculus.

The main weakness of our approach is the level of effort that is required to derive the inferences. There are two aspects to this. First, fitting the model requires nontrivial computation, with an entire data set taking on the order of a day to analyze. With parallelization and improvements in computing technology, however, we expect this problem to become less severe over time. Second, the specific form of the model depends strongly on the imaging configuration and the nature of the design, so investigators must take some care in setting up the analysis. A related problem is that using simulations to derive the results includes an extra source of uncertainty and introduces some non-determinism. In practice, this is a small effect and can be reduced with longer simulation runs, albeit at a cost of further computing time. The dependence of the results on the choice of priors might also cause concern for some. We have found that specific shape of the priors does not have a large impact on the results as long as the basic range of the parameters is suitably constrained. Our goal has been to include generally accepted information about the processes into the model, so the priors we use reflect reasonably uncontroversial constraints. Moreover, the shape of the priors can be changed with ease. Finally, our current implementation does not account for all of the complexity in the data but can be easily extended to incorporate new features. We describe some of these below.

8.2. Extensions

In this paper, we have presented a basic implementation of our modeling framework that accounts for the main sources of variation in the data. However, there are several directions in which our implementation can be extended to deal with more of the complexity.

Noise Model. Adding more general noise processes to the model poses no difficulty as long as the likelihood can be reasonably approximated. For example, with ARMA models, we can use a conditional likelihood or compute the likelihood in the spectral domain [10]. To incorporate physiological noise, we can add particular spectral components to the noise autocorrelation function whose intensities are included as parameters in the model.

Response Variations. Our parameterization of responsiveness assumes that the response for a particular condition is constant across time, but in the data, there tends to be a component of within-condition variation. It is straightforward to incorporate these variations into our model and to estimate the corresponding variance components. We add a level to the hierarchy that includes responsiveness parameters $\gamma_{v,c,e}$ for epoch e of task condition c at voxel v. Those $\gamma_{v,c,e}$ in the same voxel and condition are then taken as being drawn from a single distribution that depends on the overall responsiveness $\gamma_{v,c}$ and a variance component $\tau^2_{v,c}$, which is itself constrained. Beyond extending the model, this formulation has further application to the problem of assessing reliability for fMRI methods; see [35].

Dependence Between Shape and Responsiveness. Our parameterization of the shape of the hemodynamic response curve does not depend explicitly on the intensity of the response; rather, the shape is scaled appropriately by $\boldsymbol{\gamma}$ in the activation profile. There is some recent evidence [72], however, that the shape varies with amplitude in a non-additive way. For instance, large responses appear to exhibit a broader plateau and later decay. As such relationships are clarified by further research, we can adjust the model to include this dependence. The main cost of this is computational, so we have thus far maintained the independent parameterization.

Spatial Structure. A common statistical approach for spatial modeling is to put a Markov Random Field (MRF) prior on the underlying structure [7, 31, 44, 32]. Besides efficiency of sampling, an advantage of this approach is that the global relationships are specified only through local dependence. Unfortunately, the isotropic neighborhoods typically chosen to define an MRF prior are not sensitive to changes in the dependence relationships across boundaries, and thus MRF priors tend to blur the boundaries in an image. Geman and Geman [31] show how adding another layer to the model—a process that specifies connections between sites and thus determines regions—can improve estimation of boundaries. Johnson *et al.* [42] generalize this "line process" to allow for more efficient estimation. Johnson *et al.* [43] develop an effective approach for segmenting and recovering images with large-scale boundaries that is based on a hierarchical model.

In fMRI, however, the role of the spatial model is not image reconstruction or segmentation *per se*; rather, it is identifying localized regions with consistent physiologic properties. These regions of dependence form the sets over which the model can justifiably combine information about the parameters. A particular challenge here is that tissue boundaries in the brain are convoluted and piecemeal. The cerebral cortex, a thin layer of cells where most of the interesting neural action takes place, is an intricately folded surface which appears and disappears in the images as it meanders through the slices. There are many cases in which tissue type, vasculature, and functional response change on a millimeter scale. For all of these reasons, the regions of dependence are plausibly local, and must be able to take on varied and often non-convex shapes.

A complete model for fMRI data needs to account for the spatial relationships among the fundamental processes that generate the data. The shape of the hemodynamic response function, the magnitude of the response, the impact of physiological variations, and other such features vary in their structure across the brain. Modeling these relationships increases the precision of inferences because multiple voxels contribute information about features they have in common. We are currently exploring a new approach to spatial modeling that is adapted to fMRI. This involves adding new layers to the hierarchy in our model to relate the model parameters across voxels. The spatial model uses a mixture of MRF distributions in which the cliques and potentials defining the MRF are allowed to vary locally. In this way, the parameters adapt to borrow strength across local regions when there is sufficiently common structure.

8.3. Inferences Across Subjects.

As discussed in section 2.3, an important issue is the need to combine inferences across subjects. One approach to this problem is to abstract away from the specifics of the anatomy and to find a basis for comparison that internalizes anatomical differences. For example, the anatomically specified regions of interest defined in section 6 provide reasonable functional units for comparison. The integrated response measures $\Gamma_c(R)$ over such regions can then be plausibly compared across subjects; see [36]. More generally, let G be some functional on the parameter space that does not depend on the explicit coordinate system of the image for a given subject. G might be the integrated response over a pre-specified region of interest or the indicator that there is a location dissociation between two tasks (c.f., section 2.3). 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.

A similar problem involves relating the fMRI results from multiple subjects to behavioral covariates. For example, one question of interest in the example of section 6 is how the relationship between response and task difficulty depends on the individual subject's working memory capacity. Here, a variety of behavioral tests can be used to measure working memory capacity independently of the fMRI data; the measured capacities serve as a covariate for exploring the impact of resource limitations on the changes in response across sentence difficulty. A basic analysis here would be to bin subjects according to the measured capacity and compare the shape of the $(\gamma_{\mathrm{T}_1}, \gamma_{\mathrm{T}_2}, \gamma_{\mathrm{T}_3})$ curves across and within groups.

9. 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] H. Akaike. Time series analysis and control through parametric models. In D.F. Findley, editor, *Time Series Analysis*. Academic Press, New York, 1978.

[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] T.S. Braver, J.D. Cohen, J. Jonides, E.E. Smith, and D. C. Noll. A parametric study of pre-frontal cortex involvement in human working memory. *NeuroImage*, 5(1):49–62, 1997.

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

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

[12] P. A. Carpenter. personal communication.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] F.C. Donders. On the speed of mental processes. *Acta Psych.*, 30:412–431, 1869.

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

[24] 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).

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

[26] 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).

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

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

[29] 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.

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

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

[32] 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.

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

[34] C. R. Genovese. Comment on Lange and Zeger. *Journal of the Royal Statistical Society C*, 46:23–24, 1997.

[35] C. R. Genovese, D. C. Noll, and W. F. Eddy. Estimating test-retest reliability in fMRI I. *Magn. Res. Med.*, to appear.

[36] C. R. Genovese and J. A. Sweeney. Functional connectivity in the cortical circuits subserving eye movements. In R. E. Kass, B. P. Carlin, A. L. Carriquiry, C. Catsonis, A. Gelman, I. Verdinelli, and M. West, editors, *Case Studies in Bayesian Statistics*, volume 3. Springer Verlag, 1997.

[37] 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.

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

[39] P. J. Green. Reversible jump mcmc computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

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

[41] 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.

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

[43] 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.

[44] 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.

[45] M. A. Just and P. A. Carpenter. Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psych. Rev.*, 92:137–172, 1985.

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

[47] 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.

[48] R. E. Kass and A. E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.*, 90:773–795, 1995.

[49] 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.

[50] 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.

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

[52] 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*, submitted.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

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

[59] P. Shah and A. Miyake. The separability of working memory resources for spatial thinking and language processing: an individual differences approach. *J. Expt. Psych.: General*, 125:4–27, 1996.

[60] N. Z. Shor. *Minimization Methods for Non-differentiable functions.* Springer Verlag, 1985.

[61] 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.

[62] E.E. Smith, J. Jonides, R. A. Koeppe, and E. Awh. Spatial versus object working memory: Pet investigations. *J. Cog. Neuro. Sci.*, 7:337–356, 1995.

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

[64] S. Sternberg. Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist*, 57:421–457, 1969.

[65] J.A Sweeney, M.A. Mintun, S. Kwee, M. B. Wiseman, D. L. Brown, D. R. Rosenberg, and J. R. Carl. Positron emission tomography study of voluntary saccadic eye movements and spatial working memory. *J. Neurophysiology*, 75(1):454–468, 1996.

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

[67] K. R. Thulborn. personal communication.

[68] 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.

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

[70] L. Tierney and J. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, 81:82–86, 1986.

[71] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.*, 80:8–20, 1985.

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

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

[74] 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.

[75] 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.

[76] 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.

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

[78] K.J. Worsley. Boundary corrections for the expected euler characteristic of excursion sets of random fields, with an application to astrophysics. *Adv. Appl. Prob.*, (to appear), 1994.

[79] K.J. Worsley. Local maxima and the expected Euler characteristic of excursion sets of χ^2, f, and t fields. *Adv. Appl. Prob.*, 26:13–42, 1994.

[80] 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.*, (to appear).

[81] K.J. Worsley, A.C. Evans, S. Marrett, and P. Neelin. Detecting changes in random fields and applications to medical images. *JASA*, (to appear), 1994.

[82] 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 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 2. Illustration of two 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 separate the conditions. Panel (a) describes a simple, alternating two-condition design, using the finger tapping experiment as an example. Panel (b) shows a more complicated design with six conditions (A–F). The epochs here are shown at two separate heights to make the divisions more salient.

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 correlation map showing the correlation coefficient of every voxel time series in the image with one specific voxel time series. The grey scale ranges from -1 (white) to 1 (black). Note the range and non-locality of the spatial correlations.

Figure 6. 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 7. A voxel time series with an apparently large movement artifact near the 425th image. In this case, the movement could be clearly detected visually by examining the sequence of images in an animated loop.

Figure 8. 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) A normalized contrast map comparing conditions T_3 and Tr. This map is not thresholded, but since the normalized contrasts include a potential contribution from sub-models with no responsiveness, as described in the text, the resulting shrinkage has much the same effect. The gray-scale for this panel shows the values of the normalized contrast.

(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) 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.

(e) Binned values of the regression slope coefficients for voxels with domination probabilities

$$\mathrm{P}\{\, \max(\gamma_{T_1}, \gamma_{T_2}, \gamma_{T_3}) > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\} > 0.01$$

as described in the text. To make the figure visually interpretable, the values were binned by the t-statistic of the slope coefficient (used as benchmark only). The pixels are assigned values 1, 0, and -1, where the sign indicates the direction of the slope and nonzero values were "significant" and the 0's were not. The gray-scale serves to distinguish these three values.

(f) 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 9. 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 10. 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 8f, 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 T_1 to T_2 to 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 T_1 and T_2 than between T_2 and T_3. The marked voxel on the left tends to show a very small change between T_1 and T_2 but a large change between T_2 and T_3.

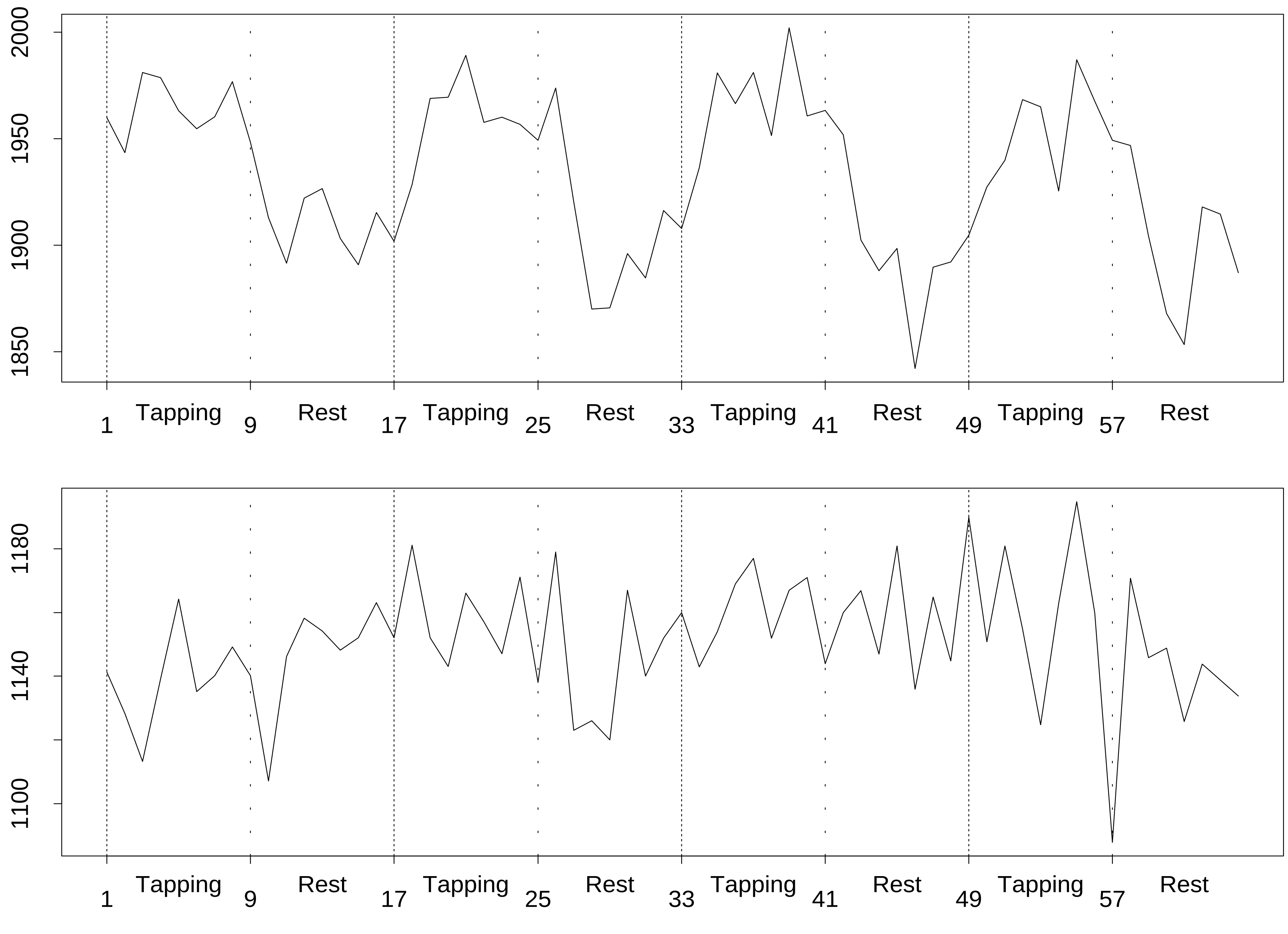
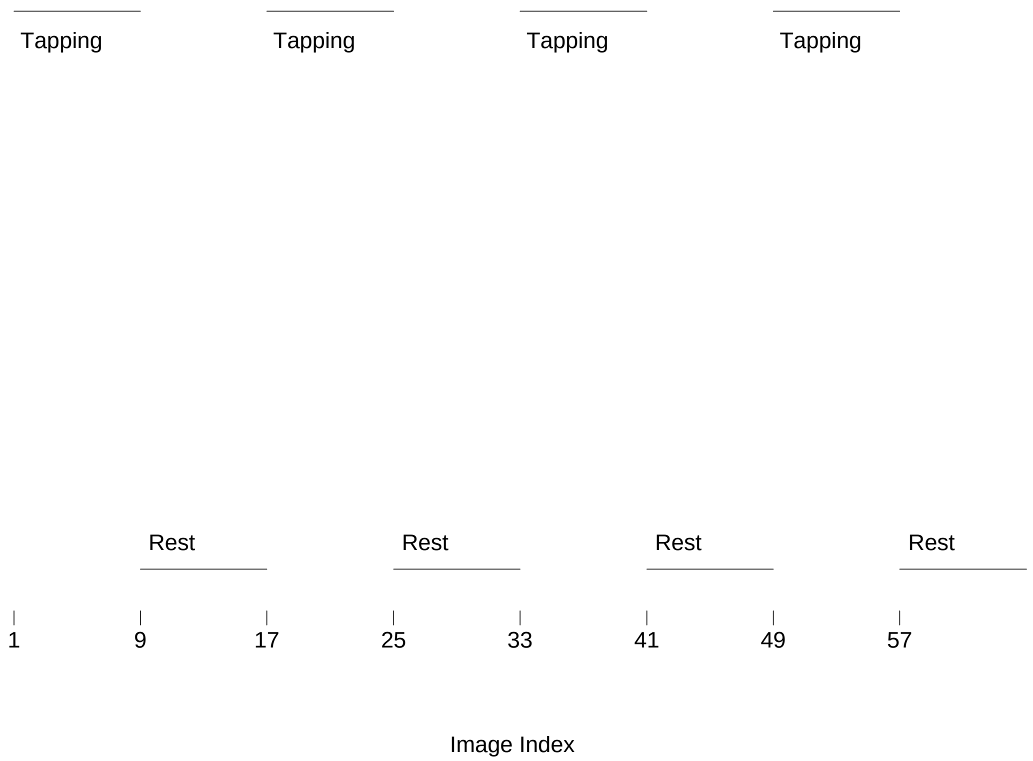

2000
1950
1900
1850
1
Tapping
9
Rest
17
Tapping
25
Rest
33
Tapping
41
Rest
49
Tapping
57
Rest
1180
1140
1100
1
Tapping
9
Rest
17
Tapping
25
Rest
33
Tapping
41
Rest
49
Tapping
57
Rest

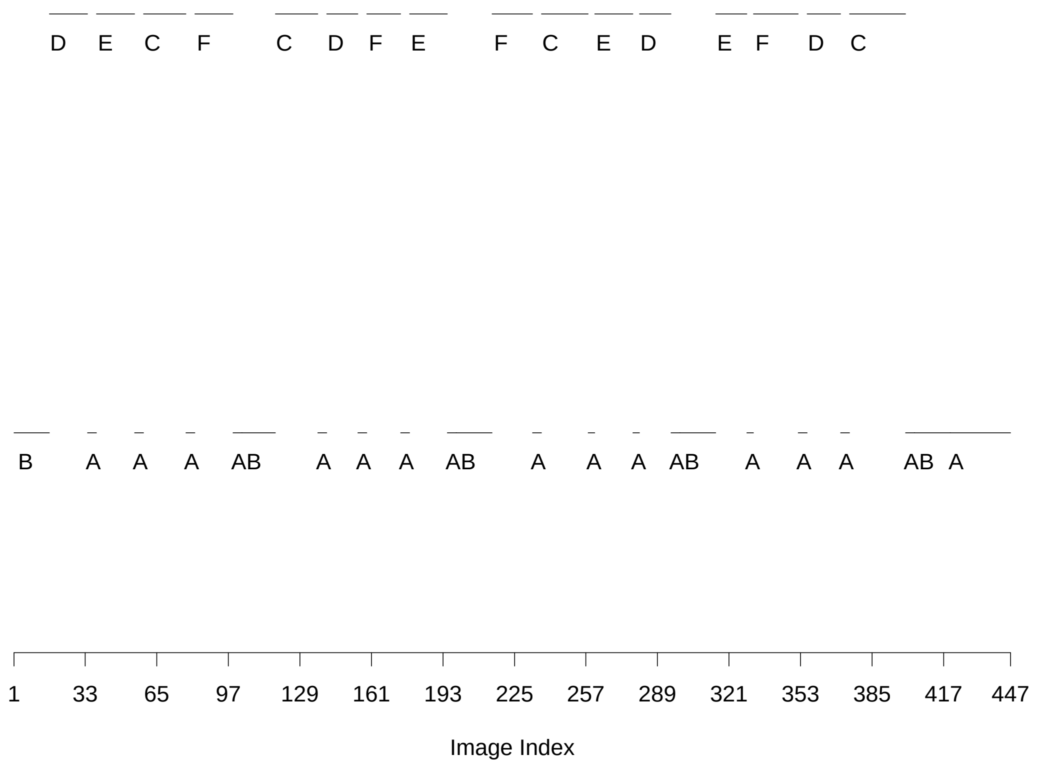
Figure 1

Figure 2



Tapping
Tapping
Tapping
Tapping
Rest
Rest
Rest
Rest
1
9
17
25
33
41
49
57
Image Index

(a)



D E C F C D F E F C E D E F D C
B A A A AB A A A AB A A A AB A A A AB A
1
33
65
97
129
161
193
225
257
289
321
353
385
417
447
Image Index

(b)

Figure 3

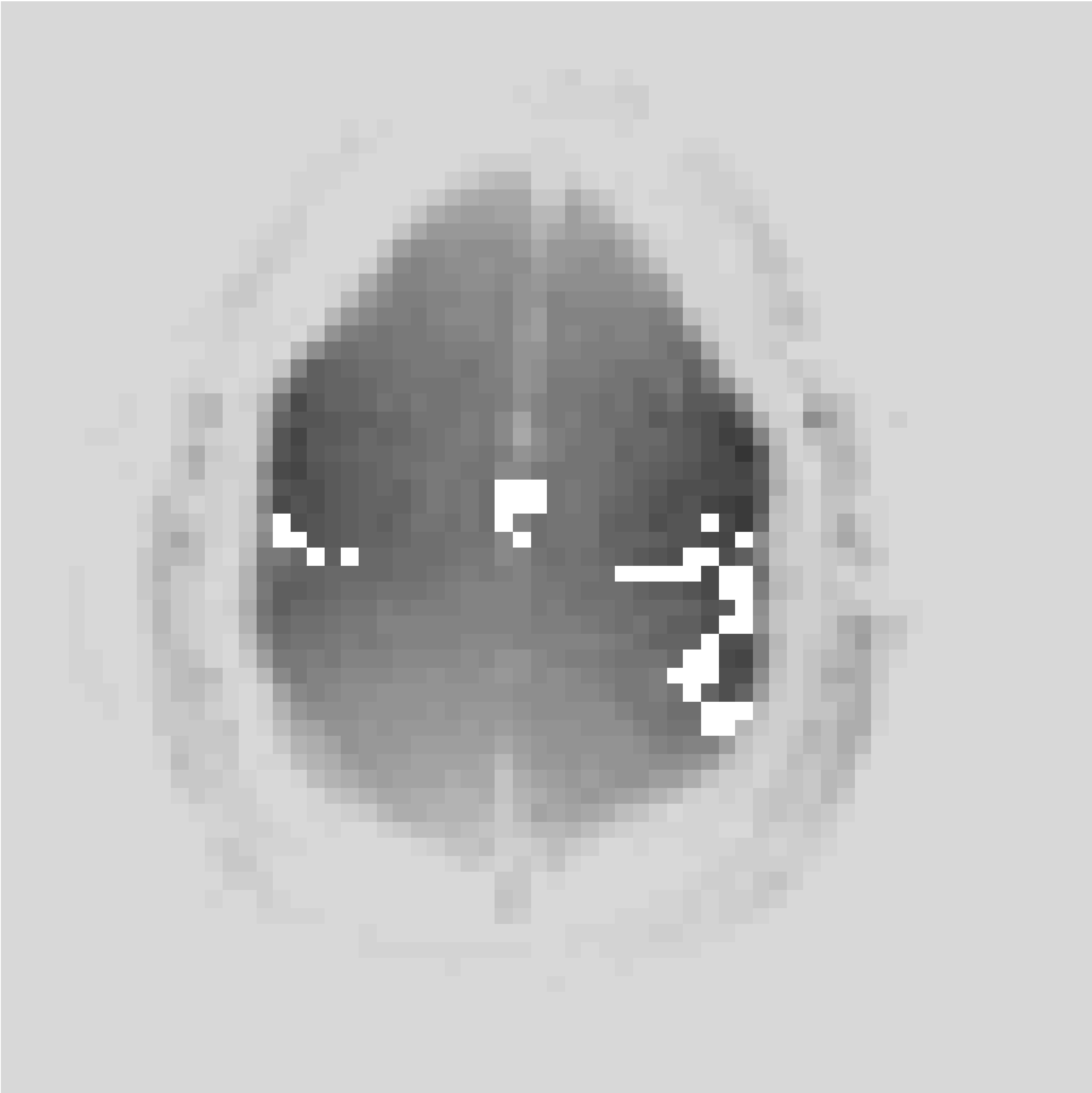
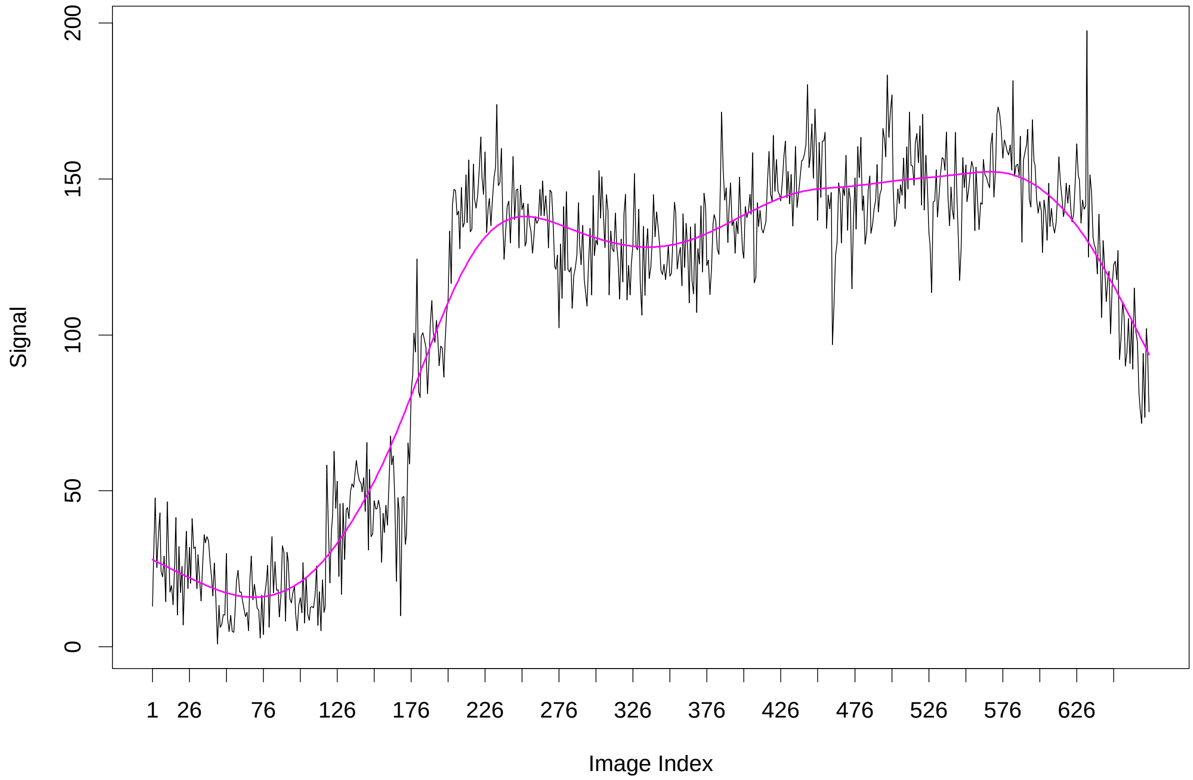
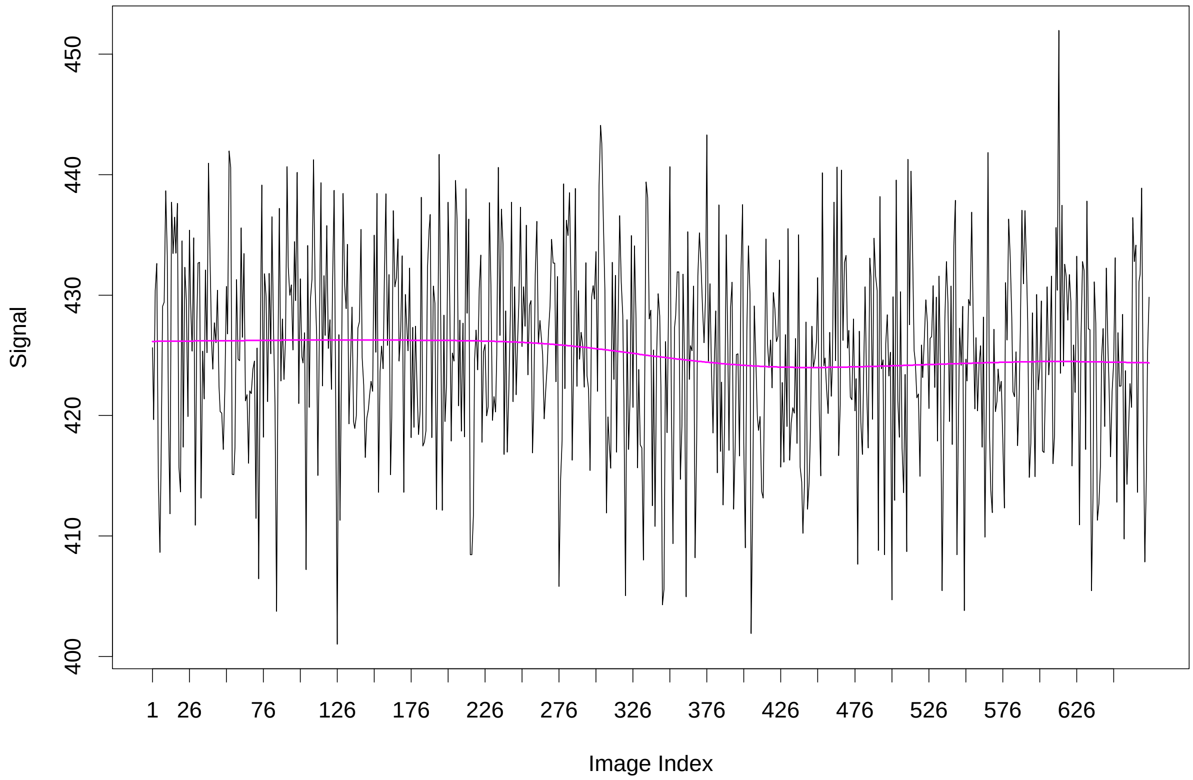


Figure 4



Signal
0
50
100
150
200
1
26
76
126
176
226
276
326
376
426
476
526
576
626
Image Index

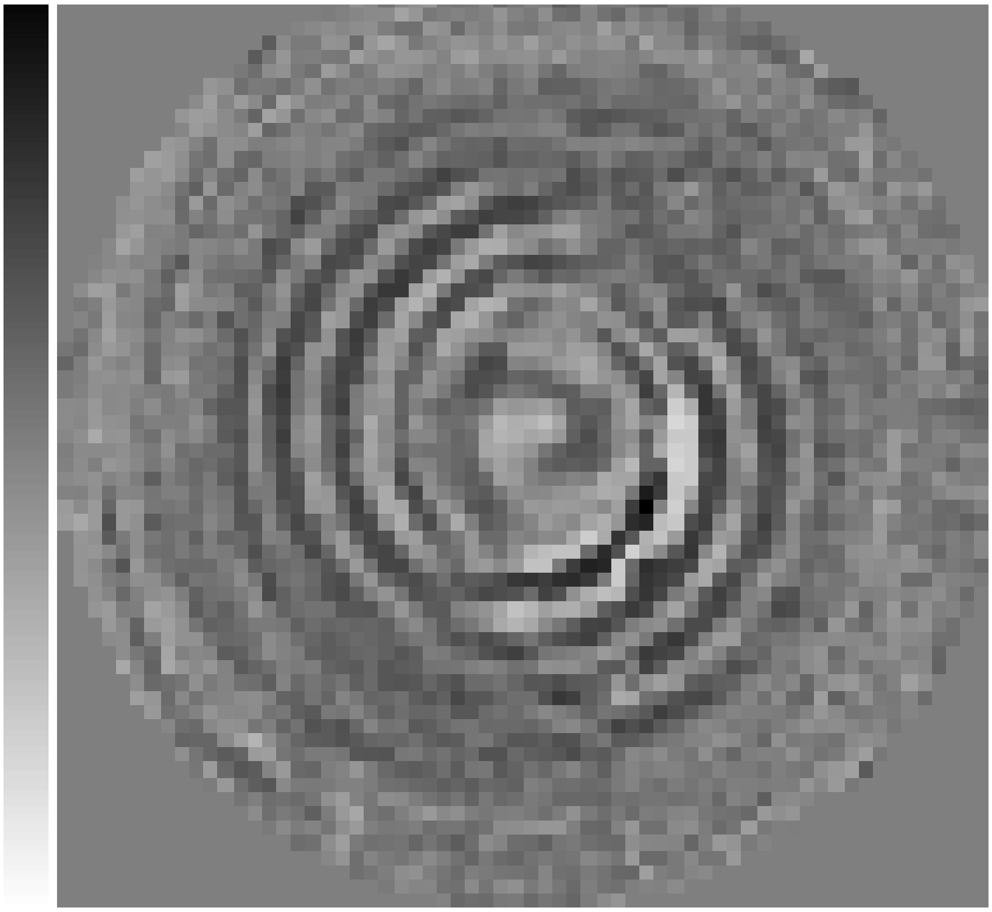
(a)



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

(b)

Figure 5



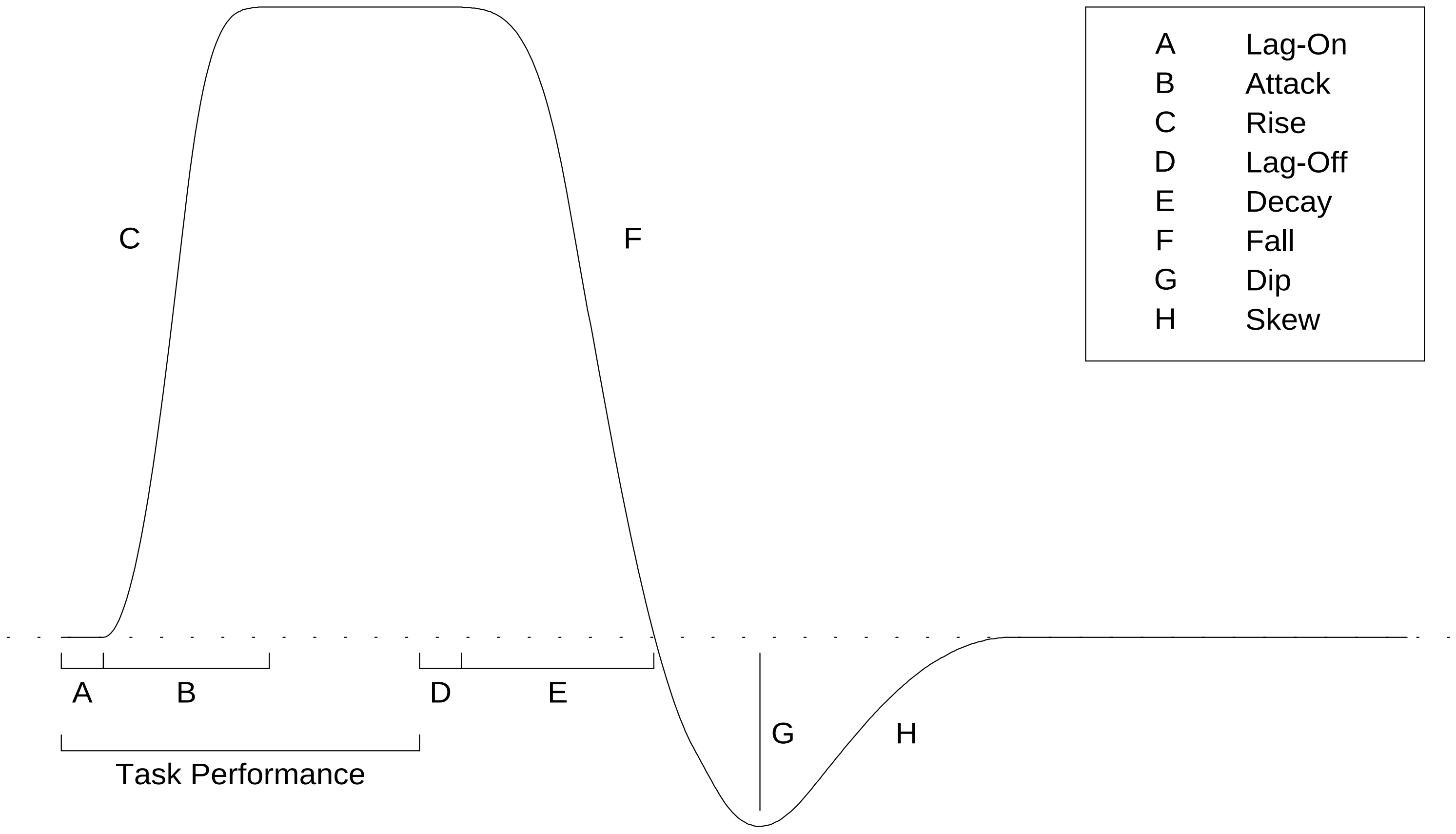

A Lag-On
B Attack
C Rise
D Lag-Off
E Decay
F Fall
G Dip
H Skew
C
F
A
B
D
E
G
H
Task Performance

Figure 6

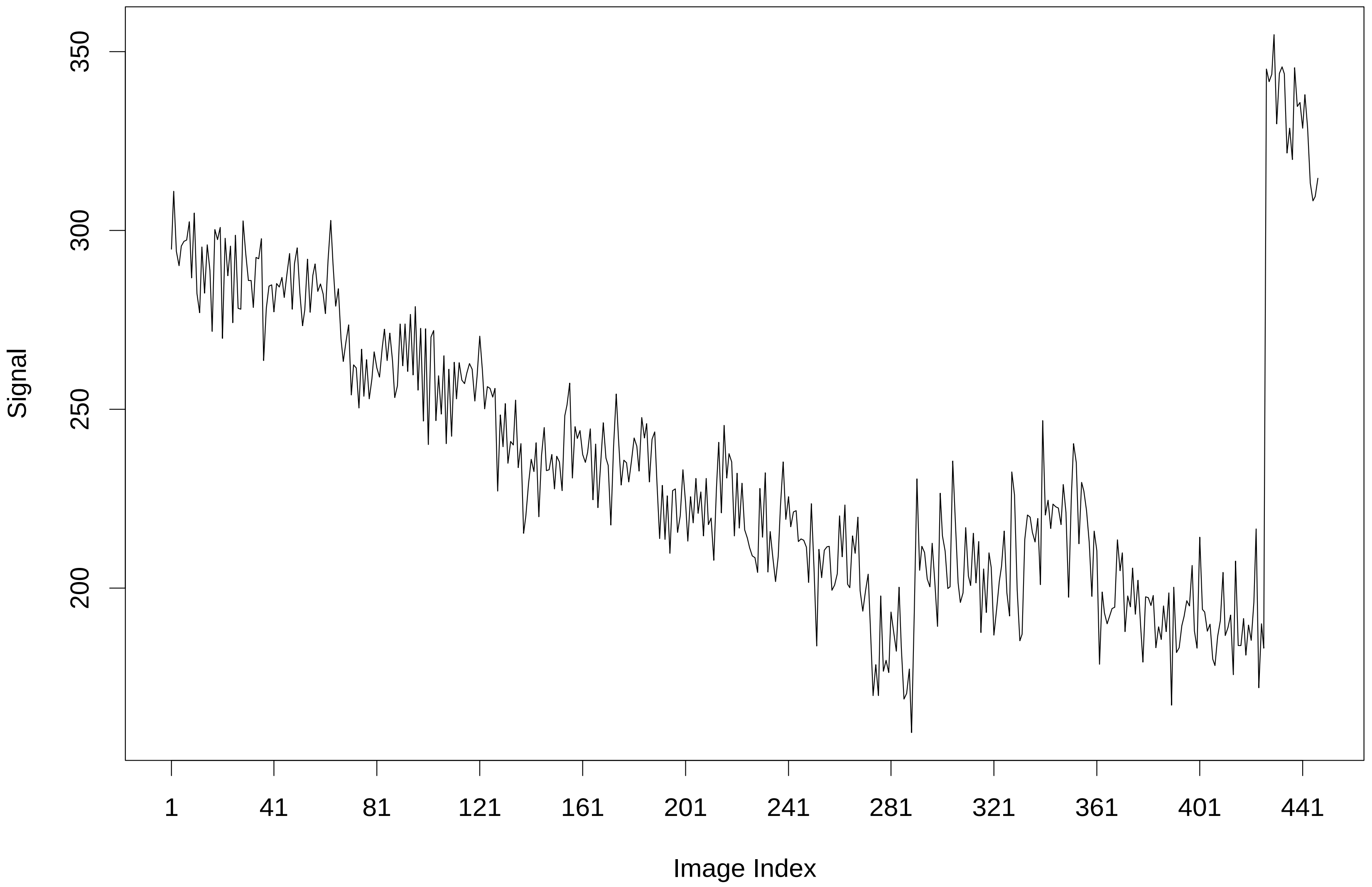
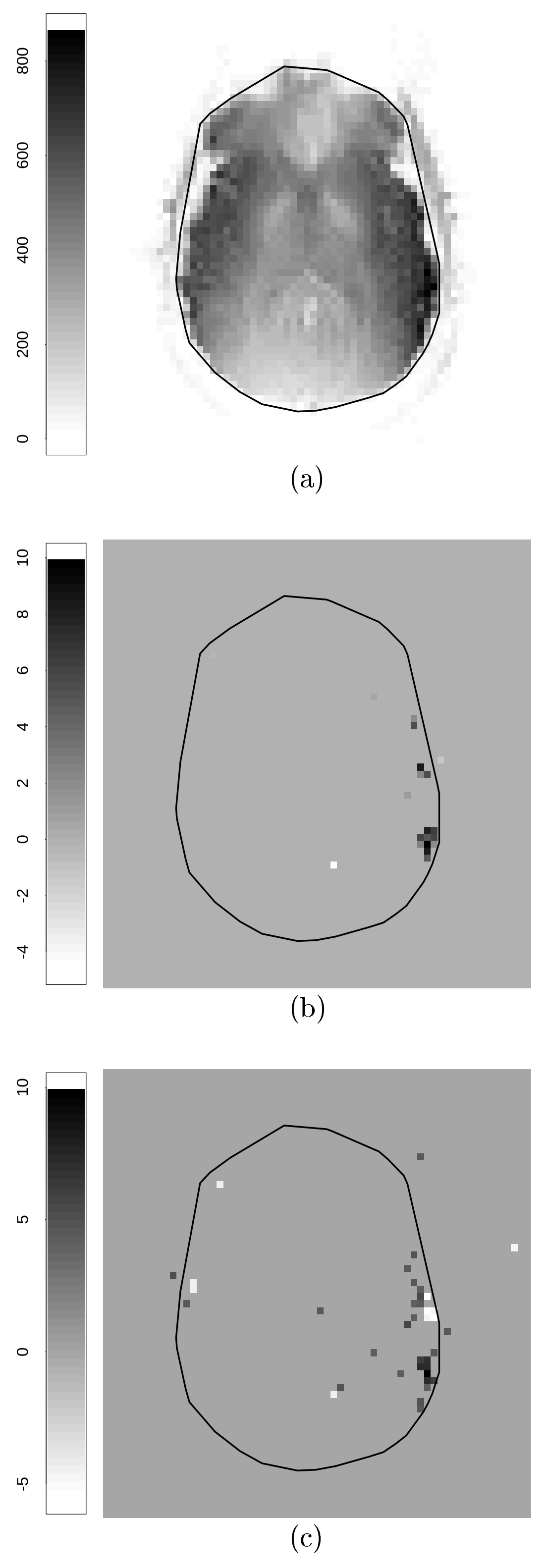

Signal
350
300
250
200
1
41
81
121
161
201
241
281
321
361
401
441
Image Index

Figure 7

Figure 8



0
200
400
600
800
(a)
0.0
0.2
0.4
0.6
0.8
1.0
(d)
-4
-2
0
2
4
6
8
10
(b)
-1.0
-0.5
0.0
0.5
1.0
(e)
-5
0
5
10
(c)
0.0
0.2
0.4
0.6
0.8
(f)

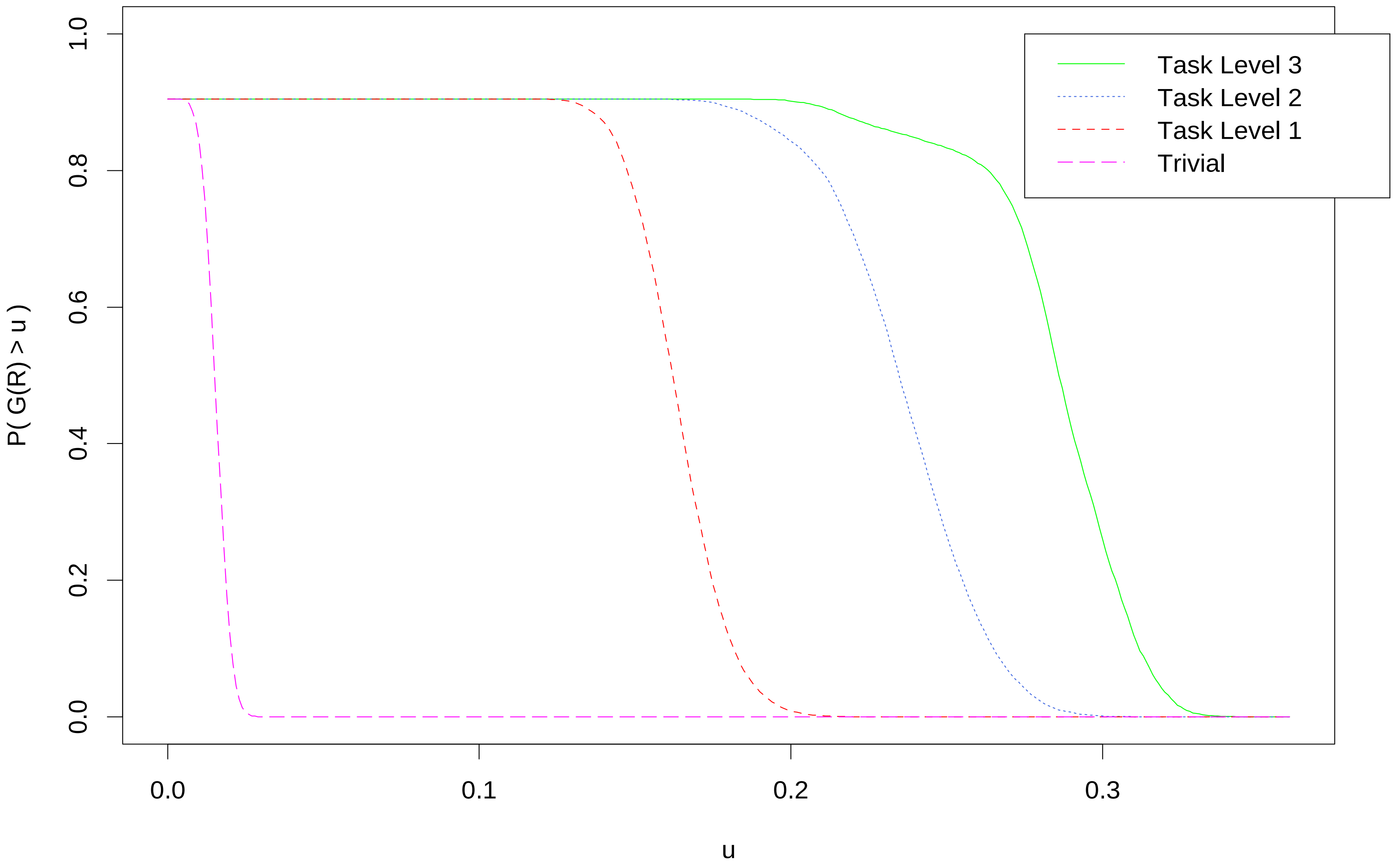

Task Level 3
Task Level 2
Task Level 1
Trivial
P(G(R) > u)
u
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3

Figure 9

Figure 10