Nonparametric Hierarchical Bayesian Model for Functional Brain Parcellation

Danial Lashkari[†] Ramesh Sridharan[†] Edward Vul[‡] Po-Jang Hsieh[‡]
Nancy Kanwisher[‡] Polina Golland[†]

[†]Computer Science and Artificial Intelligence Laboratory, MIT
[‡] Department of Brain and Cognitive Sciences, MIT
77 Massachusetts Avenue, Cambridge, MA 02139

Abstract

We develop a method for unsupervised analysis of functional brain images that learns group-level patterns of functional response. Our algorithm is based on a generative model that comprises two main layers. At the lower level, we express the functional brain response to each stimulus as a binary activation variable. At the next level, we define a prior over the sets of activation variables in all subjects. We use a Hierarchical Dirichlet Process as the prior in order to simultaneously learn the patterns of response that are shared across the group, and to estimate the number of these patterns supported by data. Inference based on this model enables automatic discovery and characterization of salient and consistent patterns in functional signals. We apply our method to data from a study that explores the response of the visual cortex to a collection of images. The discovered profiles of activation correspond to selectivity to a number of image categories such as faces, bodies, and scenes. More generally, our results appear superior to the results of alternative data-driven methods in capturing the category structure in the space of stimuli.

1. Introduction

Functional MRI studies are typically driven by a priori hypotheses. Typically, an experiment is designed based on a hypothesis and significance tests are used to localize the relevant functional regions in the brain. However, finding a good functional hypothesis is not always straightforward, especially in the presence of multiple patterns of functional specificity. For instance, the studies of visual object recognition assume that a certain category of images activates selective areas in the brain, and look for such regions using significance tests. This approach has been successfully used to discover areas selective for categories such as faces, bodies, and scenes in the cortex [5]. Considering the number of possible image categories, however, we face an exceedingly large space of likely selectivity patterns if we go beyond the obvious categories studied before. The same problem arises in other domains, e.g., the studies of language or auditory networks, if we aim to investigate the functional specificity structure in detail.

An alternative approach is to present a variety of stimuli relevant to the network under study and to apply data-driven fMRI analysis to generate appropriate hypotheses. Data-driven methods decompose the data into a number of components, each describing one temporal (functional) pattern and its corresponding spatial extent. A popular method for this decomposition is spatial Independent Component Analysis (ICA) [2, 16] wherein the goal is to make the components spatially independent. Beyond exploratory analysis, ICA is also aimed at automatic denoising of the data. Therefore, it assumes an additive model for the data and allows spatially overlapping components. However, neither of these assumptions is appropriate for studying functional specificity. For instance, an fMRI response that is a weighted combination of a component selective for scissor images and another selective for hammer images may be better described by selectivity for tools. In this case, rather than adding several components, we should associate each voxel with only one pattern of response. Moreover, common extensions of ICA to fMRI analysis require voxel-wise spatial normalization while some functional areas appear in highly variable locations across subjects [24].

Clustering, another data-driven method, has also been used to segment the fMRI data based on the time courses or protocol-related features [1, 8, 9]. Clustering is more naturally suited to the studies of functional specificity since it assigns each voxel to only one cluster. Recent work has shown that, by representing the data as a set of voxel responses to different conditions, clustering can reveal meaningful functional patterns in the brain [13, 14, 23]. However, the current methods for functional segmentation either lack a proper model for group analysis, or do not determine how to choose the number of clusters.

In this paper, we present a novel nonparametric hierarchical model for functional brain segmentation that applies to group fMRI data and automatically determines the num-

978-1-4244-7028-0/10/$26.00 ©2010 IEEE

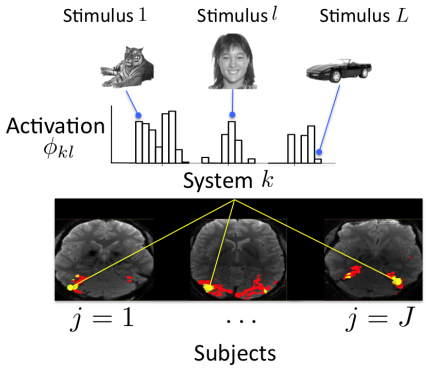

Figure 1. Schematic diagram illustrating the concept of a system. System k is characterized by vector $[\phi_{k1}, \cdots, \phi_{kL}]^T$ that specifies the level of activation induced in the system by each of the L stimuli. This system describes a pattern of response demonstrated by collections of voxels in all J subjects in the group.

ber of clusters. Our model builds upon the basic framework of Hierarchical Dirichlet Processes (HDP) [20] for sharing structure in a group of datasets. Our structures of interest are salient patterns of functional specificity, i.e., groups of voxels with similar functional responses. The nonparametric aspect of the model allows automatic search in the space of models with different numbers of systems. Moreover, we provide a model for fMRI responses that removes the need for the heuristic normalization schemes commonly used as a preprocessing step [14]. Notably, this model transforms fMRI responses into a set of activation probabilities. Therefore, the activation profiles of systems can be naturally interpreted as signatures of functional specificity: they describe the probability that any stimulus or task activates a given functional parcel. 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. Based on the model, we derive a scalable algorithm using a variational Bayesian approximation.

Nonparametric Bayesian models have been previously employed in fMRI data analysis, particularly in modeling the spatial structure in localization maps found via significance tests [12, 25]. Our probabilistic model is more closely related to a recent application of HDPs to DTI data where anatomical connectivity profiles of voxels are clustered across subjects [11]. In contrast to all these 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 [21].

This paper is organized as follows. We begin by describing the two layers of the model and our variational inference procedure in Sec. 2. We present experimental results in Sec. 3 and compare them with results found by tensorial group ICA [3] and a finite mixture-model clustering model [14]. Finally, we conclude in Sec. 4.

2. Nonparametric Hierarchical Bayesian Model for fMRI Data

Consider an fMRI experiment with a relatively large number of different tasks or stimuli, e.g., passive viewing of L distinct images in a visual study. The set of raw time courses $\{b_{it}\}$ describe the acquired BOLD signals at acquisition times t in voxels i. We can estimate the fMRI responses of voxels to each of the L experimental conditions using the standard General Linear Model for BOLD time courses [7]. As a result, the data is represented as a set of values y_{il} that describe the fMRI response of voxel i to stimulus l. Commonly, studies include data from several subjects; therefore, we sometimes denote the fMRI responses by y_{jil} where index j identifies different subjects in the group.

Our generative model explains the fMRI responses of voxels assuming a clustering structure that is shared across subjects. To distinguish these functionally-defined clusters from the traditional cluster analysis used for the correction of significance maps [6], we follow the terminology used in [14] and refer to these clusters as functional systems. Fig. 1 illustrates the idea of a system as a collection of voxels across all subjects that share a coherent pattern of functional response. We assume an infinite number of group level systems; each system k comes with a weight π_k that specifies the prior probability that it includes any given voxel. In this way, different draws from the same distribution potentially yield different finite numbers of clusters. Due to inter-subject variability and noise, the group-level system weight π is independently perturbed for each subject j to generate a subject-specific weight vector β_j. System k is also characterized by a vector $[\phi_{k1}, \cdots, \phi_{kL}]^T$ where L is the number of stimuli. Here, $\phi_{kl} \in [0, 1]$ is the probability that system k is activated by stimulus l. Based on the weight β_j and the system probabilities $\boldsymbol{\phi}$, we generate binary activation variables $x_{jil} \in \{0, 1\}$ that express whether or not voxel i in subject j is activated by stimulus l.

Up to this point, our model has the structure of a standard HDP. The next layer of this hierarchical model defines how activation variables x_{jil} generate response values y_{jil}. Our model for the fMRI response employs a set of voxel-specific response parameters $\boldsymbol{\mu}_{ji} = (\mu_{ji}, a_{ji}, \lambda_{ji})$ to express this connection. Table 1 presents a summary of all the variables and parameters and Fig. 2 shows the structure of our graphical model. Next, we discuss the details of the two layers in our model.

2.1. Model of fMRI Responses

We assume that response y_{jil} in each voxel to each stimulus can be described as a mixture of two modes: active and non-active. The expected response of the active mode is strictly greater than the expected response of the non-active

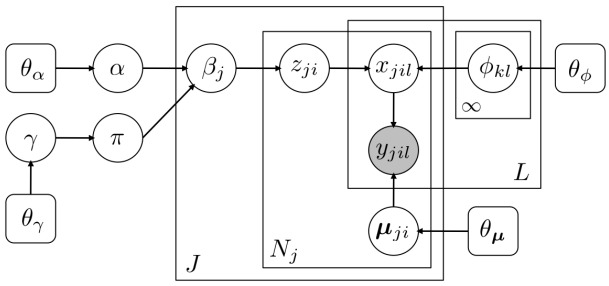

Figure 2. The graphical model for the joint distribution implied by the fMRI response model in Sec. 2.1 and the HDP prior in Sec. 2.2.

mode. This mixture distribution depends on the binary activation variable x_{jil} that specifies the mode of the response, and variables $\boldsymbol{\mu}_{ji} = (\mu_{ji}, a_{ji}, \lambda_{ji})$ that describe the parameters of the mixture. Formally, we have

$$y_{jil} \mid (x_{jil} = 0, \boldsymbol{\mu}_{ji}) \sim \text{Normal}\left(\mu_{ji}, \lambda_{ji}^{-1}\right), \quad (1)$$

$$y_{jil} \mid (x_{jil} = 1, \boldsymbol{\mu}_{ji}) \sim \text{Normal}\left(\mu_{ji} + a_{ji}, \lambda_{ji}^{-1}\right), \quad (2)$$

where μ_{ji} is the expected value of the non-active response, $a_{ji} > 0$ is the increase in the expected response due to activation, and λ_{ji} is the reciprocal of the variance of the *i.i.d.* white noise.

The goal of this layer of the model is to transform the response values so that we can effectively compare them across voxels. To show why this matters, Fig. 3 (top) shows the elements of the fMRI response vectors $[y_{i1}, \cdots, y_{iL}]^T$ for a number of voxels detected to be selective for the same category of stimuli through conventional analysis. In order to detect these voxels, we form a contrast that assumes that the responses to the preferred stimuli are on average higher than the responses to the rest. However, since the responses of different voxels have different ranges, we cannot directly compare response vectors. This phenomenon can be explained by the effect of excessive noise or the inaccuracy of the models used in the preprocessing stage. Our model encodes the relevant information as binary activation variables that describe the states of voxels relative to their own dynamic ranges. We note that the idea of describing the response of active and non-active voxels by a mixture of distributions has been used before in the conventional detection framework [15]. Here, we couple this signal model with the HDP prior on activation variables as described later in this section.

We also assume the following priors on the distribution of voxel response variables parameterized by $\theta_{\boldsymbol{\mu}} = (\theta_\mu, \theta_a, \theta_\lambda)$:

$$\mu_{ji} \mid \theta_\mu \sim \text{Normal}\left(\theta_{\mu,1}, \theta_{\mu,2}^{-1}, \theta_{\mu,2}^{-1}\right), \quad (3)$$

$$a_{ji} \mid \theta_a \sim \text{Normal}_+\left(\theta_{a,1}\theta_{a,2}^{-1}, \theta_{a,2}^{-1}\right), \quad (4)$$

$$\lambda_{ji} \mid \theta_\lambda \sim \text{Gamma}\left(\theta_{\lambda,1}, \theta_{\lambda,2}\right), \quad (5)$$

y_{jil}	fMRI response of voxel i in subject j to stimulus l
x_{jil}	binary fMRI activation of voxel i in subject j for stimulus l
μ_{ji}	non-activated fMRI response of voxel i in subject j
a_{ji}	activation increase in fMRI response of voxel i in subject j
λ_{ji}	variance reciprocal of fMRI response of voxel i in subject j
z_{ji}	system membership of voxel i in subject j
ϕ_{kl}	Bernoulli activation parameters of system k for stimulus l
β_j	system prior weight in subject j
π	group-level system prior weight
α, γ	HDP scale parameters
θ_ϕ	parameters of the prior over system parameters ϕ
$\theta_{\boldsymbol{\mu}}$	parameters of the priors on voxel response variables (μ, a, λ)
$\theta_\alpha, \theta_\gamma$	parameters of the priors over scale parameters α and γ

Table 1. Variables and parameters in the model.

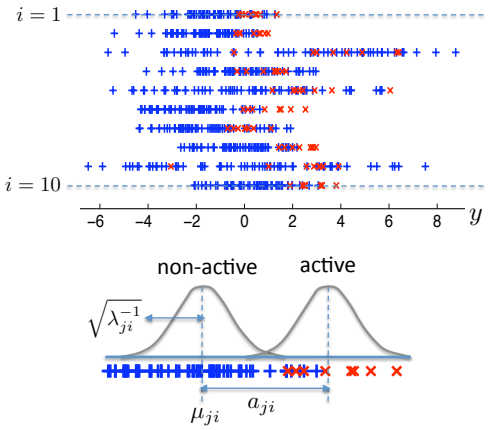

Figure 3. (Top) The values of fMRI responses y_{il} for 10 different voxels detected to be face-selective in the conventional analysis of a visual experiment in one subject. Each row contains the responses of one voxel. We show the responses to stimuli l that correspond to faces in red and the responses to all non-face stimuli in blue. (Bottom) Schematic diagram illustrating the meaning of each of voxel response parameters $\boldsymbol{\mu}_{ji} = (\mu_{ji}, a_{ji}, \lambda_{ji})$ in the model.

where $\text{Normal}_+(\theta_{a,1}\theta_{a,2}^{-1}, \theta_{a,2}^{-1})$ is the conjugate prior defined as a normal distribution restricted to positive real values:

$$p(a) \propto e^{-a^2\,\theta_{a,2} + a\,\theta_{a,1}}, \text{ for } a \geq 0. \quad (6)$$

Positivity of 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.

2.2. HDP Prior for fMRI Activations

We choose a model to simultaneously infer the number and characterization of group-level systems using Hierarchical Dirichlet Processes (HDP) [20]. More specifically, given the set of probabilities of system activation for different stimuli $\boldsymbol{\phi} = \{\phi_{kl}\}$ and system memberships of voxels

$\boldsymbol{z} = \{z_{ji}\}, z_{ji} \in \{1, 2, \cdots\}$, the model assumes

$$x_{jil} \mid z_{ji}, \boldsymbol{\phi} \overset{i.i.d.}{\sim} \text{Bernoulli}(\phi_{z_{ji}l}). \tag{7}$$

This implies that all voxels within a system have the same probability of being activated by a particular stimulus l. We use the stick-breaking formulation of HDP and define the prior for system memberships as follows:

$$z_{ji} \mid \beta_j \overset{i.i.d.}{\sim} \text{Mult}(\beta_j), \tag{8}$$

$$\beta_j \mid \pi \overset{i.i.d.}{\sim} \text{Dir}(\alpha\pi), \tag{9}$$

$$\pi \mid \gamma \sim \text{GEM}(\gamma), \tag{10}$$

where $\text{GEM}(\gamma)$ is a distribution over infinitely long vectors $\pi = [\pi_1, \pi_2, \cdots]^T$, named after Griffiths, Engen and McCloskey [18], defined as follows:

$$\pi_k = v_k \prod_{k'=1}^{k-1} (1 - v_{k'}) , \qquad v_k \mid \gamma \overset{i.i.d.}{\sim} \text{Beta}(1, \gamma). \tag{11}$$

The components of the generated vectors π sum to one with probability 1. Hence, they provide a meaningful interpretation as the group-level expected values for the subject-specific multinomial system weights β_j in Equation (9). With this prior over the system memberships of voxels $\boldsymbol{z}$, the model in principle allows an infinite number of systems; however, for any finite set of voxels, a finite number of systems is sufficient to include all voxels.

We let the prior distribution for system-level activation probabilities $\boldsymbol{\phi}$ be

$$\phi_{kl} \overset{i.i.d.}{\sim} \text{Beta}(\theta_{\phi,1}, \theta_{\phi,2}). \tag{12}$$

By introducing bias towards 0, the non-active state, in the parameters of this distribution, we can induce sparsity in the results. The graphical model for the joint distribution of this model and the fMRI response model of Sec. 2.1 is shown in Fig. 2.

Equations (7), (8), (9), (10), and (12) together define our HDP prior over fMRI activation variables $\boldsymbol{x}$. We further assume Gamma priors for the hyperparameters:

$$\alpha \sim \text{Gamma}(\theta_{\alpha,1}, \theta_{\alpha,2}) , \quad \gamma \sim \text{Gamma}(\theta_{\gamma,1}, \theta_{\gamma,2}). \tag{13}$$

2.3. Variational Bayesian Approximation

Several different Gibbs sampling schemes for inference in HDPs are discussed in [20]. Despite theoretical guarantees of their convergence to the true posterior, sampling techniques generally require a time-consuming burn-in phase. Because of the relatively large size of our problem, we choose an alternative variational approach to inference called Collapsed Variational HDP approximation [21], which is known to yield faster algorithms. Due to space constraints, we cannot provide the derivations in this paper,

$\zeta_j \sim \text{Beta}\left(E[\alpha], N_j\right)$
$\tilde{\theta}_{r_{jk}} = \exp\left(E[\log \alpha] + E[\log v_k] + \sum_{k'<k} E[\log(1 - v_{k'})]\right)$
$E[r_{jk}] = \tilde{\theta}_{r_{jk}} E_{\boldsymbol{z}}[\Psi(\tilde{\theta}_{r_{jk}} + n_{jk}) - \Psi(\tilde{\theta}_{r_{jk}})]$

$\alpha \sim \text{Gamma}(\tilde{\theta}_{\alpha,1}, \tilde{\theta}_{\alpha,2})$ $\quad$ $\gamma \sim \text{Gamma}(\tilde{\theta}_{\gamma,1}, \tilde{\theta}_{\gamma,2})$
$\tilde{\theta}_{\alpha,1} = \theta_{\alpha,1} + \sum_{jk} E[r_{jk}]$ $\quad$ $\tilde{\theta}_{\gamma,1} = \theta_{\gamma,1} + K$
$\tilde{\theta}_{\alpha,2} = \theta_{\alpha,2} - \sum_j E[\log \zeta_j]$ $\quad$ $\tilde{\theta}_{\gamma,2} = \theta_{\gamma,2} - \sum_k E\left[\log\left(1 - v_k\right)\right]$

$v_k \sim \text{Beta}(\tilde{\theta}_{v_k,1}, \tilde{\theta}_{v_k,2})$
$\tilde{\theta}_{v_k,1} = 1 + \sum_j E[r_{jk}]$ $\quad$ $\tilde{\theta}_{v_k,2} = E[\gamma] + \sum_{j,k'>k} E[r_{jk}]$

$\phi_{kl} \sim \text{Beta}(\tilde{\theta}_{\phi_{kl},1}, \tilde{\theta}_{\phi_{kl},2})$
$\tilde{\theta}_{\phi_{kl},1} = \theta_{\phi,1} + \sum_{j,i} q(z_{ji} = k) q(x_{jil} = 1)$
$\tilde{\theta}_{\phi_{kl},2} = \theta_{\phi,1} + \sum_{j,i} q(z_{ji} = k) q(x_{jil} = 0)$

$\mu_{ji} \sim \text{Normal}(\tilde{\theta}_{\mu_{ji},1}\, \tilde{\theta}^{-1}_{\mu_{ji},2}, \tilde{\theta}^{-1}_{\mu_{ji},2})$
$\tilde{\theta}_{\mu_{ji},1} = \theta_{\mu,1} + E[\lambda_{ji}] \left(\sum_l y_{jil} - E[a_{ji}] \sum_l q(x_{jil} = 1)\right)$
$\tilde{\theta}_{\mu_{ji},2} = \theta_{\mu,2} + E[\lambda_{ji}]\, L$

$a_{ji} \sim \text{Normal}_+(\tilde{\theta}_{a_{ji},1}\, \tilde{\theta}^{-1}_{a_{ji},2}, \tilde{\theta}^{-1}_{a_{ji},2})$
$\tilde{\theta}_{a_{ji},1} = \theta_{a,1} + E[\lambda_{ji}] \sum_l q(x_{jil} = 1)\, (y_{jil} - E[\mu_{ji}])$
$\tilde{\theta}_{a_{ji},2} = \theta_{a,2} + E[\lambda_{ji}] \sum_l q(x_{jil} = 1)$

$\lambda_{ji} \sim \text{Gamma}(\tilde{\theta}_{\lambda_{ji},1}, \tilde{\theta}_{\lambda_{ji},2})$
$\tilde{\theta}_{\lambda_{ji},1} = \theta_{\lambda,1} + \frac{1}{2} \sum_l \left(y_{jil} - E[\mu_{ji}] - q(x_{jil} = 1) E[a_{ji}] \right)^2$
$\quad + \frac{L}{2} V[\mu_{ji}] + V[a_{ji}] \sum_l q(x_{jil} = 1) + E[a_{ji}]^2 \sum_l V[x_{jil}]$
$\tilde{\theta}_{\lambda_{ji},2} = \theta_{\lambda,2} + \frac{L}{2}$

Table 2. Update rules for computing the posterior q over the unobserved variables.

but for completeness, here we provide a brief overview of the steps and the final update rules of the resulting algorithm (see the Supplementary Material for details).

We first marginalize our distribution over the subject-specific system weights $\boldsymbol{\beta} = \{\beta_j\}$ and add a set of auxiliary variables $\boldsymbol{r} = \{r_{ji}\}$ and $\zeta = \{\zeta_j\}$ to the model to find closed-form solutions for the inference update rules. Let $\boldsymbol{h} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{r}, \zeta, \boldsymbol{\phi}, \boldsymbol{\mu}, \pi, \alpha, \gamma\}$ denote the set of all unobserved variables. 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 $\mathcal{F}[q] = E[\log q(\boldsymbol{h})] - E[\log p(\boldsymbol{y}, \boldsymbol{h})]$. 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:

$$q(\boldsymbol{h}) = q(\alpha) q(\gamma) q(\boldsymbol{r}, \zeta | \boldsymbol{z}) \cdot \prod_{j,i} \left[q(\mu_{ji}) q(a_{ji}) q(\lambda_{ji}) q(z_{ji}) \prod_l q(x_{jil}) \right], \tag{14}$$

where we explicitly account for the dependency of the auxiliary variables on the system memberships. Including this structure maintains the quality of the approximation despite the introduction of the auxiliary variables [22].

Minimizing the Gibbs free energy function in terms of each component of q, assuming all the rest as constant, we find the update rules in Table 2. We define $n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji}, k)$ as the number of voxels in subject j that are members of system k. In addition, for the system membership updates, we find

$$q(z_{ji} = k) \propto \exp\Big\{ E_{\boldsymbol{z}^{\neg ji}}[\log(\tilde{\theta}_{r_{jk}} + n_{jk}^{\neg ji})] + \sum_l \Big(q(x_{jil} = 1) E[\log \phi_{kl}] + q(x_{jil} = 0) E[\log(1 - \phi_{kl})] \Big) \Big\}, \quad (15)$$

where $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$, and $n_{jk}^{\neg ji}$ and $\boldsymbol{z}^{\neg ji}$ indicate the exclusion of voxel i in subject j. Moreover, the posterior over activation variables can be described as

$$q(x_{jil} = 1) \propto \exp\Big\{ \sum_l q(z_{ji} = k) E[\log \phi_{kl}] - \frac{E[\lambda_{ji}]}{2} \Big[\Big(y_{jil} - E[\mu_{ji}] - E[a_{ji}] \Big)^2 + V[a_{ji}] \Big] \Big\}, \quad (16)$$

$$q(x_{jil} = 0) \propto \exp\Big\{ \sum_l q(z_{ji} = k) E[\log(1 - \phi_{kl})] - \frac{E[\lambda_{ji}]}{2} \Big(y_{jil} - E[\mu_{ji}] \Big)^2 \Big\}. \quad (17)$$

Note that under $q(\boldsymbol{z})$, each variable n_{jk} is the sum of N_j independent binary variables. Therefore, we can use the Central Limit Theorem to approximate terms of the form $E_{\boldsymbol{z}}[f(n_{jk})]$ by assuming a Gaussian distribution for n_{jk} [21].

2.4. Initialization

By iterative application of the update rules in Sec. 2.3, we can find 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 use the following approach in order to provide reasonable starting points for the algorithm.We first cluster the L values of y_{jil} corresponding to each voxel into two clusters, active and non-active, to form initial values for the activation $\boldsymbol{x}$ and voxel response $\boldsymbol{\mu}$. Then, for initializing $q(\boldsymbol{z})$, we sample the voxel memberships by introducing the voxels one by one and in random order to the collapsed Gibbs sampling scheme [20] constructed for our HDP.

3. Results and Discussion

We applied our method to data from an event-related experiment where subjects view 69 distinct images over the course of two 2-hour scanning sessions. Fig. 4 shows these images; they include 8 images from each category of animals, bodies, cars, faces, shoes, scenes, tools, trees, along with 5 vase images. The study includes 10 subjects.

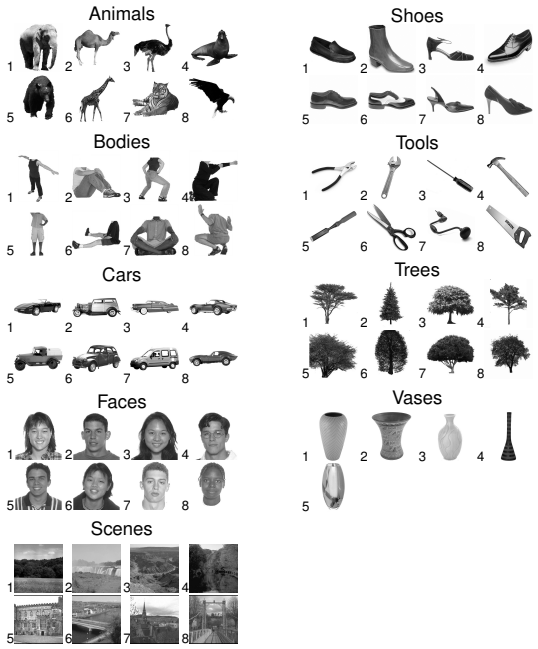

Figure 4. The set of the 69 images used as stimuli in our experiment.

The data was first motion-corrected separately for the two sessions[4], and then spatially smoothed with a Gaussian kernel of 3mm width. The BOLD time course data includes about 3,000 volumes per subject. We applied the standard General Linear Model [7] to estimate the response of each voxel to each image. We then registered the data from the two sessions to the subject's native anatomical space [10]. We applied our algorithm, the finite mixture-model clustering [14], and the tensorial ICA algorithm [3] to the same set of estimated image responses for all subjects. We used FSL/Melodic implementation of tensorial ICA.[1]

For our algorithm and the finite mixture-modeling, we removed noisy voxels from the analysis. To this end, we performed an ANOVA for all stimulus regressors and included only voxels with the F-test significance value below $p = 10^{-4}$ (uncorrected). This procedure yielded about 64,000 voxels across all subjects. We then applied the analysis in each subject's own native space. For the visualization of the resulting spatial maps, we aligned all the brains with the Montreal Neurological Institute (MNI) coordinate space [19]. Applying group ICA requires spatial normalization of all subjects' brains prior to the analysis. Since the masks generated by the ANOVA tests do not necessarily overlap after normalization, we only excluded the voxels outside the brain for this analysis. To remove any stimulus-specific pattern in the responses in all the above analyses, we subtracted from the estimated voxel response to each stimulus the average response across all voxels in that sub-

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

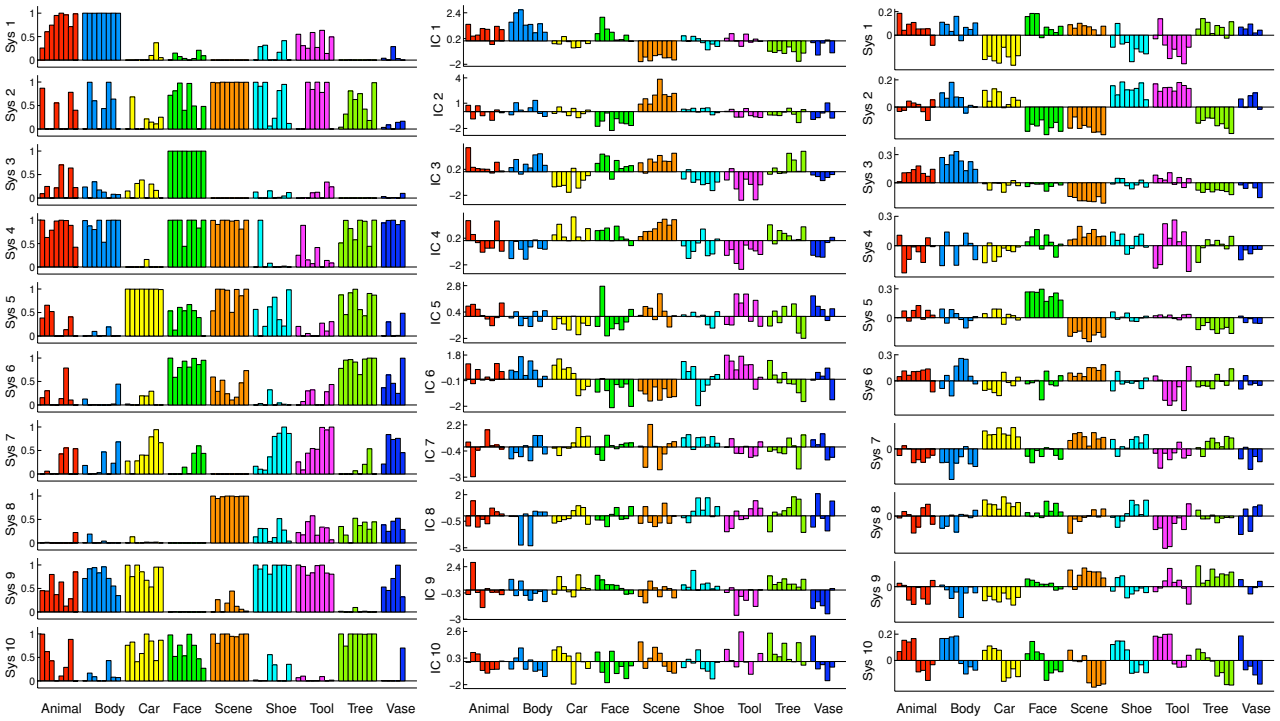

Figure 5. The profiles for the 10 most consistent discovered systems and components found by our nonparametric method, tensorial ICA, and the finite mixture model. The profiles of our systems represent the activation probabilities $E[\phi_{kl}]$ of different systems k.

ject.

We also generated the conventional significance maps for the three well-known types of category selectivity, i.e., selectivity for bodies, faces, and scenes. We formed contrasts by subtracting from the average response to each of these categories the average response to all cars, shoes, tools, and vase stimuli. We then thresholded the maps at $p = 10^{-4}$ (uncorrected).

In our method, we set the hyperparameter values $(\theta_{\gamma,1}, \theta_{\gamma,2}) = (1, 1)$, $(\theta_{\mu,1}, \theta_{\mu,2}) = (0, 0.1)$, $(\theta_{a,1}, \theta_{a,2}) = (0, 0.1)$, and $(\theta_{\lambda,1}, \theta_{\lambda,2}) = (1, 0.1)$. We also choose $(\theta_{\tau,1}, \theta_{\tau,2}) = (1, 3)$ to encourage *sparsity* in the results by biasing the activation prior towards being nonactive. Furthermore, we select $(\theta_{\alpha,1}, \theta_{\alpha,2}) = (0.01, 0.0001)$ to increase both the expected value and the variance of the scale parameter α in order to put emphasis on finding systems that are representative of the entire group (see Eq. (9)). We run the algorithm 30 times with different random initializations as described in Sec. 2.4 and select the solution with the lowest Gibbs free energy.

Our algorithm finds 22 systems; applying the automatic model selection algorithm [17] along with tensorial ICA yields 27 components. Tensorial ICA results include a set of variables that describe the contribution of each component to the results in each subject. We can create a similar measure for the consistency of each system in clustering models by first computing the ratio of all voxels assigned to that system that belongs to each subject and then computing the standard deviation of this ratio across subjects. For both 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. Since we are interested only in the most consistent systems, we rank the resulting systems and components based on the computed measures of consistency across subjects. We apply the same procedure to the results from the finite mixture modeling with 30 systems.

Fig. 5 shows 10 profiles corresponding to the most consistent (as defined above) systems or components found by our method, ICA, and finite mixture-modeling. Our activation profiles correspond to the probabilities $E[\phi_{kl}]$ of the systems. We observe that the category information is more salient in the HDP system profiles of activation, especially when compared to the ICA results. Most of our systems demonstrate similar probabilities of activation for images that belong to the same category. This suggests that the discovered systems in fact describe coherent patterns of response relevant to the nature of the intrinsic object representation in the visual system.

More specifically, systems 1, 3, and 8 appear mostly selective for bodies, faces, and scenes, respectively. Among the ICA results, we only find two components with rela-

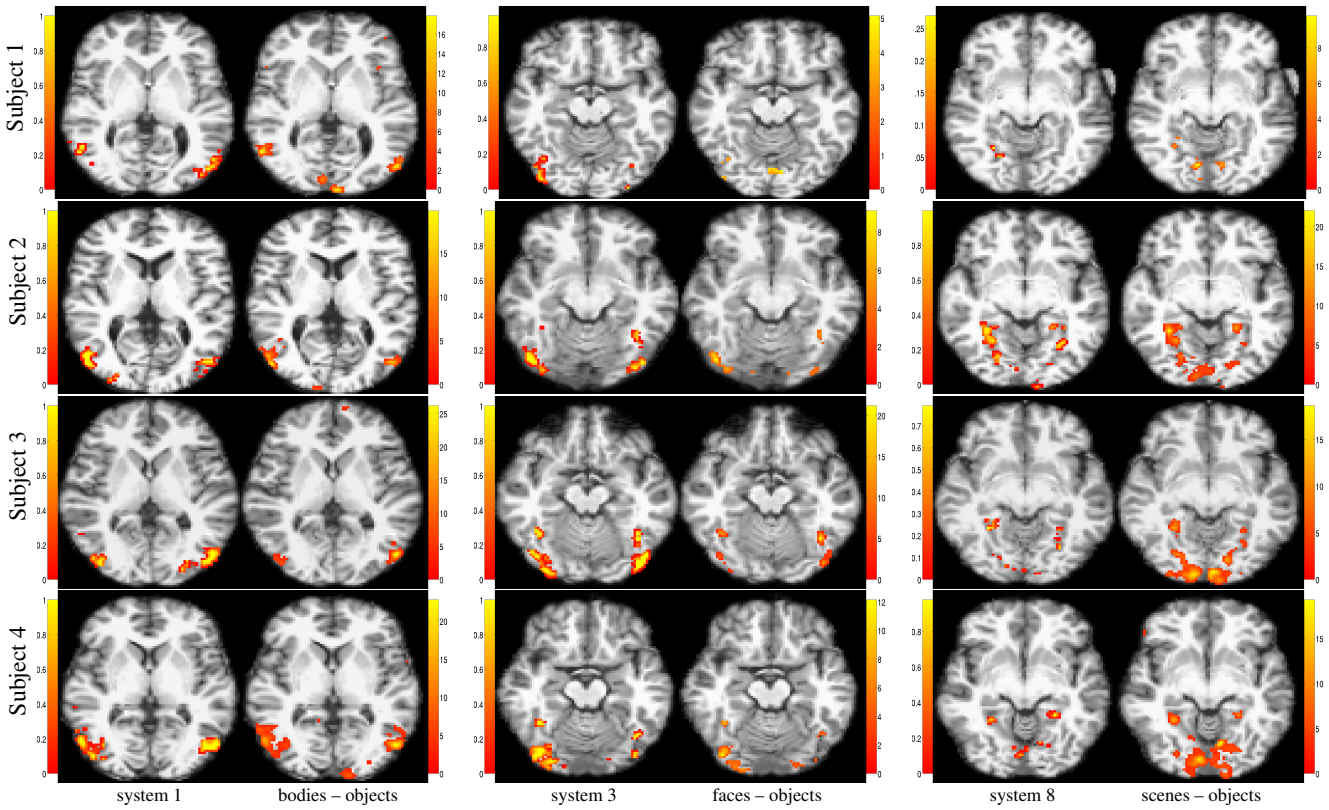

Figure 6. The body (left column set), face (middle), and scene (right) selective areas defined by our data-driven method in Fig. 5 (the right map within each set), compared to the conventional contrasts (the right map within each set) in four different subjects. Our maps present membership probabilities while the contrast maps show significance values $-\log p$ thresholded at 4. Each pair shows the maps on the same slice from one subject. We normalized all the subjects in the Talairach space so that all four slices in the same column set are aligned. Our hierarchical model is validated by the conventional contrasts: wherever a particular system is found, the conventional contrasts are significant. Moreover, our results show more consistency across subjects.

tively larger responses to bodies and scenes (components 1 and 2), while the consistency of the finite-mixture modeling systems selective for the same categories is lower than the results of our model. We emphasize the advantage of our activation profiles over the selectivity profiles of 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 a stimulus can be summarized based on our results in a vector of activations $[\,E[\phi_{1l}], \cdots, E[\phi_{Kl}]\,]^T$ that it induces over the set of all functional systems. Such a representation cannot be made from the clustering profiles in Fig. 5 since their elements do not have any clear interpretation. Inspecting the activation profiles of different systems more closely, we find other interesting patterns. For instance, we note that most images of animals can naturally activate body-selective systems as well. Even more interestingly, the two non-face images that show some probability of activating the face selective system 6, namely, animals 5 and 7 (Fig. 4) correspond to the two animals that have large faces. Another interesting case

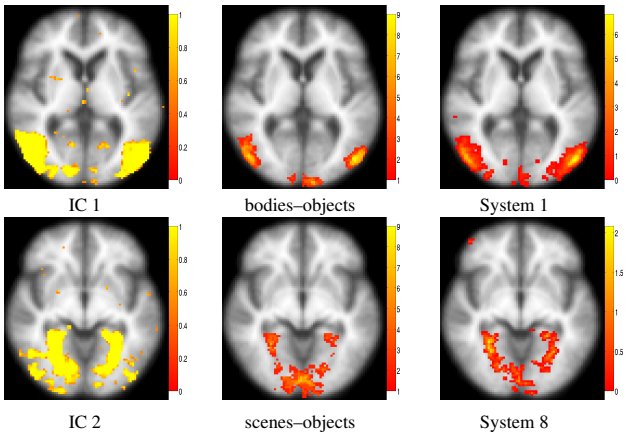

Figure 7. Comparison between the ICA maps for components 1 and 2 in Fig. 5 (left) and the group averages of contrast (middle) and HDP maps (right).

is system 4, which seems to be activated by stimuli that are larger in terms of non-background pixels (notice how only shoe 2, the largest shoe, activates this system).

Our approach also yields probabilistic spatial maps for each discovered system. Fig. 6 compares the maps of the

three systems 1, 3, and 8 with the conventional significance maps for the corresponding selective areas. The map for system k in subject j is defined by the vector of posterior probabilities $[q(z_{j1} = k), \cdots q(z_{jN_j} = k)]^T$. We have normalized the maps after the analysis and show the results on the same slice for all subjects. To choose the slice to display for any of the systems, we considered the number of voxels in each slice that were assigned to that system in more than half the subjects, and then chose the slice with the greatest such number. For the most part, voxels that are assigned to our systems also demonstrate high significance values based on the conventional statistical tests. Moreover, our results show more consistency across the subjects.We note that the characterization of the conventionally detected selective areas is less restrictive than that of our systems. While system activation probabilities completely describe the response of a system to all stimuli, conventional selectivity is defined in terms of the difference of response between two groups of stimuli. That explains the fact that, for instance, the scene contrast for the two bottom subjects include early visual areas while those voxels do not appear in our results.

Fig. 7 shows the group probability maps for the body- and scene-selective ICA components on the same slices as in Fig. 6. These maps have been computed fitting a mixture model to the z-scores from the original component maps and then thresholding at 0.5 [2]. The images clearly illustrate how the voxel-wise alignment before the analysis has blurred the results: none of the subject-specific areas in Fig.6 is as large as the maps in Fig.7. Rather, the two ICA maps seem to be driven by the variability in the anatomical locations of the areas across subjects. This becomes evident when we compare these maps in the same figure with the group sum of the thresholded contrast maps ($p = 10^{-4}$ uncorrected) and our probability maps for the corresponding forms of selectivity (both maps take values from 0 to $J = 10$). The two maps appear at the same approximate locations but with very small overlapping areas.

4. Conclusion

We presented a method that combines fMRI data from several subjects in order to identify functional systems in the brain. We defined systems as group-wide collections of voxels that demonstrate coherent patterns of response. We employed a nonparametric hierarchical Bayesian model to develop a generative model that captures shared structures in the response of a group of subjects. We further derived and implemented a fast variational algorithm for inference based on this model. We applied our method to an fMRI study of category selectivity in the visual cortex presenting a variety of distinct images. Our results demonstrate that the systems learned by the model correspond to the category-selective areas previously identified in numerous hypothesis-driven studies.

Acknowledgements This research was supported in part by the NSF IIS/CRCNS 0904625 grant, the NSF CAREER grant 0642971, the MIT McGovern Institute Neurotechnology Program, and the NIH NIBIB NAMIC U54-EB005149 and NCRR NAC P41-RR13218 grants.

References

[1] R. Baumgartner, G. Scarth, C. Teichtmeister, R. Somorjai, and E. Moser. Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: reproducibility. *J Magn Reson Imaging*, 7(6):1094–1108, 1997. 1

[2] C. Beckmann and S. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *TMI*, 23(2):137–152, 2004. 1, 8

[3] C. Beckmann and S. Smith. Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1):294–311, 2005. 2, 5

[4] R. Cox and A. Jesmanowicz. Real-time 3D image registration for functional MRI. *Magnetic Resonance in Medicine*, 42(6):1014–1018, 1999. 5

[5] P. Downing, A.-Y. Chan, M. Peelen, C. Dodds, and N. Kanwisher. Domain specificity in visual cortex. *Cerebral Cortex*, 16(10):1453–1461, 2006. 1

[6] S. Forman, J. Cohen, M. Fitzgerald, W. Eddy, M. Mintun, and D. Noll. 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, 1995. 2

[7] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, editors. *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. Academic Press, Elsevier, 2007. 2, 5

[8] P. Golland, Y. Golland, and R. Malach. Detection of spatial activation patterns as unsupervised segmentation of fMRI data. In *MICCAI*, 2007. 1

[9] C. Goutte, P. Toft, E. Rostrup, F. Nielsen, and L. Hansen. On clustering fMRI time series. *NeuroImage*, 9(3):298–310, 1999. 1

[10] D. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009. 5

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

[12] S. Kim and P. Smyth. Hierarchical Dirichlet processes with random effects. In *NIPS*, 2007. 2

[13] D. Lashkari and P. Golland. Exploratory fMRI analysis without spatial normalization. In *IPMI*, 2009. 1

[14] D. Lashkari, E. Vul, N. Kanwisher, and P. Golland. Discovering structure in the space of fMRI selectivity profiles. *NeuroImage*, 50(3):1085–1098, 2010. 1, 2, 5

[15] S. Makni, P. Ciuciu, J. Idier, and J.-B. Poline. Joint detection-estimation of brain activity in functional MRI: a multichannel deconvolution solution. *TSP*, 53(9):3488–3502, 2005. 3

[16] M. McKeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, and T. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Hum Brain Mapp*, 6(3):160–188, 1998. 1

[17] T. Minka. Automatic choice of dimensionality for PCA. *NIPS*, 2001. 6

[18] J. Pitman. Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Prob, Comput*, 11(5):501–514, 2002. 4

[19] J. Talairach and P. Tournoux. *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme, New York, 1988. 5

[20] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *JASA*, 101(476):1566–1581, 2006. 2, 3, 4, 5

[21] Y. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *NIPS*, 2008. 2, 4, 5

[22] Y. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, 2007. 4

[23] B. Thirion and O. Faugeras. Feature characterization in fMRI data: the Information Bottleneck approach. *Medical Image Analysis*, 8(4):403–419, 2004. 1

[24] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J. Poline. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum Brain Mapp*, 27(8):678–693, 2006. 1

[25] B. Thirion, A. Tucholka, M. Keller, P. Pinel, A. Roche, J. Mangin, and J. Poline. High level group analysis of FMRI data based on Dirichlet process mixture models. In *IPMI*, 2007. 2