In Search of Functional Specificity in the Brain: Generative Models for Group fMRI Data

by

Danial Lashkari

B.S., Electrical Engineering, University of Tehran, 2004
M.S., Electrical Engineering, University of Tehran, 2005

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
at the Massachusetts Institute of Technology

June 2011

Signature of Author: ____

Department of Electrical Engineering and Computer Science
May 19, 2011

Certified by: ____

Polina Golland
Associate Professor of Electrical Engineering and Computer Science
Thesis Co-Supervisor

Certified by: ____

Nancy Kanwisher
Ellen Swallow Richards Professor of Cognitive Neuroscience
Thesis Co-Supervisor

Accepted by: ____

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Committee for Graduate Students

In Search of Functional Specificity in the Brain: Generative Models for Group fMRI Data

by

Danial Lashkari

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, we develop an exploratory framework for design and analysis of fMRI studies. In our framework, the experimenter presents subjects with a broad set of stimuli/tasks relevant to the domain under study. The analysis method then automatically searches for likely patterns of functional specificity in the resulting data. This is in contrast to the traditional confirmatory approaches that require the experimenter to specify a narrow hypothesis a priori and aims to localize areas of the brain whose activation pattern agrees with the hypothesized response. To validate the hypothesis, it is usually assumed that detected areas should appear in consistent anatomical locations across subjects. Our approach relaxes the conventional anatomical consistency constraint to discover networks of functionally homogeneous but anatomically variable areas.

Our analysis method relies on generative models that explain fMRI data across the group as collections of brain locations with similar profiles of functional specificity. We refer to each such collection as a functional system and model it as a component of a mixture model for the data. The search for patterns of specificity corresponds to inference on the hidden variables of the model based on the observed fMRI data. We also develop a nonparametric hierarchical Bayesian model for group fMRI data that integrates the mixture model prior over activations with a model for fMRI signals.

We apply the algorithms in a study of high level vision where we consider a large space of patterns of category selectivity over 69 distinct images. The analysis successfully discovers previously characterized face, scene, and body selective areas, among a few others, as the most dominant patterns in the data. This finding suggests that our approach can be employed to search for novel patterns of functional specificity in high level perception and cognition.

Thesis Supervisor: Polina Golland
Title: Associate Professor

Thesis Supervisor: Nancy Kanwisher
Title: Professor

Acknowledgements

When I arrived at MIT in September 2005, I was an aspiring physicist who could not imagine himself studying anything but quantum information science. Soon, the border-less structure of MIT exposed me to an overwhelming variety of new fascinating fields. Among those was probabilistic modeling, to which I was introduced in the spring of 2006 through the newly designed course "Inference and Information." As it turned out, Polina Golland, one of the two instructors, believed she could trust a novice like me with one of her research projects. The idea was to use probabilistic modeling techniques to help Nancy Kanwisher, a neighboring MIT neuroscientist, make more sense of her functional brain images. I knew even less about the brain and brain imaging than I did about data modeling. Yet, a sense of adventure and curiosity overcame my appreciation for the beauty of quantum mechanics and I joined Polina's group in the computer vision lab that fall.

Over the past five years, I have often marveled at how lucky I have been to join this group. Besides all the joy I have had in exploring new subjects, having Polina as an advisor has undoubtedly been the most pleasant surprise of my MIT life. What I have earned working with Polina is by no means limited to a knowledge of generative models and statistical inference. Polina has taught me how to organize and present my ideas by her scrupulous edits on countless drafts of my writings. Polina's conscientious, yet laid-back, approach to academic work, which equally emphasizes excellence and fun, has become an inspiration for all my future endeavors. I am sincerely grateful for her absolute trust, ultimate support, and unimaginable patience.

The painstaking process of working with an earnest, critical scientist from a different discipline, Nancy, proved to be another exceptional learning experience for me. I thank Nancy for her enthusiasm and involvement and her group for their terrific help. Edward Vul and Po-Jang (Brown) Hsieh collected and preprocessed the data in the experiments reported in this thesis and also made major contributions to the development of the core ideas of the thesis through our discussions.

I would like to also thank all members of Polina's group who coincided with me at any point during my five-year long tenure there. They sat through my many practice presentations, provided me with great feedback, and put up with my endless rambles in our reading group. Ramesh Sridharan, in particular, collaborated with me closely in

the implementation of some of the inference algorithms presented here, and kept me company in several of the many long pre-deadline nights of work. Despite my many years in the group, I never felt quite at home dealing with computing and other related issues. As a result, fellow lab-mates had to deal with my ceaseless train of computing problems and random questions. Among others, Bo Thomas Yeo, Wanmei Ou, Mert Sabuncu, Serdar Balci, Biswajit (Biz) Bose, Gerald Dalley, Jenny Yuen, Michael Siracusa, Archana Venkataraman, and George Chen were extremely helpful in this regard. Moreover, I also had the pleasure of collaborating with Georg Langs, Bjoern Menze, and Andrew Sweet on projects that were not reported in this thesis.

On a slightly more historical note, I am greatly indebted to the faculty who provided me with support and guidance during my transition from the University of Tehran to research at MIT. In particular, I would like to express my thanks to Reza Faraji-Dana and Shamsoddin Mohajerzadeh, my advisor and my mentor at the UT, and George Verghese and Terry Orlando, from the MIT EECS department, for their tremendous assistance and advice during that process.

My experience at MIT turned out to be one of the most formative phases of my life, through which I learned much more about myself and the world around me. Weathering the challenges of this journey would not have been possible without the support and company of many friends and allies outside the lab. Although space does not allow me to thank all of them by name, I would like to specifically acknowledge fond memories that I shared with Nabil Iqbal, Christiana Athanasiou, Maria Alejandra Menchaca Brandan, and Hila Hashemi throughout all our years together at MIT.

The work that resulted in this thesis began long before I arrived at MIT. The preparation for my PhD studies may be traced back to the days when an idealist father began to enthusiastically teach his three-year old son reading and basic algebra. The academic path that brought me to MIT was meticulously engineered, years ago, by a mother’s care and attention to the early studies of her son. My father implanted in me an everlasting thirst for learning and knowing and my mother instilled in me a yearning for highest attainable levels of educational excellence. This thesis is first and foremost a tribute to my parents, to their love, and to their values, which I genuinely share and admire. I also dedicate this thesis to my sister Mitra who was very young when I left home for college studies in Tehran. Throughout my years away from home, first one thousand kilometers and then some tens of thousands, I have had far too few chances for being around her as she was growing. I hope she now accepts this dissertation as an ex post facto excuse for my long absence. My brother Nima, who will soon be writing his own dissertation in theoretical physics, has always been a source of pride for me. As the older brother, however, I admit that it came as a relief that I managed to obtain the first doctorate in the family, after all!

Cambridge, MA
May, 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

UNDERSTANDING the organization of human brain is one of the most ambitious projects undertaken by modern science. From basic sensorimotor tasks to abstract reasoning, all that we are capable of doing as human beings stems from the functioning of the brain. By explaining the brain, neuroscience unearths the underlying biological principles behind our perceptions, actions, emotions, motivations, and thoughts, as well as what differentiates between us in these activities. One can therefore argue that the success of brain science may have the most far-reaching and profound implications among sciences in different aspects of human life. Beyond obvious medical value, scholars in disciplines as varied as education, law, business, economics, politics, and philosophy have begun to actively discuss the consequences of neuroscientific discoveries in their respective fields. Expectedly, the staggering diversity of brain function broadens the range of the implications of neuroscience and also creates a major question for the field. How are all these different functions organized within the brain?

Functional specialization answers this question, at least in part. Many complex natural and artificial phenomena emerge in systems that comprise several interacting, specialized compartments. Such functional specialization is ubiquitous in many domains, particularly in biology. Bacteria are living organisms endowed with basic mechanisms of metabolism, reproduction, and even movement. In a bacterium, different cellular regions, such as nucleoid, cytoplasm, and ribosomes, each contribute to a different aspect of the cell's function. The same functional pigeonholing principle explains different organelles in more complex cells and organs in advanced macroorganisms–including, of course, the specific functionality of the brain itself as an organ within the human body. We can provide an evolutionary justification for the emergence of specialization in biological function based on an efficiency advantage. To use a metaphor, one of the hallmarks of human social evolution has been the specialization of careers, which enabled modern individuals to make more substantial progress in one specialty instead of obtaining a basic level of training in a wide variety of skills. Similarly, one can argue, once a biological unit becomes dedicated to a specific function, it can evolve to perform that task more efficiently.

Having these simple intuitions in mind, it comes as no surprise that functional specialization has been a recurring theme in neuroscience, since the inception of brain studies in the ninteenth century. Austrian physician and anatomist Joseph Gall (1758-1828) was one of the early pioneers of the doctrine of localization that posited that different cognitive functions were associated with different brain areas (Zola-Morgan, 1995). Gall's theory was not based on empirical observations and was questioned soon after by the prominent French anatomist, Pierre Flourens, who concluded based on his experiments that the cerebral cortex was an indivisible unit (Changeux and Garey, 1997). For more than 150 years, generations of neuroanatomists and neuroscientists have continued the debates started with Gall and Flourens (Kanwisher, 2010). Meanwhile, of course, much progress has been made in our understanding of the brain. This progress, however, has been rather equivocal on the question of specificity, showing evidence both for and against it on different fronts (Schiller, 1996). Most neuroscientists today may hold a moderate position that emphasizes understanding the extent of specificity in brain function while acknowledging its existence. Even if we agree with Gall on a fully compartmentalized picture of brain function, we still need to explain the necessary subsequent integration stage in which the information processed by specialized areas is combined to create a basis for coherent human beharvior. When, where, and how such an integration happens still remains a fundamental question in neuroscience.

As the title suggests, this thesis is concerned with the question of functional specificity in the brain. Broadly speaking, two important distinctions can be made in the context of the studies of functional specificity. First, we can study the brain functional specificity at varying spatial scales. At one end of the spectrum, we may distinguish specificity in the function of neurons. At the other end, we may be interested in specificity at the level of large brain areas (Schiller, 1996). For decades electrophysiology techniques have provided excellent means for probing functional specificity at the neuronal level (see, e.g., the groundbreaking work of Hubel and Wiesel, 1962, 1968). Yet studies of area-level specificity, until lately, relied primarily on patients with brain lesions. The recent advent of functional neuroimaging techniques, and in particular functional Magnetic Resonance Imaging (fMRI), revolutionized this field by offering non-invasive, large-scale observations of the brain processes in humans.

The second dimension in the study of brain specificity corresponds to specialization at different levels of functional abstraction. On one hand, we know that specific cortical patches are dedicated to the processing of low level motor and sensory inputs. The early visual cortex that includes areas V1–V5 provides a well studied example where regions with relatively distinct neuroanatomical architectures are thought to process different perceptual attributes of the visual inputs such as orientation, color, and direction of motion (Livingstone and Hubel, 1995; Zeki, 2005). When it comes to higher level functional specialization, for instance, high level perception or language, the current understanding of the functional organization is much less refined. Since the late nineteenth century lesion studies have suggested that the so-called Broca's and Wernicke's areas, in the frontal and temporal lobes of the dominant hemisphere respectively, are involved in language processing. The case of visual object recognition provides another well-studied example where category selectivity have been observed in humans and primates both at the level of neurons (Kreiman et al., 2000; Logothetis and Sheinberg, 1996; Tanaka, 1996) and brain areas (Kanwisher, 2003, 2010). Other reported examples of high level functional specialization include areas involved in the processing of other people's ideas (Saxe and Kanwisher, 2003; Saxe and Powell, 2006), cognitively demanding tasks (Duncan, 2010), as well as new language areas beyond those of Broca and Wernicke (Fedorenko et al., 2010).

This thesis develops techniques for exploratory fMRI studies of cognitive functional specificity. In terms of the classification described above, this problem translates into the search for the specialization in the brain at the area-level in spatial extent and high-level in abstraction. Originally, the work in this thesis was mainly meant to be a contribution to fMRI data analysis. However, the methods provided here open doors for performing novel studies that utilize the potential of our new approach to analysis. Hence, the work also involves an experimental aspect related to the design of fMRI studies. Section 1.1 discusses the motivations for this work in more detail and describes the particular problems addressed by methods developed in later chapters. Section 1.2 briefly explains the approach we have taken in this thesis for tackling those problems, enumerates its contributions, and provides an outline for the remainder of the thesis.

■ 1.1 Background and Motivation

Functional MRI has played a major role in the expansion of the field of cognitive neuroscience (Gazzaniga, 2004). Once fMRI allowed us to find measures of brain responses *in vivo*, we were able to employ it for investigating the organization of high level cognitive processes, phenomena that had hitherto been studied mainly using behavioral methods in psychology. Among domains where functional neuroimaging has made the most visible impact, vision undoubtedly stands out (Grill-Spector and Malach, 2004). In particular, fMRI studies have uncovered the properties of a number of regions along the ventral visual pathway that demonstrate significant selectivity for certain categories of objects (Kanwisher, 2003, 2010). Since the studies of visual category selectivity have provided the primary motivation and the main application for the fMRI models developed in this thesis, we briefly review their findings next.

■ 1.1.1 Functional Specificity and Visual Object Recognition

A ubiquitous part of our brain's daily functions, visual object recognition has motivated many neuroscientific studies in the last 50 years (Logothetis and Sheinberg, 1996). The current computational models and neuroscientific findings suggest a hierarchical processing stream in the cortex where the visual sensory input passes through different stages of representation along which the information relevant to cognitive tasks becomes more explicitly encoded (DiCarlo and Cox, 2007; Riesenhuber and Poggio, 1999; Ullman, 2003). Thus, understanding the brain visual processing system requires the characterization of different areas of the visual cortex and their corresponding representations of the visual world. In the early visual areas, the representation is generally local and encodes low level features such as orientation and color. As we progress towards high level areas, the representation becomes non-local, relatively invariant to identity-preserving transformations of the image, and predictive of image category.

Prior to the utilization of functional neuroimaging, little was known about the organization of high level visual areas in the cortex. Some studies on patients with prosopagnosia, a specific form of visual agnosia that results in impaired recognition of human faces, had suggested the existence of certain localized areas in the brain involved in face processing (Farah, 2004). Selectivity for faces had been further reported for neurons in the temporal cortex of primates (e.g., Desimone, 1991; Perrett et al., 1982), but that area was then believed to be generally involved in the processing of all complex objects (Tanaka, 1996). The discovery of a small patch of cortex on the fusiform gyrus that showed larger response to face images when compared to any other objects created a breakthrough in the studies of visual object recognition (Kanwisher et al., 1997; McCarthy et al., 1997). This region, which was named fusiform face area (FFA), can be easily and robustly identified in fMRI experiments, making it possible to use it as a region of interest (ROI) for further characterizations (Kanwisher and Yovel, 2006; Rossion et al., 2003). More recently, Tsao et al. (2006) showed that neurons in a homologous face selective area of primates, identified with a fMRI technique similar to the one used for humans, indeed demonstrate face selectivity.

Since the discovery of FFA, a handful of other category selective areas have been found in the extrastriate and temporal cortex. Notably, the parahippocampal place area (PPA) (Aguirre et al., 1998a; Burgess et al., 1999; Epstein and Kanwisher, 1998), the extrastriate body area (EBA) (Downing et al., 2001; Peelen and Downing, 2007; Schwarzlose et al., 2005), and the visual word form area (VWFA) (Baker et al., 2007; Cohen et al., 2000, 2002) exhibit high selectivity for places, body parts, and orthographically familiar letter strings, respectively. Selectivity for these categories have been also reported in other regions such as occipital face area (OFA) (Kanwisher and Yovel, 2006) and fusiform body area (FBA) (Schwarzlose et al., 2005), as well as scene-selective regions in retrosplenial cortex (RSC) and transverse occipital sulcus (TOS) (Epstein et al., 2007).

We do not expect that the brain regions engaged in object recognition consist entirely of contiguous class-selective areas. In order to efficiently employ the learning capacity of the ventral visual pathway, different objects of the same class must be represented by a pattern of response over some population of neurons. While the fMRI data essentially integrates the response of large populations of neurons within its spatial units, we can assume that the spatial patterns of fMRI response over a number of these units may still encode some information about object categories. Therefore, an alternative approach to the study of object representation attempts to find object representations distributed across large areas of cortex and overlapping across many categories of images (Cox and Savoy, 2003; Hanson et al., 2004; Haxby et al., 2001, 2000; Ishai et al., 1999). Nevertheless, a vast array of studies on the category-selective areas have yielded consistent characterizations of these areas, strongly supporting the idea that some degree of category-level functional specificity is present in the visual cortex (Op de Beeck et al., 2008). The extent of this specificity is a subject of ongoing debate in the field (Kanwisher, 2010). This thesis aims to contribute to our understanding of the question by providing a principled approach to search for functional specificity.

■ 1.1.2 Search for Specificity Using FMRI

How can we identify a pattern of functional specificity in the brain? The category-selective areas discussed so far have been discovered in the traditional confirmatory framework[1] for making inference from fMRI data. This approach first postulates a candidate pattern of functional specificity. The hypothesis may be derived from prior findings or mere intuition. Then, we design an experiment that allows us to detect brain areas that show the specificity of interest. Unfortunately, fMRI data is extremely noisy and the resulting detection map does not represent a fully faithful picture of actual brain responses. In order to confirm the hypothesis, we examine the detection maps across different subjects for contiguous areas located around the same anatomical landmarks. Finding such anatomical consistency in the activated areas attests to the validity of hypothesis.

In an admirable effort to search the space of category-selectivity beyond the current findings, Downing et al. (2006) employed this confirmatory framework to test the existence of 20 different categories. Surprisingly enough, they could not confirm even a single further type of category selectivity. Yet this outcome, as well as the failure of other efforts, raises an intriguing question about functional specificity in the visual cortex: is there really something special about the processing of faces, scenes, and bodies, or is there something wrong with our approach to searching the space?

Although it is hard to refute the first possibility given the data at hand, there are indeed enough reasons to be suspicious about the current analysis methods. We summarize the main limitations of the traditional confirmatory approach as follows:

1. *Manual Search of a Large Space:* The space of categories that may constitute a likely grouping of objects in the visual cortex is orders of magnitude larger than 20. A brute force search of such a large space does not appear to be a feasible strategy. In fact, even within the experiment performed in (Downing et al., 2006), if we consider meta-categories composed of all possible collections of their original categories, we should test for about $2^{20} \approx 10^6$ different candidates.

2. *Biased Characterization of Categories:* We usually define categories based on semantic classifications of objects. It is reasonable to expect that how the cortex groups visual stimuli for the purpose of its processing may partially reflect our conceptual abstractions of object classes. Yet, we cannot disregard the possibility that

[1] For a more detailed discussion of confirmatory analyses in fMRI, please see Section 2.2.

some cortical groupings may not exactly agree with what we think of as categories. In other words, we need to really search all possible selectivity patterns in the space of distinct objects, instead of confining the search to the object categories that make sense to us.

3. *Unproven Assumption About the Connection Between Function and Anatomy:* Is it possible that the organization of different category-selective areas varies across subjects in a way that does not respect the organization of anatomical landmarks? Is it possible that, instead of contiguous blob-like structures, category-selectivity appears in diffused networks of smaller regions? Current findings do not provide enough evidence for rejecting either of these possibilities. Yet, most fMRI analysis techniques still rely on the premise that functionally specific areas are constrained to be located around the same anatomical landmarks in all subjects.[2]

Same challenges are also present in studies of high level functional specificity beyond the visual cortex. When we get to the level of abstraction studied, say, in language, social cognition, or the theory of mind, it is never *a priori* known what types of specificity we need to consider. Moreover, our understanding of the relationship between function and anatomy generally becomes looser as we progress toward the frontal lobes of the brain.

The goal of this thesis is to devise an alternative approach to the design and analysis of fMRI studies of functional specificity, and in particular visual category selectivity, that mitigates the aforementioned problems.

■ 1.2 Contributions

In this thesis, we propose an exploratory framework for the analysis of fMRI data that enables automatic search in the space of patterns of functional specificity. We particularly focus on studies of category selectivity in high level vision as a benchmark to explain and test our framework. Accordingly, we consider a visual fMRI experiment that features a number of objects or object categories to a group of subjects. Based on the estimates of brain responses from the measured fMRI signals, we define *selectivity profiles* for different brain areas in each subject. We employ clustering to identify *functional systems* defined as collections of brain locations with similar selectivity profiles that appear consistently across subject. The methods developed here simultaneously consider all relevant brain responses to to the entire set of stimuli, and automatically learn the selectivity profiles of dominant systems from data. Crucially, our framework also avoids relying on anatomical information. Hence, they solve the most important limitations in prior studies of category selectivity discussed in the previous section.

[2] Note also that the spatial resolution of fMRI fundamentally limits the spatial extent of functional specificity that can be studied using this technique. The discussion here assumes that we seek likely patterns of specificity at a larger spatial scale than that of fMRI observations.

The methods are readily applicable to other studies of cognitive functional specificity beyond vision, as we discuss throughout the thesis.

Our basic mixture model describes the distribution of the fMRI data presented in the space of selectivity profiles. The model characterizes each functional system as a mixture component. We parametrize each component of the generative model with its center and weight, which represent the mean system selectivity profile and system size, respectively. The data is directly combined across subjects irrespective of the spatial information and fitted to the same distribution. We further provide a scheme for the statistical validation of the resulting systems based on their *functional consistency* across subjects.

The method allows us to automatically search within the set of all patterns of selectivity implicated by the experimental paradigm. Thus, the experimental design should aim to select a representative set of stimuli that provides as much richness as possible for the corresponding search. This thesis includes two experiments that aim to characterize category selective areas using this scheme. The first experiment features 2 image sets from each of 8 object categories where, in the large space of many likely patterns of selectivity, the method identifies face, scene, and body selectivity as the most consistent ones across subject. The second experiment pushes the limits of the analysis to the space of selectivities over images of distinct objects. In the result of the analysis from this data, each system selectivity profile effectively gives an account of the grouping (categorization) of objects within that system in the visual cortex. Interestingly, even in the space of 69 distinct objects presented in this experiment, the most consistent discovered selectivity profiles correspond to categories of faces, scenes, and bodies.

Our basic method includes two separate stages: we first estimate selectivity profiles from fMRI signals and then fit the model to the resulting data. Employing generative modeling for fMRI analysis enables systematic augmentation of the basic mixture model to unify the two stages and to refine the assumptions of the basic mixture model. Having established the merits of our basic framework, we further develop a nonparametric hierarchical Bayesian model for group fMRI data that integrates the mixture model prior over activations with a model for fMRI signals. The nonparametric aspect of the refined model allows the estimation of number of systems from the data. The hierarchical aspect of this model explicitly accounts for the variability in the sizes of systems among subjects. We derive an inference algorithm for the hierarchical Bayesian model and show the results of applying the algorithm to data from the visual fMRI study with 69 distinct stimuli. We show that the hierarchical model improves the accuracy of the basic model and the interpretability of the results while maintaining their favorable characteristics.

■ 1.3 Outline of the Thesis

In the next two chapters, we provide a background for the work. Chapter 2 gives a brief, general review of fMRI techniques and situates our proposed framework relative to prior work in fMRI data analysis. Chapter 3 reviews standard approaches to group analysis and explains the advantages of the choice made in this thesis for combining fMRI data across subjects.

Chapter 4 discusses the elements of our basic analysis including the basic mixture model and the consistency analysis. Chapter 5 presents the application of the basic method to data from our two visual fMRI experiments and compares them with the results of conventional confirmatory analyses.

Chapter 6 presents the nonparametric hierarchical Bayesian extension of the basic model and discusses the application of the corresponding inference algorithm to our visual fMRI study.

Finally, Chapter 7 discusses some avenues for the extension of this work in the future and concludes the thesis.

Chapter 2

Approaches to Inference from fMRI Data

FUNCTIONAL MRI yields an indirect measure of local aggregate neuronal activity based on local blood flow.[1] The close connections between the brain metabolism and its activity has been known for a long time (see, e.g. Fox and Raichle, 1986; Roy and Sherrington, 1890). Ogawa et al. (1990) were first to develop the blood-oxygenation-level dependent (BOLD) MR contrast, which was sensitive to hemodynamic variations in the Cerebral Blood Flow (CBF), and in particular to the level of deoxygenated hemoglobin in and around tissue. This development was soon followed by the application of the BOLD contrast in mapping brain activity during a task or stimulation (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992). Simultaneous electrophysiological recordings in monkeys have shown that BOLD signals correlate well with average, local measures of neuronal activity (Goense and Logothetis, 2008; Logothetis et al., 2001; Logothetis and Wandell, 2004). However, this correlation encompasses a complex relationship between the metabolism of neuronal processes (Magistretti et al., 1999) and the dynamics of the brain vascular system. The nature of this correlation is a subject of ongoing research (Attwell and Iadecola, 2002; Heeger and Ress, 2002; Raichle and Mintun, 2006; Schummers et al., 2008).

Throughout the past two decades, functional MRI has rapidly advanced to become one of the major tools available to neuroscientists (Logothetis, 2008). Unlike positron emission tomography (PET), previously used for functional brain imaging, the BOLD contrast does not require the injection of exogenous tracers and is therefore far more suitable for neuroscience research. Compared to PET and other techniques such as Electroencephalography (EEG) and Magnetoencephalography (MEG), fMRI provides a better spatial resolution while still allowing full-brain scans. At the same time, the reliance of fMRI on the coupling between neural mechanisms and the vascular phenomena creates significant drawbacks for interpretation of the data (Heeger and Ress, 2002; Logothetis, 2008). Moreover, the hemodynamics markedly limits the temporal

[1] Huettel et al. (2004) provides a detailed account of physics, physiology, and analysis of fMRI. Readers more interested in the underlying physics of MRI may refer to (Slichter, 1990). Raichle (1998) relates the early history of functional brain imaging, which might also be of interest to some.

resolution and reduces the signal to noise ratio (SNR).

We can classify the applications of fMRI into two broad categories. The first group, which mainly interests us in this thesis, involves experimental conditions associated with certain tasks or stimuli. The experimental setup for this group of studies can be devised using block or event-related design (Huettel et al., 2004). Both designs rely on averaging the BOLD signal evoked by a particular experimental condition across several trials in order to improve the signal-to-noise ratio (SNR). Block designs introduce long temporal windows presenting the same stimulus. Event-related designs rapidly present different stimuli within short periods of time, commonly in random order and with varied inter-stimulus time intervals. In addition, fMRI is extensively used in the studies of functional connectivity in the rest-state brain, which, as the name suggests, do not involve any experimental protocol (Biswal et al., 1995; Greicius et al., 2003).

An fMRI data set describes the BOLD signal in different locations in the brain volume at different acquisition times during the experiment. The size of *voxels*, i.e., units of space, defines the spatial resolution of the data, commonly around 1-5 millimeters in each dimension. The temporal sampling rate in fMRI techniques is characterized by Repetition Time (TR), that is, the time interval between two subsequent data acquisitions, which typically ranges between 1-3 seconds. The spatiotemporal properties of the data are mainly determined by the fMRI scanner and the acquisition technique used.

Once acquired, BOLD time courses are usually first corrected for the subjects' likely motion in the scanner. Still, the resulting data is an extremely noisy signature of the hemodynamic variations, and requires further processing in order to be informative about the brain function. In a typical protocol-based study, the analysis aims to employ our knowledge of the experimental paradigm, the spatiotemporal characteristics of fMRI, and the neuro-vascular coupling to make relevant inferences from the data. In parallel to the ongoing advances made in fMRI hardware and acquisition techniques, the methods of fMRI analysis have also continuously evolved, improving the quality of the inference or at times making new types of inference feasible (Logothetis, 2008).

In this chapter, we briefly review existing methods for analysis of task-based fMRI data. Here, we restrict the discussion to inference from observations in a single subject, and postpone the discussion of group inference to the next chapter.

■ 2.1 Standard Models of fMRI Data

Functional MRI data is a volumetric set of time courses; therefore, any method of analysis has to include both characterizations of space and time. In this section, we discuss different approaches to modeling of these two main aspects of the data for protocol-based fMRI studies.

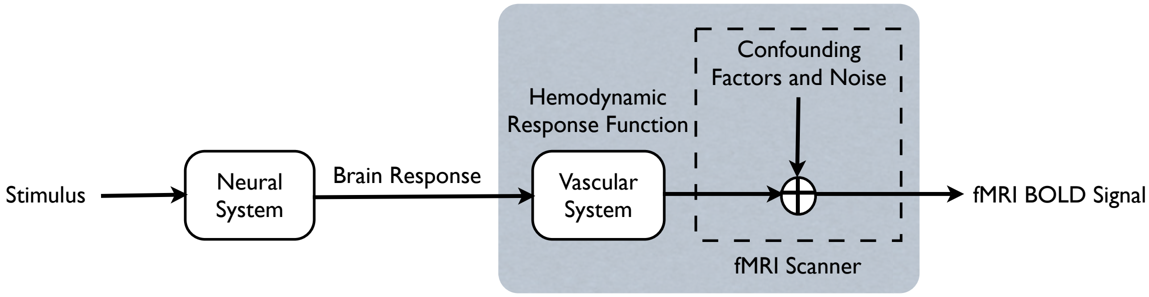


Figure 2.1. A schematic view of the relationship between the experimental protocol, the brain responses, and the BOLD measurments. The gray box indicates the parts of the process characterized by the linear time-invariant model of the response observed via fMRI.

■ 2.1.1 Temporal Models of fMRI Signals

Early fMRI experiments used basic block design paradigms, redolent of the then common PET designs, that included two different conditions such as on/off sensory stimulations or left/right motor tasks (Kwong et al., 1992). Since the blocks for each condition were long and the involved evoked hemodynamic responses relatively strong, simple subtraction of average voxel responses in one condition from that of the other yielded clear maps of areas involved in the cognitive process (Bandettini et al., 1993). Alternatively, thresholding the correlation coefficients between voxel time courses and the protocol mitigates high levels of noise in the data (Bandettini et al., 1993). As fMRI techniques found widespread applications in neuroscientific research and studies began to investigate more subtle dissociations, using, for instance, fast event-related designs (Dale, 1999; Dale and Buckner, 1997; Rosen et al., 1998; Zarahn et al., 1997a), the need for a more accurate temporal characterization of data became evident.

Linear Temporal Model of fMRI Response

Despite the complexity of the underlying physiology and physics of fMRI, Boynton et al. (1996) provided evidence that we can approximately model the relationship between behavioral stimulation and measured BOLD signals by a linear time-invariant (LTI) system. Such linear models, which also justify the correlation analyses, have proved exceptionally successful in practice (Buxton et al., 2004; Cohen, 1997; Friston et al., 1994). The linear temporal model assumes that the fMRI response results from additive contributions of stimuli,[2] confound factors, and noise, and that the fMRI response to a stimulus is fixed, irrespective of the stimulus onset and its duration. Figure 2.1 presents a schematic view of the model. We present a stimulus during the experiment and aim to find the brain response. Neural mechanisms commonly studied

[2] In this section, I interchangeably use the terms *stimulus* and *experimental condition*, by which I mean any sensory or cognitive stimulation or task presented during the experiment.

in fMRI experiments have a response time scale on the order of at most few hundred milliseconds, while typical TRs in fMRI are 1-3 seconds. We assume that the neural system, the first block in the figure, does not introduce any delay to the signal, relative to the temporal dynamics of the data. The brain response is then convolved with the hemodynamic response function (HRF), a function that peaks at about 6-9 seconds and models the intrinsic delay observed between the stimulus and the measured BOLD signal (Bandettini et al., 1993). Finally, the model assumes that a number of confounds, such as a commonly observed linear magnetic drift term (Friston et al., 1995c), is added to the signal prior to observation, along with Gaussian signal noise. The fMRI signal of any voxel during the experiment includes not only a delayed contribution from brain responses to all previous stimuli but also contributions from several such confound factors.

Let $i \in \mathcal{V}$ indicate a voxel within the set $\mathcal{V}$ of V voxels imaged in the fMRI study. In an experiment with a set $\mathcal{S}$ of S different experimental conditions, we let $\Omega_s(t)$ be a binary indicator function that shows whether stimulus $s \in \mathcal{S}$ is present during the experiment at time t. The model implies that the hemodynamic response of voxel i at time t is of the form:

$$b_{is}\ \Omega_s * \tau(t), \tag{2.1}$$

where b_{is} denotes the voxel's response to stimulus s , $\tau(t)$ is the hemodynamic response function, and $*$ indicates convolution. If we define $G_{st} = \Omega_s * \tau(t)$ to be the hemodynamic response to stimulus s at time t, we form the fMRI response y_{it} of voxel i as:

$$y_{it} = \sum_s G_{st} b_{is} + \sum_h F_{dt} e_{id} + \epsilon_{it}, \qquad 1 \leq t \leq T \tag{2.2}$$

where T is the number of points in the time course, F_{dt} represents nuisance factor d at time t, and ϵ_{it} is the Gaussian noise term. Stacking up the variables as vectors, we can rewrite (2.2) in the matrix form, as

$$\boldsymbol{y}_i = \mathbf{G}\boldsymbol{b}_i + \mathbf{F}\boldsymbol{e}_i + \boldsymbol{\epsilon}_i. \tag{2.3}$$

Vector $\boldsymbol{b}_i \in \mathbb{R}^S$ describes the response of location i in the brain to all different stimuli presented in the experiment. For now, we ignore the spatial dependencies among the variables, and consider equation (2.2) independently for each voxel. If we further assume that the Gaussian noise is white, i.e., $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \text{Normal}(\mathbf{0}, \lambda_i^{-1}\mathbf{I})$, we can estimate $\boldsymbol{b}_i$ using a simple least squares regression:[3]

$$[\hat{\boldsymbol{b}}_i^t\ \hat{\boldsymbol{e}}_i^t] = \boldsymbol{y}_i^t\,\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}, \tag{2.4}$$

where we have defined the *design matrix* $\mathbf{A} = [\mathbf{G}\ \mathbf{F}]$, and $\mathbf{A}^t$ denotes the transpose of matrix $\mathbf{A}$. The estimation procedure is the same for block design and event-related

[3] Of course, the derivation is possible only as long as matrix $\mathbf{A}$ remains nonsingular, a condition that generally holds in fMRI designs.

experiments since as far as the model is concerned all differences between protocols are contained within functions $I_s(t)$.

The above treatment assumes a known, a priori given HRF. Boynton et al. (1996) suggested a two parameter gamma function as an empirical approximation for the HRF, characterized by a time constant and a phase delay, and shifted by an overall temporal delay. Aguirre et al. (1998b) further demonstrated evidence for relative robustness of the estimated HRF across experimental sessions for the same subjects, but suggested that variability across subjects may be considerable. To address the variability concerns, several methods have been proposed for simultaneous estimation of the HRF and brain responses. One approach is to estimate the HRF as a linear combination of basis functions, e.g., Fourier basis functions (Josephs et al., 1997) or gamma functions and their derivatives (Friston et al., 1998a). More recent models employ Bayesian techniques for the estimation of HRF from data (Ciuciu et al., 2003; Genovese, 2000; Gössl et al., 2001b; Makni et al., 2005; Marrelec et al., 2003; Woolrich et al., 2004b).

Equation (2.4) makes a simplifying assumption that the temporal components of noise are independent of each other. In reality however, unwanted physiological and physical contributions in the signal exhibit non-zero temporal autocorrelations (Friston et al., 1994). Aguirre et al. (1997) and Zarahn et al. (1997b) empirically demonstrated that the power spectrum of the noise can be characterized by the inverse of frequency, and therefore does not comply with a white noise model. Bullmore et al. (1996) suggested using a first order autoregressive model AR(1) instead to directly model noise autocorrelations. Purdon and Weisskoff (1998) further modified this AR model by adding to it a white noise component.

Some methods handle the autocorrelations prior to the main linear modeling as a preprocessing step. For instance, *coloring* may be used by further temporal smoothing of the signal to the point that the original autocorrelations are overwhelmed and can be ignored (Friston et al., 1995b, 2000a; Worsley and Friston, 1995). Alternatively, *prewhitening* first estimates noise autocorrelations and uses the estimates to whiten the noise (Bullmore et al., 2001; Burock and Dale, 2000; Woolrich et al., 2001). Similar to the estimation of the HRF, Bayesian methods can also be used here to jointly estimate the brain response and noise characteristics (Friston et al., 2002a). Indeed, many recent analysis techniques take this approach in conjunction with AR models (see, e.g., Makni et al., 2008; Penny et al., 2005; Woolrich et al., 2004c).

Other Models

As already mentioned, the standard linear model of fMRI signal comes with well-known limitations. Deviations from linearity are observed especially when stimuli are not well separated (Glover, 1999). Friston et al. (1998b) suggested using a Volterra series expansion of a nonlinear response kernel to account for such nonlinearities in the response. Physiologically-informed models such as the so-called *balloon model* attempt to explain the nonlinear dynamics of BOLD signals using the specifics of the neuro-vascular coupling in terms of relevant vascular variables such as cerebral blood volume (CBV) and cerebral metabolic rate of oxygen (CMRO$_2$) (Buxton et al., 2004, 1998; Friston et al., 2000b).

■ 2.1.2 Spatial Models of fMRI Signals

Since the very early days, much of the excitement about fMRI has been commonly ascribed to its potential for *localization* of different functions in the brain (e.g., Cohen and Bookheimer, 1994). In the same way that structural MRI earlier provided us with *in vivo* maps of brain anatomy, fMRI now enables us to catch a glimpse of maps of brain function, or so the argument goes. The presence of the temporal dimension in fMRI allows us to change an experimental variable and observe how different brain locations respond to that change. If we observe a large response in any location (voxel), we declare that brain area active. In reality, the actual picture of functional organization in the brain may be more complex and involve interactions that evoke responses based on the context, prior stimulation, etc. Nevertheless, considering the rudimentary stages of our knowledge about the brain function, the localization picture can still help us gain important insights. The localization view has had great success in popularizing fMRI research through its renderings of intuitive maps of brain activation. The detection of active areas through statistical tests and their visualization as corresponding activation maps form the core of the traditional confirmatory (hypothesis-driven) analysis, which we will discuss in more detail in Section 2.2. The majority of fMRI analysis techniques have been developed with the aim of improving the quality of activation maps, as the end result of the analysis.

Smoothing of Spatial Maps

In its most basic format, the detection of activated areas is performed through a *mass univariate* analysis that tests the response of each voxel separately (Friston et al., 1994). In the temporal model discussed in Section 2.1.1, this basic approach corresponds to the assumption of independence among voxels. The analysis is simple, fast, and is widely used in practice. Yet if applied to raw time courses, the method usually creates grainy maps due to the excessive levels of noise. To further mitigate the noise and create maps that agree with our intuition about the contiguity of activated areas, it is common to apply spatial Gaussian smoothing to the data prior to analysis. Depending on the noise properties of the scanner and data, Gaussian filters with full width at half maximum (FWHM) of 3-9mm are typically used, the larger filters being more suited to data with lower SNR. Ideally, the filter size should closely match the size of activated regions we are seeking (Worsley et al., 1996a). Obviously, we do not always have good *a priori* estimates of the size of activations. As a result, although common in practice, the application of isotropic Gaussian filters may in fact cause blurring of the data and loss of fine spatial information. Alternative spatial denoising schemes include wavelet-based methods (Wink and Roerdink, 2004) or nonlinear smoothing that preserves edges (Smith and Brady, 1997). Other methods account for the structure of cortical surface in the smoothing of images (Andrade et al., 2001), or explicitly project the data onto the flattened map of the cortical surface before smoothing (Fischl et al., 1999).

Instead of smoothing the data as a preprocessing step, spatiotemporal analysis schemes explicitly encode spatial contiguity assumptions in their models of fMRI time courses. A simple approach is that of Katanoda et al. (2002) who include the neighboring time courses in the estimation of brain response of each voxel. In the system identification setting, Purdon et al. (2001) and Solo et al. (2001) estimate brain responses by adding a spatial regularization term to their model. From the Bayesian perspective, such a regularization corresponds to a prior distribution on brain responses that encourages contiguity (Gössl et al., 2001a). This approach is further extended to also include spatial priors for the parameters of the autoregressive noise model in (Penny et al., 2005), as well as the estimation of HRF in (Woolrich et al., 2004c). In earlier work, Descombes et al. (1998) suggested applying 4-dimensional continuous Markov random fields (MRF) in both space and time for a Bayesian restoration of brain responses.

Discrete Activation Models

Instead of smoothing the data or the brain responses, MRFs can be alternatively used in a discrete setting to encourage contiguity in the final maps of activated areas (Cosman et al., 2004; Kim et al., 2000; Ou and Golland, 2005; Ou et al., 2010; Svensén et al., 2002b). Conceptually, the method is also closely related to the mixture model detection (Everitt and Bullmore, 1999; Makni et al., 2005) in that both approaches explicitly define binary latent activation variables $\boldsymbol{x}$ to explain voxel brain responses. The distribution of observed fMRI signals $\boldsymbol{y}$ in each voxel (or some other statistics that is a function of $\boldsymbol{y}$) is described while conditioning on $\boldsymbol{x}$, which indicates whether or not the voxel is activated by a particular stimulus. By adding the discrete MRF as a prior on the distribution of activation variables over the image, we penalize neighboring voxels with mismatched activation labels in the resulting activation map (Cosman et al., 2004; Rajapakse and Piyaratna, 2001; Salli et al., 2002). Since the observed data does not get artificially blurred, the method may be more sensitive in detecting the edges of activated areas (Salli et al., 2001). As another example, Hartvig and Jensen (2000) employ a similar mixture model but describe the signal statistics in each voxel only in terms of its direct neighbors, independently of the rest of the image. Woolrich et al. (2005) add a third label representing de-activation, and further provide a method for estimating the discrete MRF parameter that controls the amount of spatial regularization.

Parcel-based Models of Spatial Maps

While encouraging contiguity, the methods described so far do not actually provide any model for the spatial configurations of activation areas. Simple modeling of such configurations can be achieved using extensions of mixture models. Penny and Friston (2003), for instance, assume that activated areas are composed of a number of *parcels*, each characterized by a mixture label and a corresponding Gaussian-shaped spatial blob. They parametrize the geometry of each blob with the center and spatial covariance of the Gaussian, and its function by the average and variance of the parameters of the linear response model (brain responses and the contribution of linear confounding factors) in that blob. The weight of each component in the response of any given voxel is proportional to the corresponding Gaussian function evaluated at that voxel. Hartvig (2002) defines a stochastic point process based on the same idea, although the inference for this model is sophisticated and prohibitively slow (for a very similar model based on mixtures of Gaussian experts, see Kim et al., 2006). Instead of relying on nonparametric models for finding the number of parcels, Thyreau et al. (2006) and Tucholka et al. (2008) use a cross-validation scheme in a volume-based and surface-based setting, respectively. Lukic et al. (2007) extend the framework to include more general non-Gaussian spatial kernels and provides a more efficient scheme for inference. Other closely related work include those of Friman et al. (2003) and Flandin and Penny (2007) who propose employing steerable filters and sparse wavelet priors, respectively, as spatial basis functions for response variables.

In the earlier days of fMRI, due to low spatial resolution of the acquired data and subsequent heavy spatial smoothing, activation maps indeed seemed a lot like Gaussian blobs. Nowadays, the spatial resolution of fMRI data is reaching the submillimeter range where sinuous warps of the cortical surface can be distinguished in the volume. As shape descriptors for activated areas, the only remaining appeal of Gaussian blob models may be in the algorithmic simplicity that they offer for inference. Instead of using shape descriptors, we can use a labeling map for activated areas where each label corresponds to a parcel of arbitrary shape. Voxels within each parcel may share specific properties in their fMRI responses, such as a distinct form of HRF (Makni et al., 2008; Svensén et al., 2002b). Vincent et al. (2010) recently suggested a Bayesian scheme for estimating these parcels from data where parcel contiguity is enforced through a discrete MRF, the parameter of which is also estimated from data. Bowman (2005) uses clustering of the data to create parcels that are then used for regional smoothing of brain responses.

■ 2.2 Confirmatory Analyses

As we mentioned in Section 2.1.2, the ultimate goal of functional brain imaging is traditionally considered to be the creation of activation maps, that is, the detection of voxels active to a given stimulus. Consider a simple on/off block design experimental paradigm where a stimulus is presented only in odd-numbered blocks. The *confirmatory*[4] approach treats the problem of analysis as an instance of classical statistical test-

[4] The approach is more commonly termed *hypothesis-driven* in the field, as opposed to a data-driven analysis. Here, following the terminology of (Tukey, 1977), we refer to it as confirmatory, to be contrasted from exploratory, simply because we find the more common terms slightly misleading. Any analysis,

ing. The experimental condition is viewed as an independent variable and the brain response in each voxel is treated as the dependent variable. We observe the response both during treatment (on blocks) and control (off blocks) and aim to confirm that the effect size of treatment relative to control is statistically significant.

Naturally, the null hypothesis assumes that the presence of stimulus does not alter the response. Under the null hypothesis, the probability that the effect size may be as large or larger than the observed value constitutes the statistical significance of the effect, the so-called p-value. The p-value indicates the probability that the null hypothesis may be rejected based on the observation. Thresholding the significance value at a desired level yields the detection results; thresholds of about $p = 10^{-4}$ are common in fMRI analysis. Piecing together the results of detection, or the significance values, across voxels in space provides us with the activation map for the stimulus of interest.

One of the main challenges for this simple framework stems from the fact that it treats the essentially multivariate dimension of space in a mass univariate fashion. Since the map is created from a large number of separate statistical tests, tens of thousands in a typical data set, the classical multiple comparison problem arises. If the noise properties of data exactly match our assumptions, the significance value corresponds to the probability of type-I error, i.e., the probability that a non-active voxel is declared otherwise. Among 50,000 samples from the null distribution, there are on average 5 voxels that pass the extremely conservative threshold of $p = 10^{-4}$. In other words, the probability that the collective result of our tests on such a sample is accurate will be minuscule: $(1 - 10^{-4})^{50000} \approx 0.007$. The two main approaches to confirmatory analysis, parametric and nonparametric tests, both provide techniques for dealing with the multiple comparison problem as we shall describe next.

■ 2.2.1 Statistical Parametric Mapping

Parametric tests explicitly formulate the null hypothesis for fMRI data. In the framework of the General Linear Models (GLM), Equation (2.3) can be used to derive a variety of different statistical tests for fMRI observations (Friston et al., 1995c). For instance, consider an experiment of visual object recognition where experimental conditions 1 and 2 correspond to presentations of face and object images, respectively. Say, we are interested in detecting brain areas that demonstrate significantly larger responses to face images compared to images of objects. We may form a linear *contrast* $\boldsymbol{c}^t\hat{\boldsymbol{b}}_i$ ($\boldsymbol{c} = [1, -1, 0, \ldots, 0]^t$) to compute the effect size in each voxel. Assuming white gaussian noise $\boldsymbol{\epsilon}_i$ in voxel i, under the null hypothesis that $\boldsymbol{c}^t\boldsymbol{b}_i = 0$, we can show

whether exploratory or confirmatory, is based on a hypothesis of some kind and also uses the data in some fashion to make the inference. The hypothesis-driven versus data-driven dichotomy creates the impression that the the latter avoids making any hypotheses about the data. Yet, no inference is possible unless we assume some structure in the data.

(Friston et al., 2007):

$$\frac{\boldsymbol{c}^t \hat{\boldsymbol{b}}_i}{\sqrt{\left(\frac{1}{T-D}\hat{\boldsymbol{\epsilon}}_i^t \hat{\boldsymbol{\epsilon}}_i\right) \tilde{\boldsymbol{c}}^t (\mathbf{A}^t \mathbf{A})^{-1} \tilde{\boldsymbol{c}}}} \sim \text{Student}(T-D), \tag{2.5}$$

where T is the number of points in each time course, $\hat{\boldsymbol{\epsilon}}_i$ is the residual, D is the overall number of conditions and confounds (columns of $\mathbf{A}$), Student$(T-D)$ is a Student's t-distribution with $T-D$ degrees of freedom, and $\tilde{\boldsymbol{c}} \in \mathbb{R}^D$ is a zero-padded extension of the contrast vector $\boldsymbol{c}$. This expression defines the t statistics and provides the grounds for the t-test, which is probably the most widely used test in fMRI analysis. We can now compute the significance of the face-object contrast in each voxel by computing the p-value for the observed summary statistics based on the null Students's t-distribution.

F-test is an alternative parametric test that has applications in the Analysis of Variance (ANOVA). This test aims to answer the question whether a number of experimental conditions significantly contribute to the fMRI signal in a given voxel. For example, consider a case where we are interested in detecting voxels where any of the experimental conditions elicit a significant brain response. In other words, we would like to ask whether the first term of the right hand side of Equation (2.3) is needed for explaining the observed variance in the signal at voxel i. The null distribution in this case corresponds to $\boldsymbol{b}_i = \mathbf{0}$. Let $\mathbf{P} = \mathbf{I}_{T\times T} - \mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t$ and $\mathbf{P}_0 = \mathbf{I}_{T\times T} - \mathbf{F}(\mathbf{F}^t\mathbf{F})^{-1}\mathbf{F}^t$ be projection matrices onto subspaces orthogonal to the spaces of full and reduced models, respectively. Once again, assuming white Guassian noise, we find (Friston et al., 2007):

$$\frac{T-D}{S}\frac{\boldsymbol{y}_i^t(\mathbf{P}_0 - \mathbf{P})\boldsymbol{y}_i}{\boldsymbol{y}_i^t \mathbf{P}_0 \boldsymbol{y}_i} \sim \mathrm{F}(S, T-D), \tag{2.6}$$

where S is the number of experimental conditions (regressors) and $\mathrm{F}(S, T-D)$ denotes an F-distribution with degrees of freedom S and $T-D$. The denominator of the F statistics is the residual of the entire model while its numerator is the difference between the residual of the reduced and full models. We can use the F-distribution to compute a p-value for the observed summary statistics in each voxel to test whether the variance in the response is due to any of the experiment conditions or not (omnibus F-test). More generally, an F-test can be used with any arbitrary partitioning $\mathbf{G}$ and $\mathbf{F}$ of the design matrix $\mathbf{A}$. Note that when an F-test is applied to only one regressor, it becomes equivalent to a t-test (Friston et al., 2007).

Due to the multiple comparison problem, thresholding raw significance maps computed from the above voxel-wise tests results in a large number of false positives. The Bonferroni correction is a simple way to handle this problem by multiplying the desired threshold on the false positive rate with the number of voxels tested. Since this correction may be overly conservative, a more moderate technique is to use the false discovery rate (FDR) control (Benjamini and Hochberg, 1995) for adjusting the threshold on the significance maps (Genovese et al., 2002). Worsley et al. (1996b) proposed an alternative method based on random field theory for thresholded smooth statistical maps. This theory may be further used to derive methods for inference on the number and size of connected components within the activation maps (see e.g., Chumbley and Friston, 2009). An empirical evaluation of some of the thresholding methods can be found in (Logan and Rowe, 2004).

All in all, statistical parametric tests are the most mature methods of analysis. Most spatial and temporal models discussed in Section 2.1 can be used to improve the quality of inference in this framework, including joint estimation of HRF and noise autocorrelations. The close connection between the classical and Bayesian inference (see Cosman et al., 2004) suggests that we can create posterior probability maps of activation and use them in place of significance maps (Friston et al., 2002b). Open availability of packages such as SPM[5], FSL[6], and FsFAST[7] has helped to some extent standardize the confirmatory analyses across different studies.

2.2.2 Nonparametric Methods

Nonparametric confirmatory methods avoid assuming a parametric null distribution for the data. Instead, they empirically derive the null distribution from the observed data (Hollander and Wolfe, 1999). Nonparametric permutation tests, which have recently become more popular in the neuroimaging community (Nichols and Holmes, 2002), assume certain symmetries for the null distribution and derive corresponding permutations of the data under which the statistics of interest remains invariant (Good and Good, 1994). For instance, if we consider the example faces vs. objects from Section 2.2.1, the null hypothesis of a corresponding permutation test will be: "the distribution of fMRI signal in voxel i is the same in the blocks that correspond to conditions 1 and 2." Under such a constraint on the null distribution, the distribution of the difference in the average response of the two conditions remains invariant to all permutations of the condition labels of blocks. Therefore, we can randomly permute the labels and generate samples from the null distribution based on the observed data. This distribution provides the significance measures for the statistics actually observed. Extensions have been proposed to this approach for the validation of the size of connected components (Hayasaka and Nichols, 2003, 2004).

2.3 Exploratory Analyses

Any functional MRI data set contains a wealth of information about brain activities in tens of thousands of brain locations during the experiment. A typical confirmatory analysis summarizes this entire data as a significance map. Disguised in this data, there might be much more information about the brain function, for instance, about local functional connectivity within and long-range functional connectivity between

[5] http://www.fil.ion.ucl.ac.uk/spm/

[6] http://www.fmrib.ox.ac.uk/fsl/

[7] http://surfer.nmr.mgh.harvard.edu/fswiki/FsFast

brain areas, how they change with different experimental conditions, and the impact of prior conditions on the responses to current ones. If we are to study all these phenomena, as well as the likely interactions among them, the space of effects we need to test becomes overwhelmingly large. The main motivation behind exploratory analyses is to uncover facts about brain function beyond what we can envision beforehand. Compared to confirmatory analyses, exploratory analyses treat the data more agnostically and look for interesting patterns in the data. Of course, what seems interesting to the method depends heavily on its underlying assumptions; so in this sense, compared to confirmatory analyses, there is more variety in the types of questions we can ask and patterns we may reveal using exploratory methods. On the other hand, the heterogeneity of methods makes it important to gain a deeper understanding of their workings in order to be able to correctly interpret the results.

From a machine learning viewpoint, fMRI exploratory techniques fit into the category of *unsupervised learning*, a general framework for providing compact summaries of data that capture major structures of interest. In this section, I discuss two broad classes of unsupervised techniques used in fMRI data analysis, namely, component analyses (principal and independent) and clustering. Other methods such as replicator dynamics (e.g., Neumann et al., 2006), canonical correlation analysis (e.g., Hardoon et al., 2007), and multidimensional scaling (e.g., Friston et al., 1996; Kriegeskorte et al., 2008) have also been applied to functional neuroimaging data. My focus on the two aforementioned groups of methods is due to the popularity of the first in the fMRI community and the applications of the second in the current thesis.

■ 2.3.1 Component Analyses

Principal component analysis (PCA) and independent component analysis (ICA) are two well-known unsupervised learning techniques that have found broad applications in fMRI data analysis. They can be both readily explained in a matrix factorization framework, which is a general template for viewing a wide array of learning schemes that approximate a data matrix as a product of two lower-ranked matrices. Let $\mathbf{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_V]$ be the matrix containing all measured voxel time courses of interest. Matrix factorization seeks a linear decomposition of the data:

$$\mathbf{Y} \approx \mathbf{MZ}, \tag{2.7}$$

where $\mathbf{M}$ and $\mathbf{Z}$ are $T \times K$ and $K \times V$ matrices, respectively, with $K \leq T, V$. Most extensions of ICA and PCA assume real-valued $\mathbf{M}$ and $\mathbf{Z}$ matrices but further constraints such as non-negativity (Lee and Seung, 1999) can also be applied. In general, different matrix factorization techniques are distinguished by their assumptions about matrices $\mathbf{M}$ and $\mathbf{Z}$ and the nature of the approximation.

In their application to fMRI analysis, PCA, ICA, and most of their variations assume a linear model:

$$\boldsymbol{y}_i = \mathbf{M}\boldsymbol{z}_i + \boldsymbol{\epsilon}_i, \qquad i \in \mathcal{V}, \tag{2.8}$$

where $\boldsymbol{z}_i \in \boldsymbol{R}^K$ and, similar to the linear model in Equation (2.3), different voxels' signals are assumed to be independent of each other. Comparing the role of the design matrix in Equation (2.3) with that of $\mathbf{M}$ in Equation (2.8) shows the difference between the confirmatory GLM and the exploratory analysis: while the former assumes known temporal factors, the latter treats them as model parameters and attempts to learn them from data. The result of the method can be correspondingly interpreted: column k of matrix $\mathbf{M}$, $\boldsymbol{m}_k$, corresponds to either an interesting time course or a temporal confound, while the corresponding latent variable z_{ik} represents the contribution of this factor to the fMRI signal observed in voxel i.

In its probabilistic formulation, PCA assumes independent unit-variance normal priors for latent variables $\boldsymbol{z}_i \sim \text{Normal}(\mathbf{0}, \mathbf{I})$ and white Gaussian noise (in both space and time) $\boldsymbol{\epsilon}_i \sim \text{Normal}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. PCA's main appeal may be that it has a simple linear algebraic solution. Assuming the data is zero-centered across space, the estimated column $\hat{\boldsymbol{m}}_k$ turns out to be a scaled version of the temporal component associated with the k-th largest singular value, in a singular value decomposition (SVD) of the data matrix $\mathbf{Y}$ (Tipping and Bishop, 1999). PCA was one of the very first exploratory methods investigated in fMRI analysis (e.g., Friston, 1994). Kernel PCA, which is a nonlinear extension of PCA, has been also applied to fMRI data (Thirion and Faugeras, 2003a).

PCA assumes independent Gaussian distributed latent variables. The multivariate Gaussian distribution is the only distribution where uncorrelatedness is both necessary and sufficient for independence. Obviously, assuming non-Gaussian distributions for latent variables will make the independence condition stronger. Basic ICA algorithms, sometimes also called blind separation or deconvolution, are based on this principle. They formulate and solve optimization problems that encourage the prior distribution to be both independent and non-Gaussian (Bell and Sejnowski, 1995; Comon, 1994; Hyvärinen and Oja, 2000). McKeown et al. (1998b) employed a basic noiseless ICA algorithm for the analysis of fMRI data, where $\boldsymbol{\epsilon}_i = \mathbf{0}$ in Equation (2.8), and demonstrated improved results compared to PCA (see also Biswal and Ulmer, 1999; McKeown and Sejnowski, 1998). This formulation of ICA for fMRI is sometimes referred to as spatial ICA (sICA) since, like in Equation (2.8), it considers the independence to be across space. Spatial ICA is more widely used than an alternative temporal formulation (tICA) due mainly to empirical reasons (Calhoun et al., 2001c).

ICA is used in conjunction with both studies of task related activations (e.g., Calhoun et al., 2001a; McKeown et al., 1998a) and studies of resting state (e.g., Greicius et al., 2003; van de Ven et al., 2004) and anesthetized (e.g., Kiviniemi et al., 2003) brain. Calhoun et al. (2001a) compared the results of ICA with those of GLM-based significance tests in a visual-perception task and showed that, although there are variations between the two, the methods provide complementary information about the processes of interest. A similar comparison in a clinical setting, a pre-surgical mapping study on patients with focal cerebral legions, also validated the results of ICA (Quigley et al., 2002). As examples of further applications, Duann et al. (2002) used protocol-related temporal factors given by ICA to provide estimates of the HRF, while Arfanakis et al. (2000) showed that removing such terms may allow estimation of intrinsic functional connectivity networks from data in a protocol-based experiment. ICA can be also used for removing unwanted temporal factors such as noise and motion counfounds (Beckmann et al., 2000; McKeown et al., 2003; Thomas et al., 2002).

In contrast to PCA that has a unique basic formulation, there are different ways to formulate ICA even in its basic noiseless format. Departing from its original motivation, ICA can also be described from a likelihood maximization viewpoint where varying the choice of non-Gaussian prior distribution $p(z_{ik})$ results in different optimization problems (Hyvärinen, 1999a,b; MacKay, 2003). Even for the same optimization problem, different algorithms may be derived; Esposito et al. (2002) compared two popular algorithms InfoMax (Bell and Sejnowski, 1995) and Fixed-Point (Hyvärinen, 1999a) and empirically illustrated the advantages of each in different settings. Beckmann and Smith (2004) further suggested a probabilistic formulation including Gaussian distributed noise ϵ_i. Bayesian extensions of ICA have been also derived (Chan et al., 2003; Højen-Sørensen et al., 2002b; Winther and Petersen, 2007) and applied to fMRI data with an extra assumption that latent variables are binary (Højen-Sørensen et al., 2002a).

■ 2.3.2 Clustering

Among unsupervised learning techniques, clustering may be the most intuitive and the closest to the notion of learning in our daily lives: it groups similar data points together to form data clusters. Central to the idea of clustering, of course, is the definition of a measure for pairwise similarity (or distance) of data points, simple examples of which include the Euclidean distance and the correlation coefficient. In application to fMRI, clustering has an obvious use in partitioning the set of voxels to clusters with similar time courses. A host of different clustering techniques exist that each formulate the problem in a different way (Jain, 2010). The K-means algorithm, perhaps the most basic clustering technique, iteratively assigns data points to one of K labels such that the sum of squared Euclidean distance between points and the mean of their respective cluster is minimized.

Early work in exploratory fMRI analysis focused on employing fuzzy clustering, a close relative of K-means that allows non-binary partial assignments, and comparing it with confirmatory correlation analysis in simple fMRI experimental paradigms activating early visual areas (see Baumgartner et al., 1997, 1998; Golay et al., 1998; Moser et al., 1997, and references therein). Baumgartner et al. (2000a) reported superior performance for this method compared to PCA in similar empirical settings, while Moser et al. (1999) suggested that it can be also used for removing motion confounds. Variants of fuzzy clustering (e.g., Chuang et al., 1999; Fadili et al., 2000; Jarmasz and Somorjai, 2003), K-means (e.g., Filzmoser et al., 1999), and other heuristic clustering techniques (e.g., Baune et al., 1999) have been applied to fMRI data, but little evidence exists for likely advantages of any of these methods beyond the experimental settings where they were first reported.

While K-means was not originally devised in a probabilistic modeling framework, we can in fact motivate it from that perspective. Consider Equation (2.8) with a white Gaussian noise term $\boldsymbol{\epsilon}_i$. Moreover, let us assume that $\boldsymbol{z}_i \in \{0,1\}^K$ corresponds to a binary indicator variable for the cluster assignment of voxel i such that:

$$z_{ik} = \begin{cases} 1, & \text{if voxel } i \text{ is assigned to cluster } k, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Maximum likelihood estimation of vector $\boldsymbol{m}_k$ in this case will give the mean time course of all voxels assigned to cluster k. Therefore, K-means attempts to find the assignment configuration $\mathbf{Z}$ that maximizes the joint likelihood of observed time-series and latent variables (cluster assigments). This view of clustering can be extended to include a variety of distance measures beyond the Euclidean distance (Banerjee et al., 2005). In this framework, we can further devise a soft-assignment algorithm, in the spirit of the fuzzy algorithms, by treating cluster assignments as random variables with a multinomial prior distribution (MacKay, 2003). In order to avoid introducing too many variable names, we will abuse the notation slightly and denote the cluster assignment of voxel i by $z_i \in \{1, \cdots, K\}$. This means that $\boldsymbol{z}_i$ and z_i convey exactly the same information, the latter being a binary vector with a 1 in its component z_i. Now, we assume a model of the form

$$z_i \sim \text{Mult}(\boldsymbol{\pi}), \tag{2.10}$$

$$\boldsymbol{y}_i | (z_i = k) \sim \text{Distr}(\boldsymbol{\theta}_k), \tag{2.11}$$

where $\boldsymbol{\pi} = [\pi_1, \cdots, \pi_K]^t$ is the weight parameter of the multinomial distribution, and conditioned on the assignment of voxel i to cluster k, the time course $\boldsymbol{y}_i$ has a distribution characterized by parameter vector $\boldsymbol{\theta}_k$. In a simple soft extension of K-means, the distribution $\text{Distr}(\theta_k)$ is the isotropic Gaussian and the parameter $\boldsymbol{\theta}_k$ contains the distribution mean $\boldsymbol{m}_k$ and variance ζ_k^{-1}. Note that marginalizing latent assignments yields a mixture model distribution for the time courses:

$$p(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{y}_i; \boldsymbol{\theta}_k), \tag{2.12}$$

where $f(\cdot)$ indicates the probability density function (pdf) for the distribution $\text{Distr}(\theta_k)$. This mixture model formulation of clustering has been employed in (Golland et al., 2007) for the analysis of resting state fMRI data, recovering a hierarchy of large-scale brain networks (Golland et al., 2008).

Clustering in the Feature Space

Applying clustering directly to the time course data, as described above, may not be the best strategy when it comes to discovering task-related patterns. First, the high dimensionality of fMRI time courses makes the problem hard, especially since in this case a large proportion of the variability in signal is due to noise. Second, the spatially varying properties of noise may increase the dissimilarity between the time courses of different activated areas. Third, in order to interpret the results, we need to determine the relevance of each estimated cluster mean time course to different experimental conditions. This usually involves a post hoc stage of regression or correlation that may become complicated as the number of experimental conditions and discovered components increase (see McKeown, 2000, for an example of handling the same issue in ICA). Therefore, ignoring the protocol-related information and applying clustering to time courses may be better suited to protocol-free studies of functional connectivity or experiments with simple setups.

Alternatively, some clustering methods use information from the experimental paradigm to define a measure of similarity between voxels, effectively projecting the original high-dimensional time courses onto a low dimensional feature space, and then perform clustering in the new space (Goutte et al., 2001, 1999; Thirion and Faugeras, 2003b, 2004). Compared with the original time courses, this feature space more readily represents the dimensions of interest in fMRI measurements. For instance, if the experiment involves a paradigm that is rich enough, we can simply cluster vectors of brain response $\boldsymbol{b}_i \in \mathbb{R}^S$.

Finally, we note that despite the sizable literature on clustering as an fMRI analysis method, there are few studies that have applied the technique in neuroscientific studies to provide insights about the brain organization (as an example of an exception, see e.g., Balslev et al., 2002).

2.3.3 Model Selection

One of the main challenges in unsupervised learning is the problem of model selection, which in the case of models discussed so far translates into the question of how to determine the number of components K. A wide variety of model selection techniques exist that generally aim to find the model size that yields the best generalization performance to cases beyond the data already observed (Friedman et al., 2001). While some methods empirically estimate this performance using resampling techniques (see Baumgartner et al., 2000b; Möller et al., 2002, for examples of applying these techniques to fMRI data), other approaches effectively include a prior on the number of components encouraging smaller model sizes (examples of application to fMRI data include Beckmann and Smith, 2004; Liu et al., 2010).

2.4 Multi-Voxel Pattern Analyses

Recently, a new approach, often referred to as multivariate pattern analysis (MVPA), has emerged that uses the framework of *supervised learning* to consider patterns of responses across voxels that carry information about different experimental conditions (Haxby et al., 2001). In MVPA, the response of each voxel is considered relevant to the experimental variables not only on its own, but also in conjunction with the responses of other spatial locations in the brain. Most MVPA methods train a classifier on a subset of fMRI images and use the classifier to predict the experimental conditions in the unseen subset. This approach has proved successful in a variety of applications (Norman et al., 2006; O'Toole et al., 2007), including decoding cognitive and mental states (Haynes and Rees, 2006; Mitchell et al., 2004), lie detection (Davatzikos et al., 2005), and low level vision (Haynes and Rees, 2005; Kamitani and Tong, 2005).

In the studies of visual object recognition, the localization approach had been used to identify category-selective functional regions, such as the fusiform face area (FFA) and the parahippocampal place area (PPA) in the ventral visual pathway (Epstein and Kanwisher, 1998; Kanwisher, 2003; Kanwisher et al., 1997). In contrast, applying MVPA in the same studies yields a distributed pattern in the visual cortex as an alternative to the localized representations implied by category-selective areas such as FFA and PPA (Carlson et al., 2003; Cox and Savoy, 2003; Haxby et al., 2001).

One of the major challenges of MVPA is that fMRI images contain a large number of uninformative, noisy voxels that carry no useful information about the category label. At the same time, voxels that do contain information are often strongly correlated. When trained with a relatively small number of examples, the resulting classifier is likely to capture irrelevant patterns and suffer from poor generalization performance. To mitigate the first problem, feature selection must be performed prior to, or in conjunction with, training (De Martino et al., 2008; Pereira et al., 2009).

2.5 Discussion

In light of the discussions of this chapter, we can now situate the technical approach taken in this thesis within the broad spectrum of different fMRI analysis methods. As we explained in Section 1.1.2, this thesis aims to provide a solution to the problem of how to search in the space of patterns of category selectivity. We explained that the traditional confirmatory approach to this search requires separately testing for each pattern of selectivity. The exponentially large size of the space of such patterns, in terms of the number of images or image categories considered, makes the confirmatory search infeasible in practice.

In addition, formulating category selectivity as linear contrast functions of brain responses is accompanied with an arbitrary choice of alternative category. For instance, the common definition of face selectivity contrast subtracts the response to objects from the response to faces in each voxel. The significance map corresponding to this contrast detects voxels where the face response is significantly larger than the *average* responses to other objects. However, this constraint is much looser than the more intuitive definition stating that face selective voxels respond at least twice as highly to faces as to the *maximum* response to all other stimuli (Op de Beeck et al., 2008). The latter definition cannot be formulated as a linear contrast of brain responses.

Since a brute force confirmatory search for category selectivity does not seem prac-

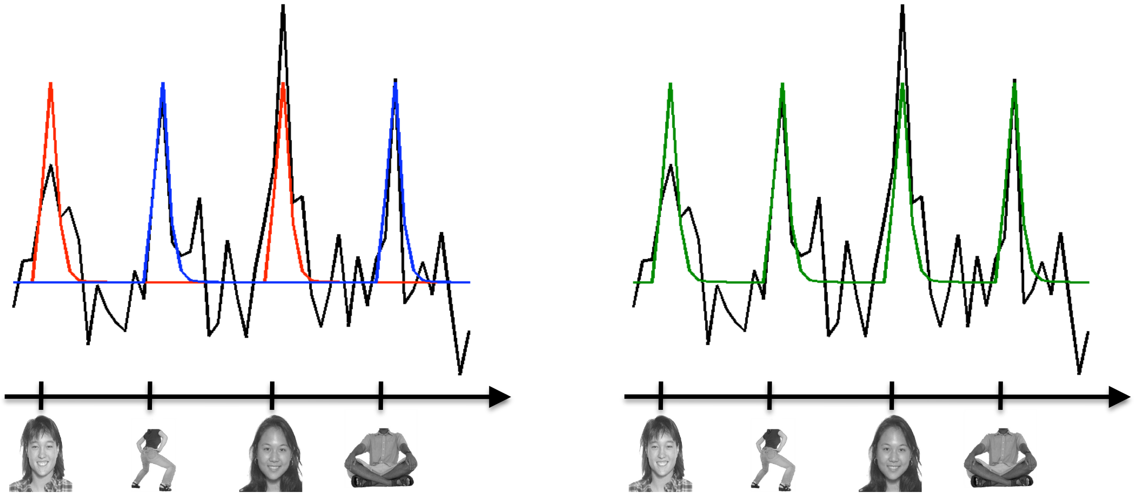


Figure 2.2. The distinction between the additive model of ICA and the clustering model. The black line indicates the hypothetical time course of a voxel; the axis shows likely onsets of a set of visual stimuli. Left, ICA may describe the black fMRI time course as a sum of the blue and red components, responsive to face and body images, respectively. However, since clustering does not allow summing different components, it may have to create a distinct component for describing the response of this voxel. We argue here that such a face-&-body selective pattern is indeed the description that we are seeking.

tical, we can employ an appropriate exploratory method to derive interesting candidates from the observed data. Among the exploratory techniques discussed in this chapter, clustering appears to capture the idea of specificity. In feature space clustering, we assign each voxel to a cluster-level vector of response in such a way that the collection of resulting clusters yields the best summary of the entire set of fMRI signals. If we construct feature vectors for fMRI responses such that they each represent a pattern of selectivity, the results can be readily interpreted as the dominant patterns suggested by the data. We will expand this idea further in Chapter 4.

Why should we prefer clustering over ICA, the other popular exploratory fMRI technique? As we saw in this chapter, ICA and clustering can be cast in terms of generative models with slightly different assumptions. In the matrix factorization setting of Equation (2.8), ICA and clustering are distinguished by their choices of latent vector $\mathbf{z}_i$. ICA assumes independent components z_{ik} and $z_{ik'}$ for all k and k' which essentially results in an additive model of the signal. Figure 2.2 illustrates a likely example of how ICA may explain the fMRI time course of a given voxel as a combination of two different components. Here, we assume that the data contains many other voxels with high selectivity for either faces or bodies. In this case, the response of the voxel will be best described by ICA as the sum of the face selective and the body-selective components. In the results, we will only see these two components, while they overlap in areas including the shown voxel. On the other hand, since the clustering model assigns the voxel to only one component, neither of the two exclusively face or body selective mean time courses can describe the signal in the voxel shown in the figure. We argue that if we are in search of category selectivity, and if there are enough voxels with responses similar to the one shown in Figure 2.2, we expect to see a component representing selectivity for the meta-category face-&-body in our results. We provide empirical comparisons of the results of the methods developed in this thesis and those of ICA, along with further discussion of the differences between the two methods, in later chapters.

This thesis chooses to formulate its unsupervised learning models in a probabilistic modeling framework. We already commented on some advantages of this framework in Section 1.2 (you can find a more detailed discussion in Appendix A). In short, this framework provides a flexible setting for a gradual refinement of the models throughout this thesis. Probabilistic modeling techniques are further equipped with a toolbox of different inference schemes that can be employed to fit the models of increasing complexity to the observed data. Applications of Bayesian nonparametric methods in this framework enable the integration of model selection with the inference in the main model, as we shall see in the model developed in Chapter 6.

Finally, we note that a number of other studies have used unsupervised learning methods, such as multidimensional scaling, to study the possibility of inferring object categorization of the visual cortex from data (Edelman et al., 1998; Hanson et al., 2004; Kriegeskorte et al., 2008).

Chapter 3

Function and Anatomy in Group fMRI Analysis

IN the discussion so far, we have talked about the human brain, its anatomy, and its functional organization as if it were an abstract template that contains the basic features of the brains of all of us. In reality, functional neuroimaging techniques acquire observations on the brains of specific human subjects. Different individuals exhibit considerable variability in virtually every measure of their brains: size, shape, cortical thickness, cortical folding patterns, and the like. It is also common to observe varying performance across individuals in response to a perceptual or cognitive task, which can be at least partially ascribed to the corresponding variability in their brain function. Furthermore, these anatomical and behavioral attributes vary in the same individual throughout her life. The fundamental problem of fMRI group analysis for cognitive neuroscience is how to use the data in a group study to make inferences about the brain; how to integrate data from a small group to discover principles governing the functional organization of all human brains.

Let us use a simple example to illustrate the difficulties of group analysis in fMRI (Brett et al., 2002). Consider a case where we aim to find areas in the brain activated by a certain stimulus or task in our fMRI study. We can employ one of the confirmatory analysis schemes discussed in Section 2.2 to detect activated voxels in each subject. The resulting maps represent brain locations in the space of acquired fMRI data, which usually has lower spatial resolution compared to structural MRI and comes with signal losses in certain parts of the brain (Devlin et al., 2000). To better characterize the location of the activated areas, we can establish correspondence between the collected functional data and the brain anatomy, captured through anatomical scans for each subject. For this purpose, we can use registration, a technique for aligning images, to find the correspondences between voxels in the functional image and voxels in the anatomical one (e.g., Greve and Fischl, 2009). Having registered functional MR images with the anatomical scans, we overlay the activation maps on the high resolution structural images and identify the areas responsive to the stimulus.

To further establish correspondences among activation maps in different subjects, the standard approach is to transform the data into a common anatomical coordinate frame. Talairach and Tournoux (1988) created a stereotactic atlas that defines an anatomical coordinate system with respect to a standard brain. The Montreal Neurological Institute (MNI) further developed an anatomical template brain by averaging structural images from a large population of subjects after aligning them with the Talairach brain (Evans et al., 1993). We can once again employ registration to align the structural image of each subject with this average template (Friston et al., 1995a; Gee et al., 1997). Applying the resulting transformation to the activation maps, we directly compare activation maps of different subjects within the same anatomical coordinate system. Consequently, we can describe the location of activations in the standard Talairach coordinates (Chau and McIntosh, 2005). The process of transforming functional data in a group of subjects to a common anatomical template is usually referred to as *spatial normalization*.

While the MNI and Talairach anatomical templates are defined in a 3-dimensional volume, most cognitive neuroscientists aim to study processes that happen in the cortex, which may be better represented by a 2-dimensional folded surface. Several methods have been demonstrated to extract the structure of the cortical sheet in each subject, characterize activation maps on the resulting surface, and register them with a population template for the cortical surface (Clouchoux et al., 2010; Fischl et al., 1999; Goebel et al., 2006; Joshi et al., 2007, 1997; Van Essen, 2005).

In practice, the resulting spatially normalized activation maps often look different from one subject to another (see, e.g., the figures demonstrating activation maps in different subjects in Fernandez et al., 2003; Spiridon et al., 2006; Wei et al., 2004). As it turns out, the brain anatomy is highly variable across subjects, in terms of macroscopic features such as cortical folding patterns of gyri and sulci (Ono et al., 1990), microscopic cytoarchitectonic characteristics (Amunts et al., 2000, 1999), and the relationship between these macroscopic and microscopic attributes (Fischl et al., 2008). In the face of this anatomical variability, and the resulting challenges for inter-subject registration techniques, the question remains as to how we should model and interpret the variability in functional areas.

Any group analysis inevitably makes assumptions about the relationship between function and anatomy independently of its choice of the inference style. Therefore, we will discuss the literature on the analysis of group fMRI data emphasizing for each method its underlying assumptions about the integration of information across the group.

■ 3.1 Voxel-wise Correspondences

The standard approach to making group inferences from fMRI data aims to provide an account of data in the anatomical template (atlas). To this end, it fully relies on the voxel-wise correspondences estimated by the spatial normalization procedure. This approach assumes that after spatial normalization, subject data for a given voxel in the atlas provide noisy instances of the response of the *same* location in the brain across subjects. Therefore, any variability in the response across subjects is attributed to noise. This strong assumption on the relationship between function and anatomy provides a means for creating intuitive and easy-to-interpret summary maps in group studies. The method of spatial normalization is therefore widely used in the fMRI group analysis.

The simplest scheme for testing a particular effect at a voxel in the template anatomical atlas is to average the corresponding spatially normalized contrast maps across subjects (Friston et al., 2007). This basic idea is at the core of the fixed-effects analysis and its generalizations (Friston et al., 1999). We can also treat the model discussed in Section 2.2.1 as the first layer in a multilevel GLM for fMRI signals that explicitly accounts for inter-subject variability. Ignoring the confounds for the sake of brevity, we assume

$$\boldsymbol{y}_{ji} = \mathbf{G}\boldsymbol{b}_{ji} + \boldsymbol{\epsilon}_{ji}, \tag{3.1}$$

where $\boldsymbol{y}_{ji}$ and $\boldsymbol{b}_{ji}$ indicate the fMRI time-course and the brain response of voxel i in subject j. In the second level, we consider

$$\boldsymbol{b}_{ji} = \bar{\boldsymbol{b}}_i + \boldsymbol{e}_{ji}, \tag{3.2}$$

where $\bar{\boldsymbol{b}}_i$ is the mean population-level brain response of voxel i and $\boldsymbol{e}_{ji}$ represents Gaussian noise capturing intersubject variability. We can now perform a test on the significance of the effect as a contrast function of the the population-level average brain responses $\bar{\boldsymbol{b}}_i$, using random-effects analysis (RFX) (Friston et al., 2002b; Penny and Holmes, 2003; Worsley et al., 2002), or mixed-effects analysis (Beckmann et al., 2003; Woolrich et al., 2004a).

Following the standard confirmatory techniques, most extensions of exploratory methods to multisubject data also rely on voxel-wise correspondences. For instance, Calhoun et al. (2001b) suggested concatenating the fMRI data along the temporal dimension after spatial normalization. For each component, this approach finds a spatial map that is identical across subjects, although its temporal response may differ in each individual. To allow variability across subjects, Beckmann and Smith (2005) proposed a tensorial factorization of the data within the ICA framework such that

$$\boldsymbol{y}_{ji} = \mathbf{M}\,\mathrm{diag}(\boldsymbol{c}_j)\,\boldsymbol{z}_i + \boldsymbol{\epsilon}_{ji}, \tag{3.3}$$

where $\boldsymbol{c}_j \in \mathbb{R}^K$ is a subject-specific vector that describes a contribution of each component to the data in subject j. In this method, both component maps and time-courses are common to all subjects, but they contribute differently to the measurements in different individuals.

The two main premises of group analyses that rely on voxel-wise correspondences are that 1) spatial normalization yields a perfect anatomical alignment across subjects to the level of voxels, and 2) there is no systematic variations in the functional characteristics of the same location in the stereotactic atlas across subjects. Unfortunately, there is ample evidence that contradicts both of these assumptions. First, due to the high variability in anatomical images acquired across subjects, registration errors on the order of a millimeter are common in the results of the best algorithms currently in use (Hellier et al., 2003; Klein et al., 2009, 2010). Second, when functional and cytoarchitectonically defined areas are identified on an individual basis using methods that do not rely on registration, they show substantial variations in their location in the stereotactic space (Amunts et al., 2000, 1999; Wohlschläger et al., 2005). To mitigate the impact of misalignments, group analyses that use voxel-wise correspondence usually perform heavy smoothing of data in the preprocessing stage (Gaussian kernels of FWHM of 4-9 millimeters are common). Such blurring causes a loss of spatial resolution and may suppress smaller activated areas that do not match perfectly across subjects (Brett et al., 2002; Saxe et al., 2006; Thirion et al., 2007a).

■ 3.2 Parcel-wise Correspondences

With current imaging techniques, inter-subject variability in anatomy poses major restrictions on the level of spatial details we can extract in our inference from group fMRI data. However, applying Gaussian smoothing of the data, commonly used in group analyses based on voxel-wise correspondences, ignores the available information about the nature of the variability. For instance, we know that individually-detected, thresholded activation maps usually consist of blob-like regions that vary across subjects in the location of their centers, or activation peaks. An alternative approach to making group inference is to focus the analysis to contiguous groups of voxels, or parcels, and then establish correspondences among these parcels across subjects. A wide variety of group analysis methods can be classified as parcel-based. The methods vary in the way they define parcels based on anatomical, functional, or both anatomical and functional characterizations.

As Nieto-Castanon et al. (2003) have suggested, a straightforward way to perform group analysis is to define regions of interests (ROI) based on anatomical markers prior to the analysis. These regions of interest can be identified partly using anatomical parcellation techniques (Fischl et al., 2004; Tzourio-Mazoyer et al., 2002). Bowman et al. (2008) has further provided a hierarchical model for the measured fMRI signal across a group that includes a parcel-level description. In this approach, the anatomical labeling, by construction, provides inter-subject correspondence among the parcels. The method therefore assumes that functional areas are aligned with anatomical parcellation a priori defined. The problem with this approach is that it still relies fully on anatomy in its definition of parcels. Other methods take functional data into account in their definition of parcels.

■ 3.2.1 Functional ROIs

Regions of interest for fMRI analysis can be identified using functional, instead of anatomical, landmarks. The functional ROI (fROI) approach employs independent experimental paradigms, called functional localizers, to detect areas that show a specific pattern in their functional response. Regions of interest are identified within the resulting thresholded map as parcels that appear on specific sulci or gyri. Once parcels are identified in each subject, the group analysis in the main experiment is confined to the parcels, assuming that they correspond to the *same* functional location across subjects. Parcels defined by this method are characterized both by their functional and anatomical properties. The method is flexible enough to allow variability in location of parcels as long as they maintain a basic relationship with anatomical landmarks (Saxe et al., 2006).

As an example, suppose we are interested in studying face processing in the visual cortex. The relevant functional landmark for an fROI analysis will then be the face fusiform area (FFA) that can be robustly detected using a faces-objects contrast (Kanwisher et al., 1997). We use this contrast to detect the FFA in a functional localizer experiment separately for each subject. We can now find the population-level effect of interest in an independent experiment by averaging the corresponding contrast within FFA across different subjects.

In traditional fROI analysis, the delineation of parcels in the localizer contrasts is performed manually. Recently, Fedorenko et al. (2010) have proposed an automatic scheme that uses an overlap between the localizer maps with a segmented probability map of detected areas averaged across all subjects to create subject-specific parcels and to establish correspondences among them.

■ 3.2.2 Structural Analysis

Structural analysis first identifies the parcels from a summary statistics or significance map in each subject and then find the correspondence among them (Thirion et al., 2007b). As a method for group fMRI analysis, structural analysis is similar in spirit to (Fedorenko et al., 2010) but has a broader range of applicability, since it is not specifically intended for well-characterized localizer contrasts.

In the first rendition of structural analysis, Coulon et al. (2000) used a multiscale image description to identify parcels of different scale in images. They created a population-level graph where each node corresponded to a parcel in a subject-specific image and edges expressed similarity of shapes and locations of parcels. The detection of activated areas and the correspondences among them was found by solving a graph labeling problem on this graph. Operto et al. (2008) further demonstrated this method on surface-based cortical maps instead of the 3-dimensional volume.

Instead of a multiscale description, Thirion et al. (2006) used an alternative method for identifying parcels in summary statistic images; they used spectral techniques for clustering voxels into contiguous parcels based on the similarity of their functional responses. Thirion et al. (2007c) and Thirion et al. (2007b) employed a nonparametric mixture model and a watershed algorithm, respectively, for parcel identification, and found the inter-subject correspondence using a belief-propagation algorithm on the graph of the resulting parcels, in a similar fashion to (Coulon et al., 2000).

Thirion et al. (2007a) and Pinel et al. (2007) compared the performance of structural and voxel-based group analyses in a data set including 81 subjects, and demonstrated empirically that the structural analysis is more suitable for capturing the variability in the spatial extent of activations.

3.2.3 Joint Generative Models for Group Data

The goal of group analysis is to characterize brain areas involved in a given task in the population. Group contrast maps created by voxel-wise methods provide an easy-to-interpret representation for the results of their analysis but fail to account for inter-subject variability, as was discussed before. Structural analyses, on the other hand, find correspondences among the detected parcels directly in different subjects, and do not explicitly construct a group template. The derivation of the group template is therefore added as a heuristic post-processing step (Thirion et al., 2007b). While the segmented probability map of activations in the scheme of (Fedorenko et al., 2010) provides a group summary of the results, it fails to explicitly describe the patterns of spatial variability of parcels across subjects.

Generative models of group fMRI summary statistic maps characterize the shape and location of parcels, as well as the variability in those parameters across the group. Like spatial mixture models discussed in Section 2.1.2, these generative models describe activation maps as a mixture of different spatial components, each representing a parcel in the group template. The map in each subject is generated as an independent sample from the distribution characterized by the group template. This approach also unifies the two separate stages of the structural analysis, namely identification of parcels and establishing correspondence among them.

Kim and Smyth (2007) and Kim et al. (2010) proposed a joint generative model for group fMRI data based on hierarchical Dirichlet processes (HDP). Xu et al. (2009) instead used a spatial point process but included more layers in their model to account for complex shapes of parcels. Despite the elegant mathematical foundations and the rigor of these models, making inference in these models is prohibitively slow due to the sampling approach in the inference procedure.

3.3 Other Approaches

All previous methods treat the registration of structural images to the atlas template as a preprocessing stage. Some strategies, however, integrate registration with the functional analysis while explicitly accounting for the inter-subject variability. Keller et al. (2008) explicitly include a jitter term in their registration framework that accounts for the uncertainty in the spatial normalization. Sabuncu et al. (2010) propose a *functional registration* that attempts to achieve a further alignment of the spatially normalized images with the template based on their functional responses. Finally, Yeo et al. (2010) integrate the two spatial and functional registration stages into one, where they learn the parameters of the registration cost-function based on the resulting alignment in functional responses in a training set.

A rather distinct approach to group analysis is that of Yang et al. (2004), who jointly model the variability of the summary statistic images and shapes of an anatomically defined parcel (ROI) across subjects. Although the idea seems rather promising, little empirical evidence has been provided by the authors, especially regarding the comparison between the method and other alternatives.

■ 3.4 Discussion

In their paper proposing an anatomical ROI approach to group analysis, Nieto-Castanon et al. (2003) mention:

> "Although it would be optimal to define these ROIs on the functional images, the amount of anatomical information resolvable in these images is judged to be too low for replicable ROI definition."

The approach we take in this thesis with respect to group fMRI analysis relies on the fact that such function-only derivation of parcels may indeed be possible, at least for the purpose of our exploratory search for functional specificity. Models introduced in Chapters 4 and 6 combine data from different subjects without incorporating any spatial information. Each voxel is merely represented by the profile of its functional response in the space of several different stimuli, and voxels with similar responses are clustered together across subjects. It turns out that in application to high-level object recognition in the cortex, the functional information is sufficient to recover spatial maps of several brain systems that also show clear correspondence when normalized to a common template.

We chose this simple approach for the parsimony of its underlying assumptions about the nature of the relationship between function and anatomy, and the resulting flexibility in the patterns the method can identify. Consider, for instance, a spatially distributed network of tiny brain areas scattered across the cortex in a seemingly random fashion. Although standard methods of analysis would decisively reject such a possibility based on their assumptions about contiguity and extent of activated areas, this hypothetical picture fits with the implicit assumptions of multi-voxel pattern classification approaches (Haxby et al., 2001). By being essentially blind to the spatial configuration of voxels within the network, the methods developed in this thesis are as eager to find such scattered systems, as they are to find contiguous activations of the same overall size that are in perfect anatomical alignment across subjects. In our search for functional specificity in brain, it is important to allow for the possibility of such scattered functional systems, as well as other likely scenarios where systems may be in loose spatial correspondence across subjects. The lack of spatial modeling in the models advocated by this thesis extends the space of likely discoveries that can be made based on the results of the analysis.

Nevertheless, we refrain from referring to a function-only approach to group analysis as the "optimal" scheme for group analysis. Rather, this thesis chooses the approach simply because of its exploratory advantages. Once we validate the assumptions of the standard methods about contiguity or spatial correspondence of functional systems, as the results of our method in Chapter 5 suggest, then employing the corresponding constraints as priors in our model will inevitably improve the sensitivity and robustness or the results. The literature reviewed in this section provide invaluable insights about how to approach building such models.

Lastly, we note that a function-only approach to group analysis is certainly not new. Svensén et al. (2002a) have employed a similar method for group analysis of data using ICA, where they concatenate the signals from different subjects across space. In comparison to alternative group analyses such as (Calhoun et al., 2001a,b), Schmithorst and Holland (2004) reports superior performance in group analysis for this scheme, despite the fact that it does not incorporate any spatial information. Langs et al. (2010) have recently proposed a similar framework for fusing data across subjects solely based on a low-dimensional spectral embedding of voxel time courses. Retinotopic mapping, which is a common technique in visual fMRI, further illustrates the fact that functional responses are sufficient for recovering fairly detailed spatial information. This technique uses fMRI signals elicited by a set of carefully designed low-level stimuli in order to characterize early visual areas such as V1 and V2 (DeYoe et al., 1996; Engel et al., 1997). Although the analysis is performed separately in each subject, once these areas are detected in different subjects, they are by definition assumed to be in functional correspondence. We conclude by emphasizing that the novelty of this thesis is in applying this function-only approach to group analysis for the purpose of an exploratory discovery of functional specificity patterns in the brain.

Chapter 4

Mixtures of Functional Systems in Group fMRI Data

IN this chapter, we describe the main building blocks of the proposed approach for discovering patterns of functional specificity from group fMRI data. First, we introduce the space of selectivity profiles and explain how this representation enables an automatic search for functional specificity. Second, we formulate a simple mixture model for the analysis of the data represented in the space of selectivity profiles, and derive an algorithm for estimating the model parameters from fMRI data. Third, we present a method for validating the results based on defining across-subject and within-subject consistency measures. In the last section, we discuss some further points about the application of the method as well as a number of extensions to the basic construction.

■ 4.1 Space of Selectivity Profiles

The goal of this thesis is to employ clustering ideas to automatically search for distinct forms of functional specificity in the data. Consider a study of high level object recognition in visual cortex where a number of different categories of images have been presented to subjects. Within a clustering framework, each voxel in the image can be represented by a vector that expresses how selectively it responds to different categories presented in the experiment. We may estimate the brain responses for each of the stimuli using the linear response model of Section 2.1.1 and perform a clustering on the resulting response vectors $\boldsymbol{b}_i$. However, the results of such an analysis may yield clusters of voxels with responses that only differ in their overall magnitude (as one can observe, e.g., in the results of Thirion and Faugeras, 2004). The vector of brain responses, therefore, does not directly express how *selectively* a given voxel responds to different stimuli.

Unfortunately, fMRI signals do not come in well-defined units of scale, making it hard to literally interpret the measured values. Univariate confirmatory methods avoid dealing with this issue by only assessing voxel contrasts, differences in signal evaluated separately in each voxel. Others instead express the values in terms of the percent changes in signal compared to some baseline, but then there is no consensus as how to define such a baseline (Thirion et al., 2007a). There is evidence that not only the characteristics of the linear BOLD response vary spatially within the brain (e.g., Makni et al., 2008; Miezin et al., 2000; Schacter et al., 1997), but even the neurovascular coupling itself may also change from an area to another (Ances et al., 2008). A wide array of factors can contribute to this within-subject, within-session variability in fMRI measurements, from the specifics of scanners to the local tissue properties and relative distances to major vessels. As might be expected, similar factors also contribute to within-subject, across-session, as well as across-subject variations in fMRI signals, although the latter has a more considerable extent due likely to inter-subject variability in brain function (Smith et al., 2005; Wei et al., 2004).

Given the reasoning above, prior to clustering, we aim to transform the brain responses into a space where they directly express their relative selectivity to different stimuli. Such a space allows us to compare voxel responses from different areas, and even from different subjects. Recall that, as discussed in Section 3.4, we further aim to avoid incorporating any spatial information into the analysis in the hope of validating current hypotheses about spatial properties of the functional systems based on our results. The construction of the space of selectivity profiles provides the common space for bringing together data from different subjects in a group study. This common space is purely functional, as contrasted with the anatomical common space used in group analyses based on spatial alignment and assumed voxel-wise correspondence.

We will describe the construction of the space of selectivity profiles using the example of a high level vision experiment. We define the *voxel selectivity profile*, as a vector containing the estimated brain responses to different experimental conditions and normalized to unit magnitude, that is,

$$\boldsymbol{x}_i = \frac{\hat{\boldsymbol{b}}_i}{\|\hat{\boldsymbol{b}}_i\|}, \tag{4.1}$$

where $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$ with $\langle \cdot, \cdot \rangle$ denoting the inner product. Selectivity profiles lie on a hyper-sphere S^{S-1} and imply a pattern of selectivity to the S experimental conditions defined by a direction in the corresponding S-dimensional space. Normalization removes the contribution of overall magnitude of response and presents the estimated response as a ratio with respect to this overall response. Furthermore, this removes the magnitude of overall BOLD response of the voxel, which is in part a byproduct of irrelevant variables such as distance from major vessels or general response to the type of stimuli used in the experiment. This provides another justification for the normalization of response vectors in addition to our interest in representing selectivity as a relative measure of response.

Figure 4.1(A) illustrates the population of unnormalized estimated vectors of brain response $\hat{\boldsymbol{b}}_i$ for all the voxels identified as selective for one of the three different conditions by a conventional t test. The data is from an experiment with a paradigm containing 8 different categories (for details see Section 5.1.1), but for visualization purposes we will only focus on three categories of faces, scenes, and human bodies. The differ-

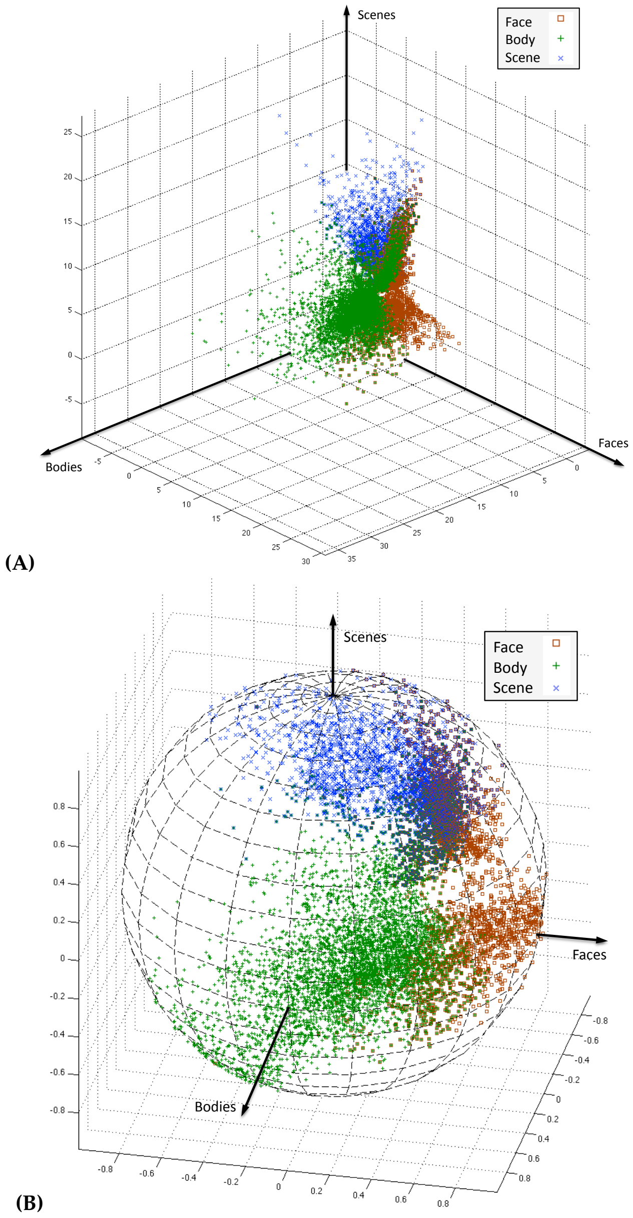


Figure 4.1. An example of voxel selectivity profiles in the context of a study of visual category selectivity. The block design experiment included several categories of visual stimuli such as faces, bodies, scenes, and objects, defined as different experimental conditions. (A) Vectors of estimated brain responses $\hat{\boldsymbol{b}} = [b_{\text{Faces}}, b_{\text{Bodies}}, b_{\text{Scenes}}]^t$ for the voxels detected as selective to bodies, faces, and scenes in one subject. As is common in the field, the conventional method detects these voxels by performing significance tests comparing voxel's response to the category of interest and its response to objects. (B) The corresponding selectivity profiles $\boldsymbol{x}$ formed for the same group of voxels.

ences between voxels with different types of selectivity are not very well expressed in this representation; there is no apparent separation between different groups of voxels. We also note that there is an evident overlap between the sets of voxels assigned to these different patterns of selectivity. The standard analysis in this case uses a contrast comparing each of the three conditions of interest with a fourth experimental condition (images of objects); therefore, it is possible for a voxel to appear selective for all three of these contrasts. In order to explain the selectivity of such a voxel, we can define a novel type of selectivity towards a meta-category which is composed of the three categories represented by these contrasts. The same argument can be applied to any combinations of the categories presented in the experiment to form various, new candidates as possible types of selectivity.

Figure 4.1(B) shows the selectivity profiles $\boldsymbol{x}_i$ formed for the same data set. We observe that the voxels associated with different types of activation become more separated, exhibiting an arrangement that is similar to a clustering structure. Furthermore, it is easy to see that the set of voxels shared among all three patterns of selectivity has a distinct structure of its own, mainly concentrated around a direction close to $[1\ 1\ 1]^t/\sqrt{3}$ on the sphere. We interpret the center of a cluster of selectivity profiles as a representative for the type of selectivity shared among the neighboring profiles on the sphere. Although the profile clusters are not well separated, the arrangement of concentrations of profiles on the sphere can carry important information about the types of selectivity more represented in the data. This information becomes more interesting as the number of dimensions (experimental conditions) grows and the overall density of profiles on the sphere decreases. This motivates us to consider application of mixture-model density estimation, the probabilistic modeling formulation of clustering (McLachlan and Peel, 2000), to the set of selectivity profiles. Each component in the mixture-model represents a cluster of voxels, i.e., a *functional system*, concentrated around a central direction on the sphere. The corresponding cluster center, which we will call *system selectivity profile*, specifies that system's type of selectivity. we use the term system, instead of cluster, mainly to distinguish these functionally defined clusters from the traditional cluster analysis used for the correction of significance maps (Forman et al., 1995).

Now consider a group study where a group of subjects take part in the same fMRI experiment. We denote the selectivity profile of a voxel in a study with J subjects by $\boldsymbol{x}_{ji}$, where $j \in \{1, \cdots, J\}$ is the subject index, and i is the voxel index as before. We aim to discover the brain systems with distinct profiles of selectivity that are shared among all subjects. Let us assume that two selectivity profiles $\boldsymbol{x}_{ji}$ and $\boldsymbol{x}_{j'i'}$, corresponding to voxel i of subject j and voxel i' of subject j', belong to the same selective system in the two brains. The overall magnitude of response of these two voxels can be different but the two profile vectors have to still reflect the corresponding type of selectivity. Therefore, they should resemble each other as well as the selectivity profile of the corresponding system. This suggests that we can fuse data from different subjects and cluster them all together in order to improve the estimates of system selectivity profiles. This approach can be thought of as a simple model that ignores possible small variability in subject-specific selectivity profiles of the same system, similar to the way that fixed effect analysis simplifies the more elaborate hierarchical model of random effect analysis in the hypothesis-driven framework. At this stage, we choose to work with this simpler model and defer the development of a corresponding hierarchical model to Chapter 6.

4.2 Basic Generative Model and Inference Scheme

4.2.1 Model

Let $\{\boldsymbol{x}_i\}_{i=1}^V$ be a set of selectivity profiles of V brain voxels. Like the mixture model introduced in Section 2.3.2, we assume that data vectors are generated *i.i.d.* by a mixture distribution

$$p(\boldsymbol{x}; \{w_k, \boldsymbol{m}_k\}_{k=1}^K, \zeta) = \sum_{k=1}^K w_k f(\boldsymbol{x}; \boldsymbol{m}_k, \zeta), \tag{4.2}$$

where $\{w_k\}_{k=1}^K$ are the weights of K components and $f(\cdot, \boldsymbol{m}, \zeta)$ is the likelihood of the data parametrized by $\boldsymbol{m}$ and $\zeta \in \boldsymbol{R}^S$. We assume that the likelihood model describes simple directional distribution on the hyper-sphere and choose the *von Mises-Fisher distribution* (Mardia, 1975) for the mixture components:

$$f(\boldsymbol{x}; \boldsymbol{m}, \zeta) = C_S(\zeta) e^{\zeta \langle \boldsymbol{x}, \boldsymbol{m} \rangle}, \tag{4.3}$$

where inner product corresponds to the correlation of the two vectors on the sphere. Note that this model is in agreement with the notion that on a hyper-sphere, correlation is the natural measure of similarity between two vectors. The distribution is an exponential function of the correlation between the vector $\boldsymbol{x}$ (voxel selectivity profile) and the *mean direction* $\boldsymbol{m}$ (system selectivity profile). The normalizing constant $C_S(\zeta)$ is defined in terms of the γ-th order modified Bessel function of the first kind I_γ:

$$C_S(\zeta) = \frac{\zeta^{S/2-1}}{(2\pi)^{S/2} I_{S/2-1}(\zeta)}. \tag{4.4}$$

The *concentration parameter* ζ controls the concentration of the distribution around the mean direction $\boldsymbol{m}$ similar to the inverse of variance for Gaussian models. In general, mixture components can have distinct concentration parameters but in this work, we use the same parameter for all the clusters to ensure a more robust estimation. This model has been previously employed in the context of clustering text data (Banerjee et al., 2006).

The problem is then formulated as a maximum likelihood estimation:

$$(\{w_k^*, \boldsymbol{m}_k^*\}_{k=1}^K, \zeta^*) = \underset{\{w_k, \boldsymbol{m}_k\}_{k=1}^K, \zeta}{\operatorname{argmax}} \sum_{i=1}^V \log p(\boldsymbol{x}_i; \{w_k, \boldsymbol{m}_k\}_{k=1}^K, \zeta). \tag{4.5}$$

Employing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to solve the problem involves introducing membership variables z_i to the problem according to Equation (2.10).

■ 4.2.2 EM Algorithm

The EM algorithm yields a set of update rules for model parameters and the posterior probability $p(z_i = k|\boldsymbol{x}_i)$ that voxel i is associated with the mixture component k, which we will henceforth denote instead by $p(k|\boldsymbol{x}_i)$ for the ease of notation. Starting with initial values $\{w_k^{(0)}, \boldsymbol{m}_k^{(0)}\}_{k=1}^K$ and $\zeta^{(0)}$ for the model parameters, we iteratively compute the the posterior assignment probabilities $p(k|\boldsymbol{x}_i)$ and then update the parameters $\{w_k, \boldsymbol{m}_k\}_{k=1}^K$ and ζ.

In the *E-step*, we fix the model parameters and update the system memberships:

$$p^{(t)}(k|\boldsymbol{x}_i) = \frac{w_k^{(t)} e^{\zeta^{(t)}\langle \boldsymbol{x}_i, \boldsymbol{m}_k^{(t)}\rangle}}{\sum_{k'=1}^K w_{k'}^{(t)} e^{\zeta^{(t)}\langle \boldsymbol{x}_i, \boldsymbol{m}_{k'}^{(t)}\rangle}} \,. \tag{4.6}$$

In the *M-step*, we update the model parameters:

$$w_k^{(t+1)} = \frac{1}{V}\sum_{i=1}^V p^{(t)}(k|\boldsymbol{x}_i)\,, \tag{4.7}$$

$$\boldsymbol{m}_k^{(t+1)} = \frac{\sum_{i=1}^V \boldsymbol{x}_i p^{(t)}(k|\boldsymbol{x}_i)}{\|\sum_{i=1}^V \boldsymbol{x}_i p^{(t)}(k'|\boldsymbol{x}_i)\|} \,. \tag{4.8}$$

After computing the updated cluster centers $\boldsymbol{m}_k^{(t+1)}$, the new concentration parameter $\zeta^{(t+1)}$ is found by solving the nonlinear equation

$$A_S(\zeta^{(t+1)}) = \Gamma^{(t+1)} \tag{4.9}$$

for positive values of $\zeta^{(t+1)}$, where

$$\Gamma^{(t+1)} = \frac{1}{V}\sum_{l=1}^K \sum_{v=1}^V p^{(t)}(k|\boldsymbol{x}_i)\langle \boldsymbol{m}_k^{(t+1)},\ \boldsymbol{x}_i\rangle \tag{4.10}$$

and the function $A_S(\cdot)$ is defined as

$$A_S(\zeta) = \frac{I_{S/2}(\zeta)}{I_{S/2-1}(\zeta)}. \tag{4.11}$$

The details of the algorithm used for solving this equation, along with the derivations of the update rules, are presented in Appendix B.1. Iterating the set of E-step and M-step updates until convergence, we find K system selectivity profiles $\boldsymbol{m}_k$ and a set of

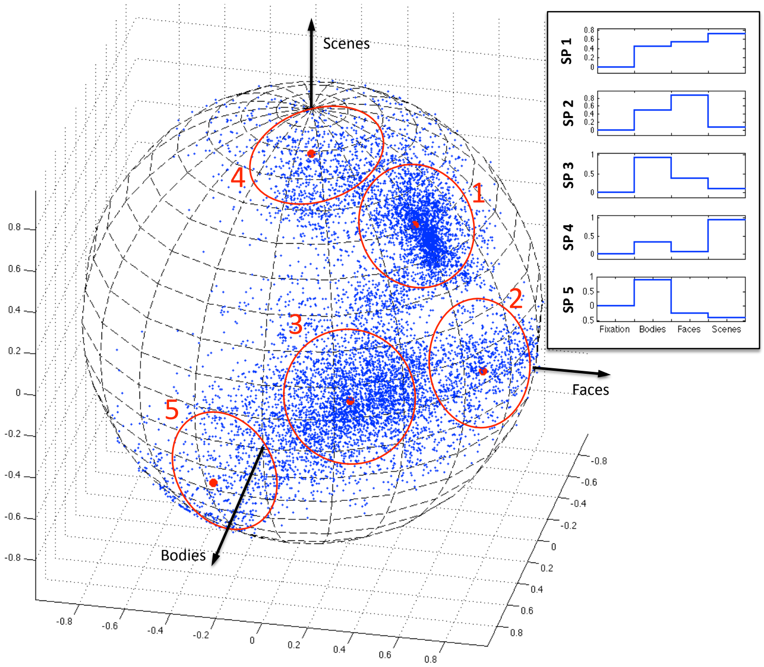


Figure 4.2. The results of mixture model density estimation with 5 components for the set of selectivity profiles in Figure 4.1(B). The resulting system selectivity profiles (cluster centers) are denoted by the red dots; circles around them indicate the size of the corresponding clusters. The box shows an alternative presentation of the selectivity profiles where the values of their components are shown along with zero for fixation. Since this format allows presentation of the selectivity profiles in general cases with $S > 3$, we adopt this way of illustration throughout the paper. The first selectivity profile, whose cluster includes most of the voxels in the overlapping region, does not show a differential response to our three categories of interest. Selectivity profiles 2, 3, and 4 correspond to the three original types of activation preferring faces, bodies, and scenes, respectively. Selectivity profile 5 shows exclusive selectivity for bodies along with a slightly negative response to other categories.

soft assignments $p(k|\boldsymbol{x}_i)$ for $k = 1, \cdots, K$. The assignments $p(k|\boldsymbol{x}_i)$, when projected to the anatomical locations of voxels, define the spatial maps of the discovered systems.

Figure 4.2 illustrates 5 systems and the corresponding profiles of selectivity found by this algorithm for the population of voxels shown in Figure 4.1(B). As expected, the analysis identifies clusters of voxels exclusively selective for one of the three conditions, but also finds a cluster selective for all three conditions along with a group of body selective voxels that show inhibition towards other categories. More complex profiles of selectivity such as the two latter cases cannot be easily detected with the conventional method.

4.3 Statistical Validation

As we argued before, the space of selectivity profiles can act as a common space for representation of data from different subjects in a group study. If the set of vectors $\{\boldsymbol{m}_k\}_{k=1}^K$ describes all relevant selectivity profiles in the brain system of interest, each voxel $\boldsymbol{x}_{ji}$ can be thought of as an independent sample from the same distribution introduced by Equation (4.2). Thus, we combine the data from several subjects to form the *group data*, i.e., $\{\{\boldsymbol{x}_{ji}\}_{i=1}^{V_j}\}_{j=1}^J$, to perform our analysis across subjects. Applying our algorithm to the group data, the resulting set of assignments $\{p(k|\boldsymbol{x}_{ji})\}_{i=1}^{V_j}$ defines the spatial map of system k in subject j.

In conventional group data analysis, spatial consistency of the activation maps across subjects provides a measure for the evaluation of the results. For the method introduced here, we focus on the functional consistency of the discovered system selectivity profiles. To quantify this consistency, we define a *consistency score* (cs) for each selectivity profile found in a group analysis. Let $\{\{\boldsymbol{x}_{ji}\}_{i=1}^{V_j}\}_{j=1}^J$ be the group data including voxel profiles from J different subjects, K be the number of desired systems, and $\{\boldsymbol{m}_k^G\}_{k=1}^K$ be the final set of system selectivity profiles found by the algorithm in the group data. We also apply the algorithm to the J individual subject data sets $\{\boldsymbol{x}_{ji}\}_{i=1}^{V_j}$ separately to find their corresponding J sets of subject-specific systems $\{\boldsymbol{m}_k^j\}_{k=1}^K$. We can then match the selectivity profile of each group system to its most similar system profile in each of the J individual data sets.

4.3.1 Matching Profiles Across Subjects

The matching between the group and individual selectivity profiles is equivalent to finding J one-to-one functions $\omega_j : \{1, \cdots, K\} \to \{1, \cdots, K\}$ which assign system profile $\boldsymbol{m}_{\omega_j(k)}^j$ in subject j to the group system profile $\boldsymbol{m}_k^G$. We select the function ω_s such that it maximizes the overall similarity between the matched selectivity profiles:

$$\omega_j^*(\cdot) = \underset{\omega(\cdot)}{\operatorname{argmax}} \sum_{k=1}^K \rho(\boldsymbol{m}_k^G, \boldsymbol{m}_{\omega(k)}^j) \,. \tag{4.12}$$

Here, $\rho(\cdot,\cdot)$ denotes the correlation coefficient between two vectors:

$$\rho(\boldsymbol{a}_1, \boldsymbol{a}_2) = \langle \boldsymbol{a}_1 - \tfrac{1}{S}\langle \boldsymbol{a}_1, \mathbf{1}\rangle, \boldsymbol{a}_2 - \tfrac{1}{S}\langle \boldsymbol{a}_2, \mathbf{1}\rangle\rangle, \tag{4.13}$$

where $\mathbf{1}$ is the S-dimensional vector of unit components. The maximization in Equation (4.12) is performed over all possible one-to-one functions ω. Finding this matching is an instance of graph matching problems for a bipartite graph (Diestel, 2005). The graph is composed of two sets of nodes, corresponding to the group and the individual system profiles, and the weights of the edges between the nodes are defined by the correlation coefficients. we can employ the well-known Hungarian algorithm (Kuhn, 1955) to solve this problem for each subject.[1]

Having matched each group system with a distinct system within each individual subject result, consistency score cs_k for group system k can be computed as the average correlation of its selectivity profile with the corresponding subject-specific system profiles:

$$\mathrm{cs}_k = \frac{1}{S}\sum_{j=1}^{J} \rho(\boldsymbol{m}_k^G, \boldsymbol{m}^j_{\omega_j^*(k)}). \tag{4.14}$$

Consistency score values measure how closely a particular type of selectivity repeats across subjects. Clearly, $\mathrm{cs} = 1$ is the most consistent case where the corresponding profile identically appears in all subjects. Because of the similarity-maximizing matching performed in the process of computing the scores, even a random data set would yield non-zero consistency score values. We employ permutation testing to establish the null hypothesis distribution for the consistency score.

■ 4.3.2 Permutation Test

To construct the baseline distribution for the consistency scores under the null hypothesis, we can make random modifications to the data in such a way that the correspondence between the components of the selectivity profiles and the experimental conditions is removed. Specifically, we may randomize the condition labels before the regression step so that the individual regression coefficients do not correspond to any non-random distinctions in the task. One way to implement such a randomization is to manipulate the linear analysis stage. If for instance, we are dealing with a block design experiment, each temporal block in the experiment has a category label that determines its corresponding regressor in the design matrix. We can randomly shuffle these labels and, as a result, the regressors in the design matrix include blocks of images from random categories. The resulting estimated regression coefficients do not correspond to any coherent set of stimuli. Applying the analysis to this modified data set still yields a set of group and individual system selectivity profiles and their corresponding cs values. Since there is no real structure in the modified data, all cs values obtained in this manner can serve as samples from the desired null hypothesis.

We evaluate statistical significance of the cs value of each system selectivity profile found in the actual data based on this null distribution. In practice, for up to 10,000 shuffled data sets, the consistency scores of most system selectivity profiles in the real data exceed all the cs values estimated from the shuffled data, implying the same empirical significance of $p = 10^{-4}$. To distinguish the significance of these different profiles through our p-value, we fit a Beta distribution to the null-hypothesis samples and compute the significance from the fitted distribution. Using a linear transformation to

[1] We used the open source matlab implementation of the Hungarian algorithm available at `http://www.mathworks.com/matlabcentral/fileexchange/11609`.

match the range $[-1, 1]$ of cs values to the support $[0, 1]$ of the Beta distribution, we obtain the pdf of the null distribution

$$f_{\text{Null}}(\text{cs}; \alpha, \beta) = \frac{2}{B(\alpha, \beta)} \left(\frac{1+\text{cs}}{2}\right)^{\alpha-1} \left(\frac{1-\text{cs}}{2}\right)^{\beta-1}, \tag{4.15}$$

where $B(\alpha, \beta)$ is the beta function. We find the maximum-likelihood pair (α, β) for the observed samples in the shuffled data set. We then characterize the significance of a selectivity profile with consistency score cs via its p-value, as inferred from the parametric fit to our simulated null-hypothesis distribution: $p = \int_{\text{cs}}^{1} f_{\text{Null}}(u; \alpha, \beta) du$.

The above scheme is computationally intensive since it requires recalculating the brain responses and applying the method to both individual and group data for each sample. More importantly, it yields a liberal test since we completely remove the correspondence between different dimensions of the selectivity profiles and the experimental conditions. It is also possible to derive an alternative permutation test that is more conservative and less time consuming. We can leave the linear estimation of brain responses and the estimation of individual system selectivity profiles intact, but reshuffle the dimensions of the selectivity profiles before performing the group analysis. In this way, our concept of the null distribution maintains the stimulus-related information in the data but removes the commonality of the space of functional profiles; the same dimension does not correspond to the same experimental condition across subjects. While the previous test can be though of as producing a null distribution that lacks *within-subject* structure, the new null distribution is short of structure *across-subject*.

■ 4.4 Discussion

While the analysis is completely independent of the spatial locations associated with the selectivity profiles, we can still examine the spatial extent of the discovered systems as a way to validate the results. If the selectivity profile of a system matches a certain type of activation, i.e., demonstrates exclusive selectivity for an experimental condition, we can compare the map of its assignments with the localization map detected for that activation by the conventional method. We will use this comparison to ensure that our method yields systems with spatial extents that correspond to the perviously characterized selective areas in the brain.

Once we identify a system to be associated with a certain activation, we quantify the similarity between the spatial maps estimated by the clustering method and that obtained via the standard confirmatory method. We can employ an asymmetric overlap measure between the spatial maps, equal to the ratio of the number of voxels in the overlapping region to the number of all voxels assigned to the system in our model. The asymmetry is included since, as we saw in the example in Section 4.1, being functionally more specific, our discovered systems are usually subsets of the localization maps found via the standard statistical test.

Although extremely simple, the construction given in Section 4.1 underlines two important features that we are seeking in the signal representation for the analysis: 1) representing the response as a voxel-specific, relative measure, and 2) relying on the experimental paradigm for defining a functional space common to all subjects in the study. The simple normalization by the length of vector of brain responses satisfies both these conditions and, as we will see in the next chapter, provides relevant results. However, the projection to a hypersphere is by no means the only way to define a measure with such qualities. In particular, Chapter 6 introduces an alternative construction for selectivity profiles based on voxel-specific binary activations, similar to that described in Section 2.1.2. Activation variables are by construction normalized and can be integrated within the model as latent variables.

Throughout this chapter, we have emphasized the critical role of the experimental paradigm for creating a common functional space. Nevertheless, there is evidence that the construction of such a space might be even possible in resting-state experiments. Langs et al. (2010) used spectral embedding to find a low-dimensional representation for voxel responses from the fMRI connectivity structure, and showed the resulting representation is consistent across subjects. Models presented in this thesis can be further extended to apply to such data (Langs et al., 2011).

As with any exploratory method, clustering is sensitive to noise. Since we aim to extract the structure in the data in an unsupervised fashion, the presence of a large number of noisy responses can overwhelm the interesting patterns in our results. In the space of selectivity profiles, the set of voxel brain responses is treated as a population of vectors without any spatial characterization. In Section 3.4, we commented on the advantages of this approach as a group analysis without any bias regarding the anatomical organization of functionally specific areas. Another benefit of this construction is in the flexibility it provides in the choice of voxels to include in the analysis. Here, we can select the set of voxels to consider for the analysis based on a prior test on their level of noise. This may be achieved employing an F test that detects voxels where the fMRI signal passes a certain SNR threshold as discussed in Section 2.2.1. Prior applications of clustering to fMRI data have also employed this pre-screening of voxels in their analyses (e.g., Goutte et al., 2001, 1999; Seghier et al., 2007). In this thesis, we refer to this preprocessing stage as the choice of *analysis mask*. Another way to think about this stage is as the selection of a functionally-defined ROI based on the experimental paradigm. This analogy makes it clear that the choice of the masking test has to be orthogonal to the space of selectivity profiles under consideration.

To create a matrix of pairwise similarities for group and individual systems to perform the matching in Section 4.3.1, we used correlation coefficients between the corresponding selectivity profiles. Once again, there are also other ways of defining a similarity measure. Specifically, if we are interested in incorporating the weight of clusters in the analysis, we may use a normalized overlap measure or the Dice measure (Dice, 1945) of volume overlap. In all the experiments presented in this thesis, we have used the definition in Equation (4.13); the alternative measures will provide similar results.

In this chapter we also presented our procedure for assessing cross-subject consistency of the discovered selectivity profiles. An alternative way to assess the resulting system selectivity profiles is to examine their consistency across repetitions of the same category in different blocks. If we define two groups of blocks that present stimuli from the same category as two distinct experimental conditions, we expect the corresponding components of a consistent system selectivity profile to be similar. We will employ this method for a qualitative study of consistency.

Finally, the technique introduced here for statistical validation of clustering results across a group has applications beyond exploratory studies of functional specificity. Recently, this approach has been applied to learn robust clustering of intrinsic patterns of functional connectivity and cortical thickness in large data sets (e.g., Yeo et al., 2011).

In the next chapter, we discuss the application of our method to two fMRI studies of high level vision.

Chapter 5

Discovering Functional Systems in High Level Vision

AS we explained in Section 1.1.1, the work in this thesis has been motivated by the question of the extent of category selectivity in the visual cortex. In this chapter, we present the results of applying the analysis technique developed in Chapter 4 to the data from two visual experiments. These two experiments were designed and performed such that they particularly utilize the exploratory potential of this novel analysis.

The first experiment focuses on searching the space of predefined categories. The study aims to reconsider the setting of the confirmatory study carried out by Downing et al. (2006) in an exploratory setting, this time allowing the possibility of selectivity for meta-categories composed of more than one of the originally hypothesized categories (see the first point in the list of the limitations of the confirmatory methods in the discussion of Section 1.1.2). The study results in the rediscovery of face, body, and scene selectivity in this large space.

The second experiment extends the search to the space of selectivity profiles over objects, rather than predefined categories. In this case, system selectivity profiles express patterns in the space of 69 distinct stimuli, and each implicitly characterize a categorization of images as seen by the respective system.[1]

■ 5.1 Block-Design Experiment

The first study is a block-design experiment that investigates patterns of functional specificity in the space of 16 distinct image sets, each drawn from one of 8 candidate categories.

■ 5.1.1 Data

The fMRI data was acquired using a Siemens 3T scanner and a custom 32-channel coil (EPI, flip angle $= 90^o$, TR $= 2$ seconds, in-plane resolution $= 1.5$ mm, slice thickness

[1] The work presented in this chapter has been carried on in collaboration with Edward Vul and Po-Jang Hsieh from the Brain and Cognitive Science department at MIT.

= 2 mm, 28 axial slices). The experimental protocol included 8 categories of images: Animals, Bodies, Cars, Faces, Scenes, Shoes, Trees, and Vases. For each image category, two different sets of images were presented in separate blocks. We used this setup to test that our algorithm successfully yields profiles with similar components for different images from the same category. Each block lasted 16 seconds and contained 16 images from one image set of one category. The blocks corresponding to different categories were presented in permuted fashion so that their order and temporal spacing was counter-balanced. With this design, the temporal noise structure is shared between the real data and the random permutations constructed by the procedure of Section 4.3.2. For each subject, there were 16 to 29 runs of the experiment where each run contained one block from each category and three fixation blocks. We perform motion correction, spike detection, intensity normalization, and Gaussian smoothing with a kernel of 3-mm width using the FreeSurger Functional Analysis Stream (FsFast).[2]

By modifying the condition-related part of the design matrix $\mathbf{G}$ in Equation (2.3) and estimating the corresponding regression coefficients $\hat{\boldsymbol{b}}$, we created three different data sets for each subject:

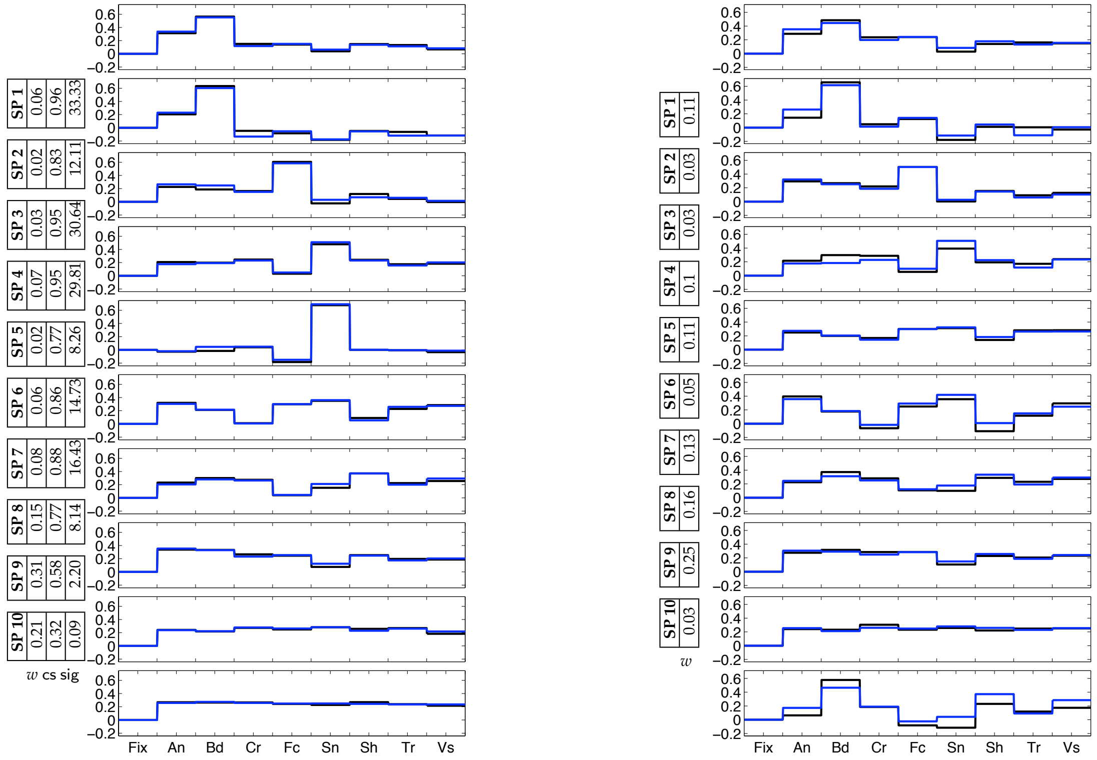
- **8-Dimensional Data:** All blocks for one category were represented as a single experimental condition by one regressor and, accordingly, one regression coefficient. The selectivity profiles were composed of 8 components each representing one category.

- **16-Dimensional Data:** The blocks associated with different image sets were represented as distinct experimental conditions. Since we had two image sets for each category, the selectivity profiles had two components for each category.

- **32-Dimensional Data:** We split the blocks for each image set into two groups and estimated one coefficient for each split group. In this data set, the selectivity profiles were 32 dimensional and each category was represented by four components.

To discard voxels with no visual activation, we formed contrasts comparing the response of voxels to each category versus fixation and applied the t-test to detect voxels that show significant levels of activation. The union of the detected voxels serves as the mask of visually responsive voxels used in our experiment. Significance thresholds were chosen to $p = 10^{-2}$, $p = 10^{-4}$, and $p = 10^{-6}$, for 32-Dimensional, 16-Dimensional, and 8-Dimensional data, respectively, so that the visually selective masks for different data sets are of comparable size. An alternative approach for selecting relevant voxels is to use an F-test considering all regressors corresponding to the visual stimuli (columns of matrix $\mathbf{G}$). We observed empirically that the results presented here are fairly robust to the choice of the mask and other details of preprocessing.

[2] http://surfer.nmr.mgh.harvard.edu/fswiki/FsFast.



(A) **(B)**

Figure 5.1. (A) A set of 10 discovered group system selectivity profiles for the 16-Dimensional group data. The colors (black, blue) represent the two distinct components of the profiles corresponding to the same category. We added zero to each vector to represent Fixation. The weight w for each selectivity profile is also reported along with the consistency scores (cs) and the significance values found in the permutation test, sig $= -\log_{10} p$. (B) A set of individual system selectivity profiles in one of the 6 subjects ordered based on matching to the group profiles in (A).

■ 5.1.2 Results

Selectivity Profiles

We apply the analysis to all three data sets. Figure 5.1(A) and Figure 5.2 show the resulting selectivity profiles of the group systems in the three data sets, where the number of clusters is $K = 10$. We also report cluster weights w_k, consistency scores cs$_k$, and their corresponding significance values sig $= -\log_{10} p$. In all data sets, the most consistent profiles are selective of only one category, similar to the previously characterized selective systems in high level vision. Moreover, their peaks match with these known areas such that in each data set, there are selectivity profiles corresponding to EBA (body-selective), FFA (face-selective), and PPA (scene-selective). For instance, in Figure 5.1(A), selectivity profiles 1 and 2 show body selectivity, selectivity profile 3

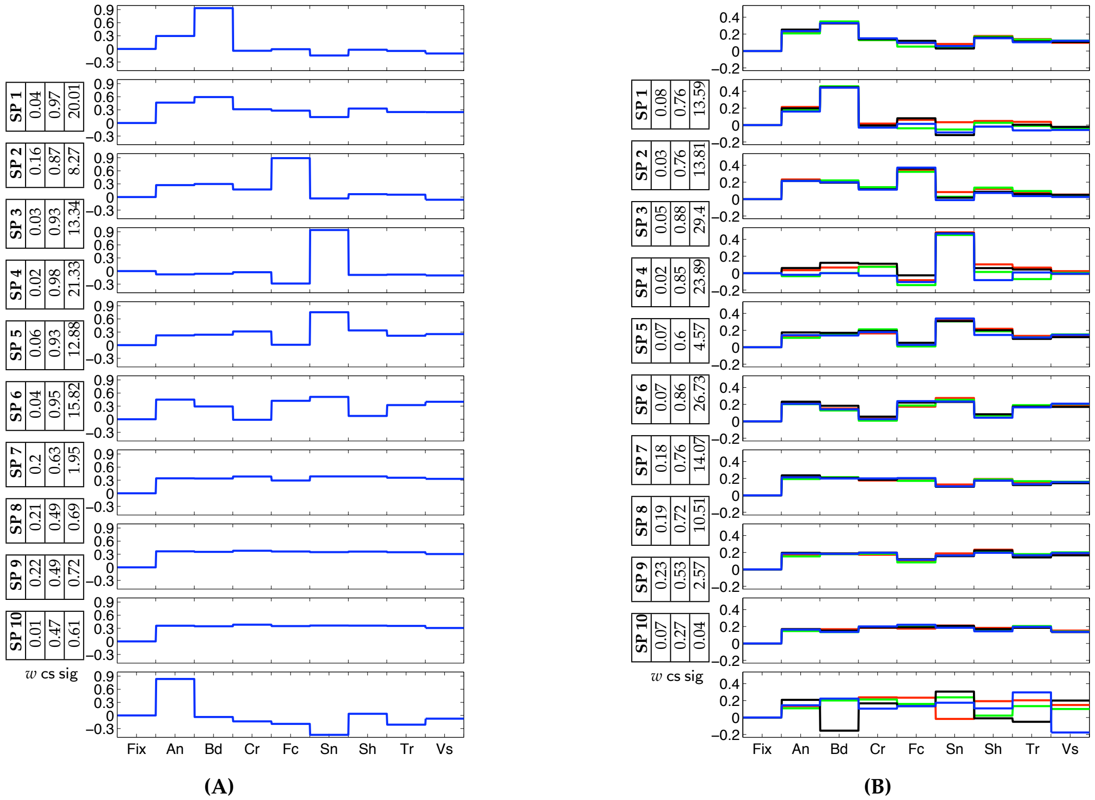
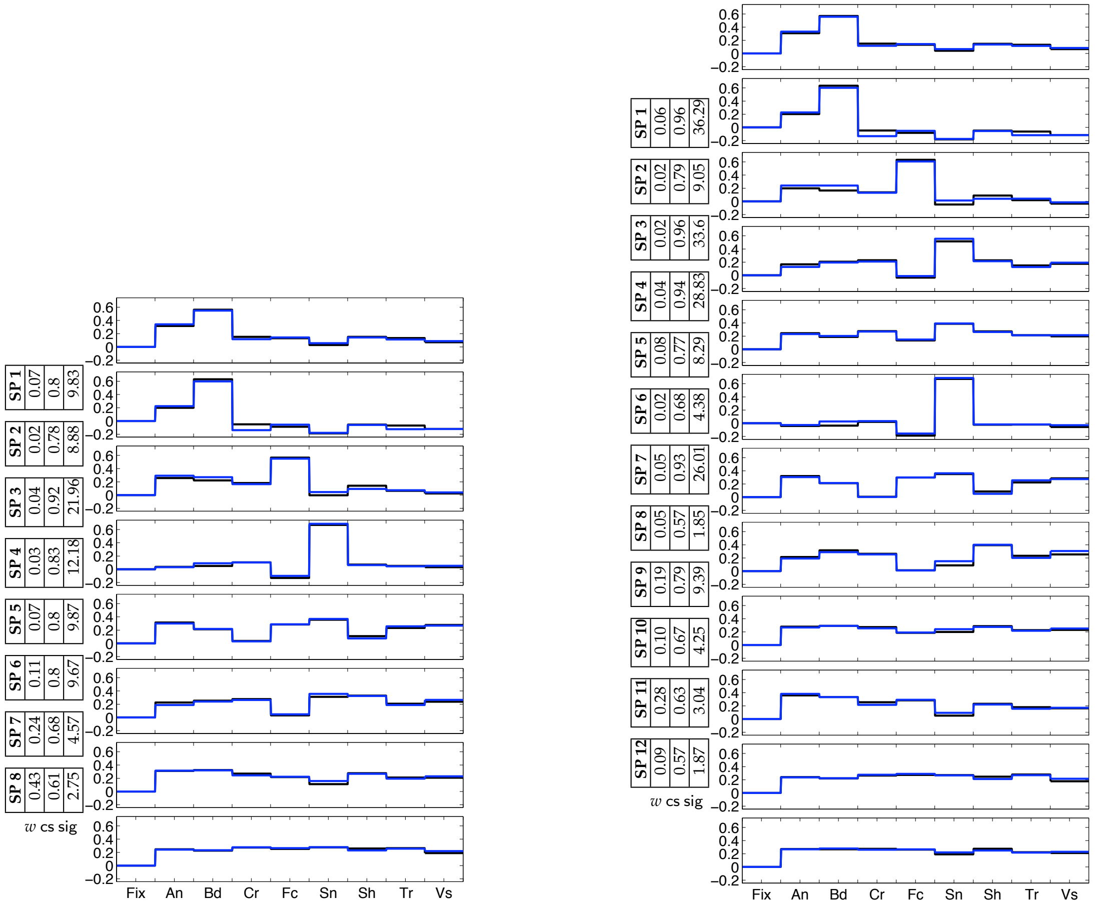


Figure 5.2. Sets of 10 discovered group system selectivity profiles for (A) 8-Dimensional, and (B) 32-Dimensional data. Different colors (blue, black, green, red) represent different components of the profiles corresponding to the same category. We added zero to each vector to represent Fixation. The weight w for each selectivity profile is also reported along with the consistency score (cs) and the significance value.

is face selective, and selectivity profiles 4 and 5 are scene selective. Similar profiles appear in the case of 8-Dimensional and 32-Dimensional data as well. Comparing the selectivity profiles found for one of the individual 16-Dimensional data sets with those of the group data in Figure 5.1(A) and (B) shows that the more consistent group profiles resemble their individual counterparts.

In each data set, our method detects systems with rather flat profiles over the entire set of presented categories. These profiles match the functional definition of early areas in the visual cortex, selective to lower level features in the visual field. Not surprisingly, there is a large number of voxels associated with these systems, as suggested by their estimated weights.

The 16-Dimensional and 32-Dimensional data sets allow us to examine the consistency of the discovered profiles across different image sets and different runs. Different components of the selectivity profiles that correspond to the same category of

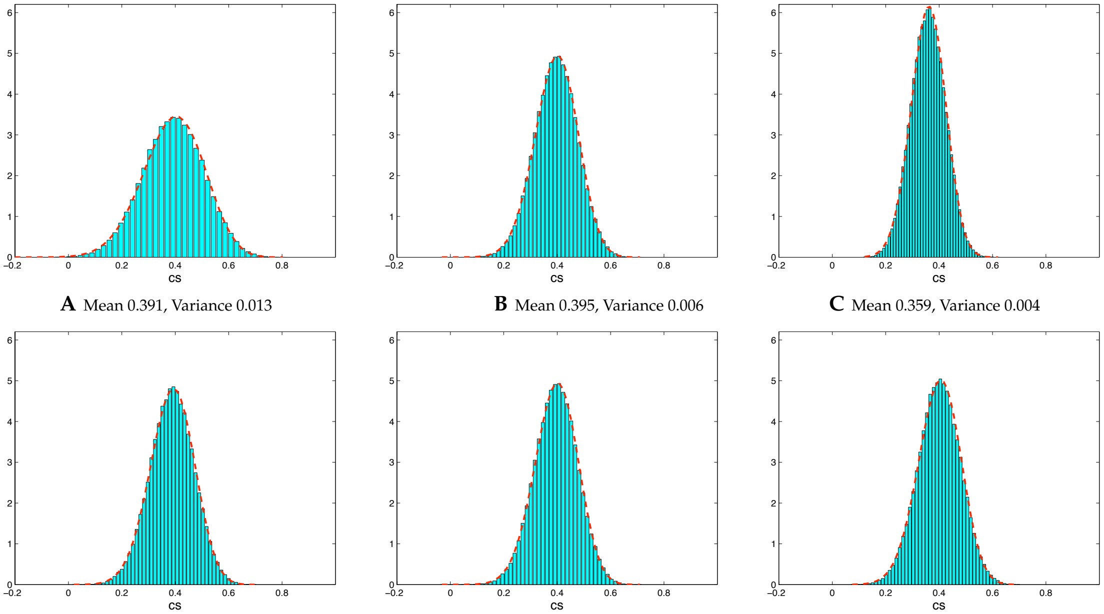


(A) **(B)**

Figure 5.3. Group system selectivity profiles in the 16-Dimensional data for (A) 8, and (B) 12 clusters. The colors (blue, black) represent the two distinct components of the profiles corresponding to the same category, and the weight w for each system is also indicated along with the consistency score (cs) and its significance value found in the permutation test.

images, illustrated with different colors in Figure 5.1, have nearly identical values. This demonstrates consistency of the estimated profiles across experimental runs and image sets. The improvement in consistency from the individual data in Figure 5.1(B) to the group data of Figure 5.1(A) further justifies our argument for fusing data from different subjects.

To examine the robustness of the discovered selectivity profiles to the change in the number of clusters, we ran the same analysis on the 16-Dimensional data for 8 and 12 clusters. Comparing the results in Figure 5.3 with those of Figure 5.1(A), we conclude that selectivity properties of the more consistent selectivity profiles remain relatively



A Mean 0.391, Variance 0.013

B Mean 0.395, Variance 0.006

C Mean 0.359, Variance 0.004

D Mean 0.387, Variance 0.007

E Mean 0.395, Variance 0.006

F Mean 0.400, Variance 0.006

Figure 5.4. Null hypothesis distributions for the consistency score values, computed from 10,000 random permutations of the data. Histograms A, B and C show the results for 8, 16, and 32-Dimensional data with 10 clusters, respectively. Histograms D, E, and F correspond to 8, 10, and 12 clusters in 16-Dimensional data (B and E are identical). We normalized the counts by the product of bin size and the overall number of samples so that they could be compared with the estimated Beta distribution, indicated by the dashed red line.

stable. In general, running the algorithm for many different values of K, we observed that increasing the number of clusters usually results in the split of some of the clusters but does not significantly alter the pattern of discovered profiles and their maps.

In order to find the significance of consistency scores achieved by each of these selectivity profiles, we performed a within-subject permutation test as described in Section 4.3. For each data set, we generated 10,000 permuted data sets by randomly shuffling labels of different experimental blocks. The resulting null hypothesis distributions are shown in Figure 5.4 for different data sets. Using these distributions, we compute the statistical significance of the consistency scores presented for the selectivity profiles in Figures 5.1, 5.2, and 5.3.

Spatial Maps

We can further examine the spatial maps associated with each system. Figure 5.5 shows the standard localization map for the face selective areas FFA and OFA in blue. This map is found by applying the t-test to identify voxels with higher response to

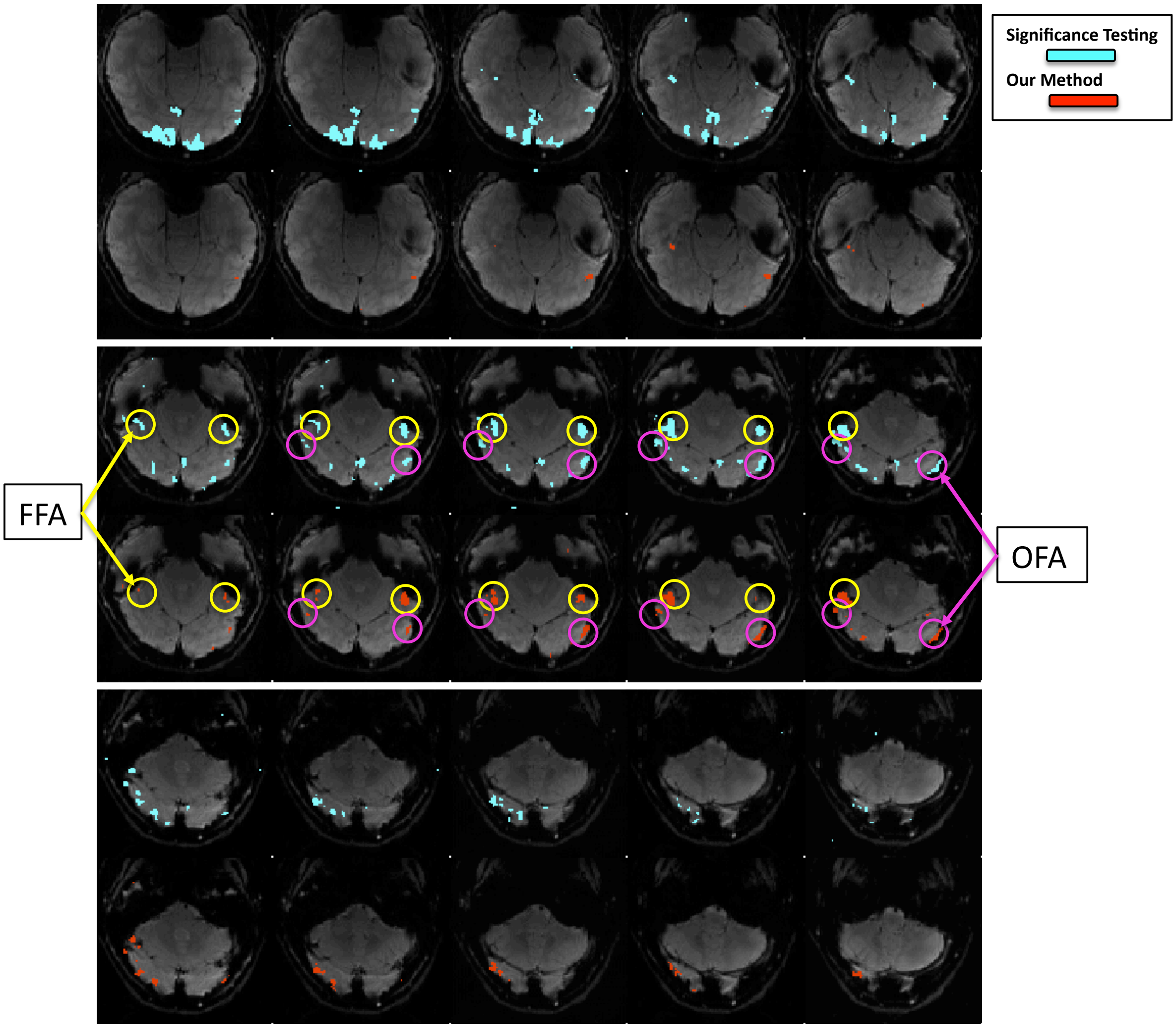


Figure 5.5. Spatial maps of the face selective regions found by the significance test (light blue) and the mixture model (red). Slices from the each map are presented in alternating rows for comparison. The approximate locations of the two face-selective regions FFA and OFA are shown with yellow and purple circles, respectively.

faces when compared to objects, with a threshold $p = 10^{-4}$, in one of the subjects. For comparison, Figure 5.5 also shows the voxels in the same slices assigned by our method to the system with the selectivity profile 3 in Figure 5.2(A) that exhibits face selectivity (red). The assignments found by our method represent probabilities over cluster labels. To generate the map, we assign each voxel to its corresponding maximum a posteriori (MAP) cluster label. We have identified on these maps the approximate locations of the two known face-selective regions, FFA and OFA, based on the result of the significance map, as it is common in the field. Figure 5.5 illustrates that, although the two maps are derived with very different assumptions, they mostly agree, especially in the more interesting areas where we expect to find face selectivity. As mentioned in Section 4.1, the conventional method identifies a much larger region as face selective, including parts in the higher slices of Figure 5.5 which we expect to be in the non-selective V1 area. Our map, on the contrary, does not include these voxels.

We compute three localization maps for face, scene, and body selective regions by applying statistical tests comparing responses of each voxel to faces, scenes, and bodies, respectively, with objects, and thresholding them at $p = 10^{-4}$ (uncorrected). To define selective systems in our results, we employ the conventional definition requiring the response to the preferred category to be at least twice the value of the response to other stimuli. We observe that the largest cluster always has a flat profile with no selectivity, e.g., Figure 5.2 (A) and (B). We form the map associated with the largest system as another case for comparison and call it the non-selective profile. Table 5.1 shows the resulting values of our overlap measure averaged across all subjects for $K = 8$, 10, and 12. We first note that the overlap between the functionally related regions is significantly higher than that of the unrelated pairs. Moreover, these results are qualitatively stable with changes in the number of clusters.

In the table, we also present the results of the algorithm applied to the data of each individual subject separately. We notice higher average overlap measures and lower standard deviations for the group data. This is due to the fact that fusing data from a cohort of subjects improves the accuracy of our estimates of the category selective profiles. As a result, we discover highly selective profiles whose response to the preferred stimuli satisfies the condition for being more than twice the other categories. On the other hand, in the results from the individual data for some noisier subjects, even the selective system does not satisfy this definition. For these subjects, no system is identified as exclusively selective of that category, degrading the average overlap measure. The improved robustness of the selectivity profile estimates in the group data prevents this effect and leads to better agreement in the spatial maps.

■ 5.2 Event-Related Experiment

In the experiment of Section 5.2, the dimensions of the space of selectivity profiles corresponded to predefined categories. However, what we really want to do in an exploratory setting is to *discover* rather than to assume the categories that are special

Gr. $K = 10$	Face	Body	Place
Face	**0.78** ± 0.08	0.14 ± 0.11	0
Body	0.07 ± 0.06	**0.94** ± 0.1	0.01 ± 0.02
Scene	0.01 ± 0.01	0.04 ± 0.04	**0.57** ± 0.19
Non-selective	0.06 ± 0.03	0.02 ± 0.02	0.1 ± 0.05

Indiv. $K = 10$	Face	Body	Place
Face	**0.28** ± 0.44	0.05 ± 0.11	0.0 ± 0.01
Body	0.1 ± 0.09	**0.65** ± 0.51	0.01 ± 0.02
Scene	0.01 ± 0.01	0.05 ± 0.08	**0.61** ± 0.47
Non-selective	0.09 ± 0.09	0.09 ± 0.08	0.13 ± 0.13

Gr. $K = 8$	Face	Body	Place
Face	**0.72** ± 0.08	0.16 ± 0.11	0
Body	0.07 ± 0.06	**0.94** ± 0.1	0.01 ± 0.02
Scene	0.02 ± 0.03	0.04 ± 0.06	**0.79** ± 0.19
Non-selective	0.05 ± 0.02	0.02 ± 0.02	0.09 ± 0.04

Gr. $K = 12$	Face	Body	Place
Face	**0.83** ± 0.06	0.15 ± 0.12	0.0 ± 0.01
Body	0.07 ± 0.06	**0.94** ± 0.09	0.01 ± 0.02
Scene	0.01 ± 0.01	0.05 ± 0.05	**0.66** ± 0.19
Non-selective	0.08 ± 0.04	0.03 ± 0.03	0.09 ± 0.05

Table 5.1. Asymmetric overlap measures between the spatial maps corresponding to our method and the conventional confirmatory approach in the block-design experiment. The exclusively selective systems for the three categories of Bodies, Faces, and Scenes, and the non-selective system (rows) are compared with the localization maps detected via traditional contrasts (columns). Values are averaged across all 6 subjects in the experiment.

in the brain. Here we tackle that goal by eliminating the assumption of which sets of images form categories constitute a category. We scan subjects while they view each of 69 unique images in an event-related design, and fit the mixture model to the selectivity profiles defined over unique images (rather than over presumed categories). The space of response profiles over the 69 images is enormous. Face, place and body selectivity represent a tiny fraction of this enormous space of possible response profiles. Here we ask whether these categories still emerge as dominant when tested against the entire space of possible response profiles. Further, do we discover new, previously unknown, response profiles that robustly arise in the ventral pathway?

■ 5.2.1 Data

Eleven subjects were scanned in the event-related experiment design. Each subject was scanned in two 2-hour scanning sessions (functional data from the two sessions were co-registered using to the subject's native anatomical space). During the scanning session the subjects saw event-related presentations of images from nine categories (animals, bodies, cars, faces, scenes, shoes, tools, trees, vases). Images were presented in a pseudo-randomized design generated by optseq 13 to optimize the efficiency of regression. During each 1.5 s presentation, the image moved slightly across the field of view either leftward or rightward, and subjects had to identify the direction of motion with a button press. Half of the images were presented in session one, and the other half were presented in session two (see Figure 5.6 for details).

Functional MRI data were collected on a 3T Siemens scanner using a Siemens 32-channel head coil. The high-resolution slices were positioned to cover the entire temporal lobe and part of the occipital lobe (gradient echo pulse sequence, TR = 2 s, TE

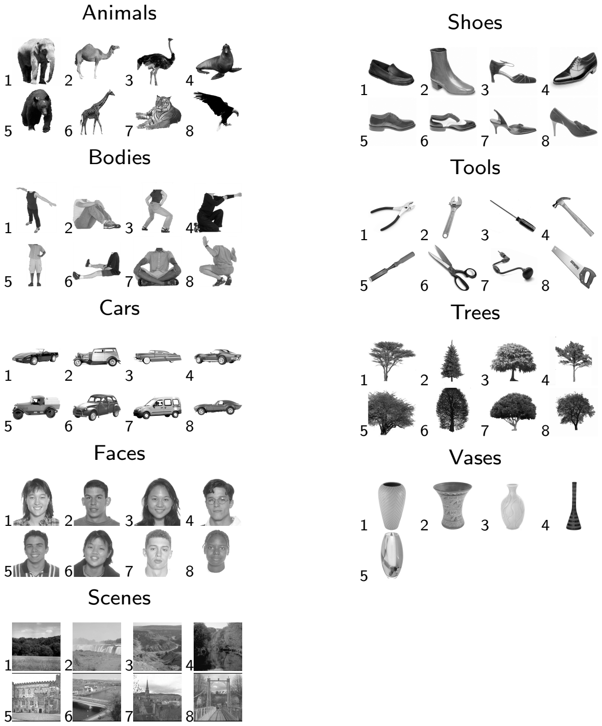


Figure 5.6. The 69 images used in the experiment.

= 30 ms, 40 slices with a 32 channel head coil, slice thickness = 2 mm, in-plane voxel dimensions = 1.6 x 1.6 mm). Data analysis was performed with FsFast,[3] fROI,[4] and custom Matlab scripts. The data was first motion-corrected separately for the two sessions (Cox and Jesmanowicz, 1999) and then spatially smoothed with a Gaussian kernel of 3mm width. The clustering analysis was run on voxels selected for each subject for responding significantly to any one stimulus (omnibus F-test). We used standard linear regression to estimate the response of voxels to each of the 69 conditions. We then registered the data from the two sessions to the subject's native anatomical space (Greve and Fischl, 2009).

■ 5.2.2 Results

First, we estimated the magnitude of response of each voxel in the ventral pathway to each of the 69 distinct images in each of 11 subjects. Then we applied our mixture of functional systems algorithm to this data set, in effect searching for the ten most prominent response profiles over the 69 images in the ventral visual pathway. Figure 5.7 shows the results of this analysis, with each of the ten system response profiles.

[3] http://surfer.nmr.mgh.harvard.edu/fswiki/FsFast

[4] http://froi.sourceforge.net

We assessed whether the detected systems were significant (more reliable across subjects than would be expected by chance) under the null hypothesis that assumes no shared structure across subjects (see Section 4.3.2); drawing 1,000 samples from this null distribution, Figure 5.8 shows the results that suggest systems 1 through 7 are significant at $p < 0.001$ each, but systems 8-10 are not significant ($p > 0.5$ each). The significant systems include profiles that appear upon visual inspection (see Figures 5.7 and 5.9) to be selective for bodies (System 1), faces (System 2), and scenes (System 3).

To confirm the intuition that these profiles are in fact selective for faces, bodies, and scenes, we computed ROC curves for how well each response profile picks out a preferred category. For each cluster we test selectivity for the category of the image to which that cluster is most responsive (thus, if the cluster is most responsive to a face, we propose that it is *face selective*), then we use the other 68 images to compute an ROC curve describing how precisely this profile selects the identified image category. Figure 5.7 shows the ROC curves for the 10 identified clusters–via bootstrapping we find that the areas under the curve for the face, body, and scene systems are all statistically significant (all $p < 0.01$). On the other hand, the specific rank-ordering of preferred images within the body (Figure 5.9a), face (Figure 5.9b), and scene (Figure 5.9c) systems shows substantial variation in the magnitude of response to different exemplars from these categories. While these systems are well characterized by selectivity for the a priori categories, they do not respond homogeneously to all stimuli within each category.

To quantitatively assess the robustness and reliability of identified clusters, we tested whether the selectivity is reliable when evaluated with respect to independent images that were not used for clustering. Thus, we split our images into two halves, with four images from each category in each half of the data. We then used one half of the image set for applying our mixture model and the other half to assess the stability of the selectivity of the identified functional systems. Specifically, we analyzed the response magnitude to the second half of the stimuli in voxels that were clustered into a given system using the first half of the stimuli. Figure 5.10 confirms the across-image reliability of category selectivity for the first four systems: selectivity identified by the method in one half of the images replicates in the second half of the images.

The analyses described so far demonstrate that from the huge space of possible response profiles that could be discovered in our analysis, response profiles reflecting selectivity for faces, bodies, and places emerge at the top of the stack, indicating that they are some of the most dominant response profiles in the ventral pathway.

New Selectivities?

Next we turn to the question of whether our analysis discovers any new functional systems not known previously. Beyond the systems that clearly reflect selectivity for faces, places, and bodies, there are four other significant systems (Systems 4, 5, 6, and 7 in Figure 5.7). A comparison of these profiles with those of systems derived from retinotopic regions of cortex (see Figure 5.11) shows that three of these systems resem-

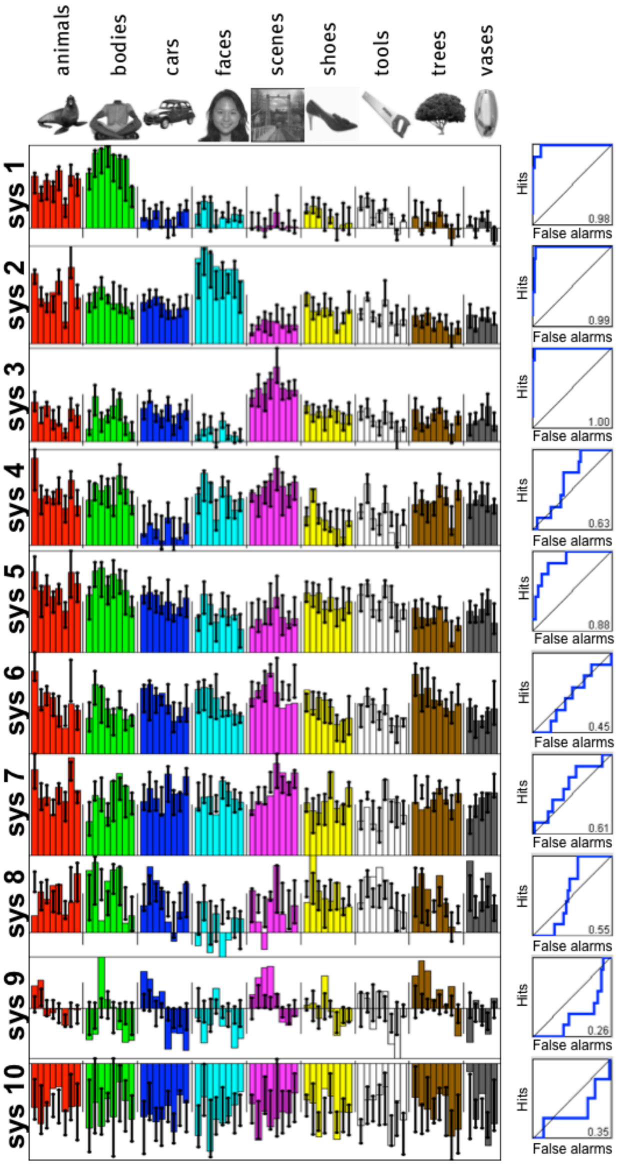


Figure 5.7. System profiles defined over 69 unique images. Each plot corresponds to one identified system (of 10 total). Plots are ordered from top based on their consistency scores. Bars represent the response profile magnitude for the group cluster, while the bars represent the 75% interquartile range of the matching individual subject clusters (see methods). ROC curves on the right show how well the system profile picks out the preferred image category (defined by the image with the highest response), as evaluated on the other 68 images. As shown in text, systems 1-7 are significantly consistent across subjects, while systems 8-10 are not.

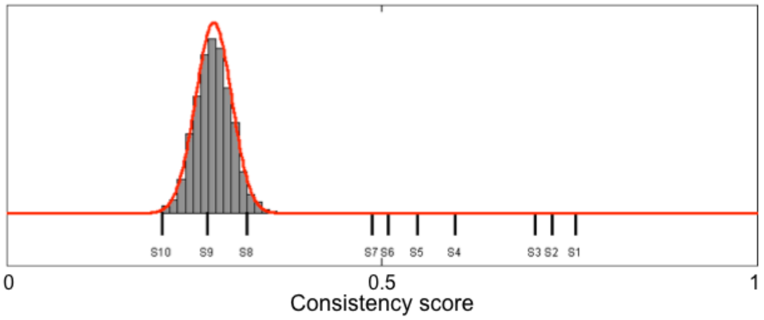


Figure 5.8. The distribution of consistency scores in the permutation analysis, along with the best-fitting beta-distribution, and markers indicating consistency scores for systems 1-10 (consistency scores for these systems, respectively: 0.76, 0.73, 0.70, 0.60, 0.55, 0.51, 0.49, 0.32, 0.27, and 0.21). Because the largest consistency score in the shuffled data is 0.36, the test suggests that profiles 1-7 are significantly consistent with a p-value of 0.001.

ble one or more of the selectivity profiles derived from retinotopic cortex: System 4 resembles System 1 from retinotopic cortex, system 6 resembles system 4 from retinotopic cortex, and system 7 resembles system 7 from retinotopic cortex ($r > 0.8$ in all cases), suggesting that these response profiles reflect the kind of basic visual properties extracted in retinotopic cortex.

Thus, the significant systems discovered by our algorithm include the three known category selectivities (Systems 1, 2, and 3), plus three systems that appear to reflect low-level visual properties (or at least resemble the selectivities that emerge from retinotopic cortex). The one remaining significant system, that does not strongly resemble any retinotopic profile, is System 5. Visual inspection of the stimuli that produce particularly high and low responses in this system (see Figure 5.12) do not lead to obvious interpretations of the function of this system. (It is tempting to label this system a selectivity for *animate objects* or *living things*, because of the high responses to bodies and animals, but neither classification can explain the low response of this system to faces and trees.) This situation reveals both the strength of our hypothesis-neutral structure discovery method (its ability to discover novel, unpredicted response profiles) and its weakness (what are we to say about these response profiles once we find them?). A full understanding of the robustness and functional significance of System 5 will have to await further investigation.

Projecting the Voxels in each System Back into the Brain

Crucially, all of the analyses described so far were blind to the anatomical location of each voxel (aside from a moderate smoothing of the data and the use of masks to select voxels in ventral pathway and retinotopic cortex for the two analyses). Thus, we made no assumptions about the spatial contiguity of voxels within a system nor did we assume that voxels within each system will be in anatomically similar locations

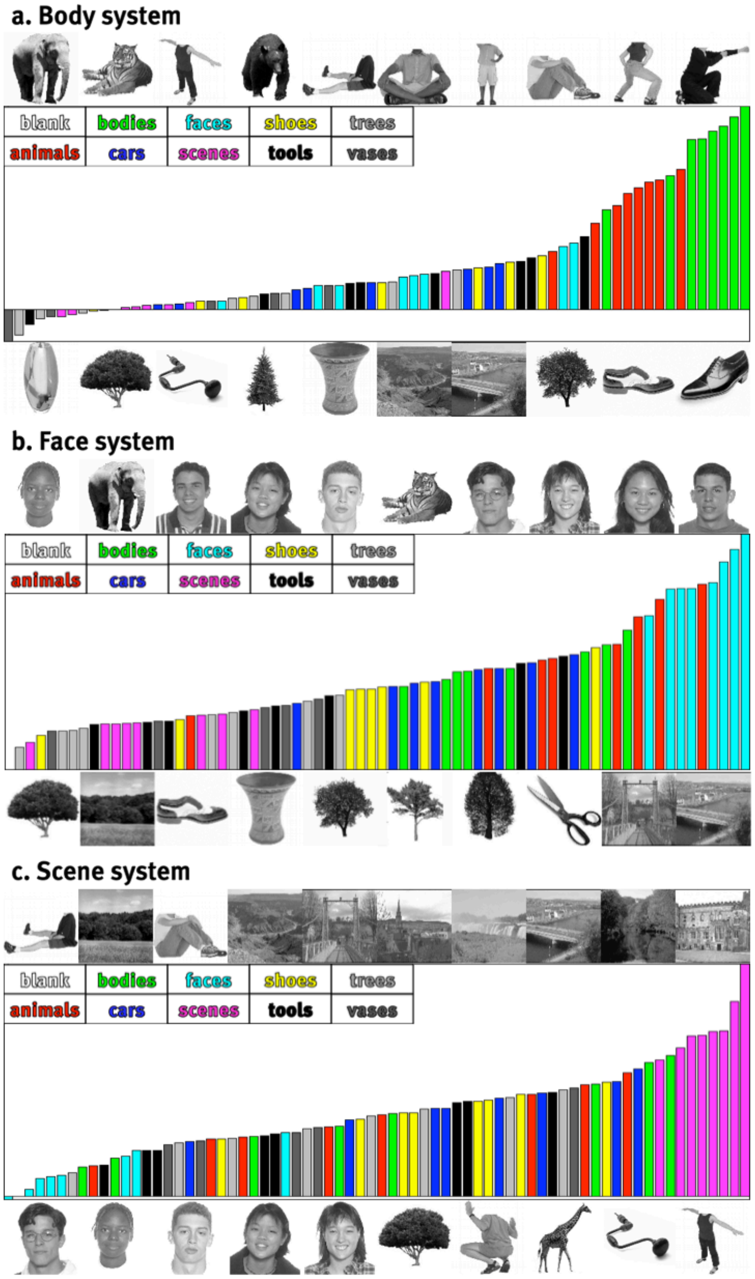


Figure 5.9. Stimulus preference for the apparent body (a; system 1), face (b; system 2), and scene (c; system 3) systems. For each system, the set of images above the rank ordered stimulus preference shows the top 10 preferred stimuli, and the images below show the 10 least preferred stimuli.

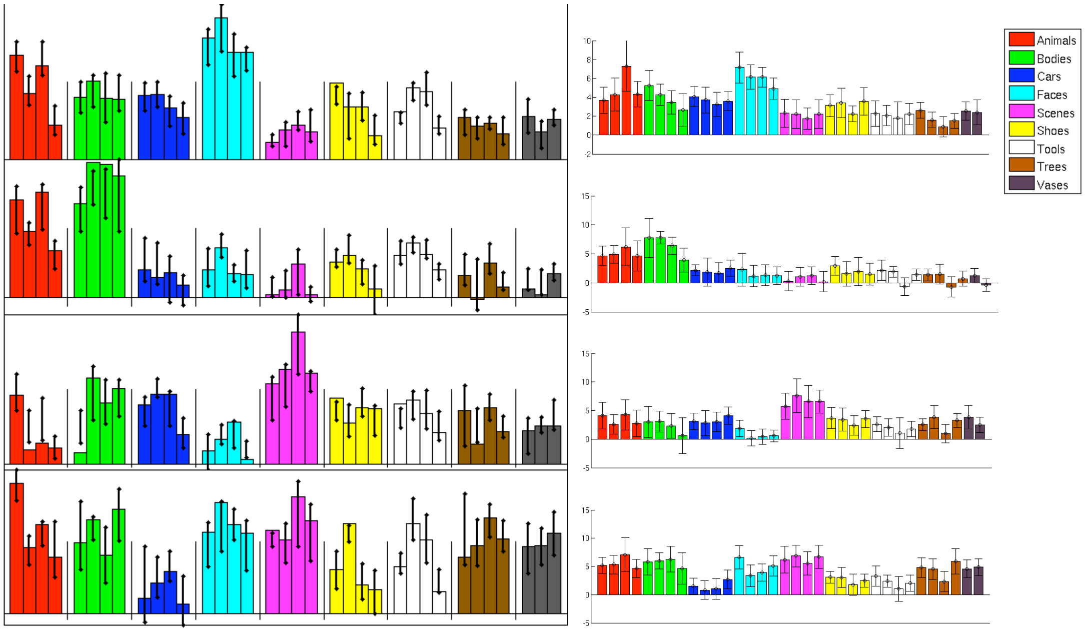


Figure 5.10. Event-related cross-validation analysis. Half of the images were used for mixture of functional systems (left), then the voxels corresponding to these clusters were selected as functional Regions of Interest, and the response of these regions were assessed in the held-out, independent half of the images (right).

across subjects. This analysis thus enables us to ask two questions that are implicitly assumed in standard group analyses (and even in most individual-subject analyses): 1) Do voxels with similar response profiles tend to be near each other in the brain, and 2) Do voxels with similar response profiles indeed land in similar anatomical locations across subjects? The answer to both questions is yes, as revealed by inspection of maps of the anatomical location of the voxels in each significant system in each subject (see Figure 5.13 and Figures C.1–C.44 in Appendix C). First, the anatomical locations of the voxels in Systems 1, 2, and 3 clearly match the well-established cortical regions selective for bodies, faces, and scenes, showing both spatial clustering within each subject for each system, and similarity in anatomical location across subjects.

Further, when the voxels in systems 4, 6, and 7, shown above to resemble profiles that emerge from retinotopic cortex, are projected back into the brain, they indeed appear mainly in posterior occipital regions known to be retinotopic (see Figures 5.13 and C.1–C.44) confirming our previous analysis that they reflect early stages of visual processing.

Finally, further evidence that the new functional profile, System 5, is indeed a novel selectivity worthy of further investigation comes from the fact that it too contains a largely contiguous cluster of voxels that is anatomically consistent across sub-

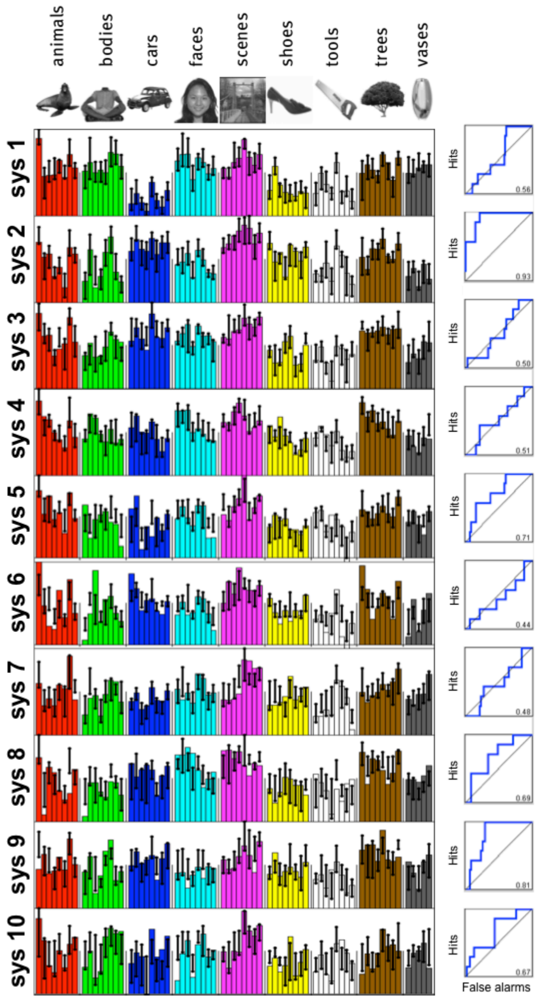


Figure 5.11. Clustering results on retinotopic occipital areas: Because the dominant selective profiles found in the ventral stream do not appear when clustering retinotopic cortex, they must reflect higher order structure rather than low-level image properties. Profiles are ordered based on the consistency scores (indicated by cs in the figure). The proportional weight of each profile (the ratio of voxels assigned to the corresponding system) is indicated by w. The thick, black line shows the group profile. To present the degree of consistency of the profile across subject, we also present the individual profiles (found independently in each of the 11 subjects) that correspond to each system (thin, colored lines).

jects. Specifically, the spatial map of System 5 consistently includes areas on the lateral surface of both hemispheres that begin inferiorly near and sometimes interdigitated within, but largely lateral to, face-selective voxels, extending up the lateral surface of the brain to more superior body-selective regions.

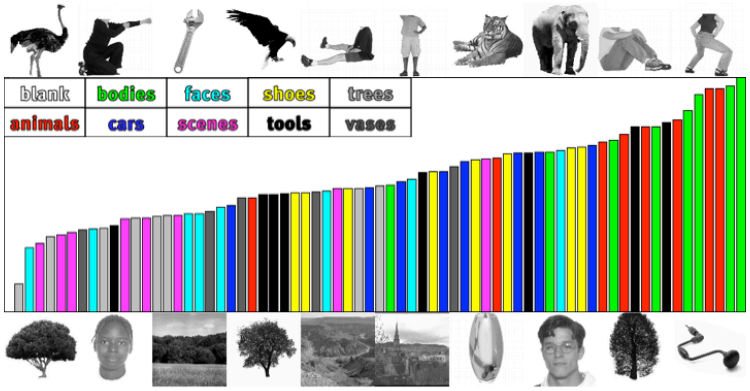


Figure 5.12. Rank ordered image preference for system 5. The set of images above the rank ordered stimulus preference shows the top 10 preferred stimuli, and the images below show the 10 least preferred stimuli.

■ 5.3 Discussion

In an initial test of our method in the block-design experiment, we applied the mixture of functional systems approach to data from a block-design experiment with eight stimulus categories. Face, place, and body selectivity popped out among the top five most consistent profiles in the ventral steam. Further, when each category was split into two non-overlapping data sets (based on blocks in which different stimulus exemplars were presented), the algorithm still produced the same face, place, and body-selective response profiles, implicitly discovering that the two different blocks within each category were the same. Although this analysis relaxed many assumptions generally made by prior work and effectively validated our analysis method, it only considered the eight categories we assumed a priori.

In the event-related experiment, we searched the enormous space of all possible response profiles over 69 stimuli, with no assumptions about 1) which of these stimuli go together to form a category, 2) what kind of response profile is expected (from an exclusive response to a single stimulus, to a broad response to many), or 3) whether voxels with similar response profiles are spatially clustered near each other in the cortex or in similar locations across subjects. Despite relaxing these assumptions, present in almost all prior work on the ventral visual pathway, we nonetheless found that three of the four most robust response profiles represent selectivity for faces, places, and bodies. Although in some sense this finding is a rediscovery of what we already knew, it is a very powerful rediscovery because it shows that even when the entire hypothesis space is tested, with all possible response profiles on equal footing, these three selectivities nonetheless emerge as the most robust. Put another way, this discovery indicates that the dominance of these response profiles in the ventral visual pathway is not due to the biases present in the way the hypothesis space has been sampled

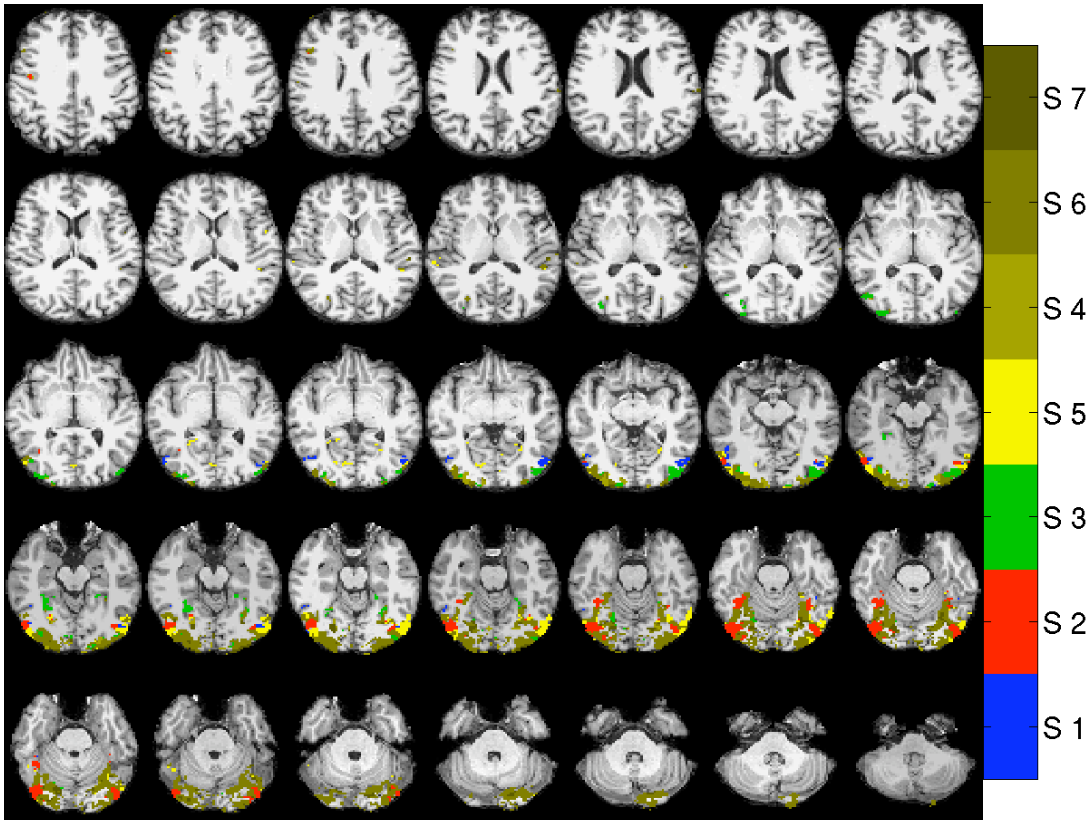


Figure 5.13. Significant systems projected back into the brain of Subject 1. As can be seen, Systems 4, 6, and 7 arise in posterior, largely retinotopic regions. Category-selective systems 1, 2, and 3 arise in their expected locations. System 5 appears to be a functionally and anatomically intermediate region between retinotopic and category-selective cortex. For analogous maps in all subjects, see Appendix C.

in the past, but instead reflects a fundamental property inherent to the ventral visual pathway.

In addition to finding face, place, and body selectivity, our clustering algorithm found four other significant systems. Three of these reflect low-level visual analyses conducted in retinotopic cortex, as evidenced both by the similarity of their response profiles to the profiles arising from retinotopic cortex, and by the anatomical location where these voxels are found – in posterior occipital cortex. Can our analysis discover any new response profiles not predicted by prior work? Indeed, one significant system (System 5) revealed a new selectivity profile that was not predicted, and that does not strongly resemble any of the profiles originating in retinotopic cortex. But the unique ability of our method to discover novel, unpredicted response profiles also raises the biggest challenge for future research: what can we say about any new response profiles we discover, if (as for System 5), they do not lend themselves to any straightforward functional hypothesis? A first question of course is whether such novel profiles will replicate in future work. If they do, their functional significance can be investigated by probing with new stimuli to test the generality and specificity of the response of these systems.

A further result of our work is to show that category selectivities in the ventral pathway cluster spatially at the grain of multiple voxels. Although this result is familiar from many prior studies, the methods used in those studies generally built in assumptions of spatial clustering – over and above smoothing during preprocessing – either explicitly (e.g., with cluster size thresholds) or implicitly (because discontiguous and scattered activations are usually discounted as noise). In contrast, our analysis was conducted without any information about the location of the voxels (see also Kriegeskorte et al., 2008), yet the resulting functional systems it discovered, when projected back into the brain, are clustered in spatially contiguous regions (see Figures 5.13 and C.1–C.44). Because the spatial clustering of these regions was not assumed either implicitly or explicitly in our analysis, the fact that those voxels are indeed spatially clustered reflects a new result.

Important caveats remain. First, although our method avoids many of the assumptions underlying conventional contrast-driven fMRI analysis, we cannot eliminate the basic experimental choice of stimuli to be tested. The set of stimuli in our experiment was designed to include images drawn from potentially novel categories as well as previously hypothesized categories, which allowed us to simultaneously validate the method on previously characterized functionally selective regions, and to potentially discover new selectivities. Moreover, although we included images of some plausible categories, we sampled each category representatively, rather than over-representing images from any one category (such as faces and animals, see Kriegeskorte et al., 2008). Nonetheless, for a completely unbiased approach, one would need to present images that were chosen independently of prior hypotheses (e.g., a representative sampling of ecologically relevant stimuli). A related caveat concerning the present results is that we do not know what other functionally defined systems may exist whose diagnostic responses concern stimuli not sampled in our experiment, and whether those systems may prove more robust than those discovered in the present analysis. Ongoing work will attempt to address these concerns by applying our methods to data obtained from a larger number of stimuli, sampled in a hypothesis-neutral fashion.

A third caveat in this work is that our analysis searches for functional profiles characterizing a large number of voxels each; therefore, we cannot rule out the hypothesis that many voxels in the ventral stream may contain idiosyncratic functional profiles, each characterizing only a small number of voxels. Thus, the fact that our analysis discovers category selectivities as the most robust profiles does not preclude the possibility that the ventral pathway also contains a large number of other voxels, each with a unique but perhaps less selective profile of response over a large number of images. Such additional voxels could collectively form a distributed code for object identity or shape, represented as a particular pattern of responses over a large set of voxels, each with a slightly different profile of response (Haxby et al., 2001). The image categories implied by such distributed codes can be assessed by methods that cluster images by the similarity of their neural response (as opposed to our approach of clustering voxels by the similarity of their responses to images). These stimulus-clustering approaches have yielded stimulus categories roughly consistent with the category-selective systems we find: a distinction between animate and inanimate images, and a further distinction between faces and bodies (Kriegeskorte et al., 2008), suggesting that the image categories defined over the whole ventral visual stream are dominated by the few category-selective areas we report here.

Despite these caveats, the current study has made important progress. Specifically, we found that even when the standard assumptions built into most imaging studies are eliminated (spatial contiguity and spatial similarity across subjects of voxels with similar functional profiles), and even when we give an equal shot to the vast number of all possible functional response profiles over the stimuli tested, we still find that selectivities for faces, places, and bodies emerge as the most robust profiles in the ventral visual pathway. Our discovery indicates that the prominence of these categories in the neuroimaging literature does not simply reflect biases in the hypotheses neuroscientists have thought to test, but rather that these categories are indeed special in the brain. Future research will conduct even more stringent tests of the dominance of these category selectivities, by testing each subject on a larger number of stimuli selected in a completely hypothesis-neutral fashion. This approach will be even more exciting if it enables us to discover new response profiles in the ventral visual pathway that were not previously known from more conventional methods.

Chapter 6

Nonparametric Bayesian Model for Group Functional Clustering

THIS chapter presents a unified model that simultaneously performs the two separate stages of the analysis in Chapter 4, namely, the estimation of selectivity profiles from time courses and the subsequent estimation of system profiles and maps. Moreover, the model refines the assumptions about the group structure of the mixture distribution by allowing variability in the size of systems across subjects, and also estimates the number of systems from data. Since all variables of interest are treated as latent random variables, the method yields posterior distributions that also encode uncertainty in the estimates.

We employ Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) to share structure across subjects. In our model, the structure shared across the group corresponds to grouping of voxels with similar functional responses. The nonparametric Bayesian aspect of HDPs allows automatic search in the space of models of different sizes.

Nonparametric Bayesian models have been previously employed in fMRI data analysis, particularly in modeling the spatial structure in the significance maps found by confirmatory analyses (Kim and Smyth, 2007; Thirion et al., 2007c). The probabilistic model introduced in this chapter is more closely related to recent applications of HDPs to DTI data where anatomical connectivity profiles of voxels are clustered across subjects (Jbabdi et al., 2009; Wang et al., 2009). In contrast to prior methods that apply stochastic sampling for inference, we take advantage of a variational scheme that is known to have faster convergence rate and greatly improves the speed of the resulting algorithm (Teh et al., 2008).

The approach introduced in this chapter also includes a model for fMRI signals that removes the need for the heuristic normalization procedure used in Section 4.1 as a preprocessing step. The model transforms fMRI responses into a set of binary activation variables. Therefore, the activation profile of a system can be naturally interpreted as a signature of functional specificity: it describes the probability that any stimulus or task activates a functional system. As before, this approach uses no spatial information other than the original smoothing of the data and therefore does not suffer from the drawbacks of voxel-wise spatial normalization.

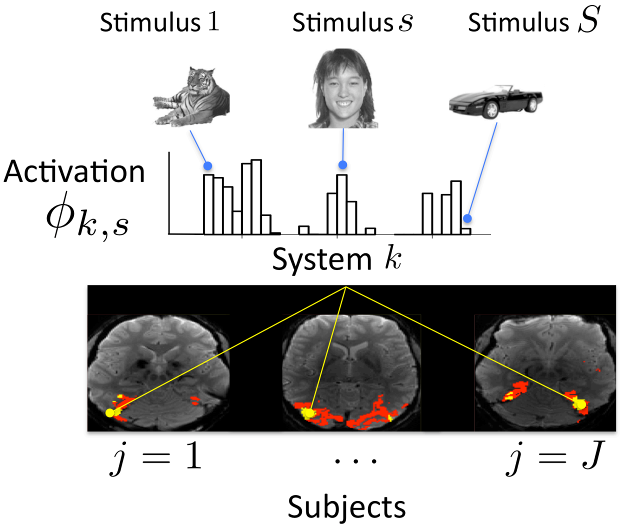


Figure 6.1. Schematic diagram illustrating the concept of a system. System k is characterized by vector $[\phi_{k1}, \cdots, \phi_{kS}]^t$ that specifies the level of activation induced in the system by each of the S stimuli. This system describes a pattern of response demonstrated by collections of voxels in all J subjects in the group.

This chapter is organized as follows. We begin by describing the two main layers of the model, the fMRI signal model and the hierarchical clustering model in Section 6.1. Then, we describe a variational inference procedure for making inference on latent variables of the model in Section 6.1.3. In Section 6.2, we present the results of applying the algorithm to data from the study of Section 5.2 and compare them with the results found by tensorial group ICA (Beckmann and Smith, 2005) and the finite mixture-model clustering model of Chapter 4. Finally, Section 6.3 discusses different aspects of the results and concludes the chapter.[1]

■ 6.1 Joint Model for Group fMRI Data

Consider an fMRI experiment with a relatively large number of different tasks or stimuli, for instance, a design that presents S distinct images in an event-related visual study. We let $\boldsymbol{y}_{ji}$ be the acquired fMRI time course of voxel i in subject j. The goal of the analysis is to identify patterns of functional specificity, i.e., distinct profiles of response that appear consistently across subjects in a large number of voxels in the fMRI time courses $\{\boldsymbol{y}_{ji}\}$. As discussed in Chapter 4, once we represent the data as a set of selectivity profiles that can be directly compared across subjects, the problem can be cast as an instance of clustering. Following the terminology introduced in Section 4.1, we refer to a cluster of voxels with similar response profiles as a functional system.

[1] The work presented in this chapter has been carried on in collaboration with Ramesh Sridharan from the Computer Science and Artificial Intelligence Laboratory at MIT.

Figure 6.1 illustrates the idea of a system as a collection of voxels that share a specific functional profile across subjects. Our model characterizes the functional profile as a vector whose components express the probability that the system is activated by different stimuli in the experiment.

To define the generative process for fMRI data, we first consider an infinite number of group-level systems. System k is assigned a prior probability π_k of including any given voxel. While the vector $\boldsymbol{\pi}$ is infinite-dimensional, any finite number of draws from the corresponding multinomial distribution will obviously yield a finite number of systems. To account for inter-subject variability and noise, we perturb the group-level system weight $\boldsymbol{\pi}$ independently for each subject j to generate a subject-specific weight vector $\boldsymbol{\beta}_j$. System k is further characterized by a vector $[\phi_{k1}, \cdots, \phi_{kS}]^t$, where $\phi_{ks} \in [0, 1]$ is the probability that system k is activated by stimulus s. Based on the weights $\boldsymbol{\beta}_j$ and the system probabilities $\boldsymbol{\phi}$, we generate binary activation variables $x_{jis} \in \{0, 1\}$ that express whether voxel i in subject j is activated by stimulus s.

So far, the model has the structure of a standard HDP. The next layer of this hierarchical model defines how activation variables x_{jis} generate observed fMRI signal values y_{jit}. If the voxel is activated ($x_{jis} = 1$), the corresponding fMRI response is characterized by a positive voxel-specific response magnitude a_{ji}; if the voxel is non-active, $x_{jis} = 0$, the response is assumed to be zero. The model otherwise follows the standard fMRI linear response model where the HRF is assumed to be variable across subjects and is estimated from the data.

Below, we present the details of the model starting with the lower level signal model to provide an intuition on the representation of the signal via activation vectors and then move on to describe the hierarchical clustering model. Table 6.1 presents the summary of all variables and parameters in the model; Figure 6.2 shows the structure of our graphical model.

6.1.1 Model for fMRI Signals

Using the same linear model for fMRI signal as in Section 2.1.1, we model measured signal $\boldsymbol{y}_{ji}$ of voxel i in subject j as a linear combination[2]

$$\boldsymbol{y}_{ji} = \boldsymbol{G}_j \boldsymbol{b}_{ji} + \boldsymbol{F}_j \boldsymbol{e}_{ji} + \boldsymbol{\epsilon}_{ji}, \tag{6.1}$$

where $\boldsymbol{G}_j$ and $\boldsymbol{F}_j$ are the stimulus and nuisance components of the design matrix for subject j, respectively, and $\boldsymbol{\epsilon}_{ji}$ is Gaussian noise. To facilitate our derivations, we rewrite this equation explicitly in terms of columns of the design matrix:

$$\boldsymbol{y}_{ji} = \sum_s b_{jis}\, \boldsymbol{g}_{js} + \sum_d e_{jid}\, \boldsymbol{f}_{jd} + \boldsymbol{\epsilon}_{ji}, \tag{6.2}$$

where $\boldsymbol{g}_{js}$ is the column of matrix $\boldsymbol{G}_j$ that corresponds to stimulus s and $\boldsymbol{f}_{jd}$ represents column d of matrix $\boldsymbol{F}_j$.

[2] See Section 2.1.1.

x_{jis}	binary activation of voxel i in subject j for stimulus s
z_{ji}	multinomial unit membership of voxel i in subject j
ϕ_{ks}	activation probability of system k for stimulus s
$\boldsymbol{\beta}_j$	system prior vector of weights in subject j
π_k	group-level prior weight for system k
α, γ	HDP scale parameters
y_{jit}	fMRI signal of voxel i in subject j at time t
e_{jid}	nuisance regressor d contribution to signal at voxel i in subject j
a_{ji}	amplitude of activation of voxel i in subject j
$\boldsymbol{\tau}_j$	a finite-time HRF vector in subject j
λ_{ji}	variance reciprocal of noise for voxel i in subject j
μ^a_j, σ^a_j	prior parameters for response amplitudes in subject j
$\mu^e_{jd}, \sigma^e_{jd}$	prior parameters for nuisance regressor d in subject j
$\omega^{\phi,1}, \omega^{\phi,2}$	prior parameters for actviation probabilities ϕ
κ_j, θ_j	prior parameters for noise variance in subject j

Table 6.1. Variables and parameters in the model.

As was explained in Section 4.1, fMRI signals do not have meaningful units and may vary greatly across trials and experiments. In order to use this data for inferences about brain function across subjects, sessions, and stimuli, we need to transform it into a standard and meaningful space. The binary *activation variables* $\boldsymbol{x}$ achieve this transformation by assuming that in response to any stimulus each voxel is either in an active or a non-active state. If voxel i in subject j is activated by stimulus s, i.e., if $x_{jis} = 1$, its response takes positive value a_{ji} that specifies a voxel-specific *amplitude of response*; otherwise, its response remains 0. Using this parameterization, $b_{jis} = a_{ji}x_{jis}$. The response amplitude a_{ji} represents uninteresting variability in fMRI signal due to physiological reasons unrelated to neural activity (examples include proximity of major blood vessels). The resulting binary activation variables for each voxel can represent the vector of selectivity profile, as was defined in Section 4.1.

As before, we have $\boldsymbol{g}_{js} = \Omega_{js} * \boldsymbol{\tau}_j$ where $\Omega_{js} \in \mathbb{R}^T$ is a binary indicator vector that shows whether stimulus $s \in \mathcal{S}$ is present during the experiment for subject j at each of the T acquisition times, and $\boldsymbol{\tau}_j \in \mathbb{R}^L$ is is a finite-time vector characterization of the hemodynamic response function (HRF) in subject j.

In our estimation of brain responses in Chapter 4, we used a canonical shape for the HRF function, letting $\boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}$ for all subjects j. However, prior research has shown that while within-subject estimates of the HRF appear consistent, there may be considerable variability in the shape of this function across subjects (Aguirre et al., 1998b;

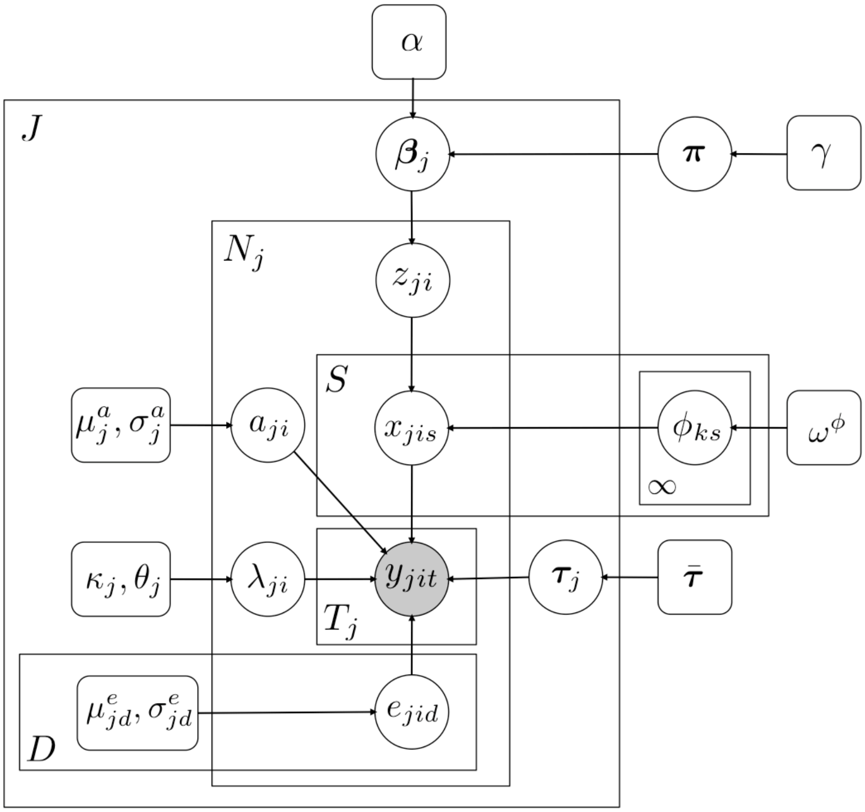


Figure 6.2. Full graphical model expressing all dependencies among different latent and observed variables in our model across subjects. Circles and squares indicate random variables and model parameters, respectively. Observed variables are denoted by a grey color. For a description of different variables in this figure, see Table 6.1.

Handwerker et al., 2004). Therefore, we consider the shape of the HRF for subject j, i.e., vector $\boldsymbol{\tau}_j$, to be an independent latent variable in our model in order to infer it from data. To simplify future derivations, we let $\boldsymbol{\Omega}_{js}$ be a $T \times L$ matrix defined based on Ω_{js} such that $\boldsymbol{g}_{js} = \boldsymbol{\Omega}_{js}\boldsymbol{\tau}_j$. Here, we assume an identical shape for the HRF within each subject since our application involves only the visual cortex. For studies that investigate responses of the entire brain or several cortical areas, we can generalize this model to include separate HRF variables for different areas (see, e.g., Makni et al., 2005).

With all the definitions above, our fMRI signal model becomes

$$\boldsymbol{y}_{ji} = a_{ji} \left(\sum_s x_{jis}\, \boldsymbol{\Omega}_{js} \right) \boldsymbol{\tau}_j + \sum_d e_{jid}\, \boldsymbol{f}_{jd} + \boldsymbol{\epsilon}_{ji}. \tag{6.3}$$

We use a simplifying assumption throughout that $\boldsymbol{\epsilon}_{ji} \overset{i.i.d.}{\sim} \text{Normal}(\mathbf{0}, \lambda_{ji}^{-1}\mathbf{I})$. In the application of this model to fMRI data, we first apply temporal filtering to the signal to decorrelate the noise in the preprocessing stage (Bullmore et al., 2001; Burock and Dale, 2000; Woolrich et al., 2001). An extension of the current model to include colored noise is possible, although it has been suggested that noise characteristics do not greatly impact the estimation of the HRF (Marrelec et al., 2002).

Priors

We assume a multivariate Gaussian prior for $\boldsymbol{\tau}_j$, with a covariance structure that encourages temporal smoothness,

$$\boldsymbol{\tau}_j \sim \text{Normal}\left(\bar{\boldsymbol{\tau}}, \boldsymbol{\Lambda}^{-1}\right), \tag{6.4}$$

$$\boldsymbol{\Lambda} = \nu\mathbf{I} + \boldsymbol{\Delta}^t\boldsymbol{\Delta} \tag{6.5}$$

where we have defined:

$$\boldsymbol{\Delta} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}, \tag{6.6}$$

and $\bar{\boldsymbol{\tau}}$ is the canonical HRF. The definition of the precision matrix above yields a prior that involves terms of the form $\sum_{l=1}^{L-1}(\tau_l - \tau_{l+1})^2$, penalizing large differences between the values of the HRF at consecutive time points.

We assume the prior distributions on the remaining voxel response variables as follows. For the response magnitude, we assume

$$a_{ji} \sim \text{Normal}_+\left(\mu_j^a, \sigma_j^a\right), \tag{6.7}$$

where $\text{Normal}_+(\eta, \rho)$ is the conjugate prior defined as a normal distribution restricted to positive real values:

$$p(a) \propto e^{-(a-\eta)^2/2\rho}, \text{ for } a \geq 0. \tag{6.8}$$

Positivity of the variable a_{ji} simply reflects the constraint that the expected value of fMRI response in the active state is greater than the expected value of response in the non-active state. For the nuisance factors, we let

$$e_{jid} \sim \text{Normal}\left(\mu_{jd}^e, \sigma_{jd}^e\right), \tag{6.9}$$

where $\text{Normal}(\eta, \rho)$ is a Gaussian distribution with mean η and variance ρ. Finally, for the noise precision parameter, we assume

$$\lambda_{ji} \sim \text{Gamma}\left(\kappa_j, \theta_j\right), \tag{6.10}$$

where Gamma (κ, θ) is a Gamma distribution parametrized by shape parameter κ and scale parameter θ^{-1}:

$$p(\lambda) = \frac{1}{\theta^{-\kappa}\Gamma(\kappa)} \lambda^{\kappa-1} e^{-\theta\lambda}. \tag{6.11}$$

■ 6.1.2 Hierarchical Dirichlet Prior for Modeling System Variability across the Group

In the group analysis of Chapter 4, when discussing the mixture model of Equation (4.2), we assumed that the same set of parameters, component centers and weights, characterize the distribution of selectivity profiles in the entire group. Component centers describe the profiles of selectivity for different functional systems that appear across the population. Component weights, which represent the proportion of voxels in each subject that belongs to each system, may well show variability across subjects due to noise, or even actual differences in the size of functional systems. The model in this section explicitly expresses this variability, in addition to including a nonparametric structure that enables estimating the number of components.

Similar to standard mixture models, we define the distribution of a voxel activation variable x_{jis} by conditioning on the system membership $z_{ji} \in \{1, 2, \cdots\}$ of the voxel and on the system probabilities of activation for different stimuli $\phi = \{\phi_{ks}\}$:

$$x_{jis} \mid z_{ji}, \phi \overset{i.i.d.}{\sim} \text{Bernoulli}(\phi_{z_{ji}s}). \tag{6.12}$$

This model implies that all voxels within a system have the same probability of being activated by a particular stimulus s.

We place a Beta prior distribution on system-level activation probabilities ϕ:

$$\phi_{ks} \overset{i.i.d.}{\sim} \text{Beta}(\omega^{\phi,1}, \omega^{\phi,2}). \tag{6.13}$$

Parameters ω^{ϕ} control the overall proportion of activated voxels across all subjects. For instance, we can induce sparsity in the results by introducing bias towards 0, i.e., the non-active state, in the parameters of this distribution.

To capture variability in system weights, we assume:

$$z_{ji} \mid \boldsymbol{\beta}_j \overset{i.i.d.}{\sim} \text{Mult}(\boldsymbol{\beta}_j), \tag{6.14}$$

$$\boldsymbol{\beta}_j \mid \boldsymbol{\pi} \overset{i.i.d.}{\sim} \text{Dir}(\alpha\boldsymbol{\pi}), \tag{6.15}$$

where $\boldsymbol{\beta}_j$ is a vector of subject-specific system weights, generated by a Dirichlet distribution centered on the population-level system weights $\boldsymbol{\pi}$. The extent of variability in the size of different systems across subjects is controlled by the concentration parameter α of the Dirichlet distribution. Finally, we place a prior on the population-level weight vector $\boldsymbol{\pi}$ that allows an infinite number of components:

$$\boldsymbol{\pi} \mid \gamma \sim \text{GEM}(\gamma), \tag{6.16}$$

where GEM(γ) is a distribution over infinitely long vectors $\pi = [\pi_1, \pi_2, \cdots]^t$, named after Griffiths, Engen and McCloskey (Pitman, 2002). Specifically,

$$\pi_k = v_k \prod_{k'=1}^{k-1} (1 - v_{k'}),$$
$$v_k \mid \gamma \overset{i.i.d.}{\sim} \text{Beta}(1, \gamma). \tag{6.17}$$

It can be shown that the components of the generated vectors π sum to 1 with probability 1. With this prior over system memberships $\boldsymbol{z} = \{z_{ji}\}$, the model in principle allows for an infinite number of functional systems; however, for any finite set of voxels, a finite number of systems is sufficient to include all voxels.

This prior for activation variables corresponds to the stick-breaking construction of HDPs (Teh et al., 2006), which is particularly suited for the variational inference scheme that we discuss in the next section.

6.1.3 Variational EM Inference

Having devised a full model for the fMRI measurements in a multi-stimulus experiment, we now provide a scheme for inference on the latent variables from the observed data. Sampling schemes are most commonly used for inference in HDPs (Teh et al., 2006). Despite theoretical guarantees of convergence to the true posterior, sampling techniques generally require a time-consuming burn-in phase. Because of the relatively large size of our problem, w will use a collapsed variational inference scheme for inference (Teh et al., 2008), which is known to yield faster algorithms. Here, we provide a brief overview of the derivation steps for the update rules. Appendix D contains the update rules and more detailed derivations.

To formulate the inference for system memberships, we integrate over the subject-specific unit weights $\boldsymbol{\beta} = \{\beta_j\}$ and introduce a set of auxiliary variables $\boldsymbol{r} = \{r_{jk}\}$ that represent the number of tables corresponding to system k in subject j according to the Chinese Restaurant Process formulation of HDP in (Teh et al., 2006). Appendix D provides some insights into the role of these auxiliary variables in our model; they allow us to find closed-form solutions for the inference update rules. We let $\boldsymbol{h} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{\phi}, \boldsymbol{\pi}, \boldsymbol{v}, \alpha, \gamma, \boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\tau}, \boldsymbol{\lambda}\}$ denote the set of all latent variables in our model. In the framework of variational inference, we approximate the model posterior on $\boldsymbol{h}$ given the observed data $p(\boldsymbol{h}|\boldsymbol{y})$ by a distribution $q(\boldsymbol{h})$. The approximation is performed through the minimization of the Gibbs free energy function:

$$\mathcal{F}[q] = E[\log q(\boldsymbol{h})] - E[\log p(\boldsymbol{y}, \boldsymbol{h})]. \tag{6.18}$$

Here, and in the remainder of the paper, $E[\cdot]$ and $V[\cdot]$ indicate expected value and variance with respect to distribution q. We assume a distribution q of the form:

$$q(\boldsymbol{h}) = q(\boldsymbol{r}|\boldsymbol{z}) \left(\prod_k q(v_k) \right) \cdot \left(\prod_{k,s} q(\phi_{ks}) \right)$$

$$\cdot \prod_j \left\{ q(\boldsymbol{\tau}_j) \prod_i \left[q(a_{ji}) q(\lambda_{ji}) q(z_{ji}) \left(\prod_s q(x_{jis}) \right) \left(\prod_d q(e_{jid}) \right) \right] \right\}, \quad (6.19)$$

where we explicitly account for the dependency of the auxiliary variables $\boldsymbol{r}$ on the system memberships $\boldsymbol{z}$. Including this structure maintains the quality of the approximation despite the introduction of the auxiliary variables (Teh et al., 2007). We use coordinate descent to solve the resulting optimization problem. Minimizing the Gibbs free energy function in terms of each component of $q(\boldsymbol{h})$ while fixing all other parameters leads to closed form update rules, provided in Appendix D.

Iterative application of the update rules leads to a local minimum of the Gibbs free energy. Since variational solutions are known to be biased toward their initial configurations, the initialization phase becomes critical to the quality of the results. We can initialize the variables in the fMRI signal model by ignoring higher level structure of the model and separately fitting the linear model of Equation (6.3) to the observed signal in each subject, starting with the canonical form of the HRF. We initialize $\boldsymbol{e}$ directly to the values of the nuisance regressor coefficients obtained via least squares estimation. Similarly, we initialize noise precision parameter $\boldsymbol{\lambda}$ to values based on the estimated residuals. Variables $\boldsymbol{a}$ and $\boldsymbol{x}$ are more difficult to initialize from the estimates of the linear model. The model produces coefficients b_{jis}, which we assume can be factored as $b_{jis} = a_{ji} x_{jis}$. Therefore, we initialize $E[a_{ji}] = \max_s b_{jis}$, and similarly $E[x_{jis}] = \left(b_{jis} - \min_s b_{jis}\right) / \left(\max_s b_{jis} - \min_s b_{jis}\right)$, where we subtract the minimum value to account for negative regression coefficient. This estimate serves merely as an initialization from which the model improves.

The update rules for each variable usually depend only on the previous values of other variables in the model. The exception to this is the update for $q(x_{jis})$, which also depends on previous estimates of $\boldsymbol{x}$. Therefore, unless we begin by updating $\boldsymbol{x}$, the first variable to be updated does not need to be initialized. Due to the coupling of the initializations for $\boldsymbol{x}$ and $\boldsymbol{a}$, we can choose to initialize either one of them first and update the other next. By performing both variants and choosing the one that provides the lower free energy after convergence, we further improve the search in the space of possible initializations and the quality of the resulting estimates.

To initialize system memberships, we introduce voxels one by one in a random order to the collapsed Gibbs sampling scheme (Teh et al., 2006) constructed for our model with each stimulus as a separate category and the initial $\boldsymbol{x}$ assumed known. Finally, we set the hyperparameters of the fMRI model to match the corresponding statistics computed via the least squares regression from the fMRI data.

■ 6.2 Results

This section presents the results of applying our method to the data from an event-related visual fMRI experiment. We compare the results of our hierarchical Bayesian method with the finite mixture model (Lashkari et al., 2010b) and the group tensorial

$$
\begin{aligned}
&\tilde{\omega}^{r}_{jk} = \exp\left(E[\log\alpha] + E[\log v_k] + \textstyle\sum_{k'<k} E[\log(1-v_{k'})]\right)\\
&E[r_{jk}] = \tilde{\omega}^{r}_{jk} E_{\boldsymbol{z}}[\Psi(\tilde{\omega}^{r}_{jk} + n_{jk}) - \Psi(\tilde{\omega}^{r}_{jk})]
\end{aligned}
$$

$$
\begin{aligned}
&v_k \sim \mathrm{Beta}(\tilde{\omega}^{v,1}_{k}, \tilde{\omega}^{v,2}_{k})\\
&\tilde{\omega}^{v,1}_{k} = 1 + \textstyle\sum_j E[r_{jk}]\\
&\tilde{\omega}^{v,2}_{k} = E[\gamma] + \textstyle\sum_{j,k'>k} E[r_{jk'}]
\end{aligned}
$$

$$
\begin{aligned}
&\phi_{k,s} \sim \mathrm{Beta}(\tilde{\omega}^{\phi,1}_{k,s}, \tilde{\omega}^{\phi,2}_{k,s})\\
&\tilde{\omega}^{\phi,1}_{k,s} = \omega^{\phi,1} + \textstyle\sum_{j,i} q(z_{ji}=k)q(x_{jis}=1)\\
&\tilde{\omega}^{\phi,2}_{k,s} = \omega^{\phi,2} + \textstyle\sum_{j,i} q(z_{ji}=k)q(x_{jis}=0)
\end{aligned}
$$

$$
\begin{aligned}
&a_{ji} \sim \mathrm{Normal}\left(\tilde{\omega}^{a,1}_{ji}\,(\tilde{\omega}^{a,2}_{ji})^{-1}, (\tilde{\omega}^{a,2}_{ji})^{-1}\right)\\
&\tilde{\omega}^{a,2}_{ji} = (\sigma^{a}_{j})^{-1} + E[\lambda_{ji}] \textstyle\sum_{s,s'} E[x_{jis}x_{jis'}]\mathrm{Tr}\left(E[\boldsymbol{\tau}_j\boldsymbol{\tau}_j^{t}]\boldsymbol{\Omega}^{t}_{js}\boldsymbol{\Omega}_{js'}\right)\\
&\tilde{\omega}^{a,1}_{ji} = \mu^{a}_{j}\left(\sigma^{a}_{j}\right)^{-1} + E[\lambda_{ji}] \textstyle\sum_{s} E[x_{jis}]E[\boldsymbol{\tau}_j]^{t}\boldsymbol{\Omega}^{t}_{js}(\boldsymbol{y}_{ji} - \textstyle\sum_d E[e_{jid}]\boldsymbol{f}_{jd})
\end{aligned}
$$

$$
\begin{aligned}
&\lambda_{ji} \sim \mathrm{Gamma}(\tilde{\omega}^{\lambda,1}_{ji}, \tilde{\omega}^{\lambda,2}_{ji})\\
&\tilde{\omega}^{\lambda,1}_{ji} = \kappa_j + \tfrac{T_j}{2}\\
&\tilde{\omega}^{\lambda,2}_{ji} = \theta_j + \|\boldsymbol{y}_{ji}\|^2 + \textstyle\sum_d \left(E[e^2_{jid}]\|\boldsymbol{f}_{jd}\|^2 + \textstyle\sum_{d'\neq d} E[e_{jid}]E[e_{jid'}]\boldsymbol{f}^{t}_{jd}\boldsymbol{f}_{jd'}\right)\\
&\qquad\quad + E[a^2_{jim}] \textstyle\sum_{s,s'} E[x_{jis}x_{jis'}]\mathrm{Tr}\left(E[\boldsymbol{\tau}_j\boldsymbol{\tau}_j^{t}]\boldsymbol{\Omega}^{t}_{js}\boldsymbol{\Omega}_{js'}\right) + E[a_{ji}] \textstyle\sum_{s,d} E[e_{jid}]E[x_{jis}]\boldsymbol{f}^{t}_{jd}\boldsymbol{\Omega}_{js}E[\boldsymbol{\tau}_j]\\
&\qquad\quad - \boldsymbol{y}^{t}_{ji}\left(\textstyle\sum_d E[e_{jid}]\boldsymbol{f}_{jd} + E[a_{ji}] \textstyle\sum_s E[x_{jis}]\boldsymbol{\Omega}_{js}E[\boldsymbol{\tau}_j]\right)
\end{aligned}
$$

$$
\begin{aligned}
&e_{jid} \sim \mathrm{Normal}\left(\tilde{\omega}^{e,1}_{jid}\,(\tilde{\omega}^{e,2}_{jid})^{-1}, (\tilde{\omega}^{e,2}_{jid})^{-1}\right)\\
&\tilde{\omega}^{e,2}_{jid} = (\sigma^{e}_{jd})^{-1} + E[\lambda_{ji}]\|\boldsymbol{f}_{jd}\|^2\\
&\tilde{\omega}^{e,1}_{jid} = \mu^{e}_{jd}(\sigma^{e}_{jd})^{-1} + E[\lambda_{ji}]\boldsymbol{f}^{t}_{jd}\left(\boldsymbol{y}_{ji} - \textstyle\sum_{d'\neq d} E[e_{jid'}]\boldsymbol{f}_{jd'} - E[a_{ji}] \textstyle\sum_s E[x_{jis}]\boldsymbol{\Omega}_{js}E[\boldsymbol{\tau}_j]\right)
\end{aligned}
$$

$$
\begin{aligned}
&\tau_j \sim \mathrm{Normal}(\Xi_j^{-1}\tilde{\boldsymbol{\omega}}^{\tau}_{j}, \Xi_j^{-1})\\
&\Xi_j = \boldsymbol{\Lambda} + \textstyle\sum_i E[\lambda_{ji}] \textstyle\sum_{s,s'} E[x_{jis}x_{jis'}]\boldsymbol{\Omega}^{t}_{js'}\boldsymbol{\Omega}_{js}\\
&\tilde{\boldsymbol{\omega}}^{\tau}_{j} = \boldsymbol{\Lambda}\bar{\boldsymbol{\tau}} + \textstyle\sum_{i,s} E[\lambda_{ji}]E[a_{ji}]E[x_{jis}]\boldsymbol{\Omega}^{t}_{js}\left(\boldsymbol{y}_{ji} - \textstyle\sum_d E[e_{jid}]\boldsymbol{f}_{jd}\right)
\end{aligned}
$$

Table 6.2. Update rules for computing the posterior q over the unobserved variables.

ICA of (Beckmann and Smith, 2005). We use the data from the study described in Section 5.2.1. To illustrate the method of this chapter and to have an even number of subjects for the later robustness analysis, we remove subject number 10 from the analysis[3] and use the remaining 10 subjects in the data set. Figure 5.6 presents the stimuli used in this study.

6.2.1 Preprocessing and Implementation Details

The data was first motion corrected separately for the two sessions (Cox and Jesmanowicz, 1999) and spatially smoothed with a Gaussian kernel of 3mm width. We then registered the two sessions to the subject's native anatomical space (Greve and Fischl, 2009). We used FMRIB's Improved Linear Model (FILM) to prewhiten the acquired fMRI time courses before applying the linear model (Woolrich et al., 2001). We created a mask for the analysis in each subject using an omnibus F-test for any stimulus regressors to significantly explain the variations in the measured fMRI time course. This step essentially removed noisy voxels from the analysis and only retained areas that are relevant for the experimental protocol at hand. Since the goal of the analysis is to study high level functional specificity in the visual cortex, we further removed from the mask the set of voxels within early visual areas. Furthermore, we included the average time course of all voxels within early visual areas as a confound factor in the design matrix of Equation (6.3).

Our method works directly on the temporally filtered time courses of all voxels within the mask. We use $\alpha = 100$, $\gamma = 5$, and $\omega^{\phi,1} = \omega^{\phi,2} = 1$ for the nonparametric prior. For the signal model, we use $\nu = 100$, and set the remaining hyperparameters using maximum likelihood estimates from the data. We run the algorithms 20 times with random initializations and choose the solution that yields the lowest Gibbs free energy function.

6.2.2 Evaluation and Comparison

We compare our results with those of the finite mixture model of (Lashkari et al., 2010b) and group tensorial ICA (Beckmann and Smith, 2005).

When evaluating the finite mixture model, we apply the standard regression analysis to find regression coefficients for each stimulus at each voxel and use the resulting vectors as input for clustering. Like the hierarchical Bayesian model, this method is also initialized with 20 random set of parameters and the best solution in terms of log-likelihood is chosen as the final result.

In Chapter 4, we provided an approach to quantifying and validating the group consistency of each profile found based on the finite mixture model. We use this method to provide an ordering of the resulting systems in terms of their consistency scores. We define the consistency scores based on the correlation coefficients between the group-wise profiles with the selectivity profiles found in each subject. We first match group-wise profiles with the set of profiles found by the algorithm in each individual subject's data separately. We employ the Hungarian algorithm (Kuhn, 1955)

[3] The conventional significance maps detect very few voxels in this subject for all three standard category selectivity contrasts (see Appendix C).

to find the matching between the two sets of profiles that maximizes the sum of edge weights (correlation coefficients in this case)[4]. The consistency score for each system is then defined as the average correlation coefficient between the corresponding group-wise profile and its matched counterparts in different subjects. We choose $K = 15$ clusters and select systems whose consistency scores are significant at threshold $p = 10^{-3}$ based on the group-wise permutation test (see Section 4.3.2).

Group tensorial ICA requires spatial normalization of the functional data from different subjects to a common spatial template. We employ FMRIB's nonlinear image registration tool[5](FNIRT) to register the structural image from each subject to the MNI template (T1 image of MNI152). As an initialization for this registration, we use FMRIB's linear image registration tool[6] (FLIRT). We create a group mask for the ICA analysis defined as the union of the masks found for different subjects by the F-test procedure above. We use the Melodic[7] implementation of the group tensorial ICA provided within the FSL package. Since the experiment includes a different number of runs for each subject, we cannot directly apply the ICA algorithm to the time courses. Instead, we use vectors of estimated regression coefficients for the 69 stimuli at each voxel as the input to ICA.

As implemented in the Melodic package, the group tensorial ICA employs the automatic model selection algorithm of Minka (2001) to estimate the number of independent components (Beckmann and Smith, 2004).

ICA provides one group spatial map for each estimated component across the entire group. In contrast, our method yields subject-specific maps in each subject's native space. In order to summarize the maps found by our method in different subjects and compare them with their ICA counterparts, we apply the same spatial normalization described above to spatial maps of the discovered systems. We then average these normalized maps across subjects to produce a group summary of the results.

As mentioned earlier, for the finite mixture model, we use the consistency scores to order the systems. For the two other methods, we use similar measures that capture the variability across subjects to provide an ordering of the profiles for their visualization. In group tensorial ICA results, variable c_{jk} expresses the contribution of component k to the fMRI data in subject j. Similarly, variable $E[n_{jk}]$ in our results denotes the number of voxels in subject j assigned to system k. We define the consistency measure for system (component) k as the standard deviation of values of $E[n_{jk}]$ (or c_{jk}) across subjects when scaled to have unit average.

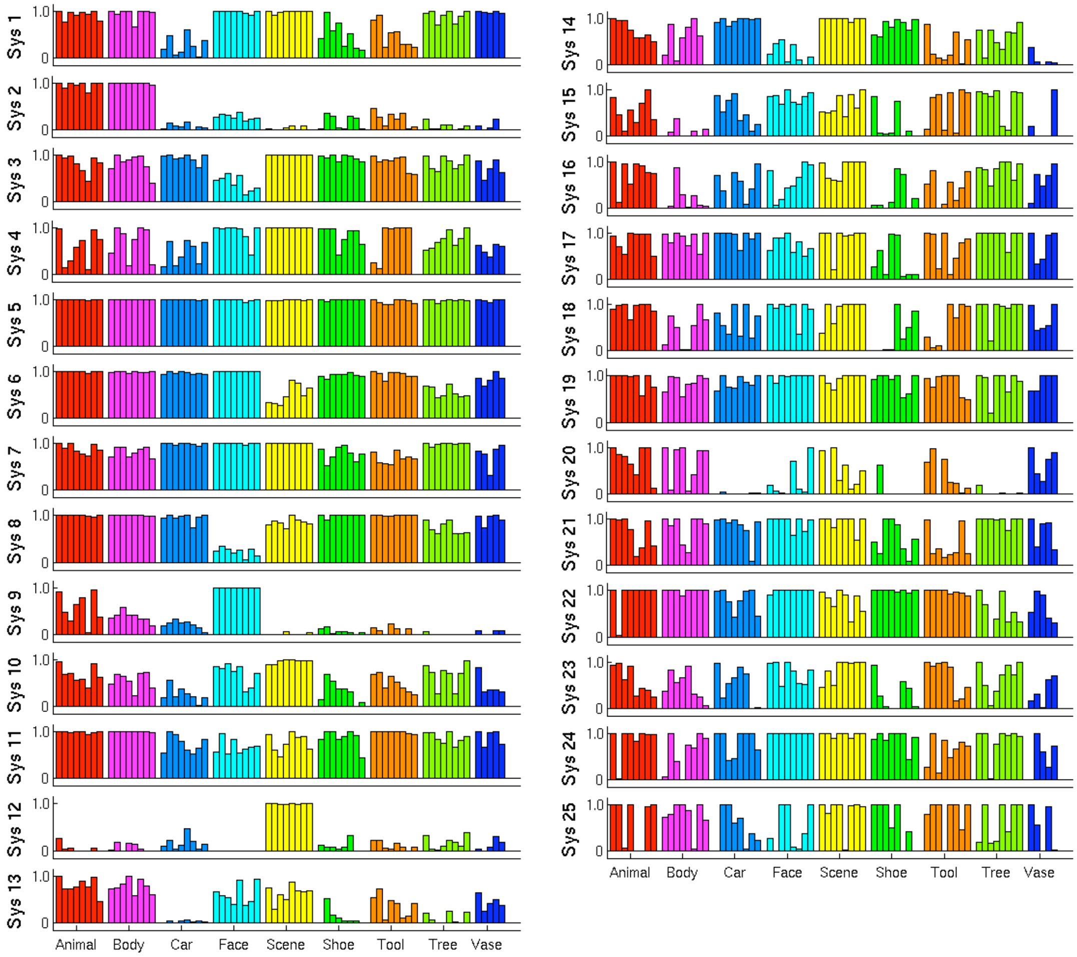


Figure 6.3. System profiles of posterior probabilities of activation for each system to different stimuli. The bar height correspond to the posterior probability of activation.

■ 6.2.3 System Functional Profiles

Our exploratory method aims to yield a characterization of likely patterns of functional specificity within the space spanned by the experimental protocol. The model expresses these patterns as activation probability profiles of functional systems extracted from data. For system k, this profile is described by the vector of posterior expected values $E[\phi_{ks}]$ for different stimuli s. In the data from ten subjects, the method

[4] We use the open source matlab implementation of the Hungarian algorithm available at http://www.mathworks.com/matlabcentral/fileexchange/11609.

[5] http://www.fmrib.ox.ac.uk/fsl/fnirt/index.html

[6] http://fsl.fmrib.ox.ac.uk/fsl/flirt/

[7] http://www.fmrib.ox.ac.uk/fsl/melodic/index.html

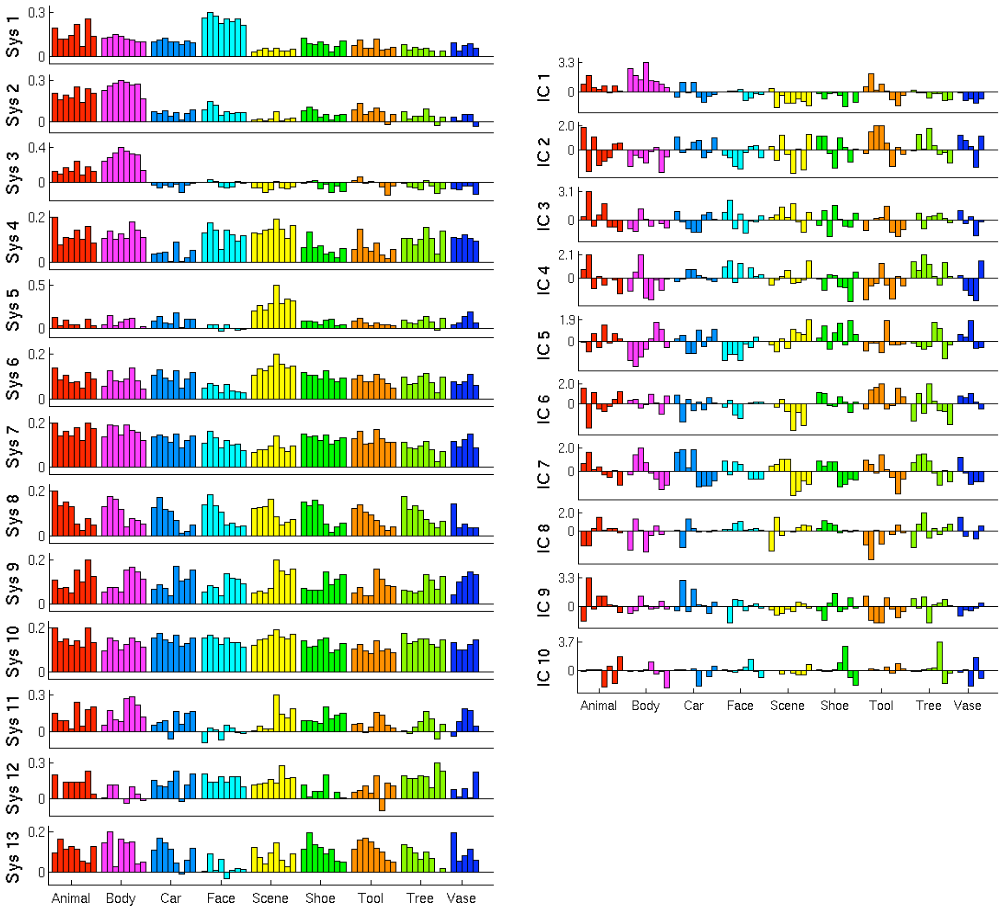


Figure 6.4. Left: system selectivity profiles estimated by the finite mixture of functional systems (Lashkari et al., 2010b). The bar height corresponds to the value of components of normalized selectivity profiles. Right: profiles of independent components found by the group tensorial ICA (Beckmann and Smith, 2005). The bar height corresponds to the value of the independent components. Both sets of profiles are defined in the space of the 69 stimuli.

finds 25 systems. Figure 6.3 presents the posterior activation probability profiles of these 25 functional systems in the space of the 69 stimuli presented in the experiment. We compare these profiles with the ones found by the finite mixture model and the group tensorial ICA, presented in Figure 6.4. ICA yields ten components. The profiles in Figure 6.3 and Figure 6.4 are presented in order of consistency. In the results from all methods, there are some systems or components that mainly contribute to the results of one or very few subjects and possibly reflect idiosyncratic characteristics of noise in those subjects.

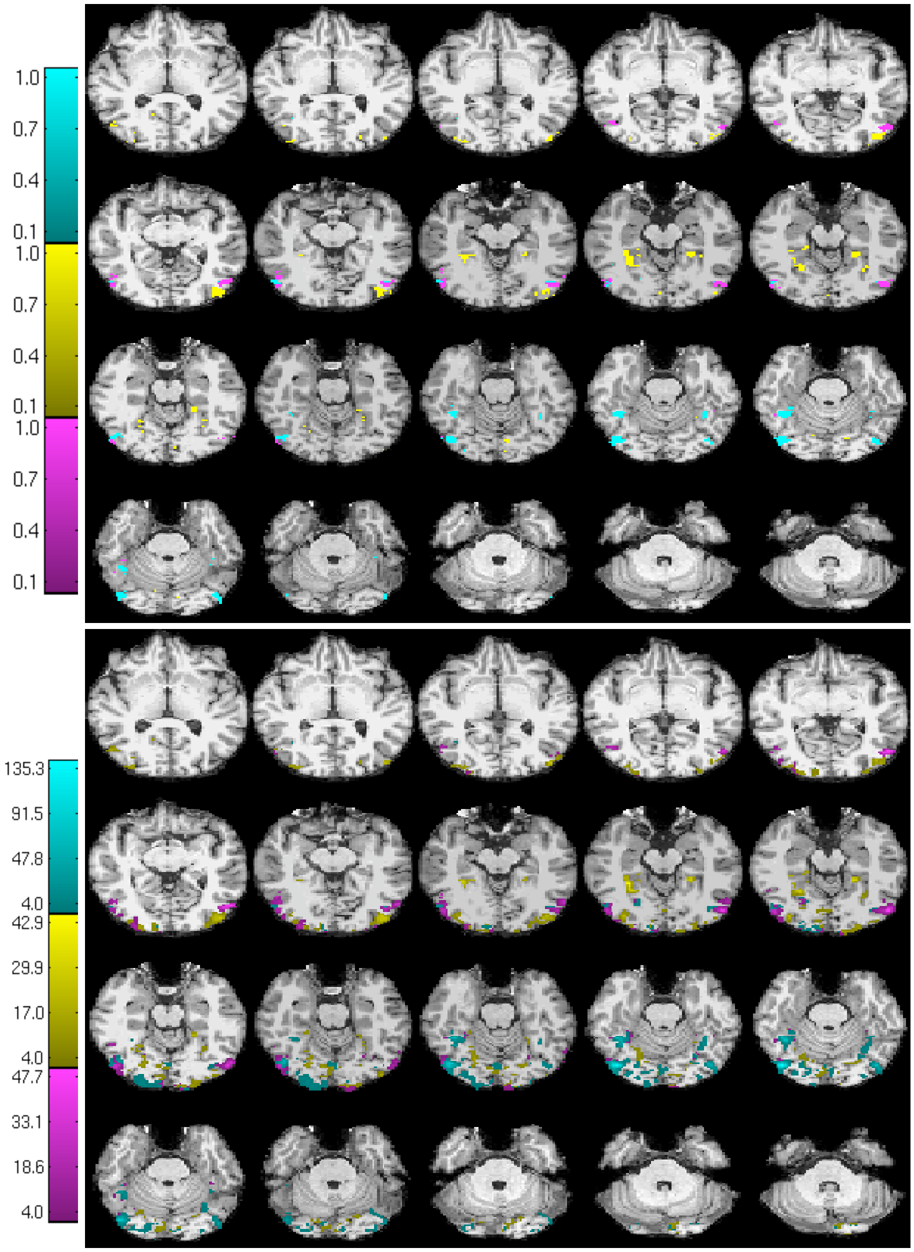


Figure 6.5. Top: membership probability maps corresponding to systems 2, 9, and 12, selective respectively for bodies (magenta), scenes (yellow), and faces (cyan) in one subject. Bottom: map representing significance values $-\log_{10} p$ for three contrasts bodies-objects (magenta), faces-objects (cyan), and scenes-objects (yellow) in the same subject.

We observe qualitatively that the category structure is more salient in the results of the model developed in this chapter. Most of our systems demonstrate similar prob-

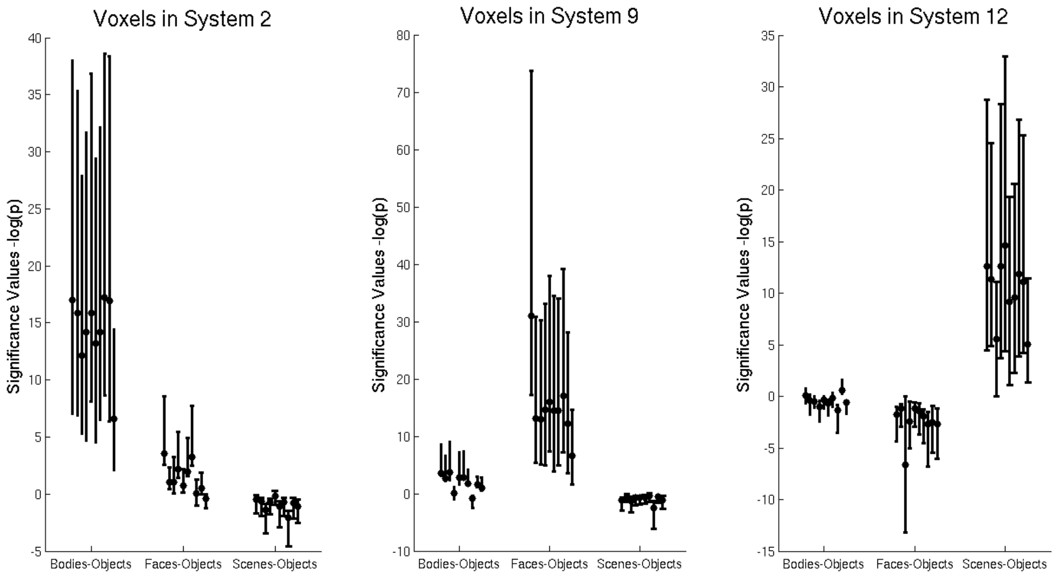


Figure 6.6. The distributions of significance values across voxels in systems 2, 9, and 12 for three different contrasts. For each system and each contrast, the plots report the distribution for each subject separately. The black circle indicates the mean significance value in the area; error bars correspond to 25th and 75th percentiles. Systems 2, 9, and 12 contain voxels with high significance values for bodies, faces, and scenes contrasts, respectively.

abilities of activation for images that belong to the same category. This structure is present to a lesser extent in the results of the finite mixture model, and is much weaker in the ICA results.

More specifically, we identify systems 2, 9, and 12 in Figure 6.3 as selective for categories of bodies, faces, and scenes, respectively (note that animals all have bodies). Among the system profiles ranked as more consistent, these profiles stand out by the sparsity in their activation probabilities. Figure 6.4 shows that similarly selective systems 1 (faces), 2 (bodies), 3 (bodies), and 5 (scenes) also appear in the results of the finite mixture model. The ICA results include only one component that seems somewhat category selective (component 1, bodies). As discussed before, previous studies have robustly localized areas such as EBA, FFA, and PPA with selectivities for the three categories above. Automatic detection of these profiles demonstrates the potential of our approach to discover novel patterns of specificity in the data.

Inspecting the activation profiles in Figure 6.3, we find other interesting patterns. For instance, the three non-face images with the highest probability of activating the face selective system 9 (animals 2, 5 and 7) correspond to the three animals that have large faces (Figure 5.6). Beyond the three known patterns of selectivity, we identify a number of other notable systems in the results of Figure 6.3. For instance, system 1

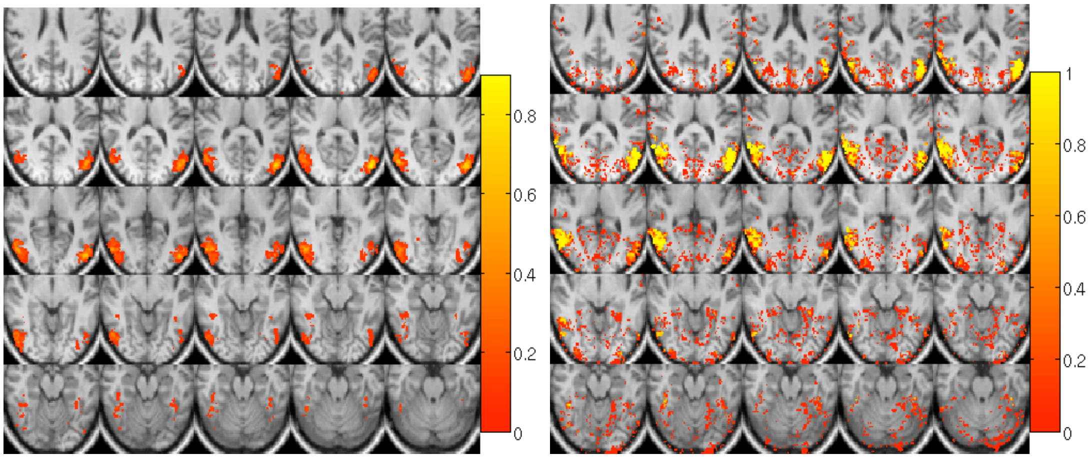


Figure 6.7. Left: the proportion of subjects with voxels in the body-selective system 2 at each location after nonlinear normalization to the MNI template. Right: the group probability map of the body-selective component 1 in the ICA results.

shows lower responses to cars, shoes, and tools compared to other stimuli. Since the images representing these three categories in our experiment are generally smaller in terms of pixel count, this system appears selective to lower level features (note that the highest probability of activation among shoes corresponds to the largest shoe 2). System 3 and system 8 seem less responsive to faces compared to all other stimuli. We investigate the spatial properties of these systems in the next section.

6.2.4 Spatial Maps

For each system k in our results, vector $\{q(z_{ji} = k)\}_{i=1}^{V_j}$ describes the posterior membership probability for all voxels in subject j. We can present these probabilities as a spatial map for the system in the subject's native space. Figure 6.5 (top) shows the membership maps for systems 2 (bodies), 9 (faces), and 12 (scenes). For comparison, Figure 6.5 (bottom) shows the significance maps found by applying the conventional confirmatory t-test to the data from the same subject. These maps present uncorrected significance values $-\log_{10}(p)$ for each of the three standard contrasts bodies-objects, faces-objects, and scenes-objects, thresholded at $p = 10^{-4}$ as is common practice in the field. While the significance maps appear to be generally spatially larger than the systems identified by our method, close inspection reveals that the system membership maps include the peak voxels for their corresponding contrasts. Figure 6.6 shows the fact that voxels within our system membership maps are generally associated with high significance values for the contrasts that correspond to their respective selectivity. The figure also clearly shows that there is considerable variability across subjects in the distribution of significance values.

Since our spatial maps are defined in each subject's native space, they yield a clear

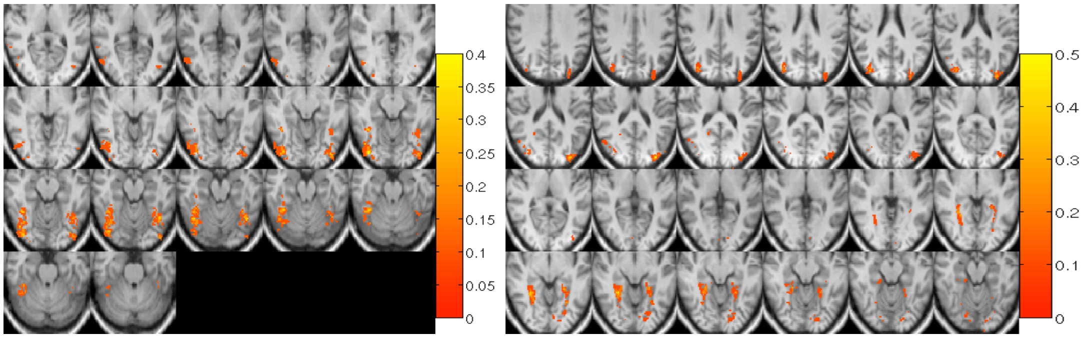


Figure 6.8. Group normalized maps for the face-selective system 9 (left), and the scene-selective system 12 (right).

picture of spatial variability of functionally specific areas. We can employ the normalization procedure described in Section 6.2.1 to summarize these results across subjects. Figure 6.7 compares the group-average of spatial maps for the body-selective system 2 with the group-level spatial map found by ICA. Although both maps cover the same approximate anatomical areas, our group map illustrates the limited voxel-wise overlap among areas associated with body-selectivity across subjects. In other words, the location of body-selective system 2 varies across subjects but generally remains at the same approximate area.

Figure 6.8 presents average normalized spatial maps for two other selective systems 9 and 12. These maps clearly contain previously identified category selective areas, such as FFA, OFA, PPA, TOS, and RSC (Epstein et al., 2007; Kanwisher and Yovel, 2006). We also examine the spatial map for system 1, which we demonstrated to be sensitive to low-level features. As Figure 6.9 (left) shows, this system resides mainly in the early visual areas. Figure 6.9 (right) shows the spatial map for system 8, which exhibits reduced activation to faces and shows a fairly consistent structure across subjects. To the best of our knowledge, selectivity similar to that of system 8 has not been reported in the literature so far.

■ 6.2.5 Reproducibility Analysis

One of the fundamental challenges in fMRI analysis is the absence of ground truth and the resulting difficulty in validation of results. In the previous two sections, we demonstrated the agreement of the results of our method with the previous findings on visual category selectivity. In this section, we validate our results based on their reproducibility across subjects. We split ten subjects into two groups of five and apply the analysis separately to each group. The method finds two sets of 17 and 23 systems in the two groups. Figure 6.10 shows the system profiles in both groups of subjects

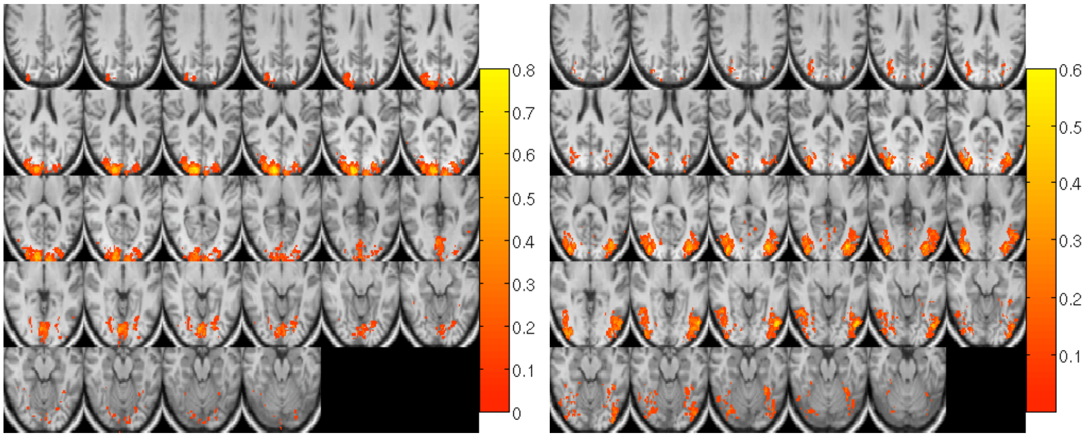


Figure 6.9. Group normalized maps for system 1 (left), and system 8 (right) across all 10 subjects.

matched[8] with the top 13 consistent profiles of Figure 6.3. Visual inspection of these activation profiles attests to the generalization of our results from one group of subjects to another. Figure 6.11 reports correlation coefficients for pairs of matched profiles from the two sets of subjects for all three methods: our hierarchical Bayesian method, the finite mixture-model, and the group tensorial ICA. This result suggests that, in terms of robustness across subjects, our unified model is more consistent than group tensorial ICA and is comparable to the finite mixture model. We note that due to the close similarity in the assumptions of the hierarchical Bayesian model and the finite mixture model, we do not expect a significant change in the robustness of the results comparing the former to the latter.

■ 6.3 Discussion

One of the crucial advantages of the model presented in this chapter over the finite mixture model of Chapter 4 is the nonparametric aspect that allows the estimation of the number of systems. As we saw in Section 5.1.2, system selectivity profiles found by the finite mixture model appear to be qualitatively consistent and robust for different number of systems K. Nevertheless, when it comes to selecting one set of selectivity profiles and their corresponding maps as the final outcome, there is no clear way how to choose one results over the other. For instance, while in the results presented in Figure 6.4 (left) found with $K = 15$ there are 13 systems with significant consistency across subjects at $p = 10^{-3}$, a similar analysis in the case of $K = 30$ yields 27 systems. On this data set, both basic model selection schemes such as BIC and computationally intensive resampling methods such as that of Lange et al. (2004) yield monotonically increasing measures for the goodness of the finite mixture model up to cluster numbers

[8] As before, we find the matching by solving a bipartite graph matching problem that maximizes the correlation coefficients between matched profiles.

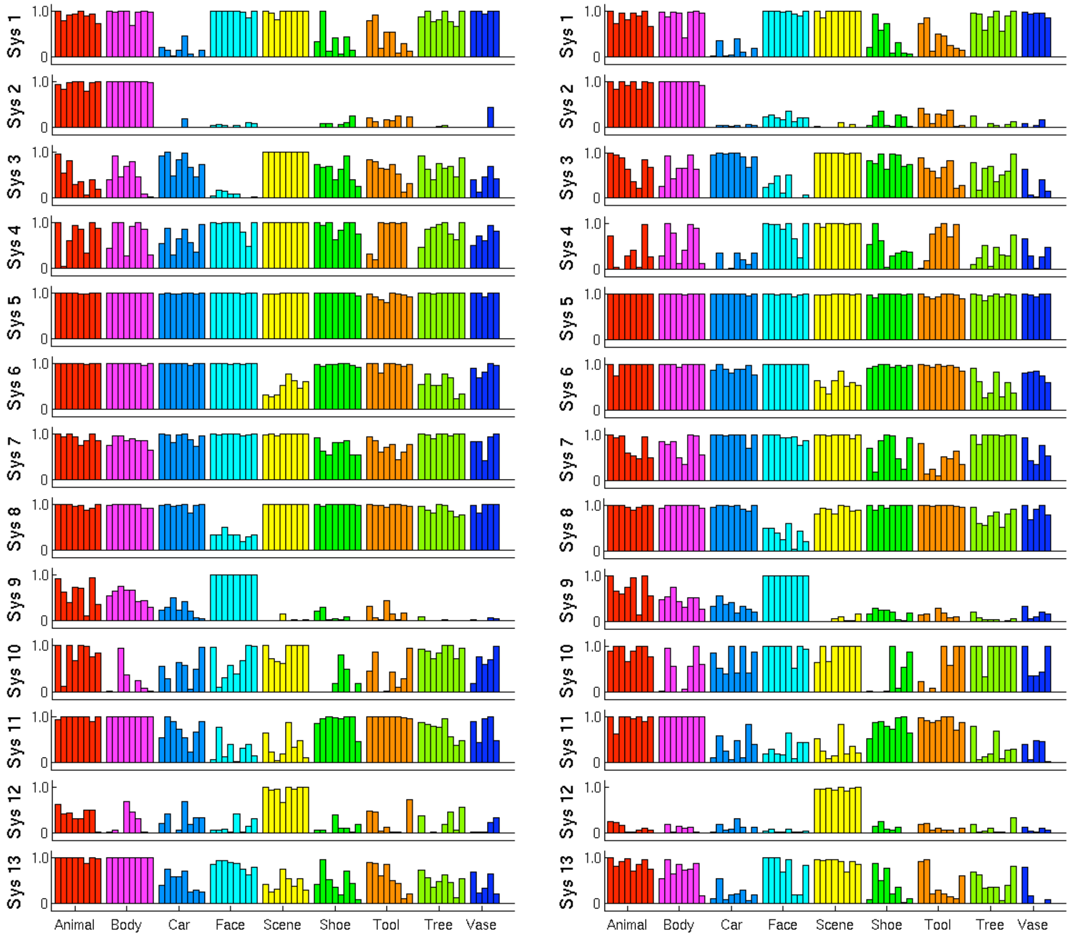


Figure 6.10. System profiles of activation probabilities found by applying the method to two independent sets of 5 subjects. The profiles were first matched across two groups using the scheme described in the text, and then matched with the system profiles found for the entire group (Figure 6.3).

of order several hundreds. In contradistinction, our nonparametric method automatically finds the number of components within the range expected based on the prior information. The estimates depend on the choice of HDP scale parameters α and γ. The results provide optimal choices within the neighborhood of model sizes allowed by these parameters. We discuss models that enable the estimation of the scale parameters elsewhere (Lashkari et al., 2010a).

Like the finite mixture model of Chapter 4, the proposed hierarchical Bayesian model avoids making assumptions about the spatial organization of functional systems across subjects. This is in contrast to group tensorial ICA, which assumes that independent components of interest are in voxel-wise correspondence across subjects. Average spatial maps presented in the previous section clearly demonstrate the ex-

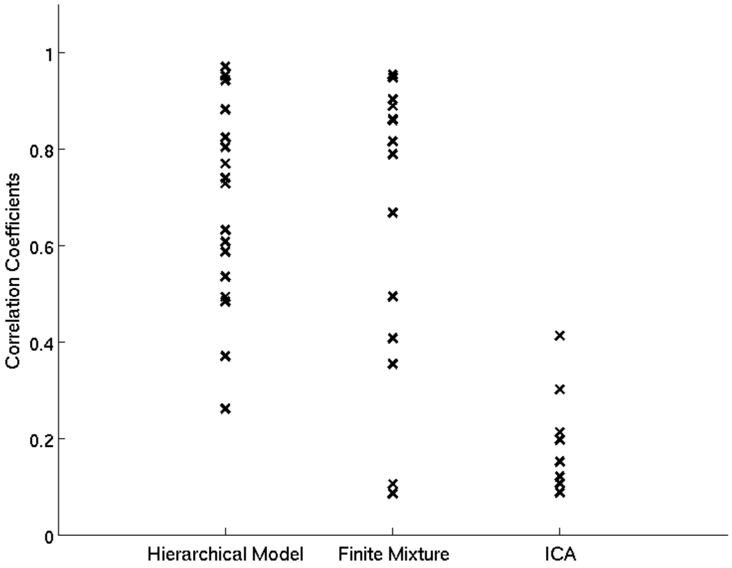


Figure 6.11. The correlation of profiles matched between the results found on the two separate sets of subjects for the three different techniques.

tent of spatial variability in functionally specific areas across subjects. This variability violates the underlying ICA assumption that independent spatial components are in perfect alignment across subjects after spatial normalization. Accordingly, ICA results are sensitive to the specifics of spatial normalization. In our experience, changing the parameters of registration algorithms can considerably alter the profiles of estimated independent components.

As mentioned earlier, Makni et al. (2005) have also employed an activation model similar to ours for expressing the relationship between fMRI activations and the measure BOLD signal. The most important distinction between the two models is that the amplitude of activations in the model of Makni et al. (2005) is assumed to be constant across all voxels. In contrast, we assume a voxel-specific response amplitude that allows us to extract activation variables as a relative measure of response in each voxel independent of the overall magnitude of the BOLD response.

A more subtle difference between the two models lies in the modeling of noise in time courses. Makni et al. (2005) assume two types of noise. First, they include the usual time course noise term ϵ_{jit} as in Equation (6.3). Moreover, they assume that the regression coefficients b_{is} are generated by Gaussian distribution whose mean is determined by whether or not voxel i is activated by stimulus s, i.e., the value of the activation variable x_{is}. This model assumes a second level of noise characterized by the uncertainty in the values of the regression coefficients conditioned on voxel activations. Our model is more parsimonious in that it does not assume any further uncertainty in brain responses conditioned on voxel activations and response amplitudes.

We emphasize the advantage of the activation profiles in our method over the cluster selectivity profiles of the finite mixture modeling in terms of *interpretability*. The elements of a system activation profile in our model represent the probabilities that different stimuli activate that system. Therefore, the brain response to stimulus s can be summarized based on our results in terms of a vector of activations $[\,E[\phi_{1s}], \cdots, E[\phi_{Ks}]\,]^t$ that it induces over the set of all functional systems. Such a representation cannot be made from the clustering profiles since the components of selectivity profiles do not have clear interpretation in terms of activation probability or magnitude.

Chapter 7

Conclusions

THIS thesis contributes to the study of functional specificity in the brain by providing a novel framework for design and analysis of fMRI studies. In our exploratory framework, the experimental design presents subjects with a broad range of stimuli/tasks related to the domain under study. Our analysis scheme then uses the resulting data to automatically search for patterns of selectivity to those experimental conditions that appear consistently across subjects. In Chapter 4, we presented a basic analysis method that relies on a finite mixture model and an effective heuristic for consistency analysis. In Chapter 6, we developed an improved method within the same framework that uses nonparametric hierarchical Bayesian techniques for modeling data in the group.

We tested our framework using two different fMRI studies of visual object recognition. Prior studies in this domain have characterized a number of regions along the ventral visual pathway associated with selectivity for categories such as faces, bodies, and scenes. In the first experiment, presented in Section 5.1, the blocked design included 16 image sets from 8 categories of visual stimuli. On the resulting data from a group of 6 subjects, our method discovered a number of patterns of category selectivity, the most consistent of which correspond to selectivity for faces, bodies, and scenes. In the second study, presented in Section 5.2, the event-related design included 69 distinct images from 9 categories. Even within this expanded space that did not explicitly encode the category structure, the most consistent selectivity profiles still agreed with the prior findings in the literature. Among the results of the analysis on this data set, both from the basic technique in Section 5.2 and the improved one in Section 6.2, we further found novel selectivity profiles that have not been characterized before.

Crucially, our framework avoids incorporating any anatomical information into the analysis. Brain locations are represented in the data only through their selectivity profiles. As a result, the method does not require establishing voxel-wise correspondence across subjects, in contrast to previously demonstrated techniques such as tensorial group ICA (Beckmann and Smith, 2005). There is in fact ample evidence for considerable variability in the location of functionally specific areas (see Chapter 3). Moreover, by ignoring spatial information our model in principle includes in its search possible networks of scattered voxels with the same pattern of selectivity.

In our studies of visual category selectivity, the resulting consistent functional systems all show contiguous spatial maps, despite the fact that our search does not constrain the systems spatially. On the one hand, this result is both surprising and reassuring. It is surprising to learn that using a more parsimonious set of assumptions about the brain function we can discover the same patterns of category selectivity known in the literature. It is also reassuring to confirm previous findings that relied on strong assumptions about contiguity and across-subject correspondence of category selective regions. On the other hand, this finding points toward a new direction for incorporating spatial information into our fMRI models. The spatial maps of our functional systems show considerable variability across subjects but are still located around the same anatomical landmarks, much like the maps found by significance tests (see Section 6.2 and Appendix C). This suggests a generalized fMRI model for functional systems that constrains activations to be in a parcel-wise, rather than voxel-wise, spatial correspondence across subjects. We emphasize however that the degree of this correspondence may strongly depend on the domain under study. Our conclusions derive exclusively from the case of visual category selectivity.[1]

Although the evidence presented in this thesis focuses mainly on the visual processing network, the analysis method is general and can be used in fMRI studies in many different domains of high level perception or cognition. In application to any new domain, we first design and perform an experiment that presents a wide variety of suitable stimlui/tasks to a number of subjects. Then, we identify which brain areas are likely to contribute to final results and include them as input to the analysis. Our method promises to discover consistent patterns of specificity in fMRI data from studies designed in this fashion. To mention a number of such applications, our collaborators have applied or are planning to apply the method to studies of processing of visual objects within the scene selective network, of nonverbal social perception in the superior temporal sulcus (STS), of moral transgressions, and of continuous visual-auditory perception (movies).

To design exploratory methods for fMRI analysis, this thesis relied on generative models that encompass explicit probabilistic assumptions connecting observed data to a number of latent random variables. One of the main advantages of this approach can be seen in the transition from the basic model of Chapter 4 to the complex model of Chapter 6. Arguably, alternative clustering algorithms may replace the finite mixture model used in Chapter 4 and yield similar results. Nevertheless, the simple mixture model setting provides the flexibility to augment the model with further latent variables to encode other known properties of the data. While such an improved model takes more known facts about the data into account, it still contains the same mixture modeling structure at its core. Consequently, in building a more accurate model for group fMRI data going from Chapter 4 to Chapter 6, our main assumptions about functional systems remain the same. The qualitative similarity in the results of the two chapters clearly attests to this fact.

[1] For a promising approach to incorporating the spatial information across the group see (Xu et al., 2009).

In summary, this thesis has attempted to employ state-of-the-art probabilistic modeling to better utilize group fMRI data for answering fundamental neuroscientific questions. The work provides a novel approach to the design and analysis of fMRI studies of functional specificity. Based on the presented evidence, we hope that researchers in different domains of neuroscience find the framework useful in their investigations of the functional organization of the brain.

Appendix A

Generative Models

This thesis chooses a generative modeling framework to design algorithms for the analysis of fMRI data. We devise graphical models that express my assumptions about the data as probabilistic relationships between a set of hidden (latent) variables and the observed data. More specifically, fMRI time courses, or brain responses estimated based on them, serve as the observed data and latent variables, for instance, correspond to system memberships, i.e., groupings of brain areas that demonstrate distinct patterns of functional specificity. Further assumptions about the data, such as voxel-level characteristics of fMRI responses and inter-subject variability, are represented by including additional hidden variables in the model. Broadly speaking, learning based on such models is achieved via inference on the set of hidden variables. Since the inference is inherently probabilistic, it allows us to represent the uncertainty in the learned structures.

There are many methodological advantages to using generative models in machine learning and pattern recognition. We can devise the model in a way that any hidden variable in the model explicitly represents a specific structure of interest in the data. *Graphical models* offer a systematic way for expressing and visualizing these structures. In this framework, it is straightforward to augment the model at any stage by adding new latent variables. A well-studied toolbox of inference techniques makes learning in resulting complex models possible (Koller and Friedman, 2009). In this thesis, I exploit the advantages of generative models through different stages of the work as we gradually include more components in the model, each capturing a different property of observed fMRI signals. Moreover, this approach accommodates different levels of supervision in learning: we can unify supervised, semi-supervised, and unsupervised learning within the same model by making different assumptions about the observability of the relevant latent variables.

■ A.1 Maximum Likelihood, the EM, and the Variational EM Algorithms

Let us denote the set of the data we intend to model by $\boldsymbol{y}$, and the set of all hidden variables by $\boldsymbol{h}$. The structure of interest is encoded in the probabilistic relationship between the data and the hidden variables expressed through a likelihood model $p_{\boldsymbol{y}|\boldsymbol{h}}(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta})$. This model is parametrized by a set of set of parameters $\boldsymbol{\theta}$. Assuming a prior distribution $p_{\boldsymbol{h}}(\boldsymbol{h};\boldsymbol{\theta})$ on the hidden variables completes the characterization of the joint model and allows us to find the marginal distribution of the observed data $p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\theta})$. The parameters sometimes act as unknown variables of interest that should be estimated from the data as well. The maximum likelihood (ML) estimator of the parameters:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \log p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \log \int_{\boldsymbol{h}} p_{\boldsymbol{y}|\boldsymbol{h}}(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta}) p_{\boldsymbol{h}}(\boldsymbol{h};\boldsymbol{\theta}) \tag{A.1}$$

is commonly used. Employing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), we define a distribution $q_{\boldsymbol{h}}$ on $\boldsymbol{h}$ that approximates $p_{\boldsymbol{h}|\boldsymbol{y}}$ through minimization of the Gibbs free energy:

$$\mathcal{F}[q_{\boldsymbol{h}},\boldsymbol{\theta}] = E_{q_{\boldsymbol{h}}} \log q_{\boldsymbol{h}}(\boldsymbol{h}) - E_{q_{\boldsymbol{h}}} \log p_{\boldsymbol{y},\boldsymbol{h}}(\boldsymbol{y},\boldsymbol{h};\boldsymbol{\theta}), \tag{A.2}$$

$$= KL\left[q_{\boldsymbol{h}}(\cdot) \| p_{\boldsymbol{h}|\boldsymbol{y}}(\cdot|\boldsymbol{y})\right] - \log p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\theta}), \tag{A.3}$$

where E_p denotes the expectation operator under the distribution p and $KL[q\|p]$ denotes the KL-divergence between distributions q and p. The first term in (A.3) is nonnegative; therefore, $\mathcal{F}$ yields an upper bound to the negative log-likelihood expression of Equation (A.1). We can solve the minimization problem by, first, letting $q_{\boldsymbol{h}}(\cdot) = p_{\boldsymbol{h}|\boldsymbol{y}}(\cdot|\boldsymbol{y})$ which makes the KL-divergence term vanish (E-step). Then, we minimize the second expression on the right hand side of Equation (A.2) with respect to $\boldsymbol{\theta}$ (M-step). We iteratively repeat the two steps until the guaranteed convergence to a local maximum of problem (A.1).

When the structure of the model is more complex, computing of the posterior distribution $p_{\boldsymbol{h}|\boldsymbol{y}}(\boldsymbol{h}|\boldsymbol{y})$ requires some approximations. The variational Bayes approximation (Jordan et al., 1999) solves this problem by adding constrains to the space of $q_{\boldsymbol{h}}$ and minimizing the Gibbs free energy $\mathcal{F}$. More specifically, we can constrain the problem to the set of distributions on the hidden variables where each variable is independent of the others. Variational approximations are usually much faster than the stochastic alternatives for large problems.

■ A.2 Bayesian Methods

In the context of Bayesian probabilistic modeling, any unknown variable is considered random, with the prior distribution encoding our initial uncertainty about the value of this variable. Making inference on these variables corresponds to computing the posterior distribution over the hidden variables conditioned on the observed data. The above ML estimates for the parameters could be understood, in this setup, as atomic approximations for the posterior distribution on these variables where the prior is assumed to be uniform. Apart from the philosophical debates about Bayesian approach (Bernardo and Smith, 1994), it has an important advantage in allowing joint model selection and learning through nonparametric Bayesian techniques such as Dirichlet Processes (Antoniak, 1974; Rasmussen, 2000; Teh et al., 2006).

Appendix B

Derivations for Chapter 4

B.1 Derivations of the Update Rules

We let $\Theta = \{\{w_k, \boldsymbol{m}_k\}_{k=1}^K, \zeta\}$ be the full set of parameters and derive the EM algorithm for maximizing the log-likelihood function

$$L(\Theta) = \sum_i^V \log p(\boldsymbol{x}_i; \Theta) \tag{B.1}$$

for a mixture model $p(\boldsymbol{x}; \Theta) = \sum_{k=1}^K w_k f(\boldsymbol{x}; \boldsymbol{m}_k, \zeta)$. The EM algorithm (Dempster et al., 1977) assumes a hidden random variable z_i that represents the assignment of each data point to its corresponding component in the model. This suggests a model in the joint space of observed and hidden variables:

$$p(\boldsymbol{x}_i, z_i = k; \Theta) = w_k f(\boldsymbol{x}_i; \boldsymbol{m}_k, \zeta), \tag{B.2}$$

where $k \in \{1, \cdots, K\}$, and the likelihood of observed data is simply

$$p(\boldsymbol{x}_i; \Theta) = \sum_{k=1}^K p(\boldsymbol{x}_i, k; \Theta). \tag{B.3}$$

.

With a given set of parameters $\Theta^{(t)}$ in step t, the E-step involves computing the posterior distribution of the hidden variable given the observed data. Since the data for each voxel is assumed to be an *i.i.d.* sample from the joint distribution (B.2), the posterior distribution for the assignment of all voxels can also be factored into terms for each voxel:

$$\begin{aligned} p^{(t)}(k|\boldsymbol{x}_i) \triangleq p(z_i = k|\boldsymbol{x}_i; \Theta^{(t)}) &= \frac{p(z_i = k, \boldsymbol{x}_i; \Theta^{(t)})}{p(\boldsymbol{x}_i; \Theta^{(t)})} \\ &= \frac{w_k^{(t)} f(\boldsymbol{x}_i; \boldsymbol{m}_k^{(t)}, \zeta^{(t)})}{\sum_{k'=1}^K w_{k'}^{(t)} f(\boldsymbol{x}_i; \boldsymbol{m}_{k'}^{(t)}, \zeta^{(t)})} \\ &= \frac{w_k^{(t)} e^{\zeta^{(t)} \langle \boldsymbol{x}_i, \boldsymbol{m}_k^{(t)} \rangle}}{\sum_{k'=1}^K w_{k'}^{(t)} e^{\zeta^{(t)} \langle \boldsymbol{x}_i, \boldsymbol{m}_{k'}^{(t)} \rangle}}. \end{aligned} \tag{B.4}$$

Using this distribution, we can express the target function of the M-step:

$$\begin{aligned}\tilde{L}(\Theta;\Theta^{(t)}) &= \sum_{i=1}^{V} E_{k|\boldsymbol{x}_i;\Theta^{(t)}}[\log p(\boldsymbol{x}_i, z_i = k;\Theta)] \\ &= \sum_{i=1}^{V}\sum_{k=1}^{K} p^{(t)}(k|\boldsymbol{x}_i)\log p(\boldsymbol{x}_i, z_i = k;\Theta) \\ &= \sum_{i=1}^{V}\sum_{k=1}^{K} p^{(t)}(k|\boldsymbol{x}_i)\log\left[\, w_k f(\boldsymbol{x}_i;\boldsymbol{m}_k,\zeta)\,\right] \\ &= \sum_{i=1}^{V}\sum_{k=1}^{K} p^{(t)}(k|\boldsymbol{x}_i)\left[\,\log w_k + \log C_S(\zeta) + \zeta\langle\boldsymbol{m}_k,\boldsymbol{x}_i\rangle\,\right]. \end{aligned} \tag{B.5}$$

Taking the derivative of this function along with the appropriate Lagrange multipliers yields the update rules for the model parameters in iteration $(t+1)$. For the cluster centers $\boldsymbol{m}_k$, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{m}_k}\left[\sum_{i=1}^{V}\sum_{k'=1}^{K} p^{(t)}(k'|\boldsymbol{x}_i)\langle\boldsymbol{m}_{k'},\boldsymbol{x}_i\rangle - \sum_{k'=1}^{K}\gamma_{k'}\Big(\langle\boldsymbol{m}_{k'},\boldsymbol{m}_{k'}\rangle - 1\Big)\right] \\ &= \sum_{i=1}^{V}\boldsymbol{x}_i p^{(t)}(k|\boldsymbol{x}_i) - 2\gamma_k\boldsymbol{m}_k, \end{aligned} \tag{B.6}$$

which implies the update rule $\boldsymbol{m}_k^{(t+1)} = \frac{1}{2\gamma_k}\sum_{i=1}^{V}\boldsymbol{x}_i p^{(t)}(k|\boldsymbol{x}_i)$. The Lagrange multiplier ensures that $\boldsymbol{m}_k$ is a unit vector, i.e.,

$$\gamma_k = \frac{1}{2}\|\sum_{i=1}^{V}\boldsymbol{x}_i p^{(t)}(k|\boldsymbol{x}_i)\|.$$

Similarly, we find the concentration parameter ζ:

$$\begin{aligned} 0 &= \frac{1}{V}\frac{\partial}{\partial\zeta}\left[V\log C_S(\zeta) + \zeta\sum_{i=1}^{V}\sum_{k=1}^{K} p^{(t)}(k|\boldsymbol{x}_i)\langle\boldsymbol{m}_k,\boldsymbol{x}_i\rangle\right], \\ &= \frac{\partial}{\partial\zeta}\left[(\tfrac{S}{2}-1)\log\zeta - \log I_{S/2-1}(\zeta) + \frac{\zeta}{V}\Gamma^{(t+1)}\right], \\ &= \frac{S/2-1}{\zeta} - \frac{I'_{S/2-1}(\zeta)}{I_{S/2-1}(\zeta)} + \frac{\Gamma^{(t+1)}}{V}, \\ &= -\frac{I_{S/2}(\zeta)}{I_{S/2-1}(\zeta)} + \frac{\Gamma^{(t+1)}}{V}, \end{aligned} \tag{B.7}$$

where we have substituted $\boldsymbol{m}_k^{(t+1)}$ in the first line, used the definition of Equation (4.10) in the second line, and the last equality follows from the properties of the modified Bessel functions. It follows then that $A_S(\zeta^{(t+1)}) = \Gamma^{(t+1)}$. Finally, for the cluster weights q_k, adding the Lagrange multiplier to guarantee that the weights sum to 1, we find

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_k}\left[\sum_{i=1}^{V}\sum_{k'=1}^{K} p^{(t)}(k'|\boldsymbol{x}_i)\log w_{k'} - \zeta\left(\sum_{k'=1}^{K} w_{k'} - 1\right)\right] \\ &= \frac{1}{w_k}\sum_{i=1}^{V} p^{(t)}(k|\boldsymbol{x}_i) - \zeta, \end{aligned} \tag{B.8}$$

which together with the normalization condition results in the update $w_k^{(t+1)} = \frac{1}{V}\sum_{i=1}^{V} p^{(t)}(k|\boldsymbol{x}_i)$.

B.2 Estimation of Concentration Parameter

In order to update the concentration parameter in the M-step using (4.9), we need to solve for ζ in the equation

$$A_S(\zeta) = \frac{I_{S/2+1}(\zeta)}{I_{S/2}(\zeta)} = \Gamma. \tag{B.9}$$

Figure B.1 shows the plot of function $A_S(\cdot)$ for several values of S. This function is smooth and monotonically increasing, taking values in the interval $[0,1)$. An approximate solution to (B.9) has been suggested in (Banerjee et al., 2006) but the proposed expression does not yield accurate values in our range of interest for S. Therefore, we derive a different approximation using the inequality

$$\frac{x}{\gamma+\frac{1}{2}+\sqrt{x^2+(\gamma+\frac{3}{2})^2}} \le \frac{I_{\gamma+1}(x)}{I_\gamma(x)} \le \frac{x}{\gamma+\frac{1}{2}+\sqrt{x^2+(\gamma+\frac{1}{2})^2}} \tag{B.10}$$

proved in (Amos, 1974). Defining $u = \frac{S-1}{2\zeta}$, it follows from Equation (B.9) and Inequality (B.10) that

$$\frac{1}{u+\sqrt{1+(1+\frac{2}{S-1})^2u^2}} \le \Gamma \le \frac{1}{u+\sqrt{1+u^2}}. \tag{B.11}$$

Due to continuity of $\left(u+\sqrt{1+\alpha^2u^2}\right)^{-1}$ as a function of $\alpha \ge 1$, this expression equals Γ for at least one value in the interval $1 \le \alpha \le 1+\frac{2}{S-1}$. For this value of α, we have

$$(\alpha^2-1)u^2+\tfrac{2}{\Gamma}u-(\tfrac{1}{\Gamma^2}-1)=0 \implies u = \frac{\sqrt{1+(\alpha^2-1)(1-\Gamma^2)}-1}{(\alpha^2-1)\Gamma}\ . \tag{B.12}$$

The expression for u is a monotonically decreasing function of α^2-1 where $0<\alpha^2-1\le\frac{4S}{(S-1)^2}$; therefore, we find

$$\frac{(S-1)^2}{4S\Gamma}\left(\sqrt{1+\frac{4S(1-\Gamma^2)}{(S-1)^2}}-1\right) \le u \le \frac{1-\Gamma^2}{2\Gamma}\ . \tag{B.13}$$

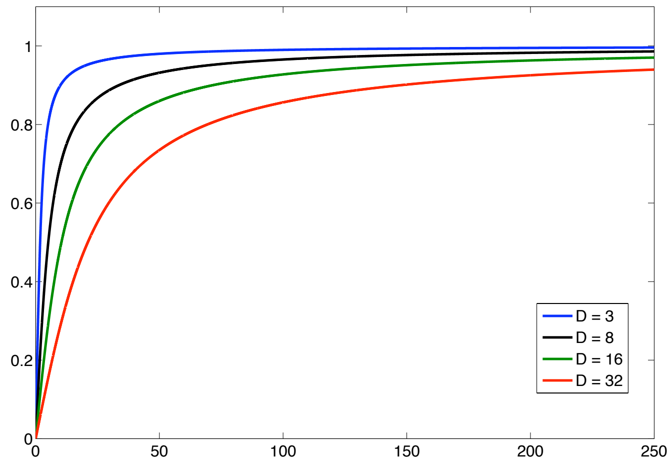


Figure B.1. Plot of function $A_S(\cdot)$ for different values of S.

Now, using the inequality $\sqrt{1+x} \geq 1 + \frac{1}{2}x - \frac{1}{8}x^2$, we find a simpler expression for the lower bound

$$\tfrac{1-\Gamma^2}{2\Gamma}\left(1 - \tfrac{S(1-\Gamma^2)}{(S-1)^2}\right) \;\leq\; u \;\leq\; \tfrac{1-\Gamma^2}{2\Gamma}\;. \tag{B.14}$$

Finally, the parameter can be bounded by

$$\tfrac{(S-1)\Gamma}{1-\Gamma^2} \;\leq\; \zeta \;\leq\; \tfrac{(S-1)\Gamma}{1-\Gamma^2}\left(1 - \tfrac{S(1-\Gamma^2)}{(S-1)^2}\right)^{-1}\;. \tag{B.15}$$

Because of the monotonicity of $A_S(\cdot)$, starting from the average of the two bounds and taking a few Newton steps towards zero of equation (B.9), we easily reach a good solution. However, when Γ is too close to 1 and, hence, ζ is large, evaluation of the function $A_S(\cdot)$ becomes challenging due to the exponential behavior of the Bessel functions. In this case, when Γ is large enough such that $\frac{S(1-\Gamma^2)}{(S-1)^2} \ll 1$ holds, we can approximate the second term in the upper bound of (B.15) as $\left(1 + \frac{S(1-\Gamma^2)}{(S-1)^2}\right)$, reaching the final approximation

$$\zeta \approx \tfrac{(S-1)\Gamma}{1-\Gamma^2} + \tfrac{S\Gamma}{2(S-1)}\;. \tag{B.16}$$

Appendix C

Spatial Maps for the Event-Related Experiment in Chapter 5

In this chapter of the appendix, we present the maps of the discovered functional systems, along with the significance maps for three different contrast bodies-objects, faces-objects, and scenes-objects for all 11 subjects in the study.

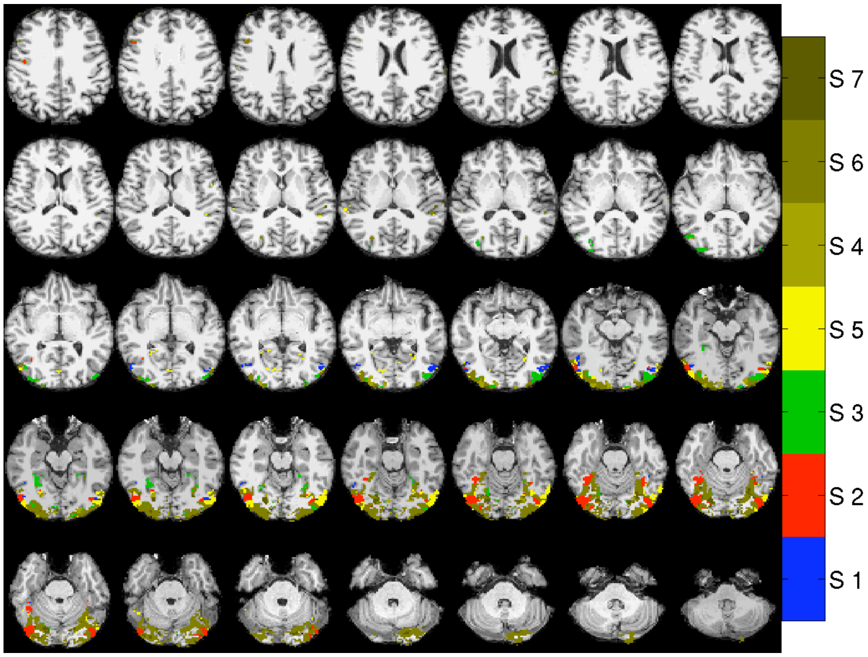


Figure C.1. Map of discovered systems for subject 1.

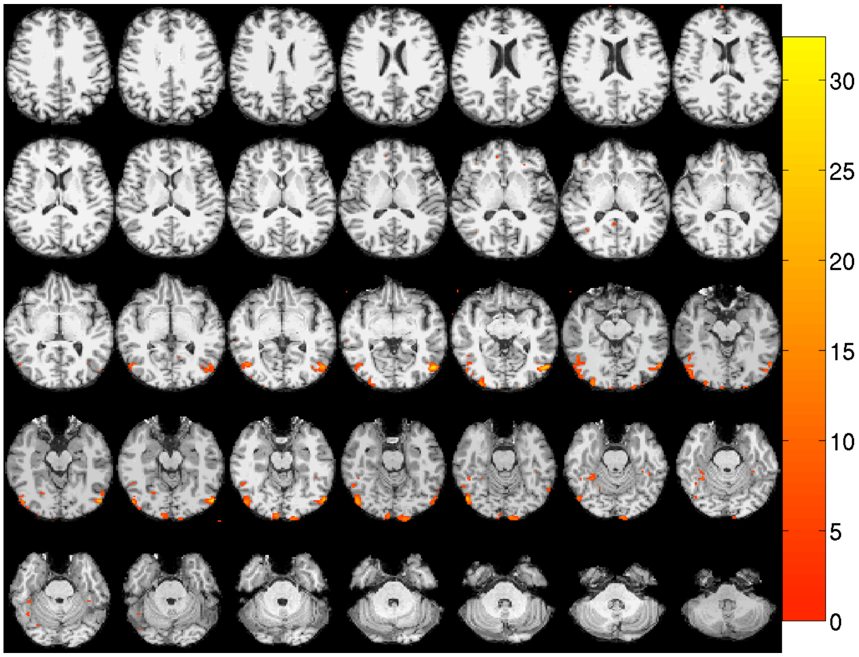


Figure C.2. Significance map for bodies–objects contrast for subject 1.

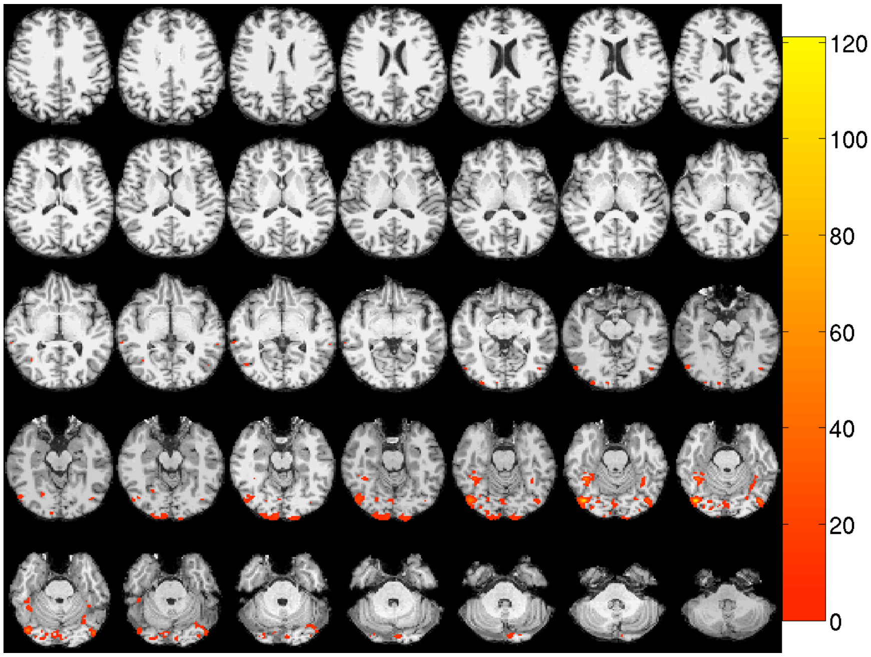


Figure C.3. Significance map for faces–objects contrast for subject 1.

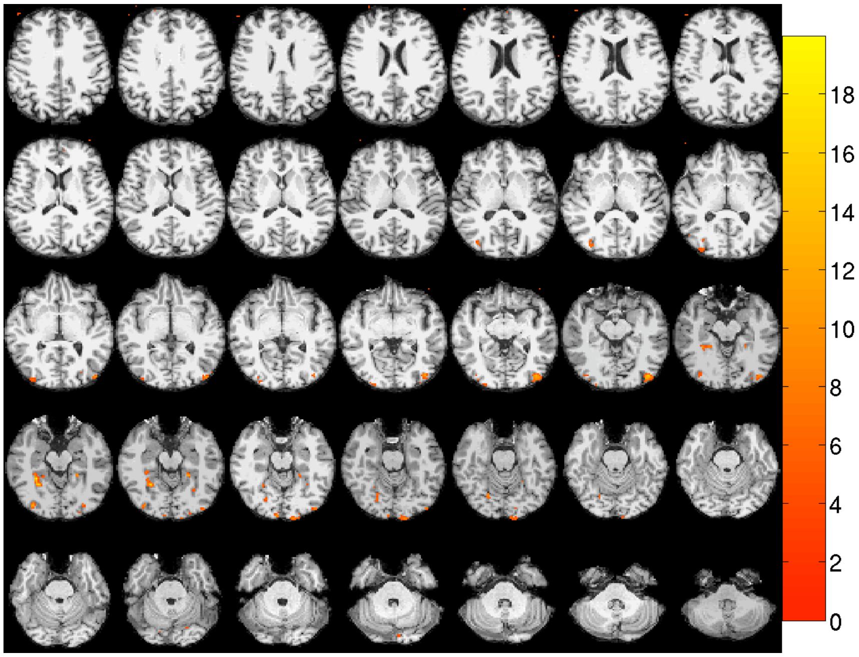


Figure C.4. Significance map for scenes–objects contrast for subject 1.

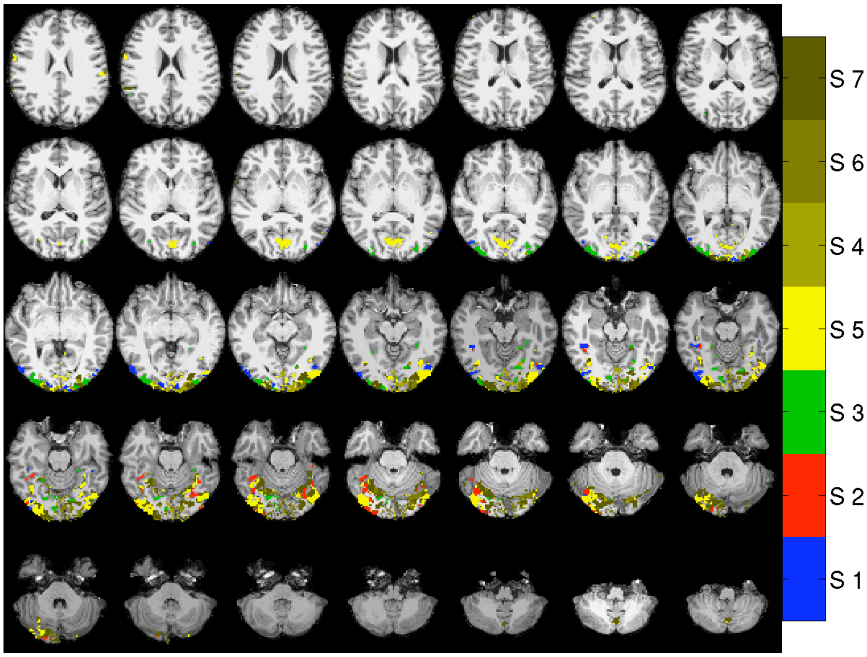


Figure C.5. Map of discovered systems for subject 2.

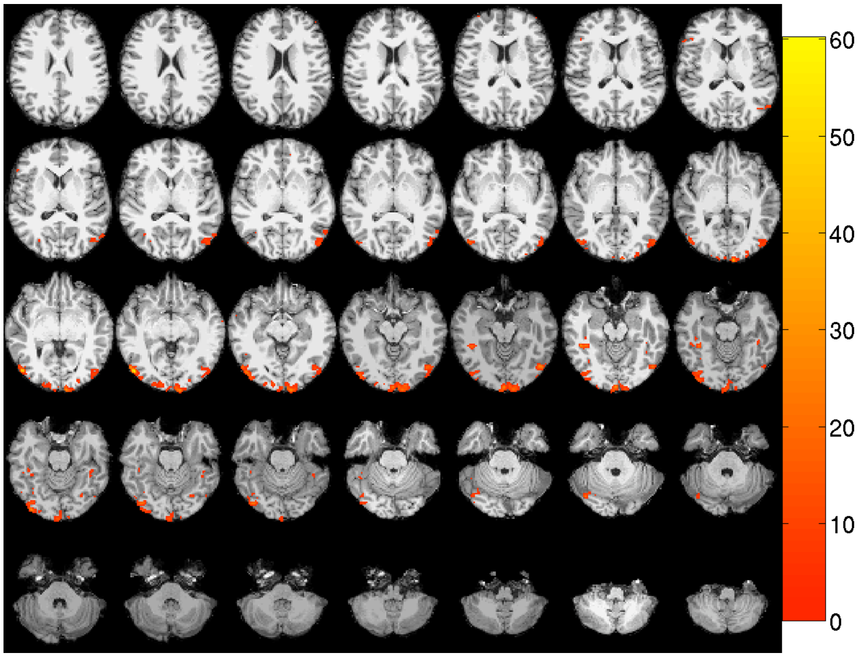


Figure C.6. Significance map for bodies–objects contrast for subject 2.

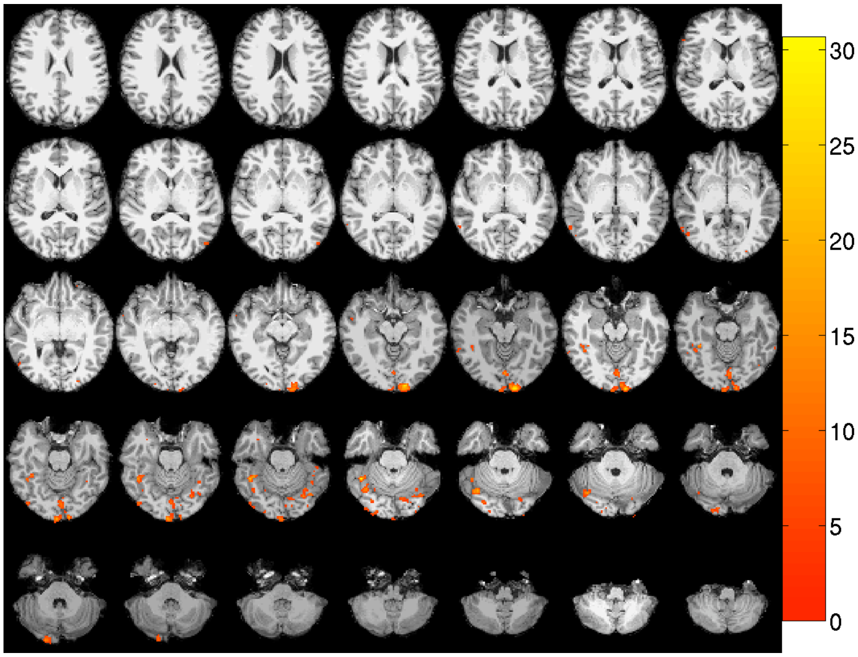


Figure C.7. Significance map for faces–objects contrast for subject 2.

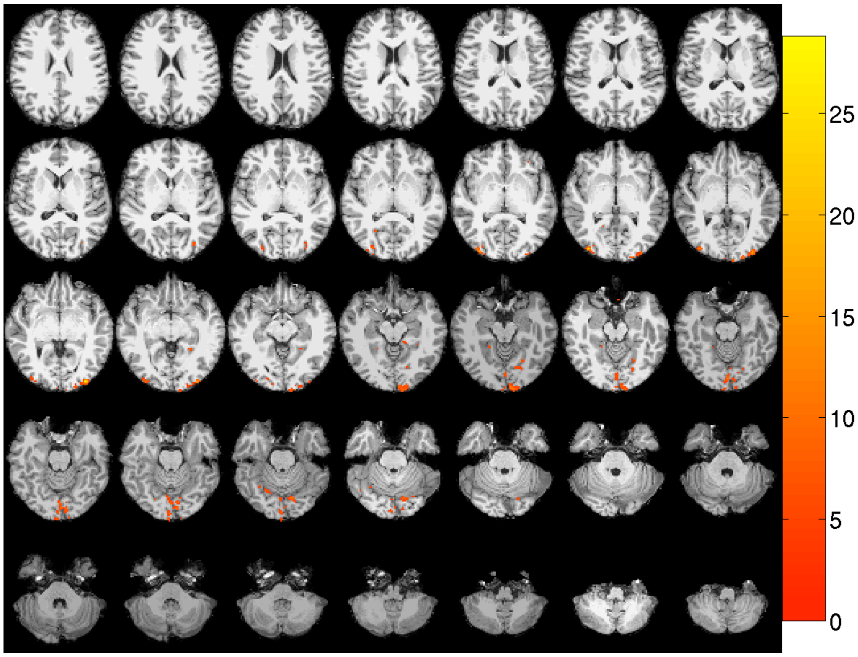


Figure C.8. Significance map for scenes–objects contrast for subject 2.

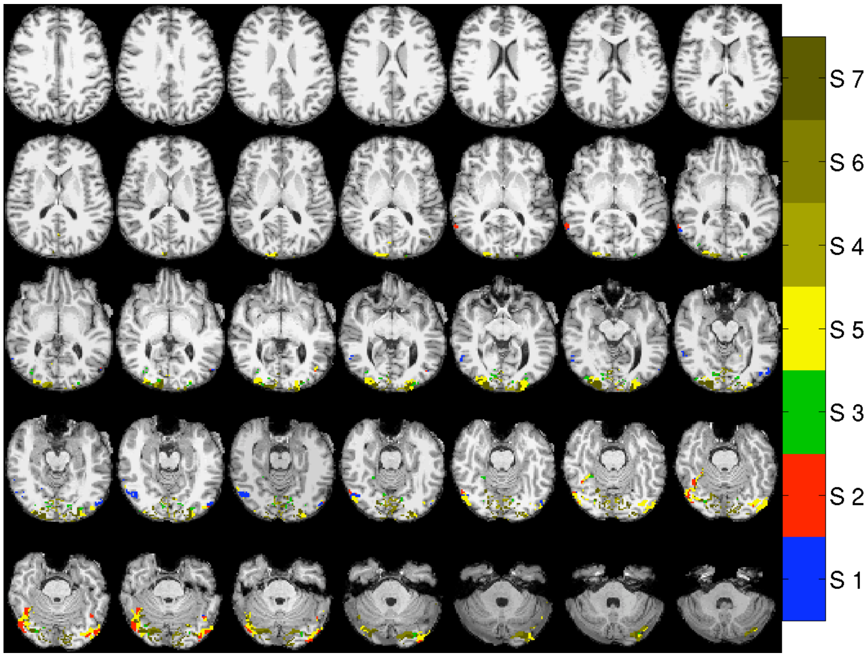


Figure C.9. Map of discovered systems for subject 3.

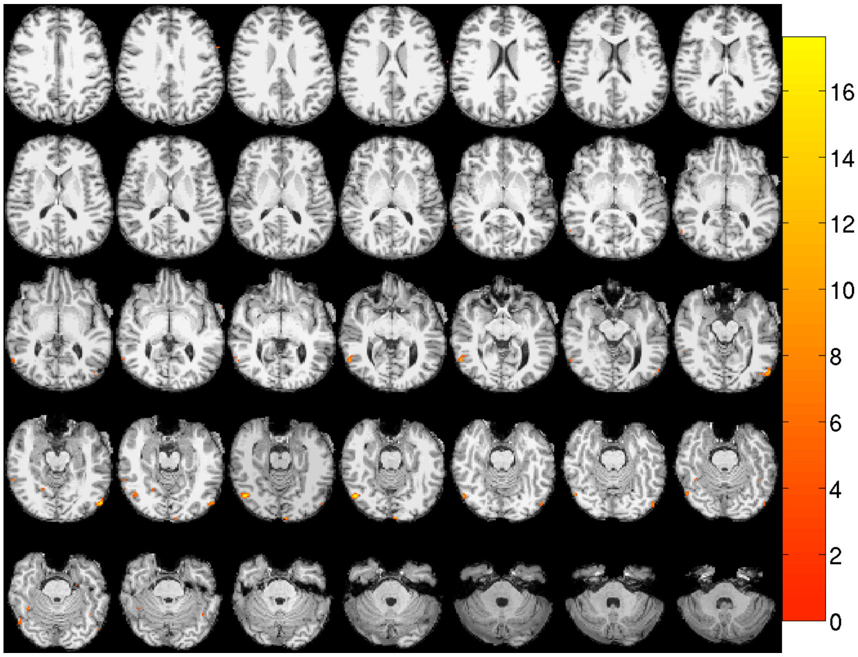


Figure C.10. Significance map for bodies–objects contrast for subject 3.

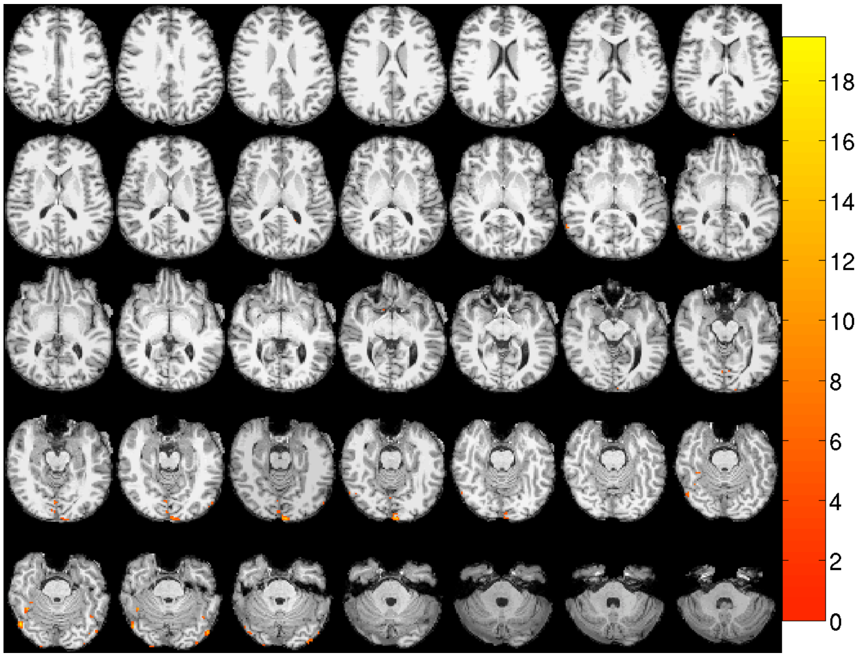


Figure C.11. Significance map for faces–objects contrast for subject 3.

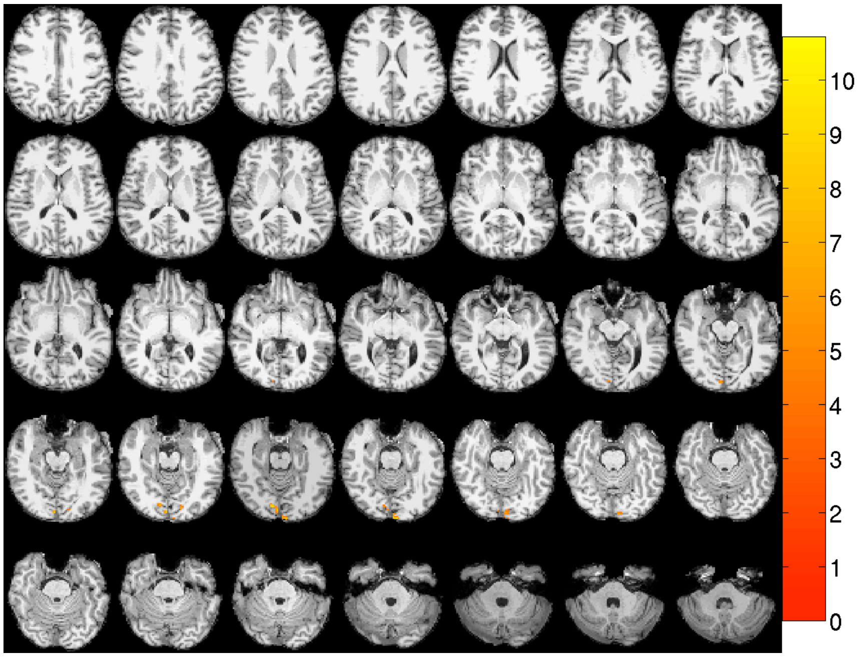


Figure C.12. Significance map for scenes–objects contrast for subject 3.

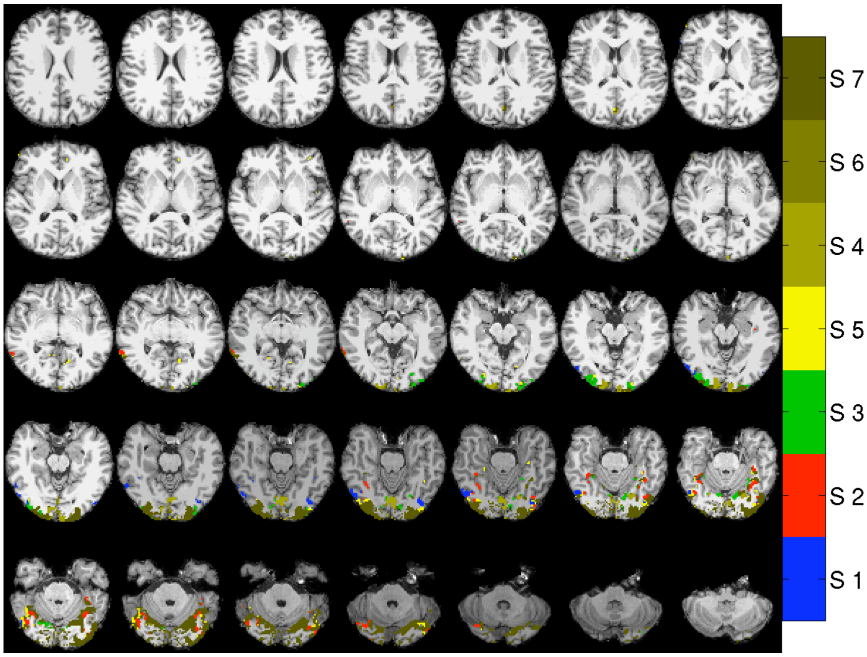


Figure C.13. Map of discovered systems for subject 4.

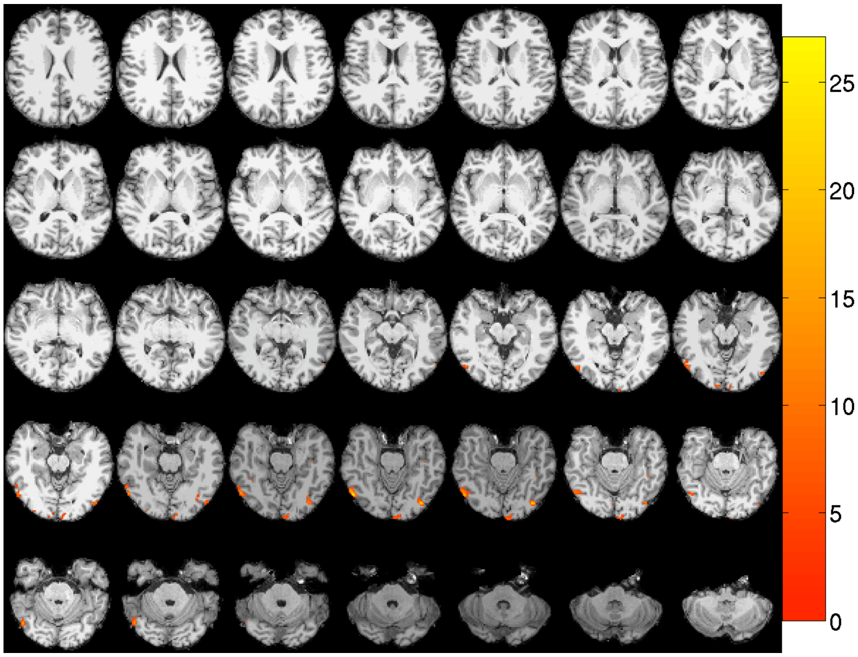


Figure C.14. Significance map for bodies–objects contrast for subject 4.

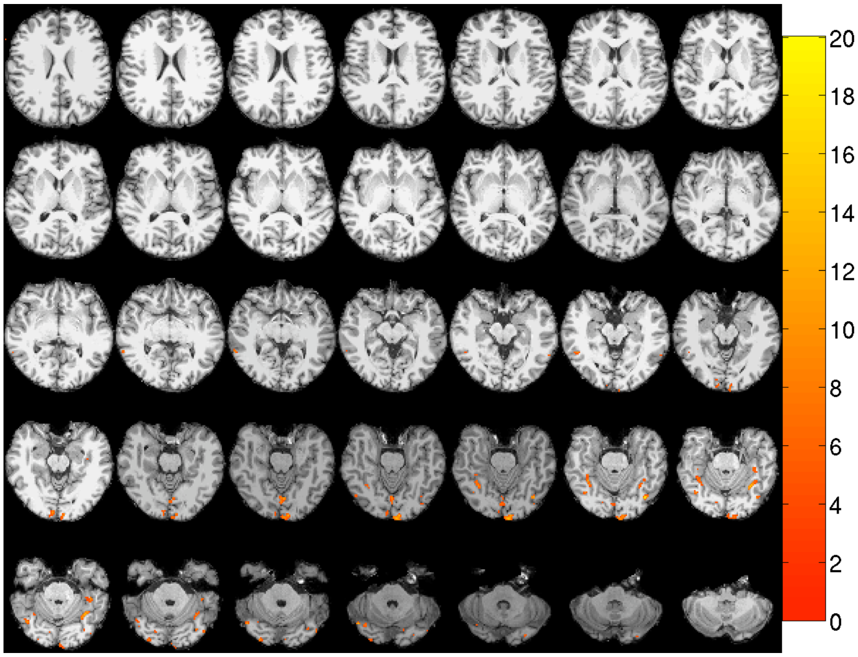


Figure C.15. Significance map for faces–objects contrast for subject 4.

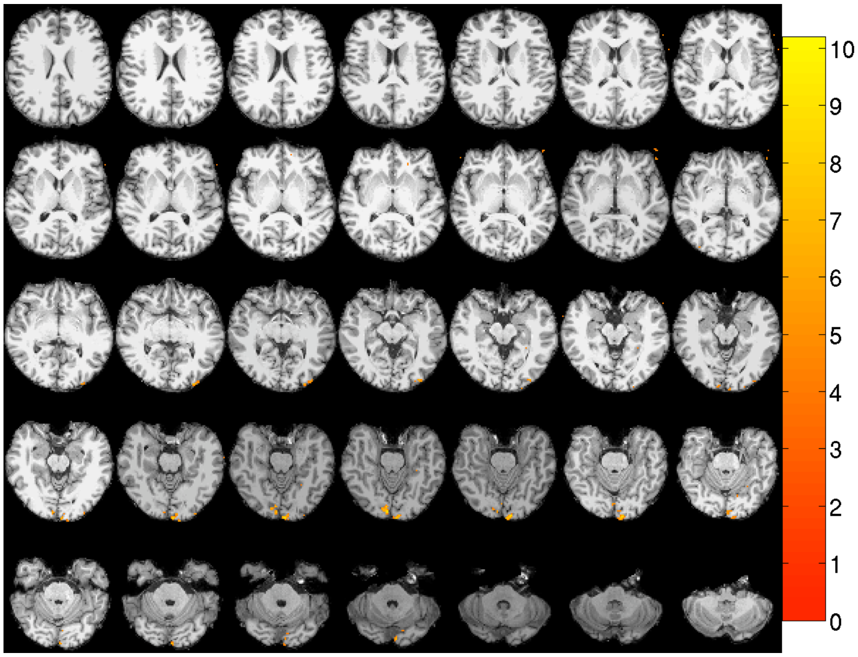


Figure C.16. Significance map for scenes–objects contrast for subject 4.

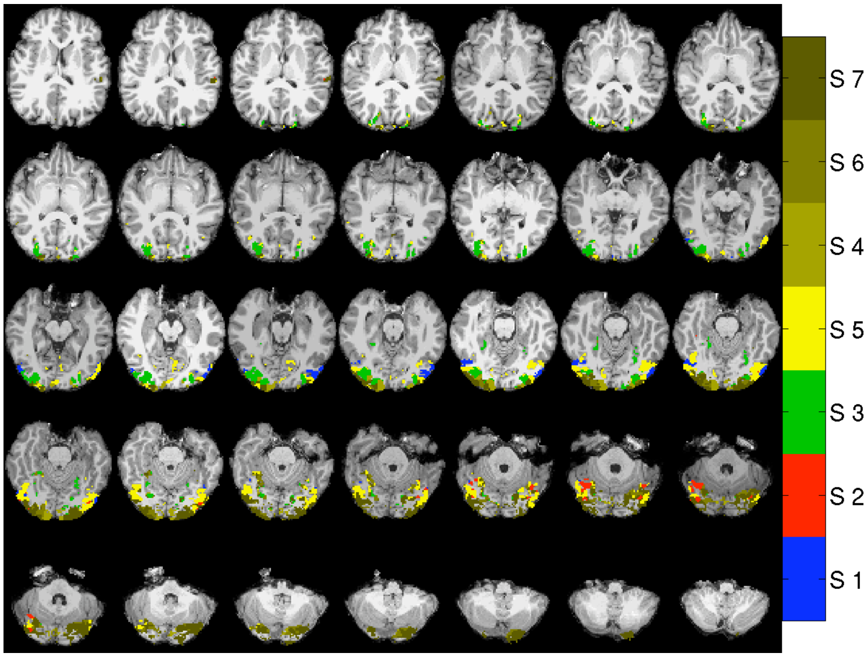


Figure C.17. Map of discovered systems for subject 5.

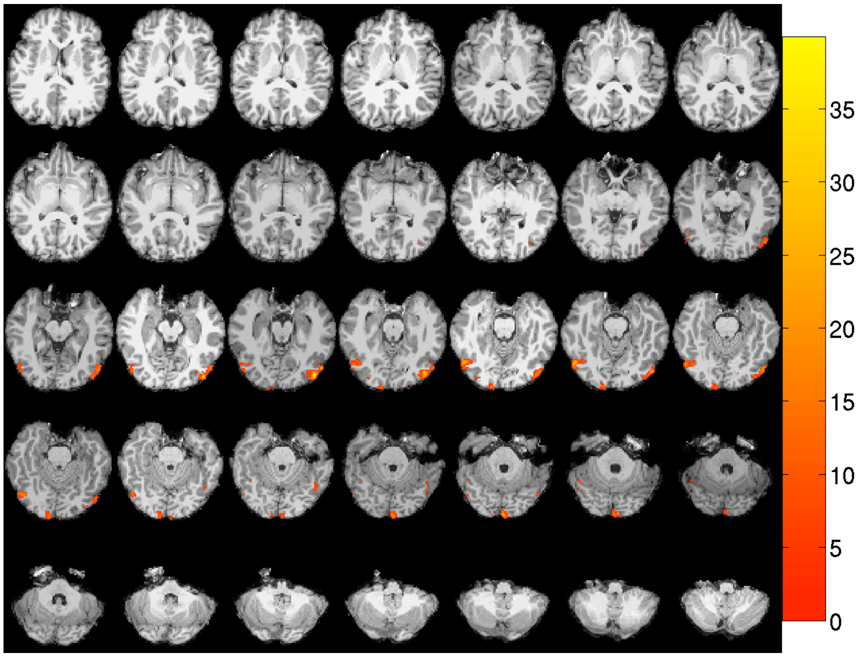


Figure C.18. Significance map for bodies–objects contrast for subject 5.

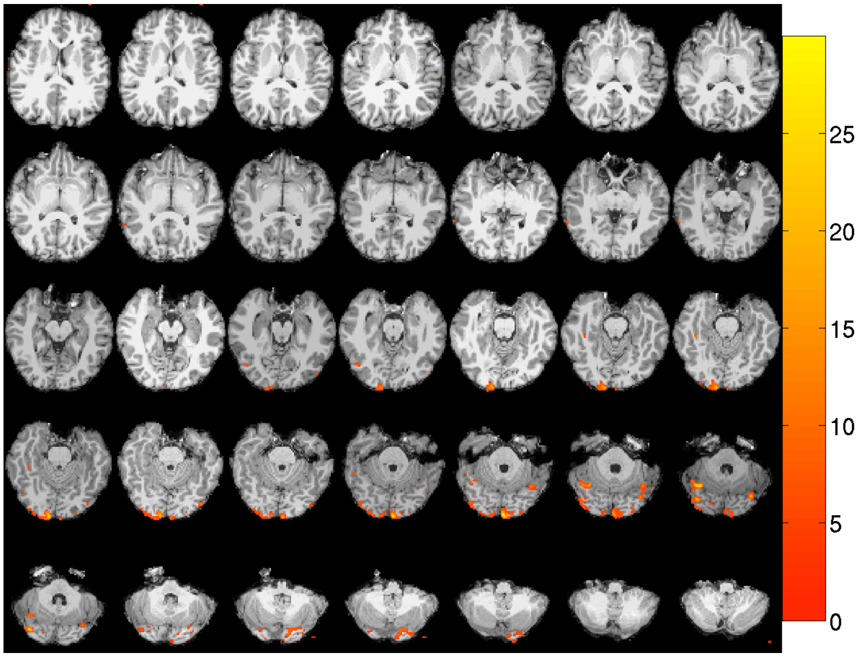


Figure C.19. Significance map for faces–objects contrast for subject 5.

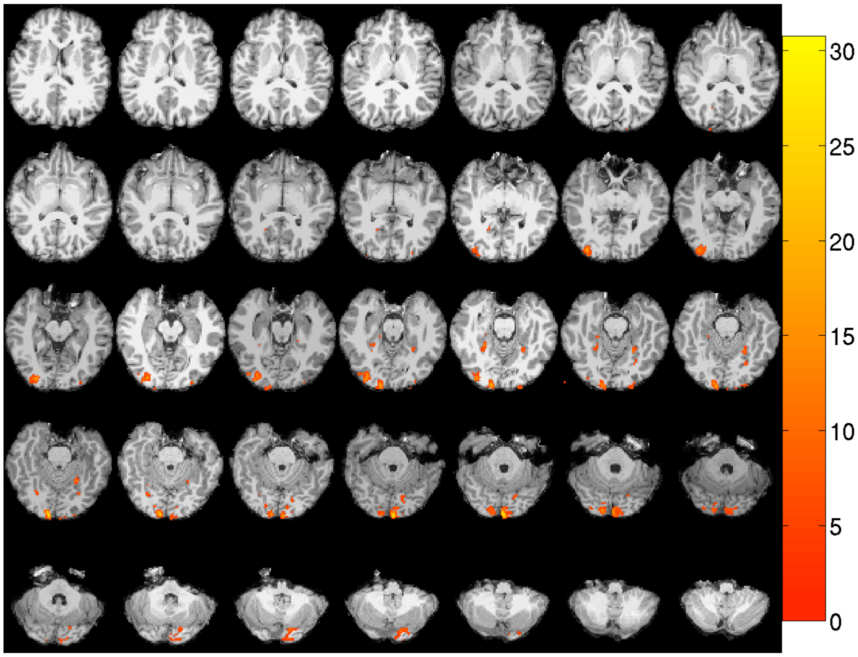


Figure C.20. Significance map for scenes–objects contrast for subject 5.

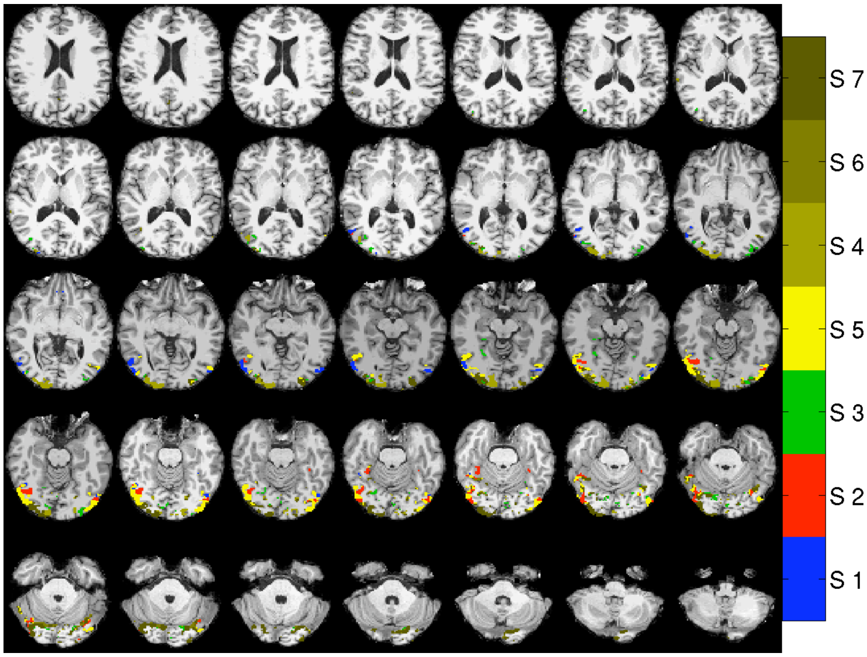


Figure C.21. Map of discovered systems for subject 6.

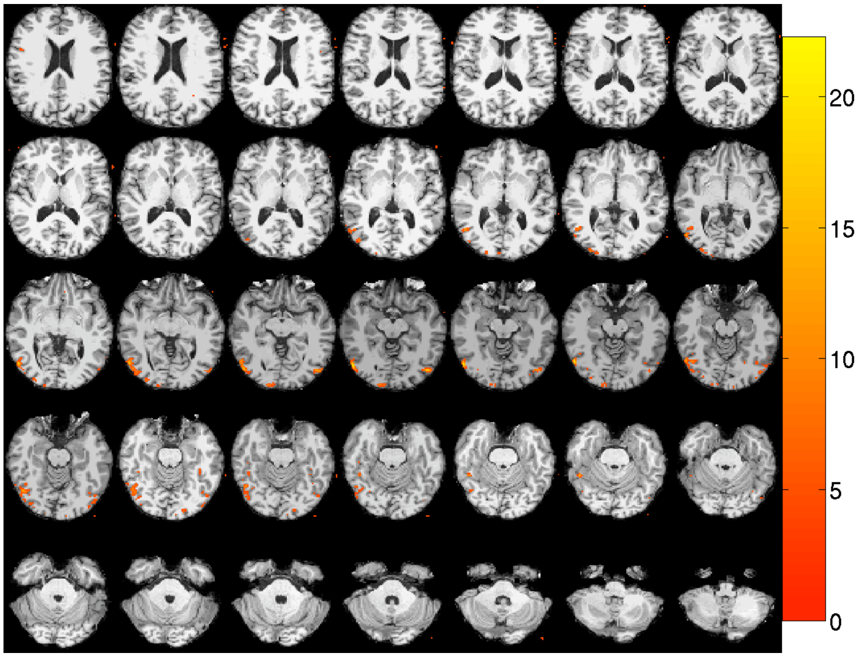


Figure C.22. Significance map for bodies–objects contrast for subject 6.

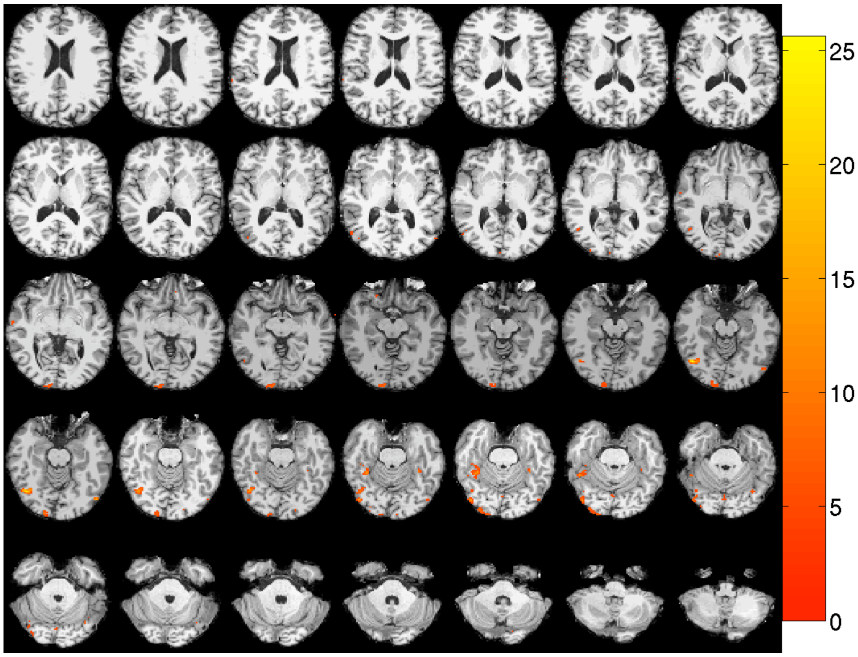


Figure C.23. Significance map for faces–objects contrast for subject 6.

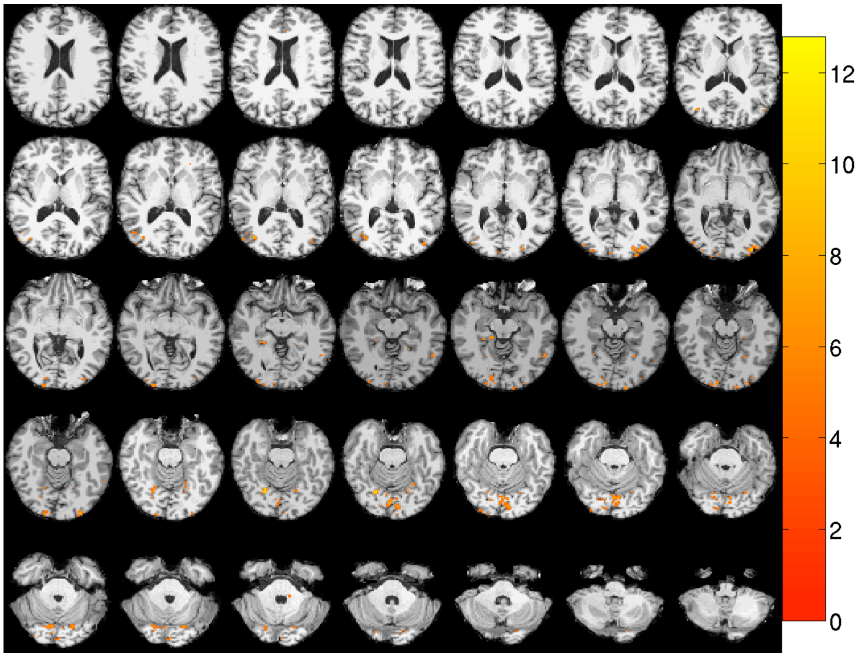


Figure C.24. Significance map for scenes–objects contrast for subject 6.

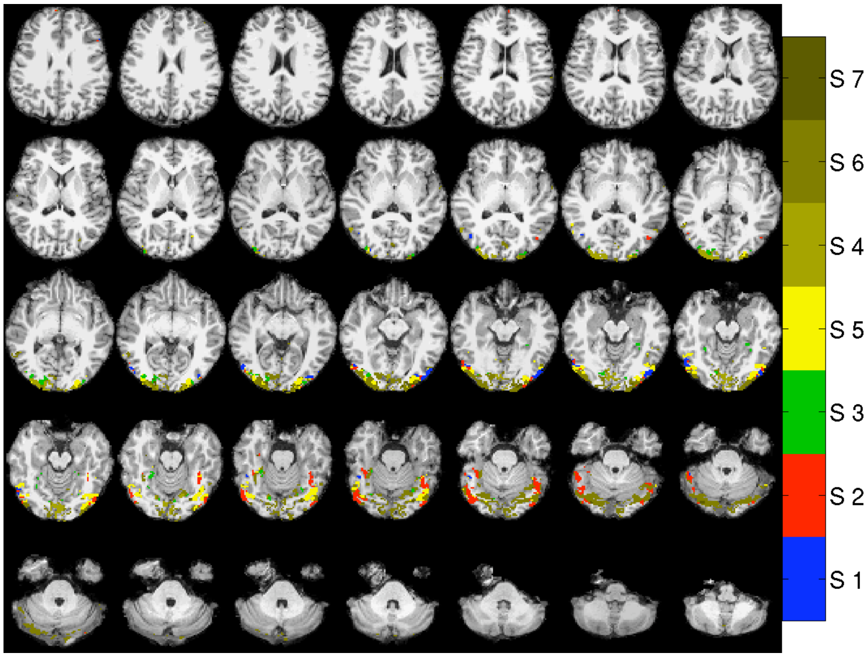


Figure C.25. Map of discovered systems for subject 7.

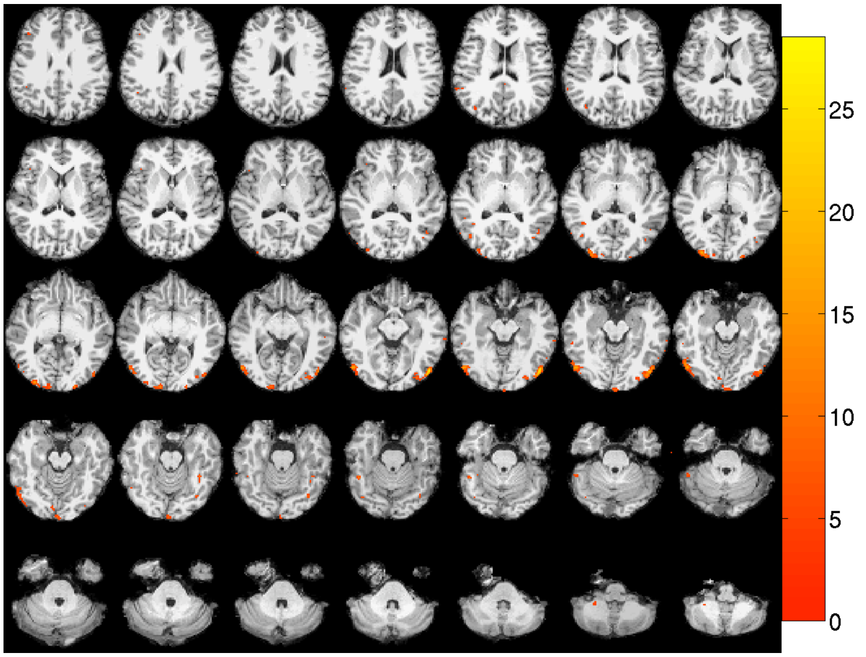


Figure C.26. Significance map for bodies–objects contrast for subject 7.

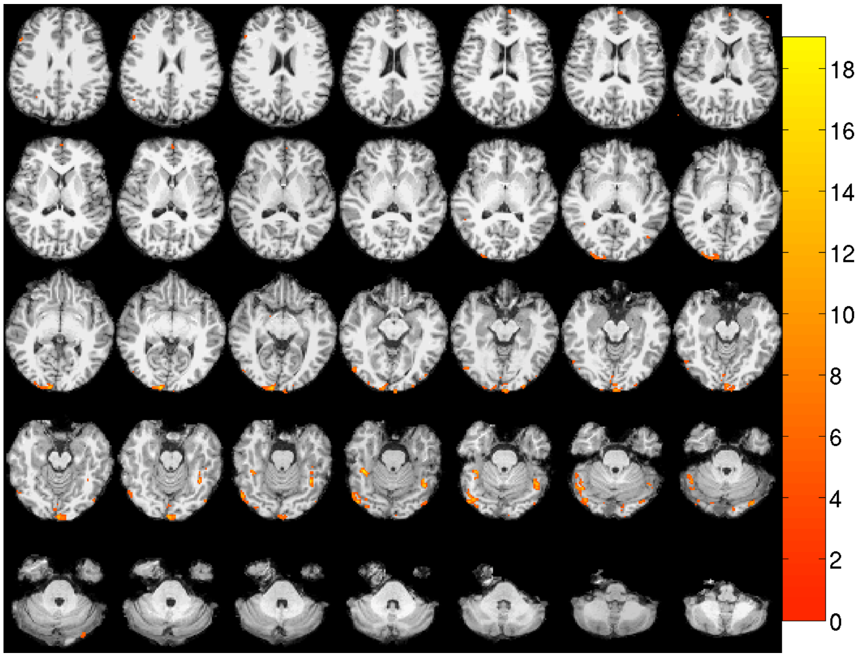


Figure C.27. Significance map for faces–objects contrast for subject 7.

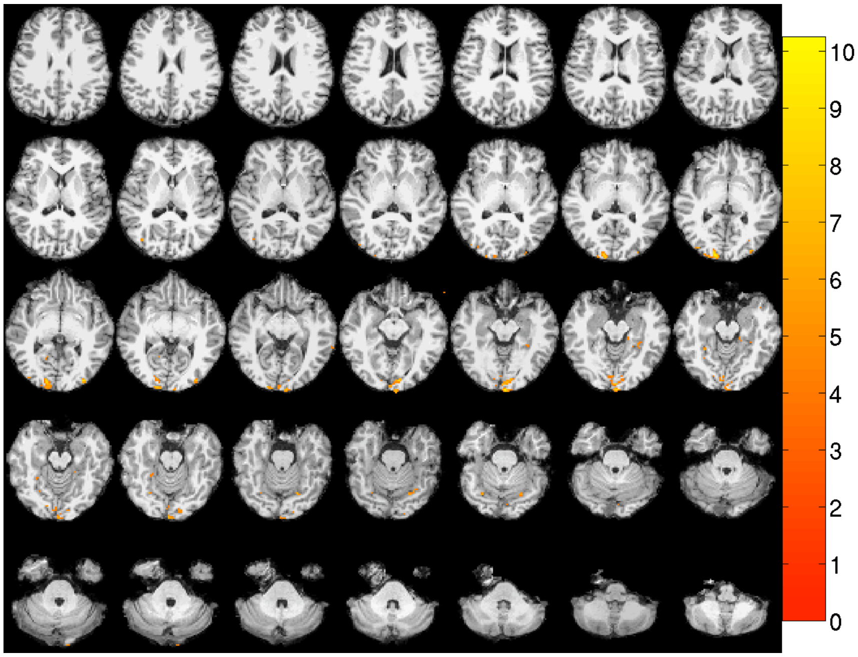


Figure C.28. Significance map for scenes–objects contrast for subject 7.

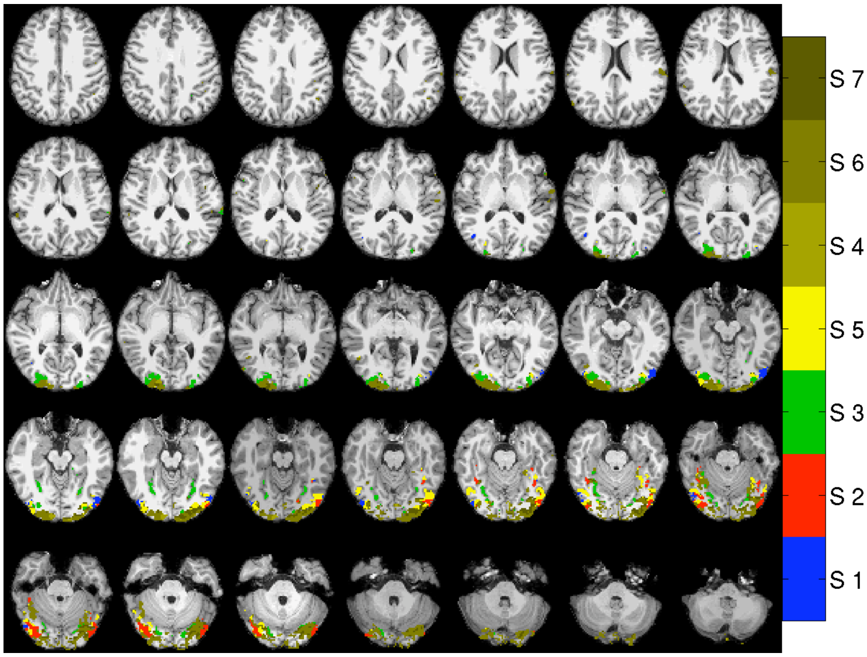


Figure C.29. Map of discovered systems for subject 9.

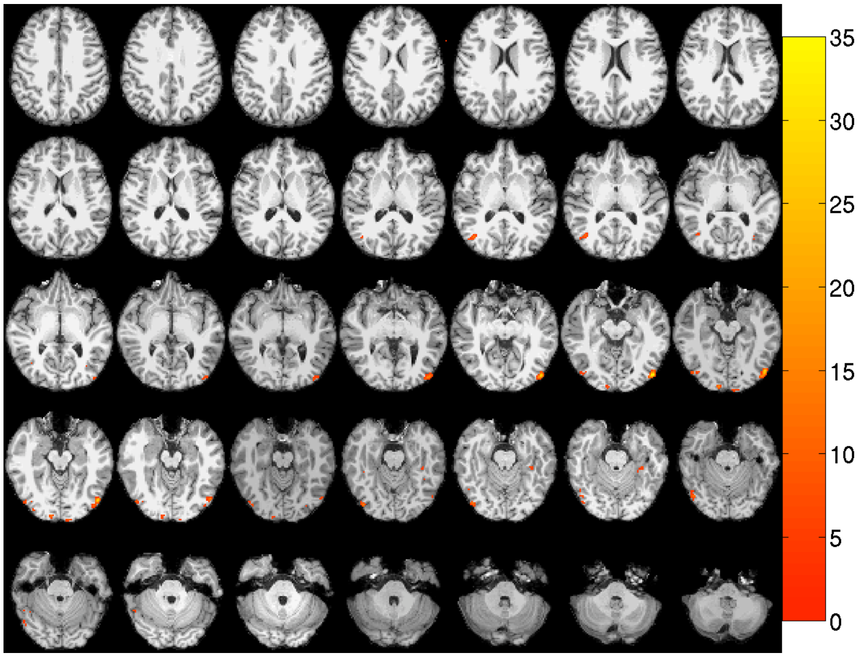


Figure C.30. Significance map for bodies–objects contrast for subject 9.

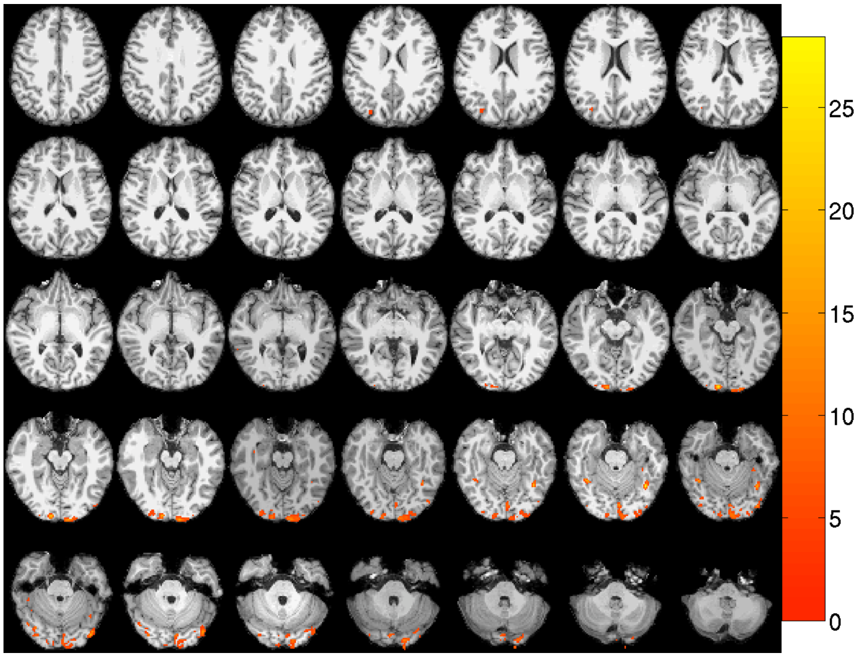


Figure C.31. Significance map for faces–objects contrast for subject 9.

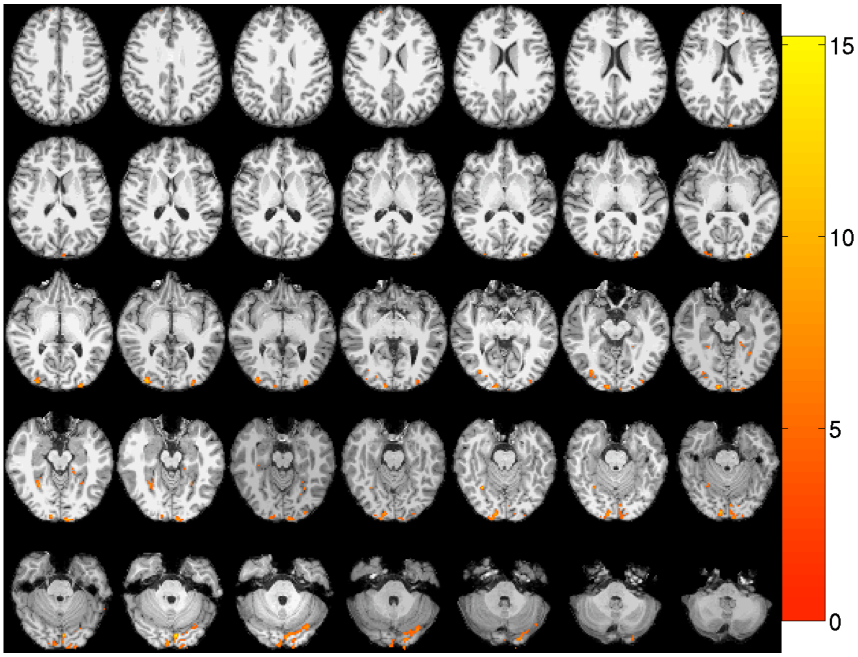


Figure C.32. Significance map for scenes–objects contrast for subject 9.

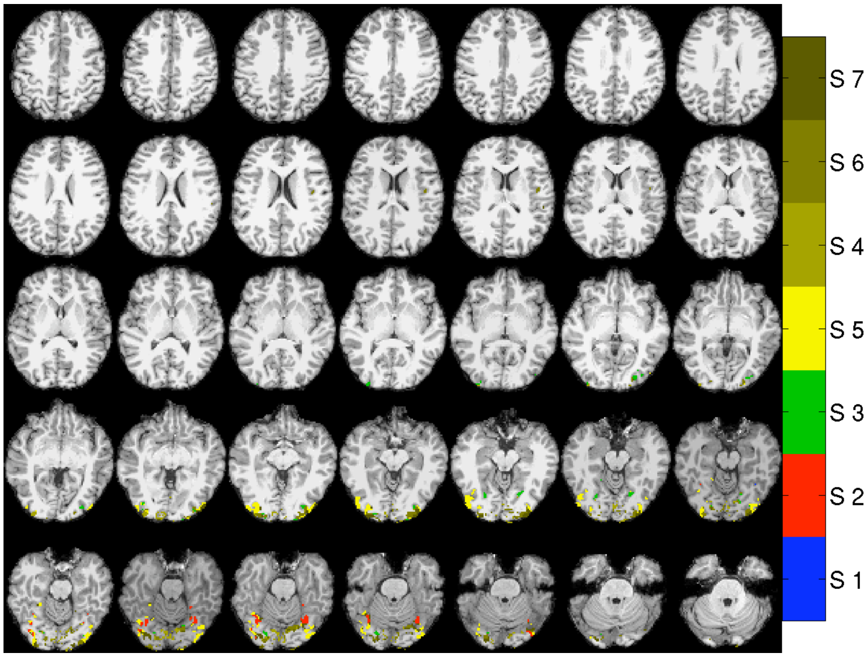


Figure C.33. Map of discovered systems for subject 10.

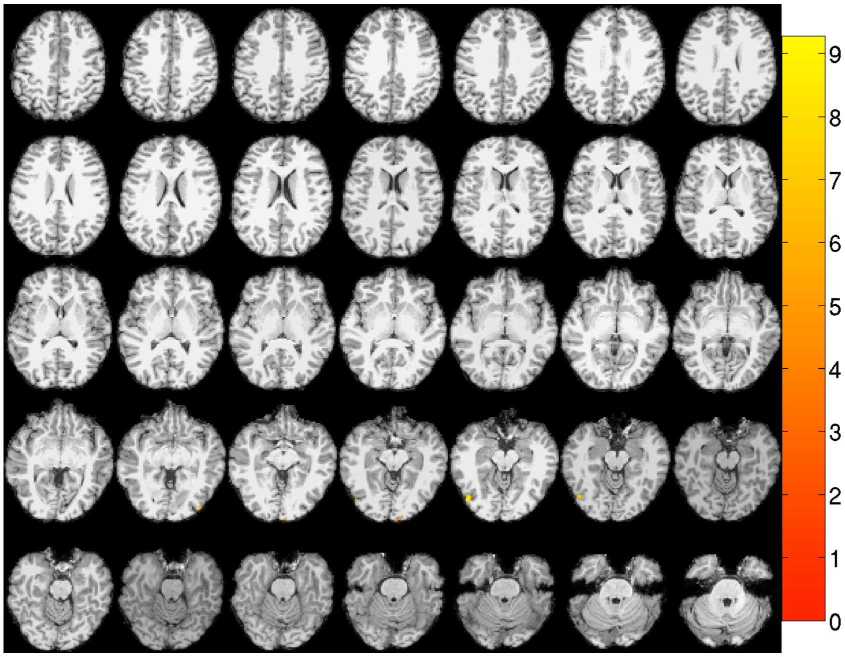


Figure C.34. Significance map for bodies–objects contrast for subject 10.

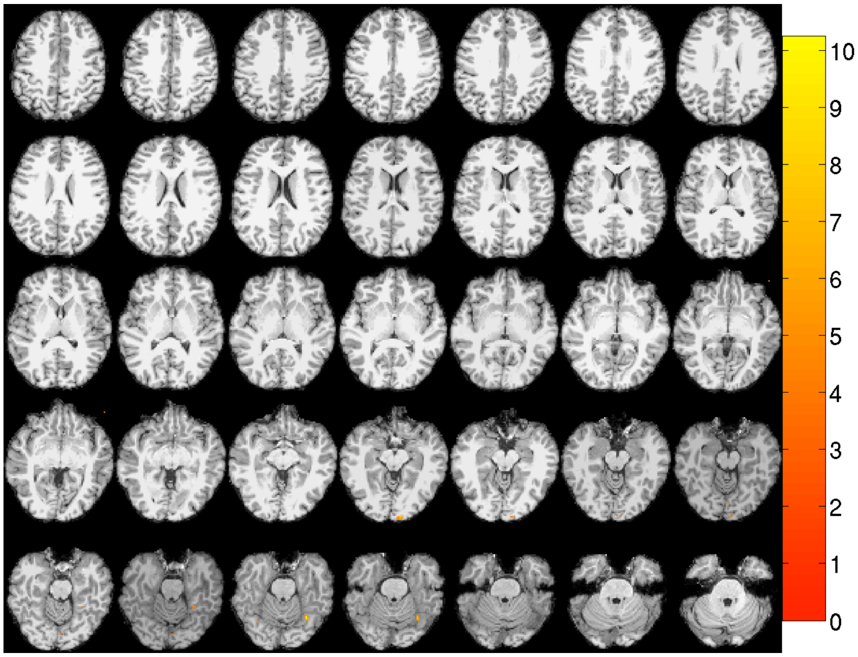


Figure C.35. Significance map for faces–objects contrast for subject 10.

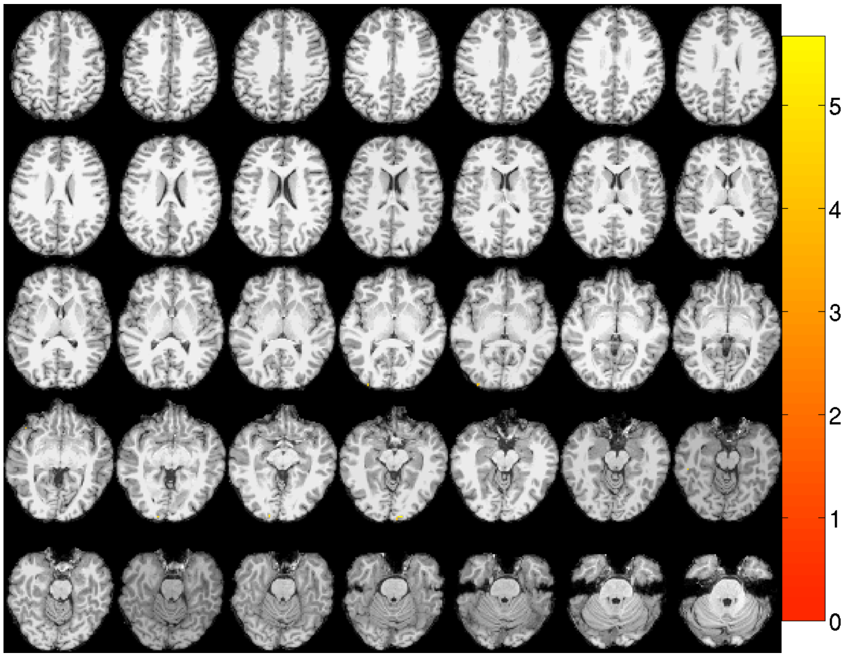


Figure C.36. Significance map for scenes–objects contrast for subject 10.

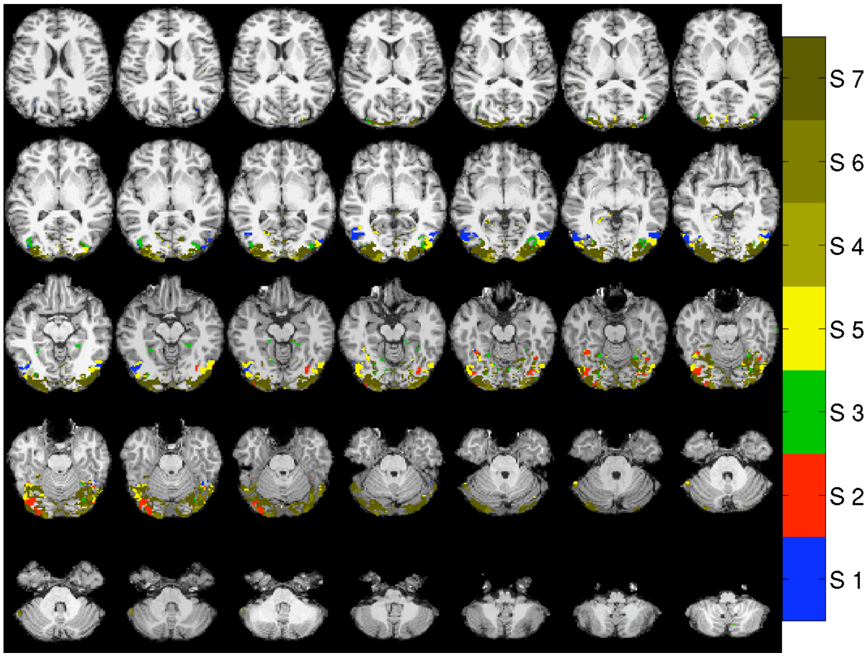


Figure C.37. Map of discovered systems for subject 11.

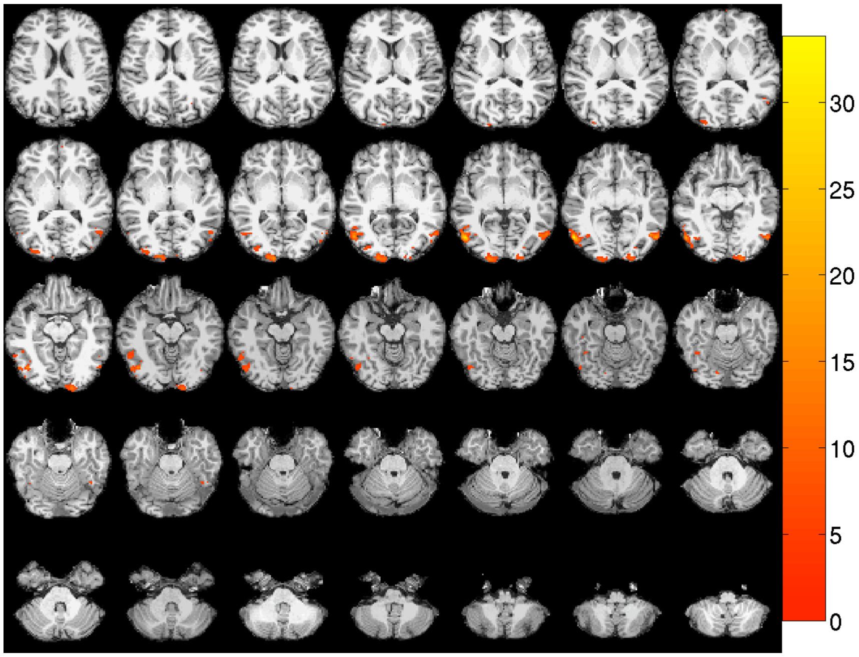


Figure C.38. Significance map for bodies–objects contrast for subject 11.

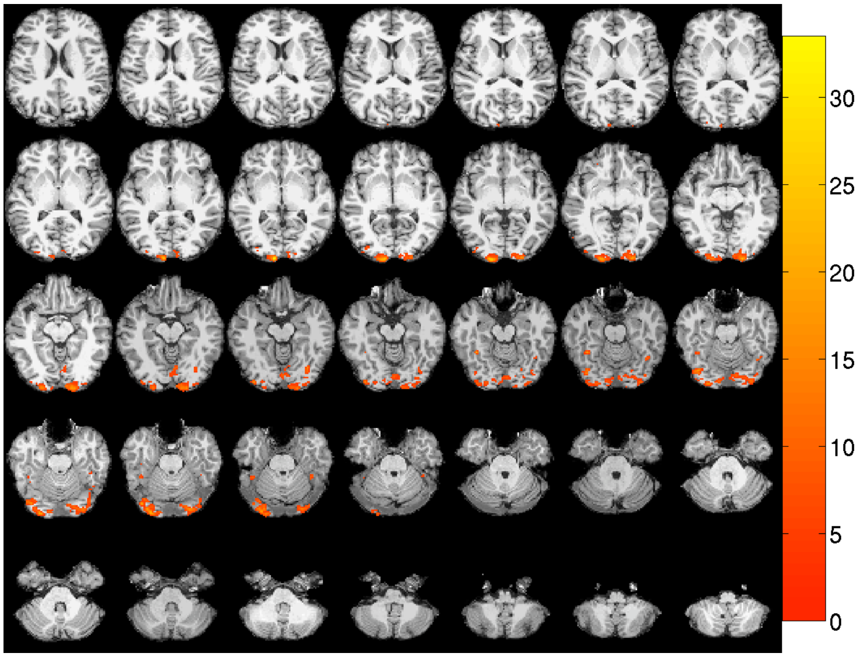


Figure C.39. Significance map for faces–objects contrast for subject 11.

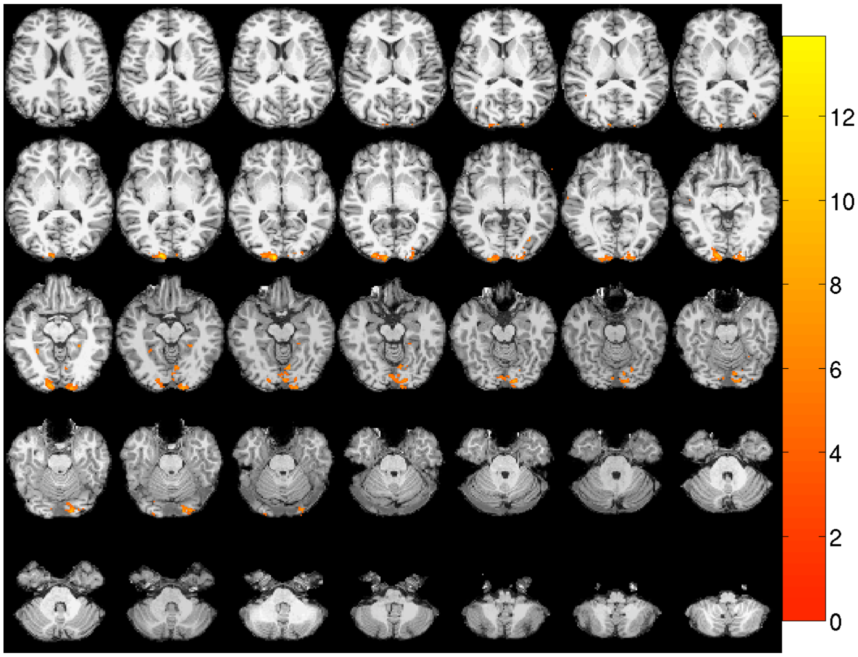


Figure C.40. Significance map for scenes–objects contrast for subject 11.

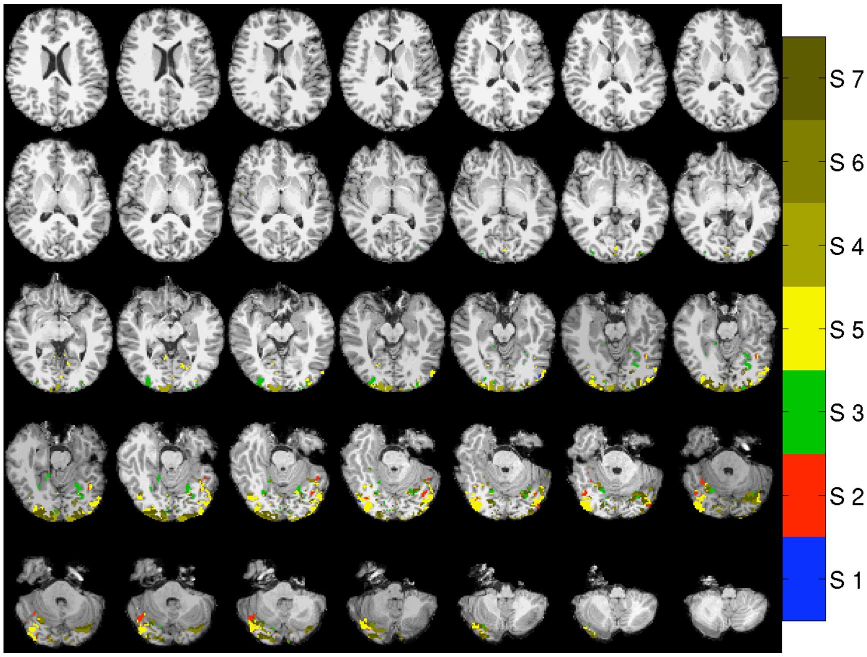


Figure C.41. Map of discovered systems for subject 13.

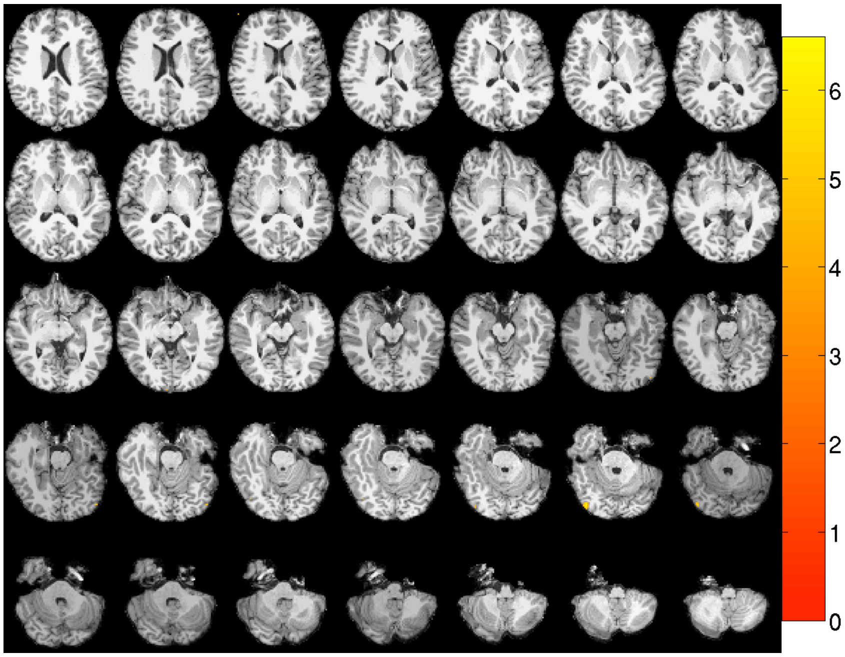


Figure C.42. Significance map for bodies–objects contrast for subject 13.

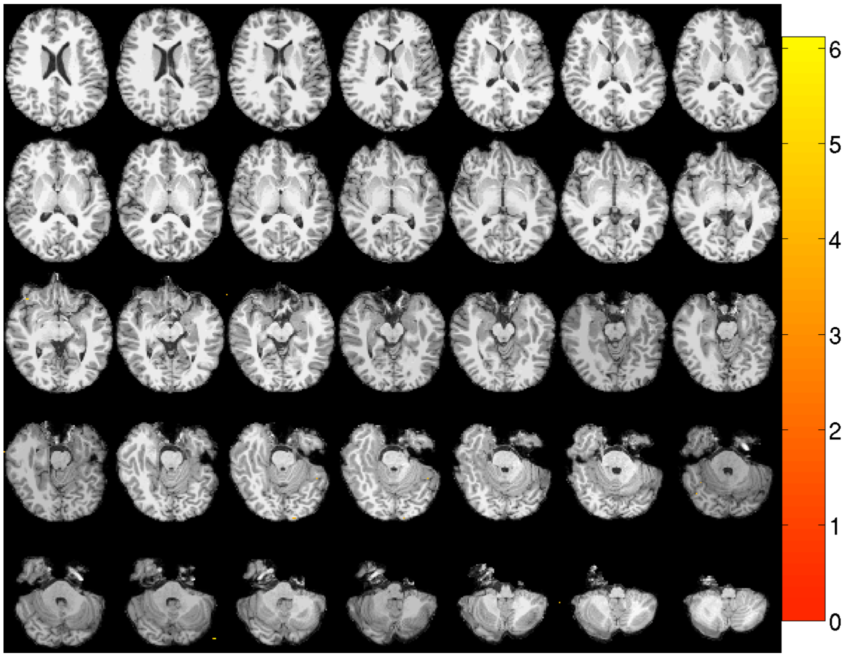


Figure C.43. Significance map for faces–objects contrast for subject 13.

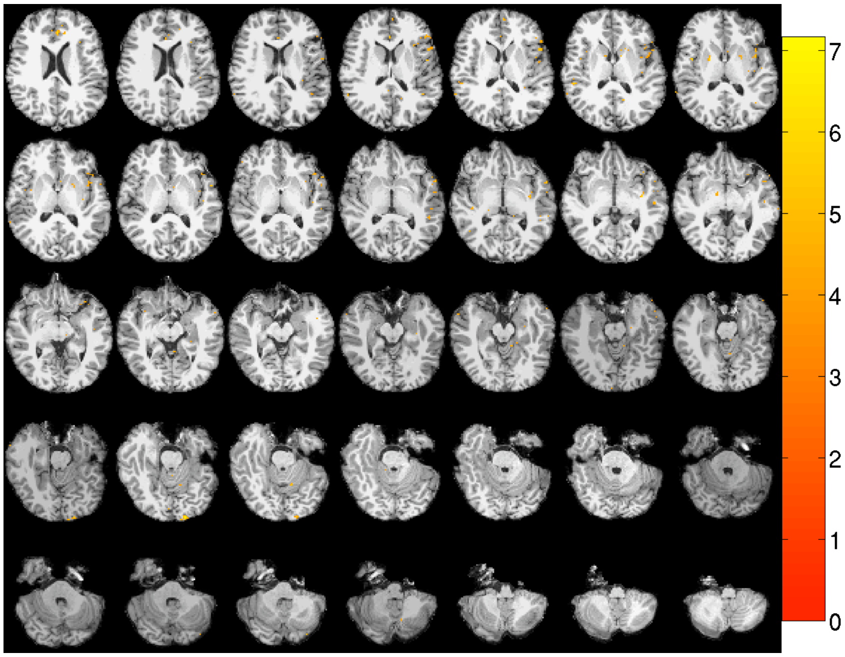


Figure C.44. Significance map for scenes–objects contrast for subject 13.

Appendix D

Derivations for Chapter 6

In this section, we derive the Gibbs free energy cost function for variational inference and derive the update rules for inference using the variational approximation.

D.1 Joint Probability Distribution

Based on the generative model described in Section 6.1, we form the full joint distribution of all the observed and unobserved variables. For each variable, we use $\omega^{\cdot}$ to denote the natural parameters of the distribution for that variable. For example, the variable e_{jid} is associated with natural parameters $\omega_{jd}^{e,1}$ and $\omega_{jd}^{e,2}$.

D.1.1 fMRI model

Given the fMRI model parameters, we can write the likelihood of the observed data $\boldsymbol{y}$:

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{\lambda},\boldsymbol{e},\boldsymbol{\tau}) = \prod_{j,i} \sqrt{\frac{\lambda_{ji}^{T_j}}{2\pi}} \exp\left\{-\frac{\lambda_{ji}}{2}\|\boldsymbol{y}_{ji} - \sum_{d} e_{jid}\boldsymbol{f}_d - a_{ji}\sum_{s} x_{jis}\boldsymbol{\Omega}_{js}\boldsymbol{\tau}_j\|^2\right\}. \tag{D.1}$$

We now express the priors on the parameters of the likelihood model defined in Section 6.1.1 in the new notation. Specifically, for the nuisance parameters $\boldsymbol{e}$, we have

$$\begin{aligned} p(e_{jid}) &= \text{Normal}(\mu_{jd}^{e}, \sigma_{jd}^{e}) &\text{(D.2)}\\ &\propto \exp\left\{-\tfrac{1}{2}(\omega_{jd}^{e,2})e_{jid}^2 + \tfrac{1}{2}(\omega_{jd}^{e,1})e_{jid}\right\}, &\text{(D.3)}\end{aligned}$$

where $\omega_{jd}^{e,2} = (\sigma_{jd}^{e})^{-1}$ and $\omega_{jd}^{e,1} = \mu_{d}^{e}(\sigma_{jd}^{e})^{-1}$.

With our definition of the Gamma distribution in Equation (6.11), the natural parameters for the noise precision variables $\boldsymbol{\lambda}$ are $\omega_{jm}^{\lambda,1} = \kappa_{jm}$ and $\omega_{jm}^{\lambda,2} = \theta_{jm}$.

The distribution over the activation heights $\boldsymbol{a}$ is given by

$$\begin{aligned} p(a_{jim}) &= \text{Normal}_{+}(\mu_{jm}^{a}, \sigma_{jm}^{a}) &\text{(D.4)}\\ &\propto \exp\left\{-\tfrac{1}{2}(\omega_{jm}^{a,2})a_{jim}^2 + \tfrac{1}{2}(\omega_{jm}^{a,1})a_{jim}\right\}, a_{jim} \geq 0 &\text{(D.5)}\end{aligned}$$

We have $\omega^{a,2}_{jm} = (\sigma^a_{jm})^{-1}$ and $\omega^{a,1}_{jm} = \mu^a_{jm}\left(\sigma^a_{jm}\right)^{-1}$.

The distribution $\text{Normal}_+(\eta, \rho^{-1})$ is a member of an exponential family of distributions and has the following properties:

$$p(a) = \sqrt{\tfrac{2\lambda}{\pi}} \left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho(a-\eta)^2/2}, \tag{D.6}$$

$$E[a] = \eta + \sqrt{\tfrac{2}{\pi\lambda}} \left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho\eta^2/2}, \tag{D.7}$$

$$E[a^2] = \eta^2 + \rho^{-1} + \eta\sqrt{\tfrac{2}{\pi\lambda}} \left[1 + \text{erf}\left(\sqrt{\tfrac{\rho}{2}}\eta\right)\right]^{-1} e^{-\rho\eta^2/2}. \tag{D.8}$$

■ D.1.2 Nonparametric Hierarchical Joint Model for Group fMRI Data

The voxel activation variables x_{jis} are binary, with prior probability ϕ_{ks} given according to cluster memberships. Since $\phi \sim \text{Beta}(\omega^{\phi,1}, \omega^{\phi,2})$, the joint density of $\boldsymbol{x}$ and $\boldsymbol{\phi}$ conditioned on the cluster memberships $\boldsymbol{z}$ is defined as follows:

$$p(\boldsymbol{x}, \boldsymbol{\phi}|\boldsymbol{z}) = \prod_{j,k,s} \left[\frac{\Gamma(\omega^{\phi,1} + \omega^{\phi,2})}{\Gamma(\omega^{\phi,1})\Gamma(\omega^{\phi,2})} \phi_{ks}^{\omega^{\phi,1} - 1 + \sum_{i,s} x_{jis}\delta(z_{ji},k)} \right.$$
$$\left. \times \left(1 - \phi_{ks}\right)^{\omega^{\phi,2} - 1 + \sum_{i,s}(1 - x_{jis})\delta(z_{ji},k)} \right].$$

We assume a hierarchical Dirichlet process prior over the functional unit memberships, with subject-level weights $\boldsymbol{\beta}$. We use a collapsed variational inference scheme (Teh et al., 2008), and therefore marginalize over these weights:

$$p(\boldsymbol{z}|\boldsymbol{\pi}, \alpha) = \int_{\boldsymbol{\beta}} p(\boldsymbol{z}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\pi}, \alpha), \tag{D.9}$$

$$= \prod_{j=1}^{J} \left[\tfrac{\Gamma(\alpha)}{\Gamma(\alpha + N_j)} \prod_{k=1}^{K} \tfrac{\Gamma(\alpha\pi_k + n_{jk})}{\Gamma(\alpha\pi_k)} \right], \tag{D.10}$$

where K is the number of non-empty functional units in the configuration and $n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji}, k)$. To provide conjugacy with the Dirichlet prior for the group-level functional unit weights $\boldsymbol{\pi}$, we prefer the terms in Equation (D.10) that include weights to appear as powers of π_k . However, the current form of the conditional distribution makes the computation of the posterior over $\boldsymbol{\pi}$ hard. To overcome this challenge, we note that for $0 \leq r \leq n$, we have $\sum_{r=0}^{n} \left[{n \atop r}\right] \vartheta^r = \Gamma(\vartheta + n)/\Gamma(\vartheta)$, where $\left[{n \atop r}\right]$ are unsigned Stirling numbers of the first kind (Antoniak, 1974). The collapsed variational approach uses this fact and the properties of the Beta distribution to add an auxiliary variable

$\boldsymbol{r} = \{r_{ji}\}$ to the model:

$$p(\boldsymbol{z}, \boldsymbol{r}, \mid \boldsymbol{\pi}, \alpha) \propto \prod_{j=1}^{J} \prod_{k=1}^{K} \left[{n_{jk} \atop r_{jk}} \right] (\alpha \pi_k)^{r_{jk}} , \tag{D.11}$$

where $r_{jk} \in \{0, 1, \cdots, n_{ji}\}$. If we marginalize the distribution (D.11) over the auxiliary variable, we obtain the expression in (D.10).

D.2 Minimization of the Gibbs Free Energy

Let $\boldsymbol{h} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{\phi}, \boldsymbol{\pi}, v, \alpha, \gamma, \boldsymbol{a}, \boldsymbol{e}, \boldsymbol{\tau}, \boldsymbol{\lambda}\}$ denote the set of all unobserved variables. In the framework of variational inference, we approximate the posterior distribution $p(\boldsymbol{h}|\boldsymbol{y})$ of the hidden variables $\boldsymbol{h}$ given the observed $\boldsymbol{y}$ by a distribution $q(\boldsymbol{h})$. The approximation is performed through minimization of the Gibbs free energy function in Equation (6.18) with an approximate posterior distribution $q(\boldsymbol{h})$ of the form in Equation (6.19). We derive a coordinate descent method where in each step we minimize the function with respect to one of the components of $q(\cdot)$, keeping the rest constant.

D.2.1 Auxiliary variables

Assuming that all the other components of the distribution q are constant, we obtain:

$$\mathcal{F}[q(\boldsymbol{r} \mid \boldsymbol{z})] = E_{\boldsymbol{z}} \Bigg[\sum_{\boldsymbol{r}} q(\boldsymbol{r}|\boldsymbol{z}) \Bigg(\log q(\boldsymbol{r}|\boldsymbol{z}) + \\ - \sum_{j,k} \Big\{ \log \left[{n_{jk} \atop r_{jk}} \right] + r_{jk} \, E[\log(\alpha \pi_k)] \Big\} \Bigg) \Bigg] + \text{const.} \tag{D.12}$$

The optimal posterior distribution on the auxiliary variables takes the form

$$q^*(\boldsymbol{r}|\boldsymbol{z}) = \prod_j \prod_k q(r_{jk}|\boldsymbol{z}). \tag{D.13}$$

Under $q*$, we have for the auxiliary variable $\boldsymbol{r}$:

$$q(r_{jk}|\boldsymbol{z}) = \frac{\Gamma(\tilde{\omega}^r_{jk})}{\Gamma(\tilde{\omega}^r_{jk} + n_{jk})} \left[{n_{jk} \atop r_{jk}} \right] (\tilde{\omega}^r_{jk})^{r_{jk}}. \tag{D.14}$$

This distribution corresponds to the probability mass function for a random variable that describes the number of tables that n_{jk} customers occupy in a Chinese Restaurant Process with parameter $\tilde{\omega}^r_{jk}$ (Antoniak, 1974). The optimal value of the parameter $\tilde{\omega}^r_{jk}$ is given by

$$\log \tilde{\omega}^r_{jk} = E[\log(\alpha \pi_k)] \tag{D.15}$$

$$= \log \alpha + E[\log v_k] + \textstyle\sum_{k'<k} E[\log(1 - v_{k'})].$$

As a distribution parameterized by $\log \tilde{\omega}^r_{jk}$, Equation (D.14) defines a member of an exponential family of distributions. The expected value of the auxiliary variable r_{jk} is therefore:

$$E[r_{jk}|\boldsymbol{z}] = \tfrac{\partial}{\partial \log \tilde{\omega}^r_{jk}} \log \tfrac{\Gamma(\tilde{\omega}^r_{jk}+n_{jk})}{\Gamma(\tilde{\omega}^r_{jk})} \tag{D.16}$$

$$= \tilde{\omega}^r_{jk}\Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \tilde{\omega}^r_{jk}\Psi(\tilde{\omega}^r_{jk}), \tag{D.17}$$

where $\Psi(\omega) = \frac{\partial}{\partial \omega} \log \Gamma(\omega)$. This expression is helpful when updating the other components of the distribution. Accordingly, we obtain expectation:

$$E[r_{jk}] = E_{\boldsymbol{z}}[E_{\boldsymbol{r}}[r_{jk}|\boldsymbol{z}]] = \tilde{\omega}^r_{jk} E_{\boldsymbol{z}}[\Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \Psi(\tilde{\omega}^r_{jk})]. \tag{D.18}$$

Under $q(\boldsymbol{z})$, each variable n_{jk} is the sum of N_j independent Bernoulli random variables $\delta(z_{ji}, k)$ for $1 \le i \le N_j$ with the probability of success $q(z_{ji} = k)$. Therefore, as suggested in (Teh et al., 2008), we can use the Central Limit Theorem and approximate this term using a Gaussian distribution for $n_{jk} > 0$. Due to the independence of these Bernoulli variables, we have

$$\Pr(n_{jk} > 0) = 1 - \prod_{i=1}^{N_j} \left(1 - q(z_{ji} = k)\right), \tag{D.19}$$

$$E[n_{jk}] = E[n_{jk}|n_{jk} > 0] \Pr(n_{jk} > 0), \tag{D.20}$$

$$E[n^2_{jk}] = E[n^2_{jk}|n_{jk} > 0] \Pr(n_{jk} > 0), \tag{D.21}$$

which we can use to easily compute $E^+[n_{jk}] = E[n_{jk}|n_{jk} > 0]$ and $V^+[n_{jk}] = V[n_{jk}|n_{jk} > 0]$. We then calculate $E[r_{jk}]$ using Equation (D.18) by noting that

$$E_{\boldsymbol{z}}[\Psi(\tilde{\omega}^r_{jk} + n_{jk}) - \Psi(\tilde{\omega}^r_{jk})] \approx$$
$$\Pr(n_{jk} > 0) \left[\Psi(\tilde{\omega}^r_{jk} + E^+[n_{jk}]) - \Psi(\tilde{\omega}^r_{jk}) + \tfrac{V^+[n_{jk}]}{2}\Psi''(\tilde{\omega}^r_{jk} + E^+[n_{jk}])\right]. \tag{D.22}$$

Lastly, based on the auxiliary variable $\boldsymbol{r}$, we find that the optimal posterior distribution of the system weight stick-breaking parameters is given by $v_k \sim \text{Beta}(\tilde{\omega}^{v,1}_k, \tilde{\omega}^{v,2}_k)$, with parameters:

$$\tilde{\omega}^{v,1}_k = 1 + \sum_j E[r_{jk}] \tag{D.23}$$

$$\tilde{\omega}^{v,2}_k = \gamma + \sum_{j,k'>k} E[r_{jk'}] \tag{D.24}$$

D.2.2 System memberships

The optimal posterior over the auxiliary variables defined in Equation (D.13) implies:

$$E[\log q^*(\boldsymbol{r}|\boldsymbol{z}) - \log p(\boldsymbol{z},\boldsymbol{r} \mid \pi,\alpha)] = \sum_j \Big(\log\Gamma(\alpha+N_j) - \log\Gamma(\alpha)\Big) + \sum_{jk} E_z\Big[\log\Gamma(\tilde{\varpi}^r_{jk}) - \log\Gamma(\tilde{\varpi}^r_{jk}+n_{jk})\Big]. \tag{D.25}$$

The Gibbs free energy as a function of the posterior distribution of a single membership variable $q(z_{ji})$ becomes

$$\mathcal{F}[q(z_{ji})] = \sum_k q(z_{ji}=k)\log q(z_{ji}=k) - \sum_k E_z\Big[\log\Gamma(\tilde{\varpi}^r_{jk}+n_{jk})\Big] - \sum_k q(z_{ji}=k)\sum_s\Big[q(x_{jis}=1)E[\log\phi_{ks}] + q(x_{jis}=0)E[\log(1-\phi_{ks})]\Big] + \text{const.} \tag{D.26}$$

We can simplify the second term on the right hand side of Equation (D.26) as:

$$E_z\Big[\log\Gamma(\tilde{\varpi}_{r_{ji}}+n_{jk})\Big] = E_z\Big[\delta(z_{ji},k)\log(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk}) + \log\Gamma(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk})\Big], \tag{D.27}$$

$$= q(z_{ji}=k)E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk})] + E_{\boldsymbol{z}^{\neg ji}}[\log\Gamma(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk})], \tag{D.28}$$

where $n^{\neg ji}_{jk}$ and $\boldsymbol{z}^{\neg ji}$ indicate the exclusion of voxel i in subject j and only the first term is a function of $q(z_{ji})$. Minimizing Equation (D.26) yields the following update for membership variables:

$$q(z_{ji}=k) \propto \exp\Big\{E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk})] + \sum_s\Big(q(x_{jis}=1)E[\log\phi_{k,l}] + q(x_{jis}=0)E[\log(1-\phi_{k,s})]\Big)\Big\},$$

In order to compute the first term on the right hand side, as with the Equation (D.22), we use a Gaussian approximation for the distribution of n_{jk}:

$$E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\varpi}^r_{jk}+n^{\neg ji}_{jk})] \approx \log(\tilde{\varpi}^r_{jk}+E[n^{\neg ji}_{jk}]) - \frac{V[n^{\neg ji}_{jk}]}{2(\tilde{\varpi}^r_{jk}+E[n^{\neg ji}_{jk}])^2}. \tag{D.29}$$

■ D.2.3 Voxel activation variables

We form the Gibbs free energy as a function only of the posterior distribution of voxel activation variables $\boldsymbol{x}$. For notational convenience, we define $\psi_{jis} = \sum_k E[\log(\phi_{ks})]q(z_{ji} = k)$ and $\bar{\psi}_{jis} = \sum_{k,l} E[\log(1-\phi_{ks})]q(z_{ji} = k)$ and obtain

$$\begin{aligned}\mathcal{F}[q(\boldsymbol{x})] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \Bigg\{ \log q(\boldsymbol{x}) - \sum_{jis} \Bigg[&(1-x_{jis})\bar{\psi}_{jis} \\ &+ x_{jis} \Bigg(\psi_{jis} + E[\lambda_{ji}] \Bigg[E[a_{ji}]E[\boldsymbol{\tau}_j]^t \boldsymbol{\Omega}_{js}^t (\boldsymbol{y}_{ji} - \sum_d E[e_{jid}]\boldsymbol{f}_{jd}) \\ - \tfrac{1}{2}E[a_{ji}^2] \Bigg(E[\boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js} \boldsymbol{\tau}_j] + 2\sum_{s' \neq s} E[x_{jis'}]E[\boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js'} \boldsymbol{\tau}_j] \Bigg) \Bigg) \Bigg] \Bigg] \Bigg\} + \text{const.} \end{aligned} \quad \text{(D.30)}$$

Minimization of this function with respect to $q(\boldsymbol{x}) = \prod_{j,i,s} q(x_{jis})$ yields the update rule:

$$\begin{aligned} q(x_{jis} = 1) \propto \exp \Bigg\{ &\psi_{jis} + E[\lambda_{ji}] \Bigg[E[a_{ji}]E[\boldsymbol{\tau}_j]^t \boldsymbol{\Omega}_{js}^t (\boldsymbol{y}_{ji} - \sum_d E[e_{jid}]\boldsymbol{f}_{jd}) \\ &- \tfrac{1}{2}E[a_{ji}^2] \operatorname{Tr} \Bigg(E[\boldsymbol{\tau}_j \boldsymbol{\tau}_j^t] \Bigg(\boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js} + 2 \sum_{s' \neq s} E[x_{jis'}] \boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js'} \Bigg) \Bigg) \Bigg] \Bigg\} \end{aligned} \quad \text{(D.31)}$$

$$q(x_{jis} = 0) \propto \exp \left\{ \bar{\psi}_{jis} \right\}, \quad \text{(D.32)}$$

where $\operatorname{Tr}(\cdot)$ is the trace operator.

■ D.2.4 fMRI model variables

We collect the free energy terms corresponding to the nuisance variables $\boldsymbol{e}$:

$$\begin{aligned} \mathcal{F}[q(\boldsymbol{e})] = \int_{\boldsymbol{e}} q(\boldsymbol{e}) \Bigg(&\log q(\boldsymbol{e}) + \tfrac{1}{2} e_{jid}^2 \omega_{jd}^{e,2} - \tfrac{1}{2} e_{jid} \omega_{jd}^{e,1} + \sum_{j,i,d} \frac{E[\lambda_{ji}]}{2} \Bigg[e_{jid}^2 \|\boldsymbol{f}_{jd}\|^2 \\ &- e_{jih} \boldsymbol{f}_{jd}^t \Bigg(\boldsymbol{y}_{ji} - \sum_{d' \neq d} E[e_{jid'}] \boldsymbol{f}_{jd'} - E[a_{ji}] \sum_s E[x_{jis}] \boldsymbol{\Omega}_{js} E[\boldsymbol{\tau}_j] \Bigg) \Bigg] \Bigg) + \text{const.} \end{aligned} \quad \text{(D.33)}$$

Recall that we assume a factored form for $q(\boldsymbol{e}) = \prod_{j,i,d} q(e_{jid})$. Minimizing with respect to this distribution yields $q(e_{jid}) \propto \exp \left\{ -\frac{1}{2} \tilde{\omega}_{jid}^{e,2} e_{jid}^2 + \frac{1}{2} \tilde{\omega}_{jid}^{e,1} e_{jid} \right\}$, with the parameters $\tilde{\omega}_{jid}^{e,1}$ and $\tilde{\omega}_{jid}^{e,2}$ given in the Table 6.2.

For the activation heights $\boldsymbol{a}$, we find

$$\mathcal{F}[q(\boldsymbol{a})] = \int_{\boldsymbol{a}} q(\boldsymbol{a}) \Bigg(\log q(\boldsymbol{a}) + \tfrac{1}{2} a_{ji}^2 \omega_j^{a,2} - \tfrac{1}{2} a_{ji} \omega_j^{a,1}$$

$$+ \sum_{j,i} \frac{E[\lambda_{ji}]}{2} \Bigg[a_{ji}^2 \sum_{s,s'} E[x_{jis} x_{jis'}] E[\boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js'} \boldsymbol{\tau}_j] - a_{ji} \sum_s E[x_{jis}] E[\boldsymbol{\tau}_j]^t \boldsymbol{\Omega}_{js}^t \big(\boldsymbol{y}_{ji} - \sum_d E[e_{jid}] \boldsymbol{f}_{jd} \big) \Bigg] \Bigg) + \text{const.} \quad \text{(D.34)}$$

Assuming a factored form, minimization yields $q(a_{ji}) \propto \exp\left\{-\frac{1}{2} a_{ji}^2 \tilde{\omega}_{ji}^{a,2} + \frac{1}{2} a_{ji} \tilde{\omega}_{ji}^{a,1}\right\}, a \geq 0$, with parameters $\tilde{\omega}_{ji}^{a,1}$ and $\tilde{\omega}_{ji}^{a,2}$ given in Table 6.2.

The terms relating to the noise precisions $\boldsymbol{\lambda}$ are computed as:

$$\begin{aligned} \mathcal{F}[q(\boldsymbol{\lambda})] = \int_{\boldsymbol{\lambda}} q(\lambda) \Bigg\{ \log q(\boldsymbol{\lambda}) - \sum_{j,i} \Bigg(& \log(\lambda_{ji})(\omega_j^{\lambda,1} - 1) + \lambda_{ji} \omega_j^{\lambda,2} \\ & - \tfrac{T_j}{2} \log(\lambda_{ji}) + \tfrac{\lambda_{ji}}{2} \Bigg[\|\boldsymbol{y}_{ji}\|^2 + E[a_{ji}^2] \sum_{s,s'} E[x_{jis} x_{jis'}] E[\boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js}^t \boldsymbol{\Omega}_{js'} \boldsymbol{\tau}_j] \\ & + \sum_d \Bigg(E[e_{jid}^2] \|\boldsymbol{f}_{jd}\|^2 + \sum_{d' \neq d} E[e_{jid}] E[e_{jid'}] \boldsymbol{f}_{jd}^t \boldsymbol{f}_{jd'} \Bigg) \\ & - \boldsymbol{y}_{ji}^t \Bigg(\sum_d E[e_{jid}] \boldsymbol{f}_{jd} + E[a_{ji}] \sum_s E[x_{jis}] \boldsymbol{\Omega}_{js} E[\boldsymbol{\tau}_j] \Bigg) \\ & + E[a_{ji}] \sum_{s,d} E[e_{jid}] E[x_{jis}] \boldsymbol{f}_{jd}^t \boldsymbol{\Omega}_{js} E[\boldsymbol{\tau}_j] \Bigg] \Bigg) \Bigg\} + \text{const.} \quad \text{(D.35)} \end{aligned}$$

Minimization with respect to $q(\lambda_{ji})$ yields $q(\lambda_{ji}) \propto \exp\left\{\log(\lambda_{ji})(\tilde{\omega}_{ji}^{\lambda,1} - 1) - \lambda_{ji} \tilde{\omega}_{ji}^{\lambda,2}\right\}$, where the parameters $\tilde{\omega}_{ji}^{\lambda,1}$ and $\tilde{\omega}_{ji}^{\lambda,2}$ are given in Table 6.2. Finally, we can write the term involving the HRF as:

$$\begin{aligned} \mathcal{F}[q(\boldsymbol{\tau})] = \int_h q(\boldsymbol{\tau}) \Bigg(\log q(\boldsymbol{\tau}) + \sum_j \Bigg[& \tfrac{1}{2} \boldsymbol{\tau}_j^t \boldsymbol{\Lambda} \boldsymbol{\tau}_j + \sum_i E[\lambda_{ji}] \sum_{s,s'} E[x_{jis} x_{jis'}] \boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js'}^t \boldsymbol{\Omega}_{js} \boldsymbol{\tau}_j \\ & - \boldsymbol{\tau}_j^t \boldsymbol{\Lambda} \bar{\boldsymbol{\tau}} - \sum_{i,s} E[\lambda_{ji}] E[a_{ji}] E[x_{jis}] \boldsymbol{\tau}_j^t \boldsymbol{\Omega}_{js}^t \Bigg(\boldsymbol{y}_{ji} - \sum_d E[e_{jid}] \boldsymbol{f}_{jd} \Bigg) \Bigg] + \text{const.} \quad \text{(D.36)} \end{aligned}$$

Assuming an approximate factored posterior distribution $q(\boldsymbol{\tau}) = \prod_j q(\boldsymbol{\tau}_j)$ and minimizing the above cost function shows that the posterior for each HRF is of the form $q(\boldsymbol{\tau}_j) \propto \exp\left\{-\frac{1}{2} \boldsymbol{\tau}_j^t \boldsymbol{\Omega}_j \boldsymbol{\tau}_j + \frac{1}{2} \boldsymbol{\tau}_j^t \tilde{\boldsymbol{\omega}}_j^h\right\}$ with parameters $\tilde{\boldsymbol{\omega}}_j^h$ and $\boldsymbol{\Omega}$ presented in Table 6.2.

D.2.5 System Activation Probabilities

For the system activation profiles, we find

$$\mathcal{F}[q(\phi_{ks})] = \int_v q(\phi_{ks}) \Bigg(\log q(\phi_{ks}) - \sum_k \Bigg[\Big\{ \omega^{\phi,1} + \sum_{j,i,s} q(z_{ji} = k) q(x_{jis} = 1) \Big\} \log \phi_{ks} + \\ \Big\{ \omega^{\phi,2} + \sum_{j,i,s} q(z_{ji} = k) \Big\} \log(1 - \phi_{ks}) \Bigg] \Bigg) + \text{const.} \quad \text{(D.37)}$$

The minimum is achieved for $\phi_{ks} \sim \text{Beta}(\tilde{\omega}_{ks}^{\phi,1}, \tilde{\omega}_{ks}^{\phi,2})$, with the following parameters:

$$\tilde{\omega}_{ks}^{\phi,1} = \omega^{\phi,1} + \sum_{j,i,s} q(z_{ji} = k) q(x_{jis} = 1) \quad \text{(D.38)}$$

$$\tilde{\omega}_{ks}^{\phi,2} = \omega^{\phi,2} + \sum_{j,i,s} q(z_{ji} = k) q(x_{jis} = 0) \quad \text{(D.39)}$$

Bibliography

Aguirre, G., Zarahn, E., and D'Esposito, M. (1997). Empirical analyses of BOLD fMRI statistics. *NeuroImage*, 5(3):199–212. 31

Aguirre, G., Zarahn, E., and D'Esposito, M. (1998a). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2):373–383. 22

Aguirre, G., Zarahn, E., and D'Esposito, M. (1998b). The Variability of Human, BOLD Hemodynamic Responses. *NeuroImage*, 8(4):360–369. 31, 90

Amos, D. (1974). Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28:239–251. 117

Amunts, K., Malikovic, A., Mohlberg, H., Schormann, T., and Zilles, K. (2000). Brodmann's areas 17 and 18 brought into stereotaxic space–where and how variable? *Neuroimage*, 11(1):66–84. 48, 50

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H., and Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *The Journal of Comparative Neurology*, 412(2):319–341. 48, 50

Ances, B., Leontiev, O., Perthen, J., Liang, C., Lansing, A., and Buxton, R. (2008). Regional differences in the coupling of cerebral blood flow and oxygen metabolism changes in response to activation: implications for BOLD-fMRI. *NeuroImage*, 39(4):1510–1521. 56

Andrade, A., Kherif, F., Mangin, J., Worsley, K., Paradis, A., Simon, O., Dehaene, S., Le Bihan, D., and Poline, J. (2001). Detection of fMRI activation using cortical surface mapping. *Human Brain Mapping*, 12(2):79–93. 33

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174. 114, 144, 145

Arfanakis, K., Cordes, D., Haughton, V., Moritz, C., Quigley, M., and Meyerand, M. (2000). Combining independent component analysis and correlation analysis to probe interregional connectivity in fMRI task activation datasets. *Magnetic Resonance Imaging*, 18(8):921–930. 40

Attwell, D. and Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends in Neurosciences*, 25(12):621–625. 27

Baker, C., Liu, J., Wald, L., Kwong, K., Benner, T., and Kanwisher, N. (2007). Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Science*, 104(21):9087–9092. 22

Balslev, D., Nielsen, F., Frutiger, S., Sidtis, J., Christiansen, T., Svarer, C., Strother, S., Rottenberg, D., Hansen, L., Paulson, O., and Law, I. (2002). Cluster analysis of activity-time series in motor learning. *Human Brain Mapping*, 15(3):135–145. 42

Bandettini, P., Jesmanowicz, A., Wong, E., and Hyde, J. (1993). Processing strategies for time-course data sets in functional mri of the human brain. *Magnetic Resonance in Medicine*, 30(2):161–173. 29, 30

Bandettini, P., Wong, E., Hinks, R., Tikofsky, R., and Hyde, J. (1992). Time course EPI of human brain function during task activation. *Magnetic Resonance in Medicine*, 25(2):390–397. 27

Banerjee, A., Dhillon, I., Ghosh, J., and Sra, S. (2006). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(2):1345–1382. 59, 117

Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. (2005). Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749. 41

Baumgartner, R., Ryner, L., Richter, W., Summers, R., Jarmasz, M., and Somorjai, R. (2000a). Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis. *Magnetic Resonance Imaging*, 18(1):89–94. 40

Baumgartner, R., Scarth, G., Teichtmeister, C., Somorjai, R., and Moser, E. (1997). Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: reproducibility. *Journal of Magnetic Resonance Imaging*, 7(6):1094–1108. 40

Baumgartner, R., Somorjai, R., Summers, R., Richter, W., Ryner, L., and Jarmasz, M. (2000b). Resampling as a cluster validation technique in fMRI. *Journal of Magnetic Resonance Imaging*, 11(2):228–231. 42

Baumgartner, R., Windischberger, C., and Moser, E. (1998). Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging*, 16(2):115–125. 40

Baune, A., Sommer, F., Erb, M., Wildgruber, D., Kardatzki, B., Palm, G., and Grodd, W. (1999). Dynamical cluster analysis of cortical fMRI activation. *NeuroImage*, 9(5):477–489. 40

Beckmann, C., Jenkinson, M., and Smith, S. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage*, 20(2):1052–1063. 49

Beckmann, C., Noble, J., and Smith, S. (2000). Artefact detection in FMRI data using independent component analysis. *NeuroImage*, 11(5):S614. 40

Beckmann, C. and Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152. 40, 42, 98

Beckmann, C. and Smith, S. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1):294–311. 14, 49, 88, 96, 97, 100, 109

Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159. 39, 40

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 57(1):289–300. 36

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley & Sons, Chicheter. 114

Biswal, B. and Ulmer, J. (1999). Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of Computer Assisted Tomography*, 23(2):265–271. 39

Biswal, B., Yetkin, F., Haughton, V., and Hyde, J. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541. 28

Bowman, F. (2005). Spatio-temporal modeling of localized brain activity. *Biostatistics*, 6(4):558–575. 34

Bowman, F., Caffo, B., Bassett, S., and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, 39(1):146–156. 50

Boynton, G., Engel, S., Glover, G., and Heeger, D. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221. 29, 31

Brett, M., Johnsrude, I., and Owen, A. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, 3(3):243–249. 47, 50

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277. 31

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T., and Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *John Wiley & Sons*, 12(2):61–78. 31, 92

Burgess, N., Maguire, E., Spiers, H., OKeefe, J., Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding. *Neuron*, 23:115–125. 22

Burock, M. and Dale, A. (2000). Estimation and detection of event-related fMRI signals with temporally correlated noise: A statistically efficient and unbiased approach. *Human Brain Mapping*, 11(4):249–260. 31, 92

Buxton, R., Uludag, K., Dubowitz, D., and Liu, T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, 23:S220–S233. 29, 32

Buxton, R., Wong, E., and Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39(6):855–864. 32

Calhoun, V., Adali, T., McGinty, V., Pekar, J., Watson, T., and Pearlson, G. (2001a). fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage*, 14(5):1080–1088. 39, 54

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001b). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14(3):140–151. 49, 54

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001c). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, 13(1):43–53. 39

Carlson, T., Schrater, P., and He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, 15(5):704–717. 43

Chan, K., Lee, T., and Sejnowski, T. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation*, 15(8):1991–2011. 40

Changeux, J. and Garey, L. (1997). *Neuronal man: The biology of mind*. Princeton University Press. 20

Chau, W. and McIntosh, A. (2005). The Talairach coordinate of a point in the MNI space: how to interpret it. *Neuroimage*, 25(2):408–416. 48

Chuang, K., Chiu, M., Lin, C., and Chen, J. (1999). Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means. *IEEE Transactions on Medical Imaging*, 18(12):1117–1128. 40

Chumbley, J. and Friston, K. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1):62–70. 37

Ciuciu, P., Poline, J., Marrelec, G., Idier, J., Pallier, C., and Benali, H. (2003). Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE Transactions on Medical Imaging*, 22(10):1235–1251. 31

Clouchoux, C., Rivière, D., Mangin, J., Operto, G., Régis, J., and Coulon, O. (2010). Model-driven parameterization of the cortical surface for localization and inter-subject matching. *NeuroImage*, 50(2):552–566. 48

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M., and Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291. 22

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., and Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain*, 125(5):1054. 22

Cohen, M. (1997). Parametric Analysis of fMRI Data Using Linear Systems Methods. *NeuroImage*, 6(2):93–103. 29

Cohen, M. and Bookheimer, S. (1994). Localization of brain function using magnetic resonance imaging. *Trends in Neurosciences*, 17(7):268–277. 32

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314. 39

Cosman, E., Fisher, J., and Wells, W. (2004). Exact map activity detection in fMRI using a GLM with an Ising spatial prior. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 3217 of *LNCS*, pages 703–710. Springer. 33, 37

Coulon, O., Zilbovicius, M., Roumenov, D., Samson, Y., Frouin, V., and Bloch, I. (2000). Structural group analysis of functional activation maps. NeuroImage. *NeuroImage*, 11(6):767–782. 51

Cox, D. and Savoy, R. (2003). Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270. 23, 43

Cox, R. and Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magnetic Resonance in Medicine*, 42(6):1014–1018. 76, 97

Dale, A. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, 8(2-3):109–114. 29

Dale, A. and Buckner, R. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, 5(5):329–340. 29

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–668. 43

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58. 43

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38. 60, 114, 115

Descombes, X., Kruggel, F., and Von Cramon, D. (1998). Spatio-temporal fMRI analysis using Markov random fields. *IEEE Transactions on Medical Imaging*, 17(6):1028–1039. 33

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3(1):1–8. 22

Devlin, J., Russell, R., Davis, M., Price, C., Wilson, J., Moss, H., Matthews, P., and Tyler, L. (2000). Susceptibility-induced loss of signal: comparing PET and fMRI on a semantic task. *Neuroimage*, 11(6):589–600. 47

DeYoe, E., Carman, G., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., and Neitz, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Science*, 93(6):2382. 54

DiCarlo, J. and Cox, D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341. 21

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302. 65

Diestel, R. (2005). *Graph Theory*. Springer-Verlag, New York. 62

Downing, P., Chan, A.-Y., Peelen, M., Dodds, C., and Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex*, 16(10):1453–1461. 23, 67

Downing, P., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473. 22

Duann, J., Jung, T., Kuo, W., Yeh, T., Makeig, S., Hsieh, J., and Sejnowski, T. (2002). Single-trial variability in event-related BOLD signals. *Neuroimage*, 15(4):823–835. 39

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4):172–179. 21

Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, 26(4):309–321. 45

Engel, S., Glover, G., and Wandell, B. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral cortex*, 7(2):181. 54

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601. 22, 43

Epstein, R., Parker, W., and Feiler, A. (2007). Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *Journal of Neuroscience*, 27(23):6141. 22, 104

Esposito, F., Formisano, E., Seifritz, E., Goebel, R., Morrone, R., Tedeschi, G., and Di Salle, F. (2002). Spatial independent component analysis of functional MRI time-series: To what extent do results depend on the algorithm used? *Human Brain Mapping*, 16(3):146–157. 40

Evans, A., Collins, D., Mills, S., Brown, E., Kelly, R., and Peters, T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *IEEE Nuclear Science Symposium and Medical Imaging Conference*, pages 1813–1817. IEEE. 48

Everitt, B. and Bullmore, E. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, 7(1):1–14. 33

Fadili, M., Ruan, S., Bloyet, D., and Mazoyer, B. (2000). A multistep unsupervised fuzzy clustering analysis of fMRI time series. *Human Brain Mapping*, 10(4):160–178. 40

Farah, M. (2004). *Visual Agnosia*. The MIT Press, Cambridge, MA. 22

Fedorenko, E., Hsieh, P., Nieto-Castanon, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, 104(2):1177. 21, 51, 52

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., and Elger, C. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, 60(6):969. 48

Filzmoser, P., Baumgartner, R., and Moser, E. (1999). A hierarchical clustering method for analyzing functional MR images. *Magnetic resonance imaging*, 17(6):817–826. 40

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B., Mohlberg, H., Amunts, K., and Zilles, K. (2008). Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, 18(8):1973–1980. 48

Fischl, B., Sereno, M., and Dale, A. (1999). Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207. 33, 48

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D., Busa, E., Seidman, L., Goldstein, J., Kennedy, D., et al. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11. 50

Flandin, G. and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125. 34

Forman, S., Cohen, J., Fitzgerald, M., Eddy, W., Mintun, M., and Noll, D. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33(5):636–647. 58

Fox, P. and Raichle, M. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Science*, 83(4):1140–1144. 27

Friedman, J., Tibshirani, R., and Hastie, T. (2001). *The elements of statistical learning*. Springer–Verlag. 42

Friman, O., Borga, M., Lundberg, P., and Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage*, 19(3):837–845. 34

Friston, K. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2(1-2):56–78. 39

Friston, K., Ashburner, J., Frith, C., Poline, J., Heather, J., and Frackowiak, R. (1995a). Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189. 48

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. Academic Press, Elsevier. 36, 49

Friston, K., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., and Turner, R. (1998a). Event-Related fMRI: Characterizing Differential Responses. *NeuroImage*, 7(1):30–40. 31

Friston, K., Frith, C., Fletcher, P., Liddle, P., and Frackowiak, R. (1996). Functional topography: multidimensional scaling and functional connectivity in the brain. *Cerebral Cortex*, 6(2):156–164. 38

Friston, K., Glaser, D., Henson, R., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: applications. *NeuroImage*, 16(2):484–512. 31

Friston, K., Holmes, A., Poline, J., Grasby, P., Williams, S., Frackowiak, R., and Turner, R. (1995b). Analysis of fMRI time-series revisited. *NeuroImage*, 2(1):45–53. 31

Friston, K., Holmes, A., Price, C., B
"uchel, C., and Worsley, K. (1999). Multisubject fMRI studies and conjunction analyses. *Neuroimage*, 10(4):385–396. 49

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., Frackowiak, R., et al. (1995c). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2(4):189–210. 30, 35

Friston, K., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171. 29, 31, 32

Friston, K., Josephs, O., Rees, G., and Turner, R. (1998b). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39(1):41–52. 31

Friston, K., Josephs, O., Zarahn, E., Holmes, A., Rouquette, S., and Poline, J. (2000a). To Smooth or Not to Smooth?:: Bias and Efficiency in fMRI Time-Series Analysis. *NeuroImage*, 12(2):196–208. 31

Friston, K., Mechelli, A., Turner, R., and Price, C. (2000b). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477. 32

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and Bayesian inference in neuroimaging: theory. *NeuroImage*, 16(2):465–483. 37, 49

Gazzaniga, M. (2004). *The cognitive neurosciences*. The MIT Press. 21

Gee, J., Alsop, D., and Aguirre, G. (1997). Effect of spatial normalization on analysis of functional data. In *Proceedings of SPIE, Medical Imaging*, volume 3034, pages 550–560. 48

Genovese, C. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data. *Journal of the American Statistical Association*, 95(451):691–703. 31

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4):870–878. 36

Glover, G. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416–429. 31

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, 27(5):392–401. 48

Goense, J. and Logothetis, N. (2008). Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current Biology*, 18(9):631–640. 27

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., and Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fMRI. *Magnetic Resonance in Medicine*, 40(2):249–260. 40

Golland, P., Golland, Y., and Malach, R. (2007). Detection of spatial activation patterns as unsupervised segmentation of fMRI data. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 4791 of *LNCS*, pages 110–118. Springer. 41

Golland, Y., Golland, P., Bentin, S., and Malach, R. (2008). Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems. *Neuropsychologia*, 46(2):540–553. 41

Good, P. and Good, P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer-Verlag. 37

Gössl, C., Auer, D., and Fahrmeir, L. (2001a). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562. 33

Gössl, C., Fahrmeir, L., and Auer, D. (2001b). Bayesian modeling of the hemodynamic response function in BOLD fMRI. *NeuroImage*, 14(1):140–148. 31

Goutte, C., Hansen, L., Liptrot, M., and Rostrup, E. (2001). Feature-space clustering for fMRI meta-analysis. *Human Brain Mapping*, 13(3):165–183. 42, 65

Goutte, C., Toft, P., Rostrup, E., Nielsen, F., and Hansen, L. (1999). On clustering fMRI time series. *NeuroImage*, 9(3):298–310. 42, 65

Greicius, M., Krasnow, B., Reiss, A., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Science*, 100(1):253–258. 28, 39

Greve, D. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72. 47, 76, 97

Grill-Spector, K. and Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27:649–677. 21

Handwerker, D., Ollinger, J., and D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651. 91

Hanson, S., Matsuka, T., and Haxby, J. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a face area? *NeuroImage*, 23(1):156–166. 23, 45

Hardoon, D., Mourao-Miranda, J., Brammer, M., and Shawe-Taylor, J. (2007). Unsupervised analysis of fmri data using kernel canonical correlation. *NeuroImage*, 37(4):1250–1259. 38

Hartvig, N. (2002). A stochastic geometry model for functional magnetic resonance images. *Scandinavian Journal of Statistics*, 29(3):333–353. 34

Hartvig, N. and Jensen, J. (2000). Spatial mixture modeling of fMRI data. *Human Brain Mapping*, 11(4):233–248. 33

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430. 23, 42, 43, 53, 86

Haxby, J., Hoffman, E., and Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–232. 23

Hayasaka, S. and Nichols, T. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, 20(4):2343–2356. 37

Hayasaka, S. and Nichols, T. (2004). Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*, 23(1):54–63. 37

Haynes, J. and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visfual cortex. *Nature Neuroscience*, 8(5):686–691. 43

Haynes, J. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534. 43

Heeger, D. and Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, 3(2):142–151. 27

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D., Evans, A., Malandain, G., Ayache, N., Christensen, G., et al. (2003). Retrospective evaluation of intersubject brain registration. *IEEE Transactions on Medical Imaging*, 22(9):1120–1130. 50

Højen-Sørensen, P. et al. (2002a). Analysis of functional neuroimages using ICA with adaptive binary sources. *Neurocomputing*, 49(1-4):213–225. 40

Højen-Sørensen, P., Winther, O., and Hansen, L. (2002b). Mean-field approaches to independent component analysis. *Neural Computation*, 14(4):889–918. 40

Hollander, M. and Wolfe, D. (1999). *Nonparametric statistical methods*. Wiley–Interscience. 37

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106. 20

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215. 20

Huettel, S., Song, A., and McCarthy, G. (2004). *Functional magnetic resonance imaging*. Sinauer Associates, Sunderland, MA. 27, 28

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634. 40

Hyvärinen, A. (1999b). The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10(1):1–5. 40

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430. 39

Ishai, A., Ungerleider, L., Martin, A., Schouten, J., and Haxby, J. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Science*, 96(16):9379–9384. 23

Jain, A. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666. 40

Jarmasz, M. and Somorjai, R. (2003). EROICA: exploring regions of interest with cluster analysis in large functional magnetic resonance imaging data sets. *Concepts in Magnetic Resonance Part A*, 16A(1):50–62. 40

Jbabdi, S., Woolrich, M., and Behrens, T. (2009). Multiple-subjects connectivity-based parcellation using hierarchical dirichlet process mixture models. *NeuroImage*, 44(2):373–384. 87

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233. 114

Josephs, O., Turner, R., and Friston, K. (1997). Event-related fMRI. *Human brain mapping*, 5(4):243–248. 31

Joshi, A., Shattuck, D., Thompson, P., and Leahy, R. (2007). Surface-constrained volumetric brain registration using harmonic mappings. *IEEE Transactions on Medical Imaging*, 26(12):1657–1669. 48

Joshi, S., Miller, M., and Grenander, U. (1997). On the geometry and shape of brain sub-manifolds. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(8):1317–1343. 48

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679. 43

Kanwisher, N. (2003). The ventral visual object pathway in humans: evidence from fMRI. In Chalupa, L. and Wener, J., editors, *The Visual Neurosciences*, pages 1179–1189. MIT Press. 21, 43

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Science*, 107(25):11163. 20, 21, 23

Kanwisher, N., McDermott, J., and Chun, M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311. 22, 43, 51

Kanwisher, N. and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical transactions of the Royal Society. Series B, Biological Sciences*, 361(1476):2109–2128. 22, 104

Katanoda, K., Matsuda, Y., and Sugishita, M. (2002). A spatio-temporal regression model for the analysis of functional MRI data. *NeuroImage*, 17(3):1415–1428. 33

Keller, M., Roche, A., Tucholka, A., and Thirion, B. (2008). Dealing with spatial normalization errors in fMRI group inference using hierarchical modeling. *Statistica Sinica*, 18(4):1357–1374. 52

Kim, J., Fisher, J., Tsai, A., Wible, C., Willsky, A., and Wells, W. (2000). Incorporating spatial priors into an information theoretic approach for fMRI data analysis. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 1935 of *LNCS*, pages 62–71. Springer. 33

Kim, S. and Smyth, P. (2007). Hierarchical Dirichlet processes with random effects. *Advances in Neural Information Processing Systems*, 19:697–704. 52, 87

Kim, S., Smyth, P., and Stern, H. (2006). A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 4191 of *LNCS*, pages 217–224. Springer. 34

Kim, S., Smyth, P., and Stern, H. (2010). A Bayesian mixture approach to modeling spatial activation patterns in multisite fMRI data. *IEEE Transactions on Medical Imaging*, 29(6):1260–1274. 52

Kiviniemi, V., Kantola, J., Jauhiainen, J., Hyvärinen, A., and Tervonen, O. (2003). Independent component analysis of nondeterministic fMRI signal sources. *Neuroimage*, 19(2):253–260. 39

Klein, A., Andersson, J., Ardekani, B., Ashburner, J., Avants, B., Chiang, M., Christensen, G., Collins, D., Gee, J., Hellier, P., et al. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802. 50

Klein, A., Ghosh, S., Avants, B., Yeo, B., Fischl, B., Ardekani, B., Gee, J., Mann, J., and Parsey, R. (2010). Evaluation of volume-based and surface-based brain image registration methods. *Neuroimage*, 51(1):214–220. 50

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press. 113

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953. 21

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141. 38, 45, 85, 86

Kuhn, H. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97. 63, 97

Kwong, K., Belliveau, J., Chesler, D., Goldberg, I., Weisskoff, R., Poncelet, B., Kennedy, D., Hoppel, B., Cohen, M., and Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Science*, 89(12):5675–5679. 27, 29

Lange, T., Roth, V., Braun, M., and Buhmann, J. (2004). Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323. 105

Langs, G., Lashkari, D., and Sweet, A. (2011). Learning the atlas of a cognitive process in its functional geometry. *manuscript under review*. 65

Langs, G., Tie, Y., Rigolo, L., Golby, A., and Golland, P. (2010). Functional geometry alignment and localization of brain areas. In *Advances in Neural Information Processing Systems*, volume 23, pages 1225–1233. MIT Press. 54, 65

Lashkari, D., Sridharan, R., Vul, E., Hsieh, P., Kanwisher, N., and Golland, P. (2010a). Nonparametric hierarchical Bayesian model for functional brain parcellation. In *Proceedings of MMBIA: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 15–22. IEEE. 106

Lashkari, D., Vul, E., Kanwisher, N., and Golland, P. (2010b). Discovering structure in the space of fMRI selectivity profiles. *NeuroImage*, 50(3):1085–1098. 14, 95, 97, 100

Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. 38

Liu, D., Lu, W., and Zhong, N. (2010). Clustering of fMRI Data Using Affinity Propagation. *Brain Informatics*, 6334:399–406. 42

Livingstone, M. and Hubel, D. (1995). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Frontiers in Cognitive Neuroscience*, 240:24. 20

Logan, B. and Rowe, D. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, 22(1):95–108. 37

Logothetis, N. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197):869–878. 27, 28

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157. 27

Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621. 21

Logothetis, N. and Wandell, B. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, 66:735–770. 27

Lukic, A., Wernick, M., Tzikas, D., Chen, X., Likas, A., Galatsanos, N., Yang, Y., Zhao, F., and Strother, S. (2007). Bayesian kernel methods for analysis of functional neuroimages. *IEEE Transactions on Medical Imaging*, 26(12):1613–1624. 34

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK. 40, 41

Magistretti, P., Pellerin, L., Rothman, D., and Shulman, R. (1999). Energy on demand. *Science*, 283(5401):496–497. 27

Makni, S., Ciuciu, P., Idier, J., and Poline, J. (2005). Joint detection-estimation of brain activity in functional MRI: a multichannel deconvolution solution. *IEEE Transactions on Signal Processing*, 53(9):3488–3502. 31, 33, 91, 107

Makni, S., Idier, J., Vincent, T., Thirion, B., Dehaene-Lambertz, G., and Ciuciu, P. (2008). A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *NeuroImage*, 41(3):941–969. 31, 34, 56

Mardia, K. (1975). Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 349–393. 59

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, 19(1):1–17. 31

Marrelec, G., Benali, H., Ciuciu, P., and Poline, J.-B. (2002). Bayesian estimation of the hemodynamic response function in functional MRI. In *Bayesian Inference and Maximum Entropy Methods Workshop*, volume 617 of *AIP*, pages 229–247. IOP Institute of Physics. 92

McCarthy, G., Puce, A., Gore, J., and Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9(5):605–610. 22

McKeown, M. (2000). Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage*, 11(1):24–35. 42

McKeown, M., Hansen, L., and Sejnowsk, T. (2003). Independent component analysis of functional MRI: What is signal and what is noise? *Current Opinion in Neurobiology*, 13(5):620–629. 40

McKeown, M., Jung, T., Makeig, S., Brown, G., Kindermann, S., Lee, T., and Sejnowski, T. (1998a). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. *Proceedings of the National Academy of Science*, 95(3):803–810. 39

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998b). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188. 39

McKeown, M. and Sejnowski, T. (1998). Independent component analysis of fMRI data: examining the assumptions. *Human Brain Mapping*, 6(5-6):368–372. 39

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. 58

Miezin, F., Maccotta, L., Ollinger, J., Petersen, S., and Buckner, R. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, 11(6):735–759. 56

Minka, T. (2001). Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems*, pages 598–604. 98

Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., and Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175. 43

Möller, U., Ligges, M., Georgiewa, P., Grünling, C., Kaiser, W., Witte, H., and Blanz, B. (2002). How to avoid spurious cluster validation? A methodological investigation on simulated and fMRI data. *NeuroImage*, 17(1):431–446. 42

Moser, E., Baumgartner, R., Barth, M., and Windischberger, C. (1999). Explorative signal processing in functional MR imaging. *International Journal of Imaging Systems and Technology*, 10(2):166–176. 40

Moser, E., Diemling, M., and Baumgartner, R. (1997). Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part II: quantification. *Journal of Magnetic Resonance Imaging*, 7(6). 40

Neumann, J., Von Cramon, D., Forstmann, B., Zysset, S., and Lohmann, G. (2006). The parcellation of cortical areas using replicator dynamics in fMRI. *Neuroimage*, 32(1):208–219. 38

Nichols, T. and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25. 37

Nieto-Castanon, A., Ghosh, S., Tourville, J., and Guenther, F. (2003). Region of interest based analysis of functional imaging data. *Neuroimage*, 19(4):1303–1316. 50, 53

Norman, K., Polyn, S., Detre, G., and Haxby, J. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430. 43

Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science*, 87(24):9868–9872. 27

Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Science*, 89(13):5951–5955. 27

Ono, M., Kubik, S., and Abernathey, C. (1990). *Atlas of the cerebral sulci*. Georg Thieme Verlag. 48

Op de Beeck, H., Haushofer, J., and Kanwisher, N. (2008). Interpreting fMRI data: maps, modules and dimensions. *Nature Reviews Neuroscience*, 9(2):123–135. 23, 43

Operto, G., Clouchoux, C., Bulot, R., Anton, J., and Coulon, O. (2008). Surface-based structural group analysis of fMRI data. *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, 5241:959–966. 51

O'Toole, A., Jiang, F., Abdi, H., Pénard, N., Dunlop, J., and Parent, M. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, 19(11):1735–1752. 43

Ou, W. and Golland, P. (2005). From spatial regularization to anatomical priors in fmri analysis. In *Proceedings of IPMI: International Conference on Information Processing in Medical Imaging*, volume 3565 of *LNCS*, pages 88–100. Springer. 33

Ou, W., Wells III, W., and Golland, P. (2010). Combining spatial priors and anatomical information for fMRI detection. *Medical Image Analysis*, 14(3):318–331. 33

Peelen, M. and Downing, P. (2007). The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648. 22

Penny, W. and Friston, K. (2003). Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22(4):504–514. 33

Penny, W. and Holmes, A. (2003). Random effects analysis. In R.S.J., F., K.J., F., and C.D., F., editors, *Human Brain Function II*, pages 843–850. Elsevier, Oxford. 49

Penny, W., Trujillo-Barreto, N., and Friston, K. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362. 31, 33

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1S1):199–209. 43

Perrett, D., Rolls, E., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47(3):329–342. 22

Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J., and Dehaene, S. (2007). Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neuroscience*, 8(1):91. 51

Pitman, J. (2002). Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5):501–514. 94

Purdon, P., Solo, V., Weisskoff, R., and Brown, E. (2001). Locally regularized spatiotemporal modeling and model comparison for functional MRI. *NeuroImage*, 14(4):912–923. 33

Purdon, P. and Weisskoff, R. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, 6(4):239–249. 31

Quigley, M., Haughton, V., Carew, J., Cordes, D., Moritz, C., and Meyerand, M. (2002). Comparison of independent component analysis and conventional hypothesis-driven analysis for clinical functional MR image processing. *American Journal of Neuroradiology*, 23(1):49. 39

Raichle, M. (1998). Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proceedings of the National Academy of Science*, 95(3):765. 27

Raichle, M. and Mintun, M. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, 29:449–476. 27

Rajapakse, J. and Piyaratna, J. (2001). Bayesian approach to segmentation of statistical parametric maps. *IEEE Transactions on Biomedical Engineering*, 48(10):1186–1194. 33

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12:554–560. 114

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025. 21

Rosen, B., Buckner, R., and Dale, A. (1998). Event-related functional MRI: past, present, and future. *Proceedings of the National Academy of Science*, 95(3):773. 29

Rossion, B., Caldara, R., Seghier, M., Schuller, A., Lazeyras, F., and Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, 126(11):2381–2395. 22

Roy, C. and Sherrington, C. (1890). On the regulation of the blood-supply of the brain. *The Journal of physiology*, 11(1-2):85. 27

Sabuncu, M., Singer, B., Conroy, B., Bryan, R., Ramadge, P., and Haxby, J. (2010). Function-based intersubject alignment of human cortical anatomy. *Cerebral Cortex*, 20(1):130. 52

Salli, E., Aronen, H., Savolainen, S., Korvenoja, A., and Visa, A. (2002). Contextual clustering for analysis of functional MRI data. *IEEE Transactions on Medical Imaging*, 20(5):403–414. 33

Salli, E., Korvenoja, A., Visa, A., Katila, T., and Aronen, H. (2001). Reproducibility of fMRI: effect of the use of contextual information. *NeuroImage*, 13(3):459–471. 33

Saxe, R., Brett, M., and Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage*, 30(4):1088–1096. 50, 51

Saxe, R. and Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in" theory of mind". *NeuroImage*, 19(4):1835–1842. 21

Saxe, R. and Powell, L. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699. 21

Schacter, D., Buckner, R., Koutstaal, W., Dale, A., and Rosen, B. (1997). Late Onset of Anterior Prefrontal Activity during True and False Recognition: An Event-Related fMRI Study. *NeuroImage*, 6(4):259–269. 56

Schiller, P. (1996). On the specificity of neurons and visual areas. *Behavioural Brain Research*, 76(1–2):21–35. 20

Schmithorst, V. and Holland, S. (2004). Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. *Journal of Magnetic Resonance Imaging*, 19(3):365–368. 54

Schummers, J., Yu, H., and Sur, M. (2008). Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. *Science*, 320(5883):1638–1643. 27

Schwarzlose, R., Baker, C., and Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience*, 25(47):11055–11059. 22

Seghier, M., Friston, K., and Price, C. (2007). Detecting subject-specific activations using fuzzy clustering. *Neuroimage*, 36(3):594–605. 65

Slichter, C. (1990). *Principles of magnetic resonance*. Springer Verlag. 27

Smith, S., Beckmann, C., Ramnani, N., Woolrich, M., Bannister, P., Jenkinson, M., Matthews, P., and McGonigle, D. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping*, 24(3):248–257. 56

Smith, S. and Brady, J. (1997). SUSANA new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78. 32

Solo, V., Purdon, P., Weisskoff, R., and Brown, E. (2001). A signal estimation approach to functional MRI. *IEEE Transactions on Medical Imaging*, 20(1):26–35. 33

Spiridon, M., Fischl, B., and Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human Brain Mapping*, 27(1):77–89. 48

Svensén, M., Kruggel, F., and Benali, H. (2002a). ICA of fMRI group study data. *NeuroImage*, 16(3):551–563. 54

Svensén, M., Kruggel, F., and Von Cramon, D. (2002b). Probabilistic modeling of single-trial fMRI data. *IEEE Transactions on Medical Imaging*, 19(1):25–35. 33, 34

Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme, New York. 48

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139. 21, 22

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 87, 94, 95, 114

Teh, Y., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for HDP. *Advances in Neural Information Processing Systems*, 20:1481–1488. 87, 94, 144, 146

Teh, Y., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 19:1353–1360. 95

Thirion, B. and Faugeras, O. (2003a). Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage*, 20(1):34–49. 39

Thirion, B. and Faugeras, O. (2003b). Feature detection in fMRI data: the information bottleneck approach. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 2879 of *LNCS*, pages 83–91. Springer. 42

Thirion, B. and Faugeras, O. (2004). Feature characterization in fMRI data: the Information Bottleneck approach. *Medical Image Analysis*, 8(4):403–419. 42, 55

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., and Poline, J. (2006). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Human Brain Mapping*, 27(8):678–693. 51

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., and Poline, J. (2007a). Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. *NeuroImage*, 35(1):105–120. 50, 51, 56

Thirion, B., Pinel, P., Tucholka, A., Roche, A., Ciuciu, P., Mangin, J., and Poline, J.-B. (2007b). Structural analysis of fMRI data revisited: Improving the sensitivity and reliability of fMRI group studies. *IEEE Transactions on Medical Imaging*, 26(9):1256–1269. 51, 52

Thirion, B., Tucholka, A., Keller, M., Pinel, P., Roche, A., Mangin, J., and Poline, J. (2007c). High level group analysis of FMRI data based on Dirichlet process mixture models. In *Proceedings of IPMI: International Conference on Information Processing in Medical Imaging*, volume 4584 of *LNCS*, pages 482–494. Springer. 51, 87

Thomas, C., Harshman, R., and Menon, R. (2002). Noise reduction in BOLD-based fMRI using component analysis. *Neuroimage*, 17(3):1521–1537. 40

Thyreau, B., Thirion, B., Flandin, G., and Poline, J. (2006). Anatomo-functional description of the brain: a probabilistic approach. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, volume 5. IEEE. 34

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622. 39

Tsao, D., Freiwald, W., Tootell, R., and Livingstone, M. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670. 22

Tucholka, A., Thirion, B., Perrot, M., Pinel, P., Mangin, J., and Poline, J. (2008). Probabilistic anatomo-functional parcellation of the cortex: how many regions? *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, 5241:399–406. 34

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley. 34

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289. 50

Ullman, S. (2003). Approaches to visual recognition. In Kanwisher, N. and Duncan, J., editors, *Functional Neuroimaging of Visual Cognition*, pages 143–167. Oxford University Press. 21

van de Ven, V., Formisano, E., Prvulovic, D., Roeder, C., and Linden, D. (2004). Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Human brain mapping*, 22(3):165–178. 39

Van Essen, D. (2005). A population-average, landmark-and surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*, 28(3):635–662. 48

Vincent, T., Risser, L., and Ciuciu, P. (2010). Spatially Adaptive Mixture Modeling for Analysis of fMRI Time Series. *IEEE Transactions on Medical Imaging*, 29(4):1059–1074. 34

Wang, X., Grimson, W., and Westin, C. (2009). Tractography segmentation using a hierarchical dirichlet processes mixture model. In *Proceedings of IPMI: International Conference on Information Processing in Medical Imaging*, volume 5636 of *LNCS*, pages 101–113. Springer. 87

Wei, X., Yoo, S., Dickey, C., Zou, K., Guttmann, C., and Panych, L. (2004). Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *NeuroImage*, 21(3):1000–1008. 48, 56

Wink, A. and Roerdink, J. (2004). Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing. *IEEE Transactions on Medical Imaging*, 23(3):374–387. 32

Winther, O. and Petersen, K. (2007). Flexible and efficient implementations of Bayesian independent component analysis. *Neurocomputing*, 71(1-3):221–233. 40

Wohlschläger, A., Specht, K., Lie, C., Mohlberg, H., Wohlschläger, A., Bente, K., Pietrzyk, U., Stöcker, T., and Zilles, K. (2005). Linking retinotopic fMRI mapping and anatomical probability maps of human occipital areas V1 and V2. *Neuroimage*, 26(1):73–82. 50

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. (2004a). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747. 49

Woolrich, M., Behrens, T., Beckmann, C., and Smith, S. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *IEEE Transactions on Medical Imaging*, 24(1):1–11. 33

Woolrich, M., Behrens, T., and Smith, S. (2004b). Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21(4):1748–1761. 31

Woolrich, M., Jenkinson, M., Brady, J., and Smith, S. (2004c). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Transactions on Medical Imaging*, 23(2):213–231. 31, 33

Woolrich, M., Ripley, B., Brady, M., and Smith, S. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14(6):1370–1386. 31, 92, 97

Worsley, K. and Friston, K. (1995). Analysis of fMRI time-series revisitedagain. *NeuroImage*, 2(3):173–181. 31

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15. 49

Worsley, K., Marrett, S., Neelin, P., and Evans, A. (1996a). Searching scale space for activation in PET images. *Human Brain Mapping*, 4(1):74–90. 32

Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., and Evans, A. (1996b). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73. 36

Xu, L., Johnson, T., Nichols, T., and Nee, D. (2009). Modeling Inter-Subject Variability in fMRI Activation Location: A Bayesian Hierarchical Spatial Model. *Biometrics*, 65(4):1041–1051. 52, 110

Yang, J., Papademetris, X., Staib, L., Schultz, R., and Duncan, J. (2004). Functional brain image analysis using joint function-structure priors. In *Proceedings of MICCAI: International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 3217 of *LNCS*, pages 736–744. Springer. 53

Yeo, B., Sabuncu, M., Vercauteren, T., Holt, D., Amunts, K., Zilles, K., Golland, P., and Fischl, B. (2010). Learning Task-Optimal Registration Cost Functions for Localizing Cytoarchitecture and Function in the Cerebral Cortex. *IEEE Transactions on Medical Imaging*, 29(7):1424–1441. 52

Yeo, B., Sepulcre, J., Sabuncu, M., Lashkari, D., Roffman, J., Smoller, J., Fischl, B., Liu, H., and Buckner, R. (2011). Estiamtes of surface-based cortical networks using intrinsic functional connectivity from 1000 subjects. *manuscript under review*. 66

Zarahn, E., Aguirre, G., and D'Esposito, M. (1997a). A trial-based experimental design for fMRI. *NeuroImage*, 6(2):122–138. 29

Zarahn, E., Aguirre, G., and D'Esposito, M. (1997b). Empirical analyses of BOLD fMRI statistics. *NeuroImage*, 5(3):179–197. 31

Zeki, S. (2005). Behind the seen: the functional specialization of the brain in space and time. *Philosophical Transactions of the Royal Society. Series B, Biological Sciences*, 360(1458):1145. 20

Zola-Morgan, S. (1995). Localization of brain function: The legacy of Franz Joseph Gall (1758-1828). *Annual Eeview of Neuroscience*, 18(1):359–383. 20