A Bayesian Mixture Approach to Modeling Spatial Activation Patterns in Multi-site fMRI Data

Seyoung Kim, Padhraic Smyth, *Member, IEEE,* and Hal Stern

***Abstract*—We propose a probabilistic model for analyzing spatial activation patterns in multiple fMRI activation images such as repeated observations on an individual or images from different individuals in a clinical study. Instead of taking the traditional approach of voxel-by-voxel analysis, we directly model the shape of activation patterns by representing each activation cluster in an image as a Gaussian-shaped surface. We assume that there is an unknown true template pattern and that each observed image is a noisy realization of this template. We model an individual image using a mixture of experts model with each component representing a spatial activation cluster. Taking a nonparametric Bayesian approach, we use a hierarchical Dirichlet process to extract common activation clusters from multiple images and estimate the number of such clusters automatically. We further extend the model by adding random effects to the shape parameters to allow for image-specific variation in the activation patterns. Using a Bayesian framework, we learn the shape parameters for both image-level activation patterns and the template for the set of images by sampling from the posterior distribution of the parameters. We demonstrate our model on a dataset collected in a large multi-site fMRI study.**

***Index Terms*—Functional magnetic resonance imaging, brain activation, hierarchical model.**

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) is widely used to study how the brain functions in response to external stimuli. In each run of an fMRI scan, data are collected as a time-series of 3-dimensional voxel images while a subject is responding to external stimuli or performing a specific cognitive task. The temporal aspect of the time-series data for a run is often summarized as a β-map, a 3-dimensional image of β coefficients that estimate the amount of activation at each voxel. An experiment often comprises multiple runs within a visit, and may also include multiple visits, and multiple subjects. There is also increasing interest in analyzing data taken at multiple different fMRI sites (machines) [1], [2].

In a typical approach to analysis of fMRI data, the activation maps are analyzed using voxel-by-voxel hypothesis testing. The set of voxels that are found to be statistically significant (e.g., based on t-statistics) are then used to define the activated region in the brain [3]. This approach assumes that the activation at each voxel is independent of the activation in neighboring voxels, and ignores the spatial information in the overall activation pattern. While spatial statistics that are derived from multiple neighboring voxels have been used to test significance of activations [4], [5], more recent approaches propose to directly take into account the spatial information in the activation pattern by modeling the shape of local activation regions explicitly. For example, Hartvig [6] represented the activation surface in fMRI as a parametric function consisting of a superposition of Gaussian-shaped bumps and a constant background level, and used a stochastic geometry model to find the number of bumps automatically. Penny and Friston [7] proposed a mixture model with each mixture component representing a local activation cluster. In earlier work we proposed a response surface model that represents an activation pattern as a superposition of Gaussian shaped parametric surfaces and demonstrated how the model could be used to characterize and quantify inter-machine variability in multi-site fMRI studies [8].

The methods discussed above on modeling spatial activation shape in fMRI data can handle only a single image. The problem of extracting spatial patterns from multiple images has not been addressed, even though detection and characterization of such patterns can in principle provide richer information (than voxel-level information) about cognitive activity and its variation across individuals, across time, and across machines. Previous approaches for spatial modeling in multiple images were based on first extracting spatial statistics from individual images, finding correspondences of those statistics across multiple images, and finding brain regions with significant activations in terms of the statistics matched across images [9]–[13]. Although the spatial statistics were extracted from multiple correlated voxels in the neighborhood of an activation region, their representation did not explicitly capture the activation shape information, and the overall approach did not provide a mechanism to systematically learn the variability across different images. A Bayesian hierarchial model has been proposed to account for spatial variability in activation across multiple images [14], but the spatial correlation was modeled through a covariance parameter between each pair of voxels, instead of using an explicit representation of activation shapes.

In this paper we propose a new statistical approach that can characterize spatial fMRI activation patterns across multiple images[1], where the multiple images could be repeated observations on an individual (as in a recent fMRI reliability study [1], [16]) or could be different individuals from a group in a

S. Kim is with Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: sssykim@cs.cmu.edu.

P. Smyth is with the Department of Computer Science, University of California, Irvine, CA 92697-3435. E-mail: smyth@ics.uci.edu.

H. Stern is with the Department of Statistics, University of California, Irvine, CA 92697-1250. E-mail: sternh@uci.edu.

Manuscript received December 19, 2007; revised December 23, 2009.

[1]This paper extends results that were presented earlier in preliminary form as a short conference paper [15].

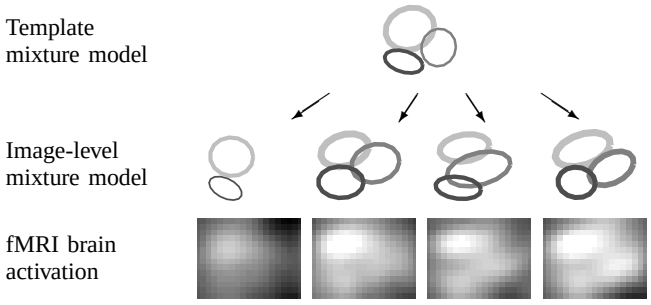


Fig. 1. Illustration of image-level variations from the template model.

clinical study. We model each activation cluster in an image as a Gaussian-shaped surface with parameters for (a) its height or peak value, representing the amount of activation, (b) the location of the cluster, modeling the center of activation in voxel-space, and (c) the width of the cluster. Given multiple activation images, we extract common activation clusters across images and learn the image-specific variation in the activation shape in each image. The general idea is illustrated in Figure 1. At the bottom of Figure 1 are fMRI activation images in the right motor region of the brain over four runs from the same subject performing a sensorimotor task. There are three activation clusters that appear in all or some of the four images, with image-specific variation in intensity and location. These types of variation are common in multi-image fMRI experiments, due to a variety of factors such as head motion and variation in the physiological and cognitive states of the subject. The underlying assumption in the model that we propose in this paper is that there is an unknown true activation pattern (as shown at the top of Figure 1) in a subject's brain given a particular stimulus, and that the activation patterns in the observed images (as shown in the middle row of Figure 1) are noisy realizations of this true activation template, with variability in the activation patterns due to various sources. Our goal is to build a probabilistic model that infers both the overall template and image-specific activation patterns given multiple observed images.

We base our probabilistic model for multiple images on a mixture of experts model with a Dirichlet process prior for a single image [8]. We model spatial activation patterns in a single activation image as a mixture of experts [17]–[19] with a constant background component and one or more activation components. Each local activation cluster is modeled using a parametric surface model with parameters for its peak value, the location of the cluster, and the width. The problem of estimating the number of such activation components is handled by combining this mixture of experts model with a Dirichlet process prior. A Dirichlet process prior is a nonparametric Bayesian prior that has been shown to be useful in learning the number of components in a mixture model from data without having to specify it *a priori* [20].

To model multiple activation images in a single probabilistic framework, we link the common components in multiple single-image models by introducing a hierarchy in the Dirichlet process prior. The hierarchical Dirichlet process is a model developed for handling multiple related mixture models with shared mixture components [21]. At the top of the hierarchy is a template mixture model with a set of components that are common across bottom-level mixture models. All of the bottom-level mixture models have components with the same fixed component parameters but are allowed to have their own mixing proportions. When we apply the hierarchical Dirichlet process to the specific problem of fMRI activation modeling, we can simultaneously infer both the top-level template and the image-level activation patterns from observed images, as well as the number of components at each level.

The hierarchical Dirichlet process assumes that the mixture component parameters (e.g, the intensities and locations of the Gaussian-shaped surface models we wish to use for fMRI activation modeling) are fixed across images. However, as can be seen in Figure 1, there is image-specific local variation in the shape of activation clusters. In this paper we introduce additional flexibility to the hierarchical Dirichlet process as follows: each image is allowed to have its own shape parameter, with a common prior distribution across images that has a mean shape in the activation template, and with a variance controlling the amount of image-specific random variation. This image-specific random variation added to the template shape parameters is sometimes referred to as *random effects* in the statistical literature [22]. By introducing random effects to the component parameters in the hierarchical Dirichlet process and estimating the random effects parameters, we can learn both the template activation shape and the image-specific random variation. In our experiments using data from a large multi-site fMRI study, we demonstrate that the hierarchical Dirichlet process with random effects leads to systematically better results when compared to alternative approaches.

The rest of the paper is organized as follows. In Section II, we introduce a mixture of experts model with a Dirichlet process prior for a single image. In Section III, we propose a model for multiple images that uses a hierarchical Dirichlet process. In Section IV, we further extend this model by introducing random effects to the shape parameters in activation components. Section V provides a demonstration of the proposed models using multi-site fMRI data. We conclude in Section VI with a brief discussion of future work.

II. A Mixture of Experts Model with Dirichlet Process Prior for a Single Image

We begin by discussing the mixture of experts model for activation patterns in a single image, which we then generalize to multiple images in later sections.

A. The Model

We develop the model for the case of 2-dimensional slices of β maps—the 3-dimensional case can be derived as an extension of the 2-dimensional case, but is not pursued in this paper. We assume that the β values $y_i, i = 1, \ldots, N$ (where N is the number of voxels) are conditionally independent of each other given the voxel position $\mathbf{x}_i = (x_{i1}, x_{i2})$ and the model parameters. We then model the activation y_i at voxel $\mathbf{x}_i$ with a mixture of experts model [18], [23]:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{z_i \in \mathscr{C}} p(y_i|z_i, \mathbf{x}_i, \boldsymbol{\theta}) P(z_i|\mathbf{x}_i, \boldsymbol{\theta}), \qquad (1)$$

where $\mathscr{C} = \{c_{\text{bg}}, c_m, m = 1, \ldots, M\}$ is a set of component labels for the background c_{bg} and the M activation components (the c_m's), z_i is the component label for the ith voxel, and $\boldsymbol{\theta} = \{\theta_{\text{bg}}, \theta_m, m = 1, \ldots, M, \sigma^2_{\text{bg}}, \sigma^2_{\text{act}}\}$ represents the component parameters. The θ_{bg} and σ^2_{bg} are parameters for the background component, and the θ_m's and σ^2_{act} are used to model activation components as described below.

The first term on the right hand side of (1) defines the spatial model or expert for a given mixture component. We model the activation component as a normal distribution having a mean that is a Gaussian-shaped surface centered at $\mathbf{b}_m$ with width Σ_m and height k_m,

$$E[y_i|\mathbf{x}_i, z_i = c_m, \theta_m, \sigma^2_{\text{act}}] = k_m \exp\left(-(\mathbf{x}_i - \mathbf{b}_m)'(\Sigma_m)^{-1}(\mathbf{x}_i - \mathbf{b}_m)\right), \quad (2)$$

where $\theta_m = \{k_m, \mathbf{b}_m, \Sigma_m\}$, and a variance σ^2_{act}. The background component is modeled as a normal distribution with mean

$$E[y_i|\mathbf{x}_i, z_i = c_{\text{bg}}, \theta_{\text{bg}}, \sigma^2_{\text{bg}}] = \mu,$$

where $\theta_{\text{bg}} = \{\mu\}$, and variance σ^2_{bg}.

The second term in (1) is known as a gate function in the mixture of experts framework—it decides which model should be used to make a prediction for the activation level at position $\mathbf{x}_i$. Using Bayes' rule we write this term as

$$P(z_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|z_i, \boldsymbol{\theta})\pi_{z_i}}{\sum_{c\in\mathscr{C}} p(\mathbf{x}_i|c, \boldsymbol{\theta})\pi_c}, \quad (3)$$

where π_{z_i} is a class prior probability $P(z_i)$ [19], [24]. $p(\mathbf{x}_i|z_i, \boldsymbol{\theta})$ is defined as follows. For the activation components with $z_i = c_m$, $p(\mathbf{x}_i|z_i, \boldsymbol{\theta})$ is a normal density with mean $\mathbf{b}_m$ and covariance Σ_m. The $\mathbf{b}_m$ and Σ_m are shared with the Gaussian surface model in (2). This implies that the probability of activating the mth model or expert is highest at the center of the activation and gradually decays as $\mathbf{x}_i$ moves away from the center. $p(\mathbf{x}_i|z_i, \boldsymbol{\theta})$ for the background component is modeled as having a uniform distribution of $1/N$ for all positions in the brain, where N is the number of voxels in the image. If $\mathbf{x}_i$ is not close to the center of any activations, the gate function selects the background expert for the voxel. The denominator of (3) provides an expression for the distribution of the input space as $p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{c\in\mathscr{C}} p(\mathbf{x}_i|c, \boldsymbol{\theta})\pi_c$.

Combining the pieces results in a full generative-model specification [18], [24],

$$\begin{aligned} p(y_i, \mathbf{x}_i|\boldsymbol{\theta}) &= p(y_i|\mathbf{x}_i, \boldsymbol{\theta})p(\mathbf{x}_i|\boldsymbol{\theta}) \\ &= \sum_{z_i\in\mathscr{C}} p(y_i|z_i, \mathbf{x}_i, \boldsymbol{\theta})P(z_i|\mathbf{x}_i, \boldsymbol{\theta})p(\mathbf{x}_i|\boldsymbol{\theta}) \\ &= \sum_{z_i\in\mathscr{C}} p(y_i|z_i, \mathbf{x}_i, \boldsymbol{\theta})p(\mathbf{x}_i|z_i, \boldsymbol{\theta})\pi_{z_i}. \end{aligned} \quad (4)$$

The second line in the above equation follows from (1) and the third line is obtained by applying (3).

Using a Bayesian modeling framework, we place prior distributions on the parameters as follows. We let the center of activation $\mathbf{b}_m$ be *a priori* uniformly distributed inside or a half voxel away from the brain region in the image. We let the height parameter k_m be *a priori* uniformly distributed between 0 and a pre-defined value K_{max}. The K_{max} is set to a value about 15-20% higher than the maximum β value in the image. For the width parameter Σ_m, we use a half-normal distribution with mean 0 and variance σ_0^2 as a prior for the variance terms in Σ_m and place a uniform prior over [-0.5, 0.5] on the correlation coefficients. This ensures that Σ_m always stays positive-definite. A normal distribution $\mathcal{N}(0, {\tau_0}^2)$ is used as a prior on the background mean μ. The variances σ^2_{act} and σ^2_{bg} are given half-normal prior distributions with mean 0 and variance σ^2_{act0} and σ^2_{bg0} respectively.

B. Dirichlet Process as a Prior

A practical issue with the finite mixture model approach is how to determine the number of components in the model [25], [26]. Nonparametric Bayesian approaches address this problem by assuming an infinite number of components *a priori* and then letting the data determine how many components exist in the posterior. In particular, Dirichlet processes are well-suited as a nonparametric Bayesian prior for mixture models [20], [27]. Using this approach, we can infer a posterior distribution over the number of components given the observed data.

A Dirichlet process [28] is a measure over probability measures denoted as $DP(\alpha, G_0)$ with a concentration parameter $\alpha > 0$ and a base distribution G_0 as its two parameters. The Dirichlet process is most easily described by associating a component parameter vector ϕ_i with each voxel $(\mathbf{x}_i, y_i)$. We reserve the notation θ_c for a distinct component parameters; several ϕ_i's may be equal to the same θ_c. We place a Dirichlet process prior $DP(\alpha, G_0)$ on the parameters $\boldsymbol{\pi} = \{\pi_c\text{'s}, c \in \mathscr{C}\}$ and $\boldsymbol{\theta}$ as follows:

$$\phi_i|G \sim G, \quad G|\alpha, G_0 \sim DP(\alpha, G_0), \quad (5)$$

where G itself is a discrete probability measure in the form of

$$G = \sum_{c=1}^{\infty} \pi_c \delta_{\theta_c},$$

generated from $DP(\alpha, G_0)$. The prior has an infinite number of components, but conditioned on the observed data, only a finite number of components exist in the posterior. We designate the first component as the only background component. All of the remaining components are activation components, and the model can have an arbitrary number of such components. A stick-breaking representation describes the construction of G from $DP(\alpha, G_0)$ [29] as follows:

$$\boldsymbol{\pi}|\alpha \sim \text{Stick}(\alpha), \quad \theta_c|G_0 \sim G_0, \quad (6)$$

where $\boldsymbol{\pi}$ is constructed from a stick-breaking process Stick(α) as described below [29]:

$$\pi_c = \pi'_c \prod_{i=1}^{c-1}(1 - \pi'_i), \quad \pi'_c \sim \text{Beta}(1, \alpha). \quad (7)$$

We can understand the process of generating π_c's via (7) as breaking a unit-length stick sequentially as follows. We

sample π'_1 from Beta$(1, \alpha)$, break the stick at π'_1, and set $\pi_1 = \pi'_1$. We take the remainder of the stick of length $(1 - \pi'_1)$, select the second break-point by sampling π'_2 from Beta$(1, \alpha)$, and set $\pi_2 = \pi'_2(1 - \pi'_1)$, and so on. It is straightforward to show that π_c's generated using (7) sum to 1. The component parameter θ_c associated with each π_c is drawn from the base distribution G_0, which is made up of the prior distributions defined in Section II-A for the activation component parameters $\{\mathbf{b}_m, k_m, \Sigma_m\}$ and the background component parameter μ. We use a Gamma(a, b) distribution as a prior for the concentration parameter α.

If we integrate out G from the model described in (5), it can be shown that the labels (the z_i's) have the following clustering property [30]:

$$z_i | z_1, \ldots, z_{(i-1)}, \alpha \sim \sum_{c=1}^{K} \frac{n_c^{-i}}{i - 1 + \alpha} \delta_c + \frac{\alpha}{i - 1 + \alpha} \delta_{c_{\text{new}}}, \quad (8)$$

where n_c^{-i} represents the number of $z_{i'}$ variables, $i' < i$, that are assigned to component c, and K is the number of components that have one or more voxels in $z_1, \ldots, z_{i-1}$ associated with them. The probability that z_i is assigned to a new component is proportional to α. Note that the component with more observations already assigned to it has a higher probability of attracting the next observation. It can be shown that $z_1, \ldots, z_N$ are exchangeable under (8) [31].

C. Markov Chain Monte Carlo (MCMC) Sampling for Learning

Because of the nonlinearity of the model and non-conjugacy in the base distribution of the Dirichlet process prior, we rely on MCMC simulation methods to obtain samples from the posterior probability distribution of the parameters given the data. In a Bayesian mixture model framework it is common to augment the unknown parameters with the unknown component labels for observations and consider the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{y}, \mathbf{X})$ where $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{z}$ represent a collection of y_i's, $\mathbf{x}_i$'s and z_i's for $i = 1, \ldots, N$ and, $\boldsymbol{\theta} = \{\mu, \sigma^2_{\text{bg}}, \sigma^2_{\text{act}}, \{\mathbf{b}_m, \Sigma_m, k_m\}, m = 1, \ldots, M, \alpha\}$. During each sampling iteration the component labels $\mathbf{z}$ and parameters $\boldsymbol{\theta}$ are sampled alternately.

1) Sampling Component Labels: For each i we sample z_i from its conditional posterior given as

$$p(z_i | \mathbf{z}_{-i}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}, \alpha) \propto p(y_i | z_i, \mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{x}_i | z_i, \boldsymbol{\theta}, \alpha) P(z_i | \mathbf{z}_{-i}, \alpha), \quad (9)$$

where the first two terms on the right-hand side of the equation are data likelihoods, and the last term is given by (8), assuming that z_i is given by (8), acting as if z_i is the last item in a sequence like (8) that contains all N voxels.

Because of the non-conjugacy of the base distribution in the Dirichlet process prior, it is not possible to analytically compute the conditional posterior in (9), so we use a sampling method based on the Metropolis-Hastings algorithm for Dirichlet process mixtures with a non-conjugate prior [31]. We briefly describe the sampling algorithm as follows. If $z_i = z_j$ for some $j \neq i$ (i.e., z_i is a non-singleton), we propose to re-assign z_i to a newly created component. If $z_i \neq z_j$ for all $j \neq i$ (i.e., z_i is a singleton), we propose to re-assign z_i to one of the existing components with some other data points assigned to it. We accept the proposal based on the Metropolis-Hastings acceptance rule. To improve the mixing in the sampling algorithm, we use an additional partial Gibbs sampling step for non-singletons [31].

As we sample from the conditional posterior in (9), the z_i's are associated with a finite number of components. Note that the number of components represented by the z_i's can change from one sampling iteration to the next. New components can get generated, and existing components can disappear if they become empty with no z_i's assigned to them. We are interested only in which observations are associated with the same component, regardless of the ordering of the components in the labeling scheme. From a Bayesian viewpoint, at the end of the sampling iterations, we obtain a posterior distribution of the number of components as a histogram of the sampled values over iterations. If we wish to report one fixed value for the number of components, we can (for example) select the sample with the highest posterior probability and use the associated assignments and components.

2) Sampling Component Parameters: Given the component labels $\mathbf{z}$, we use a Gibbs sampling algorithm with some Metropolis steps to sample the component parameters. For the background mean μ and the concentration parameter α, it is possible to use the full conditional distribution of the parameter given all others. For the remaining parameters, we use a Metropolis algorithm with Gaussian random walk proposal distributions.

3) Initialization Methods: At the start of the MCMC sampling the model is initialized to one background component and one or more activation components. To initialize the model, we use a heuristic algorithm to find candidate voxels for local activation centers and assign a mixture component to each of the candidates. The heuristic procedure is as follows. To find candidate voxels, we take all of the positive voxels in an image, and repeatedly select the largest voxel among the voxels that have not been chosen and are at least several voxels apart from the previously selected voxels until there are no voxels left. The location and height parameters of the component are set to the position and the β-value of the candidate voxel respectively.

III. Hierarchical Dirichlet Processes for Multiple fMRI Images

To generalize the method of Section II so that it can learn common activation patterns from multiple images, we model each image as a mixture of experts model as described in Section II and let the components be shared among images by combining image-level models using a hierarchical Dirichlet process. A hierarchical Dirichlet process combines Dirichlet processes in a hierarchical manner to model multiple related mixture models with shared mixture components but different mixing proportions [21].

A. The Models

Assume that we have J images and that each image is modeled with a mixture of experts with a Dirichlet process prior as in Section II-A. To allow for the sharing of components across images, we combine Dirichlet processes hierarchically and model the base distribution G_0 as coming from an upper-level Dirichlet process $DP(\gamma, H)$. Let y_{ji} be the ith voxel ($i = 1, \ldots, N_j$) in image j ($j = 1, \ldots, J$). Then the complete model for a hierarchical Dirichlet process is given as

$$\begin{aligned} \phi_{ji} \sim G_j, \quad G_j|\alpha, G_0 \sim DP(\alpha, G_0), \\ G_0|\gamma, H \sim DP(\gamma, H), \end{aligned} \tag{10}$$

where H is a base distribution for the component parameters as defined in Section II-A, ϕ_{ji} is the component parameter associated with the ith voxel in the jth image, and γ and α are concentration parameters for Dirichlet processes. The G_0 in the above equation takes the following form

$$G_0 = \sum_{c=1}^{\infty} \beta_c \delta_{\theta_c}, \tag{11}$$

where β_c's are generated from Stick(γ) and θ_c's are distinct values in ϕ_{ji}'s representing distinct component parameters sampled from the base distribution H. The G_j's are given as

$$G_j = \sum_{c=1}^{\infty} \pi_{jc} \delta_{\theta_c}, \tag{12}$$

where π_{jc}'s are mixing proportions for the mixture model of the jth image. It is important to notice that G_j comes from the Dirichlet process with a discrete distribution G_0 as its base distribution. Thus, the components represented in the G_j's are the ones present in G_0, forcing the J images to have components with the same parameters.

We can derive a similar clustering property to (8) for hierarchical Dirichlet processes at each level of the hierarchy. At the bottom level, voxels y_{ji} for $i = 1, \ldots, N_j$ are assigned to one of the T_j local clusters within image j. When we integrate out G_j in (10), we obtain

$$p(h_{ji}|\mathbf{h}_{-ji}, \alpha) \sim \sum_{t=1}^{T_j} \frac{n_t^{-ji}}{N_j - 1 + \alpha} \delta_t + \frac{\alpha}{N_j - 1 + \alpha} \delta_{t_{\text{new}}}, \tag{13}$$

where h_{ji} represents the label assignment of y_{ji} to one of the T_j local clusters, $\mathbf{h}_{-ji}$ is all of the h_{ji}'s in the jth image excluding h_{ji}, and n_t^{-ji} is the number of voxels assigned to the tth local cluster excluding y_{ji}.

At the top-level Dirichlet process, the local clusters are assigned to one of the global clusters with parameters θ_k. This assignment can be seen by integrating out G_0 in (10):

$$\begin{aligned} p(l_{jt}|\mathbf{l}_{-jt}, \gamma) \sim \sum_{k=1}^{K} \frac{m_k^{-jt}}{\sum_u m_u - 1 + \gamma} \delta_k \\ + \frac{\gamma}{\sum_u m_u - 1 + \gamma} \delta_{k_{\text{new}}}, \end{aligned} \tag{14}$$

where l_{jt} maps the tth local cluster in image j to one of the K global clusters shared by all of the J images, $\mathbf{l}_{-jt}$ represents the label assignments for all of the local clusters across J images excluding l_{jt}, m_k^{-jt} is the number of local clusters in image j assigned to the kth global cluster excluding l_{jt}, and m_k is the number of local clusters assigned to the kth global cluster.

According to (13) the probability that a new local cluster is generated within image j is proportional to α. This new cluster is generated according to (14). If a new component is selected in (14) the corresponding component parameter is drawn from the base distribution H.

Notice that more than one local cluster in image j can be linked to the same global cluster. It is the assignment of voxels to one of the K global clusters via local cluster labels that is of interest.

B. MCMC Sampling for Learning

To learn the model from a set of activation images we use an MCMC sampling algorithm to sample from the posterior distribution of the component labels and unknown parameters given a set of images. The quantities of interest are the local cluster labels h_{ji}'s, the global cluster labels l_{jt}'s, and the component parameters $\boldsymbol{\theta} = \{\mu, \sigma^2_{\text{bg}}, \sigma^2_{\text{act}}, \{\mathbf{b}_m, \Sigma_m, k_m\}, m = 1, \ldots, M, \alpha, \gamma\}$. We sample each of these in turn from its conditional posterior distribution in each iteration of Gibbs sampling. See Appendix A for more details.

IV. Hierarchical Dirichlet Processes with Random Effects for Multiple fMRI Images

To achieve the flexibility required to model the type of image-specific variation illustrated in Figure 1, we further extend the model described in Section III by introducing random effects on component parameters.

A. The Models

We take the model in Section III-A and let the jth image have its own component parameters $\mathbf{b}_{mj}$'s, $\mathbf{k}_{mj}$'s, Σ_{mj}'s, and μ_j's as follows. Let z_{ji} be the label assignment of the ith voxel in the jth image to one of the global clusters through h_{ji}'s and l_{jt}'s. Then, the observation $(y_{ji}, \mathbf{x}_{ji})$ in the jth image is modeled as

$$\begin{aligned} &E[y_{ji}|\mathbf{x}_{ji}, z_{ji} = c_m, \theta_{mj}, \sigma^2_{\text{act}}] \\ &= k_{mj} \exp\left(-(\mathbf{x}_{ji} - \mathbf{b}_{mj})'(\Sigma_{mj})^{-1}(\mathbf{x}_{ji} - \mathbf{b}_{mj})\right), \end{aligned} \tag{15}$$

and if z_{ji} is a background component, the observation $(y_{ji}, \mathbf{x}_{ji})$ in the jth image is modeled as

$$E[y_{ji}|\mathbf{x}_{ji}, z_{ji} = c_{\text{bg}}, \theta_{j,\text{bg}}, \sigma^2_{\text{bg}}] = \mu_j.$$

The image-specific parameters $\mathbf{b}_{mj}$'s and k_{mj}'s, for each of the M activation components, and the μ_j's are modeled as coming from a common prior distribution given by

$$\begin{aligned} \mathbf{b}_{mj} &\sim \mathcal{N}(\mathbf{b}_m, \boldsymbol{\Psi}_{\mathbf{b}_m}) \\ k_{mj} &\sim \mathcal{N}(k_m, \psi^2_{k_m}) \\ \mu_j &\sim \mathcal{N}(\mu, \psi^2_{\mu}), \end{aligned} \tag{16}$$

where the $\mathbf{b}_m$'s and k_m's define the unknown template activation shape for the mth component, and μ defines the overall

background mean across images. The variance parameters $\mathbf{\Psi}_{\mathbf{b}_m}$, $\psi^2_{k_m}$, and ψ^2_μ represent the amount of variation in parameters $\mathbf{b}_{mj}$, k_{mj}, and μ_j for each component across images. Thus, the $\mathbf{b}_m$'s, k_m's, and μ can be viewed as defining templates, and the $\mathbf{b}_{mj}$'s, k_{mj}'s, and μ_j's as noisy observations of the template for image j with variances $\mathbf{\Psi}_{\mathbf{b}_m}$'s, $\psi^2_{k_m}$'s, and ψ^2_μ respectively. The width parameters Σ_{mj}'s are modeled as coming from a prior of a half-normal distribution for the variance elements and a uniform distribution over $[-0.5, 0.5]$ for the correlations. We do not constrain Σ_{mj}'s across images through a population-level distribution. For $\psi^2_{k_m}$'s and ψ^2_μ in (16), we use half-normal distributions with mean 0 and predefined values for variances as priors. For $\mathbf{\Psi}_{\mathbf{b}_m}$'s, we set the covariance elements to 0, assuming that b_{m_1} and b_{m_2} in $\mathbf{b}_m = (b_{m_1}, b_{m_2})$ are independent of each other, and use a half-normal prior on the variance elements.

We extend the clustering properties in (13) and (14) for hierarchical Dirichlet processes to describe the generative process of the model described above. The only difference is that if a global cluster selected using (14) (for the assignment of l_{jt}) is a component that has not been used in the jth image, the image-specific shape parameters for the jth image needs to be generated from its prior.

B. Learning with MCMC Sampling

For inference we use an MCMC sampling scheme that is based on the clustering property of the model described in the previous section. In each iteration of the sampling algorithm we alternately sample labels $\mathbf{h} = \{h_{ji} \text{ for all } j, i\}$, $\mathbf{l} = \{l_{jt} \text{ for all } j, t\}$ and component parameters $\boldsymbol{\theta} = \{\mu, \mu_j, \sigma^2_{\text{bg}}, \sigma^2_{\text{act}}, \{\mathbf{b}_m, \mathbf{b}_{mj}, k_m, k_{mj}, \Sigma_{mj}, \mathbf{\Psi}_{\mathbf{b}_m}, \psi^2_{k_m}, \psi^2_\mu\}, m = 1, \ldots, M, j = 1, \ldots, J, \alpha, \gamma\}$ from their conditional posterior distributions. Details are given in Appendix B.

C. Demonstration of the Model on Simulated Data

We demonstrate the performance of the hierarchical Dirichlet process with random effects on simulated data, and compare the results with what we obtain from a simple thresholding scheme for a voxel-wise analysis. We assume a template activation pattern of three activation clusters of the same intensity located at $\mathbf{b}_1$=(7,7), $\mathbf{b}_2$=(7,19), and $\mathbf{b}_3$=(15,15) in a 20-by-25 region, and generate 10 observed images from this template, using 0.3, 0.8, and 1.2 as the variance element of $\Psi_{\mathbf{b}_m}$'s, and 0.3, 0.2, and 0.1 as ψ_{k_m}'s for $m = 1, 2, 3$. We further assume that the first three observed images contain all of the three activation clusters, the next three images contain two clusters, $m = 1$ and 2, and the remaining four images have a different subset of clusters, $m = 2$ and 3. Given the true parameters, the intensity value at each voxel of an image is determined by computing the intensity level for each of the activation and background components and taking the maximum value. The corresponding component with the maximum value is considered as the true cluster label for the voxel. Finally, we add noise generated from $N(0, \sigma^2_{\text{act}})$ and $N(0, \sigma^2_{\text{bg}})$ to the intensity level of each voxel.

Illustrations demonstrating the fit of the hierarchical Dirichlet process with random effects to two sets of sample images are provided in Figures 2(a) and (b). Figure 2(a) shows the single MCMC draw with the highest posterior probability when the model is fit to data with low activation intensities and high noise levels (k_m=1.0 and $\sigma^2_{\text{act}} = \sigma^2_{\text{bg}} = 1.0$). Figure 2(b) shows the single MCMC draw with the highest posterior probability when the model is fit to data with high activation intensities and low noise levels (k_m=2.0 and $\sigma^2_{\text{act}} = \sigma^2_{\text{bg}} = 0.2$). The left-most panel in Figures 2(a) and (b) shows the true locations of the template activation clusters as '+'s and the region within one standard deviation as measured by the estimated $\mathbf{\Psi}_m$'s is indicated by ellipses around each location. In the remaining panels of Figures 2(a) and (b), we show the estimated image-specific activation clusters for seven (out of the ten) images in the image set as ellipses marking the region within one standard deviations (as measured by the estimated Σ_{mj}'s) of the image-specific location; the ellipses are overlaid on the images. In the high-noise case, Figure 2(a), our model is able to determine the approximate locations of the two out of the three true activation components by combining the information across multiple images, even when the activation areas are not clear from individual images. When there is clear evidence for activation with low noise as in Figure 2(b), our model is able to correctly identify the three true activation clusters.

In order to systematically compare the performance of our method for detecting activated voxels and a benchmark threshholding approach, we simulate 20 sets of 10 images (each set of 10 generated as described above), and plot receiver operating characteristic (ROC) curves averaged over the 20 datasets. To obtain an ROC curve from the hierarchical Dirichlet process with random effects, we fit the model to each dataset, and use the following scheme to rank the voxels. Using the posterior samples for cluster labels h_{ji}'s from the MCMC sampling algorithm, we estimate the posterior probability of each voxel belonging to either background component or any of the activation components. We aggregate these probabilities so that we obtain for each voxel a single posterior probability of that voxel being part of an activation cluster (without regard to which cluster). We rank the voxels according these posterior probabilities of belonging to any of the activation components, compare the sorted list with the set of truly activated voxels, and plot type I errors and powers as an ROC curve. For the thresholding method, we simply rank the voxels according to their intensities to obtain ROC curves.

We compute ROC curves for varying values of the parameter defining activation heights (k_m=1.0, 1.5, 2.0) in Figure 2(c) with the width fixed ($\Sigma_{mj} = 3^2$), and then for varying values of the parameter defining activation widths ($\Sigma_{mj} = 2^2, 3^2, 4^2$) in Figure 2(d) with the height fixed ($k_m = 1.5$). The three sets of ROC curves from the left to the right in Figures 2(c) and (d) correspond to different levels of noise, $\sigma^2_{\text{bg}} = \sigma^2_{\text{act}} = 0.2$, 0.6, and 1.0, respectively. The results show that across different values of shape parameters and noise levels our method outperforms the thresholding method that treats voxels independently of each other.

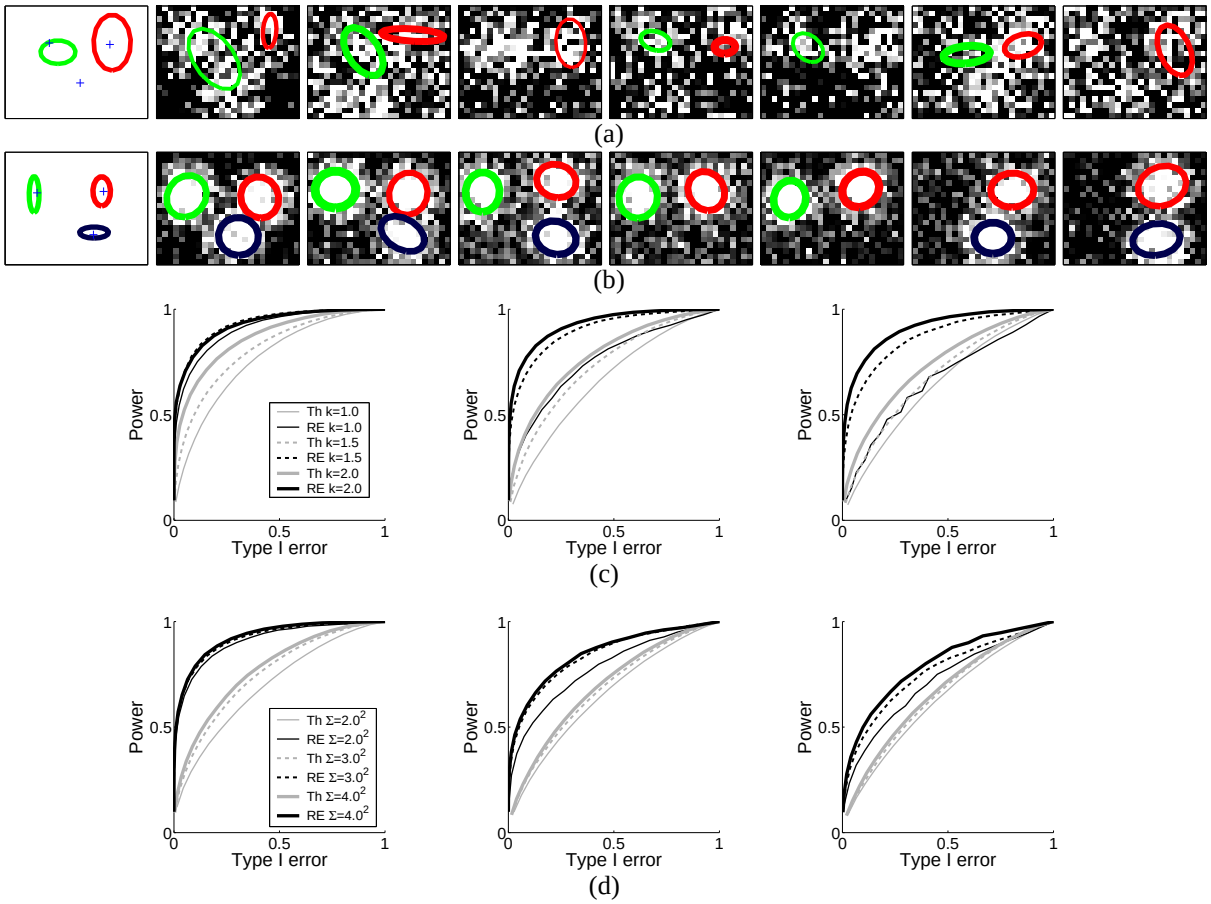


Fig. 2. Results on simulated data comparing the hierarchical Dirichlet process with random effects and a thresholding approach. The activation clusters estimated from hierarchical Dirichlet process with random effects are shown for datasets simulated with (a) low activation intensities with high noise level (k_m=1.0 and $\sigma^2_{\text{act}} = \sigma^2_{\text{bg}} = 1.0$), and (b) high activation intensities with low noise level (k_m=2.0 and $\sigma^2_{\text{act}} = \sigma^2_{\text{bg}} = 0.2$). The results are based on the single MCMC sample with the highest posterior probability. The left-most images in Panels (a) and (b) show the true locations of the template activation clusters as '+'s, and estimated regions within one standard deviation (as measured by the $\mathbf{\Psi}_m$'s) as ellipses. In the next seven images in Panels (a) and (b), the estimated image-specific activation clusters are overlaid over the raw images as ellipses of width one standard deviations (as estimated by the Σ_{mj}'s). ROC curves for detecting the true activation voxels averaged over 20 simulated datasets are shown for (c) varying heights k_m's (with fixed Σ_{mj}), and (d) varying widths, the variance elements of Σ_{mj}'s, of the activation clusters (for fixed k_m). In Panels (c) and (d), the three sets of ROC curves from the left to the right correspond to the noise levels $\sigma^2_{\text{bg}} = \sigma^2_{\text{act}} = 0.2$, 0.6, and 1.0, respectively.

V. Experiments

Using data from a multi-site fMRI study we experimentally compare the different models described in Sections II-IV, namely, the mixture of experts model with a Dirichlet process prior, the hierarchical Dirichlet process, and the hierarchical Dirichlet process with random effects.

A. Multi-Site Data Collection and Preprocessing

FMRI scans for the same five control subjects were collected from 10 different scanners as part of a multi-site study of functional brain images, known as FIRST BIRN or fBIRN (Functional Imaging Research on Schizophrenia Testbed Biomedical Informatics Research Network) [1], [32]. For each subject there were two visits to each site, and at each visit fMRI data were collected for four runs of a sensorimotor task and two runs of each of breathholding, resting, and cognitive recognition tasks, using a common protocol. The primary goal of this study is to better understand the variability of fMRI response patterns across runs, visits, scanners (sites) and subjects, so that future data collected across sites and subjects can be analyzed collectively and consistently (e.g., [16]). In the experiments below we use the data from the sensorimotor task, and focus on activation within specific regions of interest that are relevant to this task such as the left and right precentral gyri, the left and right superior temporal gyri, and the left and right occipital lobes. During the sensorimotor task subjects were presented with auditory and visual stimuli and were asked to tap fingers on each hand periodically. The brain regions analyzed correspond to areas expected to reflect visual, auditory and motor activity.

Each run of the sensorimotor task produces a series of 85 scans that can be thought of as a time-series of voxel images. The set of scans for each run is preprocessed in a standard manner using SPM99 [33] with the default settings. The preprocessing steps include correction of head motion, normalization to a common brain shape, and spatial smoothing. Using the motion correction algorithm as implemented in SPM99 all of the four runs in each visit are realigned to the first image of the first run. The head motion correction

is followed by a co-registration and normalization step that transforms the images into a standard space defined by the SPM eco-planar imaging (EPI) canonical template (Montreal Neurological Institute template, or MNI template). The normalized images are interpolated using bilinear interpolation and resliced to $2\times2\times2$mm voxels before being smoothed with an 8mm FWHM (Full Width at Half-Maximum) 3D Gaussian kernel.

A general linear model is then fit to the time-series data for each voxel, yielding a regression coefficient β that estimates the amount of activation at each voxel. The design matrix used in the analysis includes the on/off timing of the sensorimotor stimuli measured as a boxcar convolved with the canonical hemodynamic response function. A β-map is a voxel image of the regression coefficients (β's) that summarizes the activation across time as an activation map. Binary masks for regions of interest from the normalized atlas were then used to extract the β values for all voxels within a region. These β values serve as our data. The approach can be easily applied to other summary measures, e.g., t-statistics. We focus here on detecting areas of increased activation during the sensorimotor task relative to rest periods; the models outlined above could be modified to address expected decreases in activation as well.

B. Experimental Setup

In this section we describe the setup for our application of the models to the FBIRN data, including specification of the prior distributions for the models used in our experiments and initialization methods for our algorithms. We selected a 2-dimensional cross-section from each of the six regions of interest to fit the models. In all of the experiments we ran the sampling algorithm for 4000 iterations and used the samples over the last 3000 iterations after 1000 burn-in iterations to present the results. After a few initial runs of the sampling algorithm, we found that 4000 iterations with 1000 burn-in iterations were sufficient for convergence.

1) Single-image Models: We set the prior distributions for the model as follows. We *a priori* assumed a small value of the concentration paramater α that tends to encourage a relatively small number of components, and hence set the prior for α to Gamma(0.1, 1) in order to keep the mean of the gamma relatively small at 0.1. For the activation components, the priors for the width parameters Σ_m were set to a half-normal distribution with mean 0 and variance 100. We used a large variance in the half-normal prior to make the prior nearly uninformative within a fairly large range of values. We assumed an uninformative prior for the height parameters k_m, and used a uniform distribution over $[0, K_{\max}]$ with $K_{\max}$ set to a value 25% larger than the maximum intensity in the image. The initialization scheme described in Section II-C was used to initialize the model.

2) Multiple-image Models: For our experiments with the two hierarchical Dirichlet process models (with and without random effects), we analyzed a single subject at a time, using a collection of 80 images (10 sites $\times$ 8 runs per site) of each region of interest for the given subject. Thus, the model has three layers of Dirichlet processes, corresponding to (from the top) the global template, the site-specific template, and the image-specific (or run-specific) activation model. At the bottom level of the hierarchy are the eight images from the eight runs at a specific site. In the previous analysis of the same dataset [16], it was shown that the visit variability is much smaller than the run variability. Thus, in our analysis, we combined the four runs from each of the two visits, and modeled them as if they were eight runs from one visit, although a simple extension of our model with an addition of another hierarchy would allow us to explicitly model the visit variability. The model with random effects had $\mathbf{b}_m$'s and k_m's as global templates, and $\mathbf{b}_{mj}$'s and k_{mj}'s as site-specific parameters for the j-th site, modeled as coming from the distributions $\mathcal{N}(\mathbf{b}_m, \mathbf{\Psi}_{\mathbf{b}_m})$ and $\mathcal{N}(k_m, \psi^2_{k_m})$, respectively. The activation component parameters $\mathbf{b}_{mjr}$'s and k_{mjr}'s for an individual image of the rth run in the jth site were modeled as coming from the distributions $\mathcal{N}(\mathbf{b}_{mj}, \mathbf{\Psi}_{\mathbf{b}_{mj}})$ and $\mathcal{N}(k_{mj}, \psi^2_{k_{mj}})$, respectively. The model without random effects had $\mathbf{b}_m$'s and k_m's in a global template fixed across all of the images.

In all of our experiments, the prior distributions were set as follows. The priors for the concentration parameters at each of the three levels of Dirichlet processes were set, from the top level, to Gamma(0.1, 1), Gamma(2, 2), and Gamma(2, 2), respectively. For the activation components, the prior for the width parameters and the height parameters in the global template were set in the same manner as in the single-image model. For the variance parameters in the random effects model, we used a half-normal(0,4) and a half-normal(0,0.4) as priors for $\mathbf{\Psi}_{\mathbf{b}_m}$'s and $\psi^2_{k_m}$'s, respectively, at the top-level, and a half-normal(0,2) and a half-normal(0,0.2) as priors for $\mathbf{\Psi}_{\mathbf{b}_{mj}}$'s and $\psi^2_{k_{mj}}$'s, respectively, at the site-level. We *a priori* allowed the shape parameters to vary more at the top level than at the site-level by using a larger value for the variance in the half-normal priors for the top-level variance parameters. This is a reasonable assumption because intuitively we expect to see a larger variability in the activation patterns across sites, compared to across runs within a site.

We ran the heuristic clustering algorithm for initialization of the single-image model described in Section II-C on one of the 80 images (10 sites $\times$ 8 runs), and used the results as initial values for the number of components and component parameters. The output of this heuristic algorithm for the component labels of the single image was used to initialize the labels for all of the other images.

C. Illustrative Results for a Single Subject

To illustrate the methodology, we fit the two multiple-image models to the right precentral gyrus of subject 5. Using the single sample with the highest posterior probability, we show the estimated image-level (or run-level) activation patterns for the 8 runs in site 3 in Figure 3 for the non-random effects model and the random effects model. In the first column of Figure 3, we include the raw images, and in the second column in Figure 3, we highlight the top 15 % voxels with the highest intensities in each image, which would correspond to the results of single-voxel analyses. Ellipses are drawn to

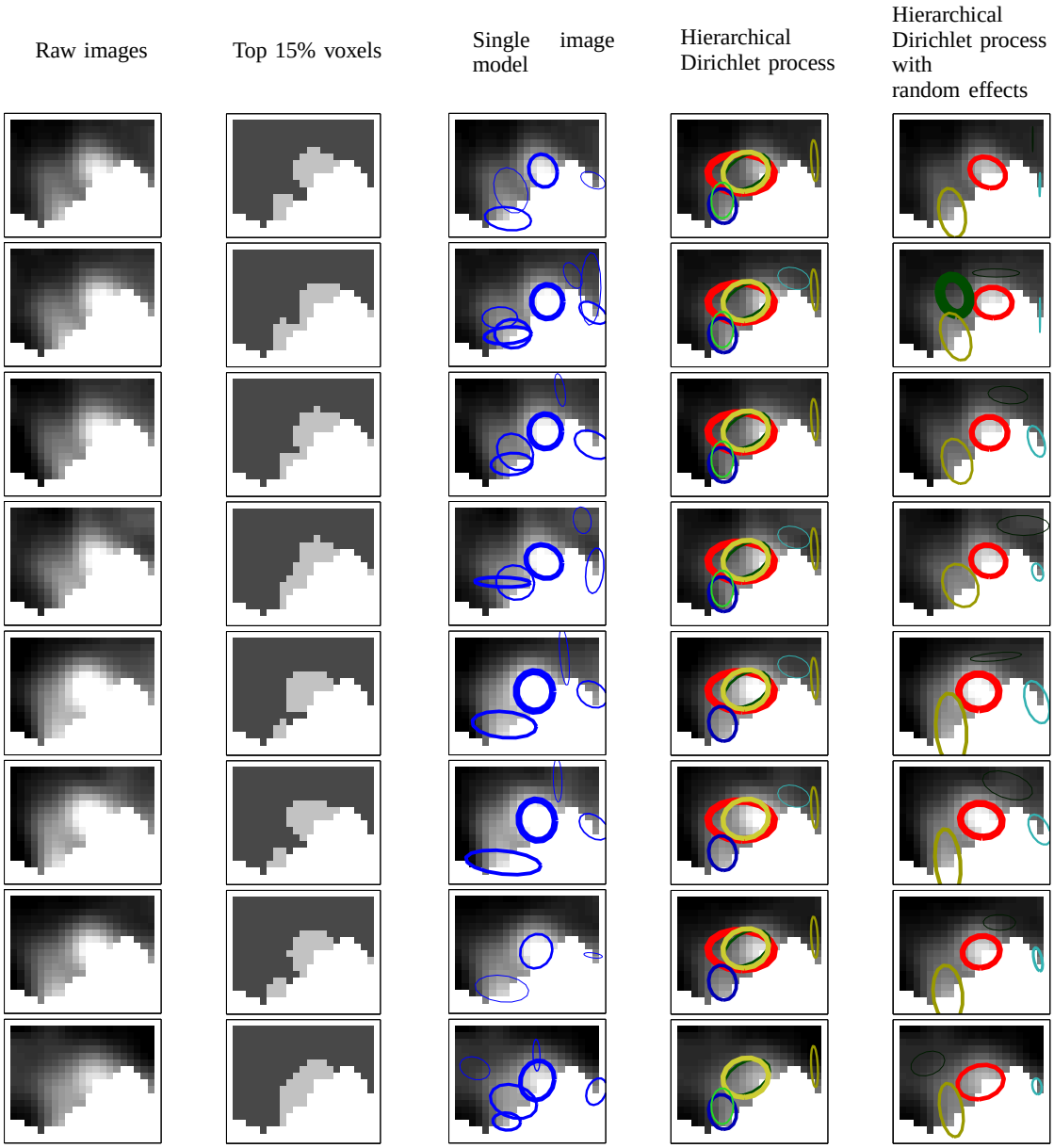


Fig. 3. Results for the right precentral gyrus for subject 5. Each row corresponds to an image from one of 8 runs for the same individual. In column 2, the top 15% voxels with the highest intensity levels are highlighted. In columns 3-5, estimated activation components are overlaid on the images. In column 3, a single-image model is fit to each of the 8 runs separately. In the case of multiple-image models (columns 4 and 5), we fit the models to all of the 80 images for subject 5 and show the run-level activation patterns, using the single sample with the highest posterior probability.

show a region of width standard deviation centered around the location parameters with the height parameters represented by the thicknesses of the ellipses. Ellipses with the same color across all sites and runs correspond to the same activation component.

For comparison, we fit the single-image model to each of the 8 images separately and show the results in the third column of figure 3. This single-image model correctly recognizes the high intensity areas as activation components, and finds other activation components with lower intensities. However, since this model analyzes each image separately it cannot link information across the eight runs even though the activation patterns are quite consistent among the images.

In the last two columns of Figure 3 we see that the hierarchical Dirichlet process with random effects captures common activation clusters better than the model without random effects, in the presence of run-specific variations in activation shapes. For example, the random effects model recognizes the activation components with a relatively high intensity in the middle of the images as realizations of the same component shared among those images, whereas the non-random effects model fits the same activation clusters with different combinations of multiple components in the different images. This shows that having a fixed set of parameters for all of the images does not give the model enough flexibility to model the variability due to sites and runs. The random effects model found a more compact summary of the site-specific activation pattern than the model without random effects.

Histograms of the number of components over the last

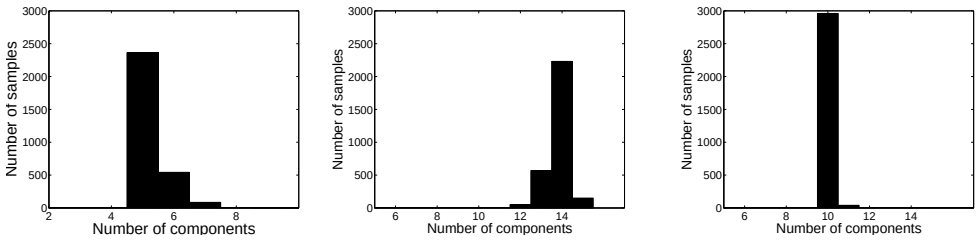


Fig. 4. Histograms of the number of components over the last 3000 iterations for the right precentral gyrus for subject 5. For the single-image model, the histogram for run 1 in visit 1 is shown.

3000 iterations are shown in Figure 4. The posterior mass is peaked around a larger number of components for the hierarchical Dirichlet process without random effects than for the model with random effects. This is because the hierarchical Dirichlet process generates a larger number of components to summarize the activation pattern in the same set of images, compared to the model with random effects. Since the hierarchical Dirichlet process has one set of fixed parameters shared across all of the images, it tries to explain the local variability in activation shape through additional activation components.

Using the same sample, we show the estimated top-level and site-level template activation patterns in Figure 5(a) for the hierarchical Dirichlet process and in Figure 5(b) for the model with random effects. For the model without random effects in Figure 5(a), the activation components are drawn as ellipses with size proportional to one standard deviation of the width parameters Σ_m's centered around the location parameters $\mathbf{b}_m$'s, and with the thickness of the ellipses proportional to the height parameters k_m. In Figure 5(a), the site-specific images on the right contain a subset of exact copies of components from the global template on the left. For the model with random effects in Figure 5(b), the global template shows the template activation components as ellipses centered around the location parameters $\mathbf{b}_m$'s with the height parameters k_m's as thicknesses. For the same model, we show the site-level templates on the right, using the site-specific shape parameters, $\mathbf{b}_{mj}$'s and k_{mj}'s, to draw the ellipses. Note that the radii of the dotted-lined ellipses for the model in Figure 5(b) are proportional to 1.5 times the standard deviation of the covariance parameters $\mathbf{\Psi}_{\mathbf{b}_m}$'s and $\mathbf{\Psi}_{\mathbf{b}_{mj}}$ that in turn represent the variation in the locations of activation components.

As we can see in Figure 5, once again, the random effects model finds a more compact summary of the activation pattern than the model without random effects by using a smaller number of components to explain the activation pattern.

We notice that the across-run variability represented as $\mathbf{\Psi}_{\mathbf{b}_{mj}}$'s in Figure 5(b) (on the right) is generally smaller than the across-site variability represented as $\mathbf{\Psi}_{\mathbf{b}_m}$'s (on the left). The frequencies of each activation component appearing in any of the ten sites for the subject are shown as numbers next to each ellipse in the global template in Figure 5. Similarly, the frequencies of each activation component appearing in any of the eight runs in each site for the subject are shown in the site-level templates. Most of the components are common across all of the eight runs within a site. This again shows that activation patterns are fairly consistent across runs within a site.

As for the computation time, it took 71 minutes to run the MCMC sampling algorithm for the hierarchical Dirichlet process model on the 80 images used in Figures 3 and 5, and 101 minutes for the model with random effects on the same set of images. These computation times could be considerably shortened by code optimization and/or by parallelizing the algorithms for execution on multi-core machines or grid architectures.

D. Comparison of Models across Subjects

We fit the model with random effects to the fMRI data for each subject for the left precentral gyrus and right superior temporal gyrus and show the estimated global activation templates in Figure 6. The estimated activation components are shown as ellipses that correspond to 1.5 times the covariance $\mathbf{\Psi}_{\mathbf{b}_m}$ in the component location parameters, centered around the location parameters $\mathbf{b}_m$, with the height parameters k_m as thicknesses. The number of times that each component appears in the 10 site-specific templates is shown next to the ellipses. We show only those components that appear in 5 or more sites.

Even though each subject is analyzed separately, there are several activation clusters in the results shown in Figure 6 that appear consistently across subjects within a region of interest. For example, in the left precentral gyrus in Figure 6, the models found an activation cluster in the upper left of the image in all of the subjects. For the right superior temporal gyrus all of the subjects show two activation clusters on the right of the images.

E. Analysis of Variability in Activation Patterns

The model with random effects can be used to estimate how much variability in the activation patterns is due to different sources, e.g., run-to-run versus site-to-site variability. The results from Figure 5(b) suggest that site variability is larger than run variability in terms of the locations of activation clusters, since for most of the clusters, the $\mathbf{\Psi}_{\mathbf{b}_m}$'s are larger than the $\mathbf{\Psi}_{\mathbf{b}_{mj}}$'s. Here we quantify more precisely the overall variation in the height and location parameters due to sites and runs on a per-subject basis. Given the estimated parameters for each region of interest (from the sample with the highest posterior probability from the model) we compute the overall

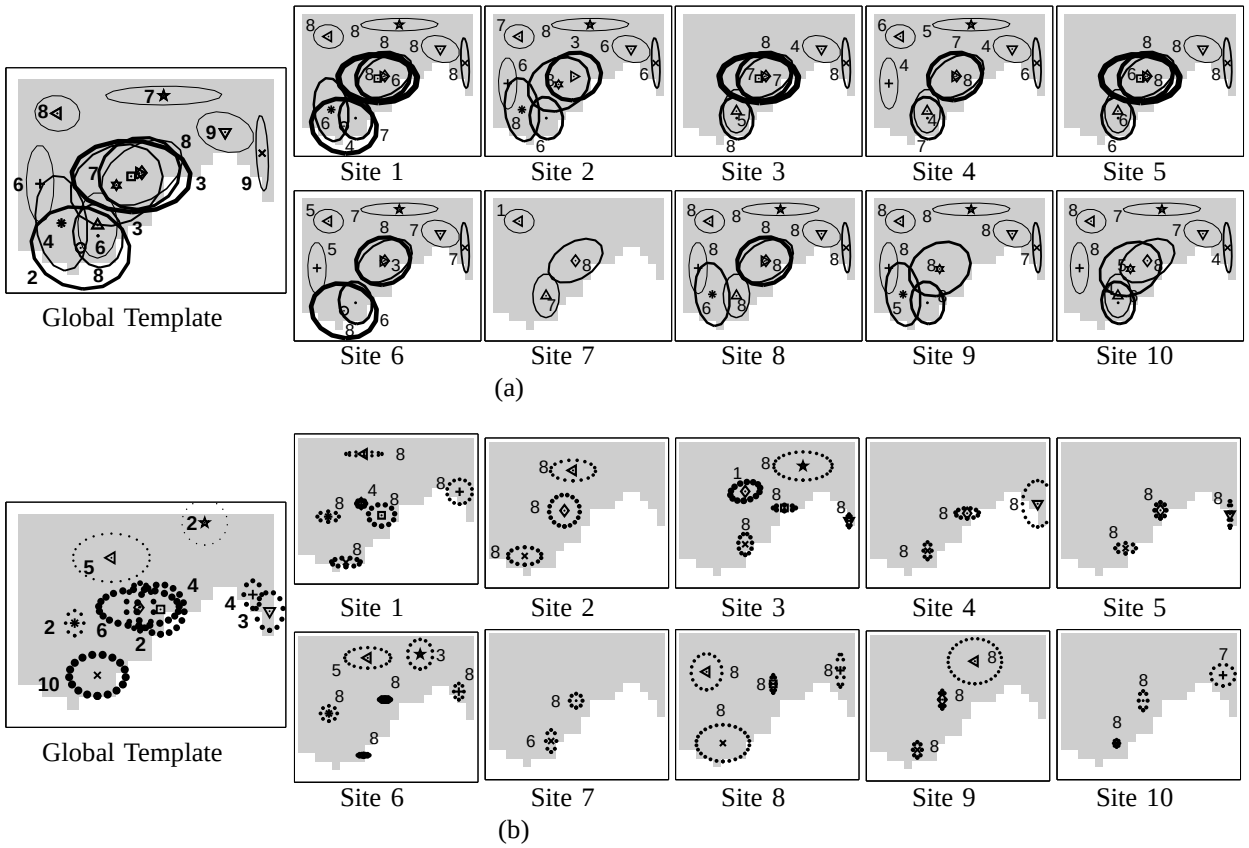


Fig. 5. The global template (on the left) and site-level templates (on the right) estimated by (a) a hierarchical Dirichlet process and (b) a hierarchical Dirichlet process with random effects, using the right precentral gyrus of subject 5. Note that the solid ellipses in (a) represent the widths of the activation components, whereas the dotted ellipses in (b) represent the variation in the location parameters of activation clusters. The single sample with the highest posterior probability is shown. The frequencies of each activation component appearing in any of the 10 sites for the subject are shown as numbers next to each ellipse in the global template. Similarly, the frequencies of each activation component appearing in any of the eight runs in each site for the subject are shown in the site-level templates.

site variability in the height parameters $\text{Var}_{\text{(height,site)}}$ by taking an average of the variance parameters $\psi^2_{k_m}$'s over all of the components as follows:

$$\text{Var}_{\text{(height,site)}} = \frac{\sum_m \psi^2_{k_m}}{\text{(Number of activation components)}}.$$

Similarly, we compute the overall run variability in the height parameters $\text{Var}_{\text{(height,run)}}$ by taking an average of the variance parameters $\psi^2_{k_{mj}}$'s over all of the sites and components as follows:

$$\text{Var}_{\text{(height,run)}} = \frac{1}{\text{(Number of sites)}} \cdot \sum_j \frac{\sum_m \psi^2_{k_{mj}}}{\text{(Number of activation components in site } j)}.$$

We plot the results for the right precentral gyrus in Figure 7(a). As we expected, for all of the subjects, the site variability is much larger than the run variability.

We perform the same analysis for the location parameters. We summarize the information in the 2×2 covariance matrices $\mathbf{\Psi}_{\mathbf{b}_m}$'s by taking the sum of the two variance elements $\mathbf{\Psi}_{\mathbf{b}_m}(1,1)$ and $\mathbf{\Psi}_{\mathbf{b}_m}(2,2)$. We can compute the overall site variability in the location parameters $\text{Var}_{\text{(loc, site)}}$ as

$$\text{Var}_{\text{(loc,site)}} = \frac{\sum_m \left(\mathbf{\Psi}_{\mathbf{b}_m}(1,1) + \mathbf{\Psi}_{\mathbf{b}_m}(2,2)\right)}{\text{(Number of activation components)}},$$

and, similarly, the overall run variability in the location parameters $\text{Var}_{\text{(loc, run)}}$ as

$$\text{Var}_{\text{(loc,run)}} = \frac{1}{\text{(Number of sites)}} \cdot \sum_j \frac{\sum_m \left(\mathbf{\Psi}_{\mathbf{b}_{mj}}(1,1) + \mathbf{\Psi}_{\mathbf{b}_{mj}}(2,2)\right)}{\text{(Number of activation components in site } j)}.$$

The results are shown in Figure 7(b) for the right precentral gyrus. Again, we see that in location parameters the site variability is larger than the run variability. These results are consistent with those of Friedman et al [16], who analyzed images from the same experiment using analysis of variance models applied to statistics such as the maximum and median values of percent signal change and contrast-to-noise ratio within each region of interest.

In Figures 7(a) and (b), the difference between the site and run variability is larger for the height parameters than for the location parameters. A plausible explanation is that there are scanner-specific characteristics such as the magnet strength that affect the heights of activation clusters more than the locations.

In Figure 5(b), we notice that most of the activation components are shared among all of the images across runs in the same site, whereas this persistency is weaker across sites.

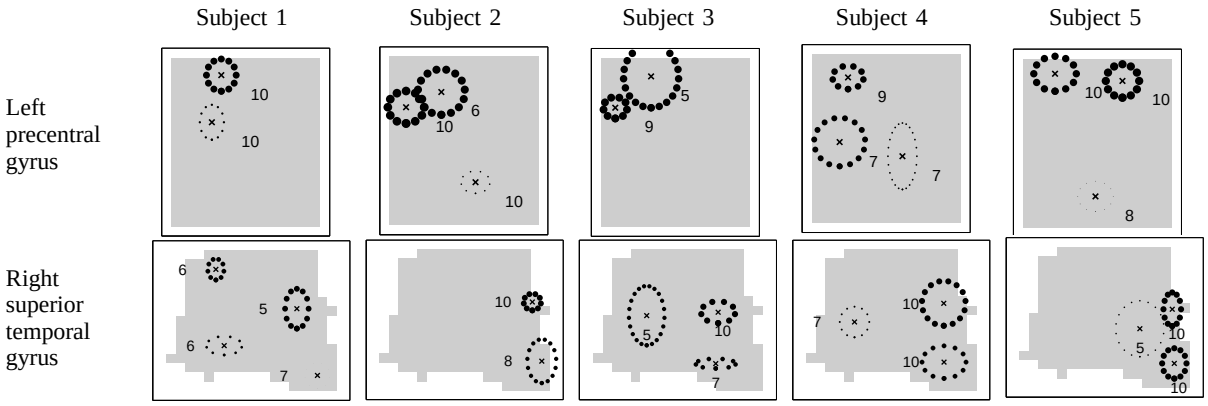


Fig. 6. Hierarchical Dirichlet processes with random effects fit to the left precentral gyrus and the right superior temporal gyrus of subjects 1, 2, 3, 4, and 5. The global template activations based on the single sample with the highest posterior probability are shown as ellipses of 1.5 standard deviation of the covariances $\mathbf{\Psi}_{\mathbf{b}_m}$ centered at the locations $\mathbf{b}_m$'s with the heights k_m's as thicknesses. The frequencies of each activation component appearing in any of the 10 sites for the subject are shown as numbers next to each ellipse.

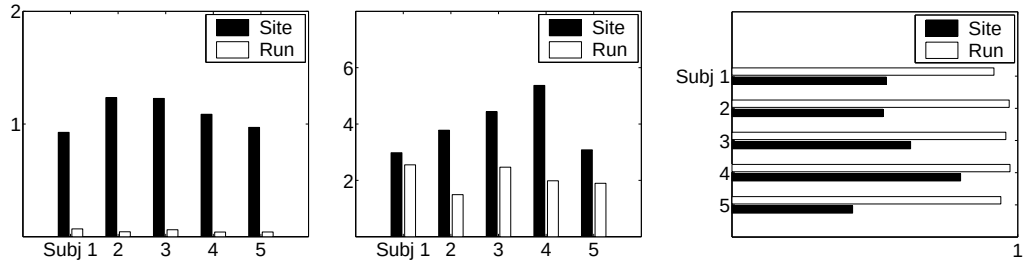


Fig. 7. Analysis of variability in activation patterns. (a) Variability in height parameters, (b) variability in location parameters, and (c) average rate of occurrence of activation components across images.

In order to quantify how persistent an activation component is across sites or runs, we compute the average rate of occurrence of an activation component among the 10 sites or the 8 runs. We compute the measure for site as follows.

$$\mathrm{F}_{\text{site}} = \frac{1}{\text{(Number of activation components)}} \cdot \sum_m \frac{\text{(Frequency of the } m\text{th component among 10 sites)}}{\text{(Number of sites)}}$$

Similarly, we use the following as the measure for runs.

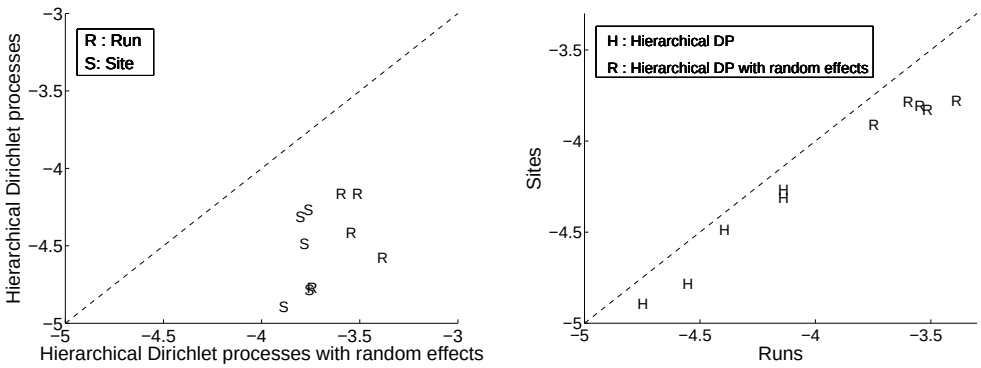
$$\mathrm{F}_{\text{run}} = \frac{1}{\text{(Number of sites)}} \cdot \sum_j \frac{\sum_m \frac{\text{(Frequency of the } m\text{th component among 8 runs in site j)}}{\text{(Number of runs)}}}{\text{(Number of activation components in site } j\text{)}}$$

If the values for F_{site} or F_{run} are close to 1, most of the images across sites or runs share common components. We plot the results in Figure 7(c). As we expected, activation components are more persistent across runs than across sites.

F. Evaluation of Predictive Performance

To evaluate the predictive benefit of adding random effects to the hierarchical Dirichlet process model, we conducted a number of cross-validation experiments. Specifically, we compute the logP score, $p(D_{\text{test}}|D_{\text{train}})$, of test data D_{test} given training data D_{train} for each model in each fold of the cross validation. The logP score is a fair estimator of the predictive power of a model (irrespective of how many parameters the model has), as it evaluates how much probability mass a model assigns to unseen test data, higher probability values being better (e.g., [34]). We compute $p(D_{\text{test}}|D_{\text{train}})$ using Monte Carlo integration over the parameters (the component labels and random effects parameters) as follows. We draw parameters from their posterior distribution given the training data, evaluate the likelihood of the test data given these parameters, and compute an average of this likelihood over multiple posterior draws of the parameters.

For a given subject and region of interest, we perform cross-validation at two different levels, one at the run level and the other at the site level. For run-level cross-validation, we leave out one run from each of the 10 sites, use those held-out 10 images as test data, and perform an 8-fold (across 8 runs) cross-validation. For each set of held-out runs, we train the model on the remaining 70 images from the 10 sites, and compute the predictive log-likelihood (or logP score) of the 10 held-out test images (one per site). In the site-level cross-validation, we leave out one site at a time, use the 8 images in the held-out site as test set, and perform a 10-fold cross-validation. We use the $9 \times 8 = 72$ images in the other 9 sites as training set, learn the model, and compute the predictive log-likelihood of the 8 images at the test site. Intuitively, there should be more uncertainty in a future observation when the same subject is scanned at a new site, compared to when the same subject is scanned for another run at the same site. Thus, we expect to see a lower predictive logP score per image



(a) (b)

Fig. 8. Results from cross-validations for right precentral gyrus. (a) With random effects vs. without random effects. Each letter corresponds to leave-one-site-out or leave-one-run-out cross validations for each of the five subjects. (b) Leave-one-site-out vs. leave-one-run-out. Each letter corresponds to the models with or without random effects for each of the five subjects.

for the leave-one-site-out than for the leave-one-run-out cross-validation.

We show the average per-voxel logP scores of test data for the right precentral gyrus in Figure 8. The figure shows the scores for the five subjects from leave-one-run-out and leave-one-site-out cross-validation. In Figure 8(a), the x-axis represents logP scores from the model with random effects, and the y-axis from the model without random effects. For all of the subjects, the model with random effects shows systematic improvement in logP scores compared to the hierarchical Dirichlet process in both leave-one-run-out and leave-one-site-out cross-validations.

In Figure 8(b), we plot the logP scores of the five subjects for both models using the x-axis as the scores from the leave-one-run-out cross-validation and the y-axis as the scores from the leave-one-site-out cross-validation. In all of the cases, the subjects shown as letters lie under the $x = y$ line, confirming our intuition that the leave-one-site-out cross-validation would give a lower logP score.

VI. Conclusions

In this paper we proposed a probabilistic framework for analyzing spatial activation patterns in multiple fMRI activation images. Each image was modeled as a mixture of a background component and a number of activation components with each activation component representing an activation cluster as a Gaussian-shaped surface. We combined multiple single-image models through a hierarchical Dirichlet process. With the hierarchical Dirichlet process we were able to infer the activation clusters that appear commonly in all or a subset of the images. The number of activation components was inferred from the data using a nonparametric Bayesian framework with a hierarchical Dirichlet process. To allow further flexibility in the model we incorporated random effects in the activation shape parameters and let each individual image have image-specific variation in the activation shape rather than forcing all images to have a fixed set of activation shape parameters as is the case in the hierarchical Dirichlet process. In this probabilistic framework we were able to learn the unknown template activation shape as well as the random effects parameters for each image, and we demonstrated this on a dataset from a multi-site fMRI study.

The model we propose in this paper assumes that the group-specific variation in parameters in any single mixture component is independent of the variation in parameters of other components. A possible extension would be to model additional systematic group variation in the mixture component parameters such as global translations of the template (or a subset of the components) in an image, e.g., due to different MRI machine characteristics or head positioning. We could also include across-subject variability in the model instead of analyzing each subject separately, and model the interaction between subjects and sites in terms of variation in the activation shape.

Other information could also be used to further enhance the model. For example, in this paper we focused on activation maps that summarize the voxel time-series into a single image. To take advantage of all of the information present in the dataset, a useful extension would be to model spatial patterns over time, e.g., combining the proposed Dirichlet process framework with the time-dependent model of Penny and Friston [7]. Furthermore, structural MRI scans collected for a subject could be used as a spatial prior to constrain modeled activation areas to gray matter regions in the brain (e.g., as proposed in [35]).

Another useful direction would be to extend the hierarchical Dirichlet process with random effects proposed in this paper to model differences between labeled groups of individuals, e.g., in studies of controls and patients for a particular disorder. This could be done by introducing a variable for a group label in the model, whose value is known in the training data, but is unknown at prediction time.

Appendix A
Sampling Algorithm for Hierarchical Dirichlet Processes

A. Sampling Component Labels

To sample the component labels we use the sampling algorithm based on the clustering properties of (13) and (14). We sample the local cluster labels h_{ji}'s by drawing each of

the h_{ji}'s in turn from the conditional posterior distribution, given as

$$p(h_{ji}=t|\mathbf{h}_{-ji},\boldsymbol{\theta},\mathbf{y},\mathbf{X}) \propto \begin{cases} n_t^{-ji}p(y_{ji},\mathbf{x}_{ji}|\theta_{jt}) & \text{if } t \text{ was used} \\ \alpha p(y_{ji},\mathbf{x}_{ji}|\mathbf{h}_{-ji},\boldsymbol{\theta},\gamma) & \text{if } t=t_{\text{new}}, \end{cases}$$

where $\boldsymbol{\theta}$ is the set of all component parameters, θ_{jt} is the parameters of one of the K components associated with the tth local cluster in image j, and

$$p(y_{ji},\mathbf{x}_{ji}|\mathbf{h}_{-ji},\boldsymbol{\theta},\gamma)=\sum_{k=1}^{K}\frac{m_k}{\sum_u m_u+\gamma}p(y_{ji},\mathbf{x}_{ji}|\theta_k) +\frac{\gamma}{\sum_u m_u+\gamma}\int p(y_{ji},\mathbf{x}_{ji}|\theta)p(\theta)\mathrm{d}\theta. \quad (17)$$

We sample the global cluster labels l_{jt}'s using the conditional posterior distribution given as

$$p(l_{jt}=k|\mathbf{l}_{-jt},\boldsymbol{\theta},\mathbf{y},\mathbf{X}) \propto \begin{cases} m_k^{-jt}\prod_{i:h_{ji}=t}p(y_{ji},\mathbf{x}_{ji}|\theta_k) \\ \qquad \text{if } k \text{ was used in image } j \\ \gamma\int\prod_{i:h_{ji}=t}p(y_{ji},\mathbf{x}_{ji}|\theta)p(\theta)\mathrm{d}\theta \\ \qquad \text{if } k \text{ is new in image } j. \end{cases} \quad (18)$$

Since we do not have conjugate priors for the component parameters in this model for fMRI data, it is not possible to evaluate the integrals in (17) and (18) analytically for a new component. We approximate the integrals by drawing a sample from the prior and evaluating the likelihood using this sample [31].

B. Sampling Component Parameters

Given the sample for component labels we sample the component parameters. We use Gibbs sampling to sample the background mean μ and the Metropolis algorithm with the normal distribution as a proposal for all of the other parameters. We place a gamma prior on α and γ and sample values for these parameters from their conditional posterior distributions [21].

APPENDIX B
SAMPLING ALGORITHM FOR HIERARCHICAL DIRICHLET PROCESSES WITH RANDOM EFFECTS

A. Sampling Component Labels

Because of the presence of image-specific shape parameters, the sampling methods for component labels for a hierarchical Dirichlet process in (17) and (18) cannot be directly applied to its extension with random effects. In a hierarchical Dirichlet process, since image-level shape parameters are exact copies of the corresponding component parameters in the template activation pattern, whenever we decide to generate a new local cluster for an image-specific activation pattern, we can simply copy the parameters from the template. However, in the model with random effects, each image inherits a perturbed version of the parameters in the template, and we should consider two separate cases of known and unknown image-specific parmaters for each of the template component, when generating a new local cluster. The known image-specific parameters indicate that the component in the template pattern has been introduced to the image before, whereas such parameters do not exist for the component being introduced to the image for the first time. We modify the sampling equations for hierarchical Dirichlet processes to take into account image-specific parameters as described below.

We sample h_{ji}'s using the following conditional distribution:

$$P(h_{ji}=t|\mathbf{h}_{-ji},\boldsymbol{\theta},\mathbf{y},\mathbf{X}) \propto \begin{cases} n_t^{-ji}p(y_{ji},\mathbf{x}_{ji}|u_{jt}) & \text{if } t \text{ was used} \\ \alpha\, p(y_{ji},\mathbf{x}_{ji}|\mathbf{h}_{-ji},\boldsymbol{\theta},\gamma) & \text{if } t=t_{\text{new}}, \end{cases}$$

where u_{jt} is the image-specific activation component parameters associated with the tth local cluster in image j and

$$p(y_{ji},\mathbf{x}_{ji}|\mathbf{h}_{-ji},\boldsymbol{\theta},\gamma)=\sum_{k\in\mathbf{A}}\frac{m_k}{\sum_u m_u+\gamma}p(y_{ji},\mathbf{x}_{ji}|u_{jk}) \quad (19a)$$

$$+\sum_{k\in\mathbf{B}}\frac{m_k}{\sum_u m_u+\gamma}\int p(y_{ji},\mathbf{x}_{ji}|u)p(u|\theta_k)\mathrm{d}u \quad (19b)$$

$$+\frac{\gamma}{\sum_u m_u+\gamma}\int\int p(y_{ji},\mathbf{x}_{ji}|u)p(u|\theta)p(\theta)\mathrm{d}u\mathrm{d}\theta. \quad (19c)$$

In (19a) the summation is over components in $\mathbf{A}=\{k|$ some $h_{ji'}$ for $i'\neq i$ is assigned to $k\}$, representing global clusters that already have some local clusters in image j assigned to them. In this case, since u_{jk} is already known, we can simply compute the likelihood $p(y_{ji},\mathbf{x}_{ji}|u_{jk})$. In (19b) the summation is over $\mathbf{B}=\{k|$ no $h_{ji'}$ for $i'\neq i$ is assigned to $k\}$ representing global clusters that have not yet been assigned in image j. In (19c) we model the case where a new global component gets generated. The integrals in (19b) and (19c) cannot be evaluated analytically, so we approximate the integral by sampling new values for u_{jk} and θ_k from their prior distributions and evaluating the likelihood $p(y_{ji},\mathbf{x}_{ji}|u_{jk})$ [31].

Samples for l_{jt}'s can be obtained from the conditional distribution given as

$$P(l_{jt}=k|\mathbf{l}_{-jt},\mathbf{u},\boldsymbol{\theta},\mathbf{y},\mathbf{X}) \propto \begin{cases} m_k^{-jt}\prod_{i:h_{ji}=t}p(y_{ji},\mathbf{x}_{ji}|u_{jk}) \\ \qquad \text{if } k \text{ was used in image } j \\ m_k^{-jt}\int\prod_{i:h_{ji}=t}p(y_{ji},\mathbf{x}_{ji}|u)p(u|\theta_k)\mathrm{d}u \\ \qquad \text{if } k \text{ is new in image } j \\ \gamma\int\int\prod_{i:h_{ji}=t}p(y_{ji},\mathbf{x}_{ji}|u)p(u|\theta)p(\theta)\ \mathrm{d}u\mathrm{d}\theta \\ \qquad \text{if } k \text{ is a new component.} \end{cases} \quad (20)$$

As in the sampling of h_{ji}, we cannot evaluate the integrals in (20) analytically. We approximate the integrals by sampling new values for u_{jk} and θ_k and from the priors and evaluating the likelihood.

B. Sampling Component Parameters

Given $\mathbf{h}$ and $\mathbf{l}$ we use Gibbs sampling to sample the background means μ and μ_j's and use the Metropolis algorithm with a normal distribution as a proposal for all of the other parameters.

In practice, this MCMC scheme for the hierarchical Dirichlet process with random effects can mix poorly and get stuck in

local maxima where the labels for two image-level components are swapped relative to the same two components in the template. To address this problem and restore the correct correspondence between template components and image-level components we propose a move that swaps the labels for two group-level components at the end of each sampling iteration and accepts the move based on a Metropolis acceptance rule.

ACKNOWLEDGMENT

The authors acknowledge the support of the following grants: the Functional Imaging Research in Schizophrenia Testbed, Biomedical Informatics Research Network (FIRST BIRN; U24RR021992, www. nbirn.net); and through the Transdisciplinary Imaging Genetics Center (P20RR020837-01) and the National Alliance for Medical Image Computing (NAMIC; Grant U54 EB005149), funded by the National Institutes of Health through the NIH Roadmap for Medical Research. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. This material is also based upon work supported by the National Science Foundation under grant No. IIS-0431085.

REFERENCES

[1] K. Zou, D. Greve, M. Wang, S. Pieper, S. Warfield, N. White, S. Manandhar, G. Brown, M. Vangel, R. Kikinis, and W. Wells, "Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network," *Radiology*, vol. 237, pp. 781–789, 2005.

[2] L. Friedman, G. Glover, and T. fBIRN Consortium, "Reducing scanner-to-scanner variation of activation in multi-center fMRI studies: adjustment for signal-to-fluctuation-noise-ratio," *NeuroImage*, vol. 33, no. 2, pp. 471–481, 2006.

[3] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995.

[4] S. Hayasaka and T. E. Nichols, "Validating cluster size inference: random field and permutation methods," *NeuroImage*, vol. 20, pp. 2343–2356, 2003.

[5] G. Flandin and W. Penny, "Bayesian fMRI data analysis with sparse spatial basis function priors," *NeuroImage*, vol. 34, no. 3, pp. 1108–1125, 2007.

[6] N. Hartvig, "A stochastic geometry model for fMRI data," University of Aarhus, Department of Theoretical Statistics, Research Report 410, 1999.

[7] W. Penny and K. Friston, "Mixtures of general linear models for functional neuroimaging," *IEEE Transactions on Medical Imaging*, vol. 22, no. 4, pp. 504–514, 2003.

[8] S. Kim, P. Smyth, and H. Stern, "Parametric response surface models for analysis of multi-site fMRI data," in *Proceedings of the 8th International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 2. Germany: Springer Verlag, 2005, pp. 217–224.

[9] B. Thirion, P. Pine, A. Tucholka, A. Roche, P. Ciuciu, J.-F. Mangin, and J.-B. Poline, "Structural analysis of fMRI data revisited: Improving the sensitivity and reliability of fMRI group studies," *IEEE Transactions on Medical Imaging*, vol. 26, no. 9, pp. 1256–69, 2007.

[10] G. Operto1, C. Clouchoux, R. Bulot, J. Anton, and O. Coulon, "Surface-based structural group analysis of fMRI data," in *Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Verlag, 2008, pp. 959–966.

[11] A. Uthama, R. Abugharbieh, S. J. Palmer, A. Traboulsee, and M. J. McKeown, "SPHARM-based spatial fMRI characterization with intersubject anatomical variability reduction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 907–18, 2008.

[12] B. Ng, R. Abugharbieh, X. Huang, and M. J. McKeown, "Spatial characterization of fMRI activation maps using invariant 3-D moment descriptors," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 261–68, 2009.

[13] B. Ng, R. Abugharbieh, G. Hamarneh, and M. J. McKeown, "Random walker based estimation and spatial analysis of probabilistic fmri activation maps," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention Workshop on Statistical modeling and detection issues in intra- and inter-subject functional MRI data analysis*, 2009, pp. 37–44.

[14] D. Bowman, B. Caffo, S. S. Bassett, and C. Kilts, "A Bayesian hierarchical framework for spatial modeling of fMRI data," *NeuroImage*, vol. 39, pp. 146–56, 2008.

[15] S. Kim and P. Smyth, "Hierarchical Dirichlet processes with random effects," in *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, 2007, pp. 697–704.

[16] L. Friedman, H. Stern, G. Brown, D. Mathalon, J. Turner, G. Glover, R. Gollub, J. Lauriello, K. Lim, T. Cannon, D. Greve, H. Bockholt, A. Belger, B. Mueller, M. Doty, J. He, W. Wells, P. Smyth, S. Pieper, S. Kim, M. Kubicki, M. Vangel, and S. Potkin, "Test-retest and between-site reliability in a multicenter fMRI study," *Human Brain Mapping*, vol. 29, no. 8, pp. 958–972, 2008.

[17] C. Rasmussen and Z. Ghahramani, "Infinite mixtures of Gaussian process experts," in *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2002, pp. 881–888.

[18] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[19] L. Xu, M. Jordan, and G. Hinton, "An alternative model for mixtures of experts," in *Advances in Neural Information Processing Systems 7*. Cambridge, MA: MIT Press, 1995, pp. 633–640.

[20] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[21] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of American Statistical Asslciation*, vol. 101, no. 476, pp. 1566–1581, 2006.

[22] N. Laird and J. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[23] S. Kim, P. Smyth, and H. Stern, "A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data," in *Proceedings of the 9th International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 2. Germany: Springer Verlag, 2006, pp. 217–224.

[24] E. Meeds and S. Osindero, "An alternative infinite mixture of Gaussian process experts," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2006, pp. 883–890.

[25] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley-Interscience, 2000.

[26] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[27] C. Rasmussen, "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press, 2000, pp. 554–560.

[28] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[29] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[30] D. Blackwell and J. MacQueen, "Ferguson distribution via Polya urn schemes," *Annals of Statistics*, vol. 1, pp. 353–355, 1973.

[31] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[32] S. Potkin, N. Andreasen, G. Brown, G. Glover, R. Kikinis, J. Lauriello, J. Lieberman, K. Lim, R. McCarley, G. McCarthy, B. Rosen, and A. Toga, "Multi-site brain fMRI imaging studies in schizophrenia using the BIRN methodology and federated database approach," in *The 41st Annual Meeting of the American College of Neuropsychopharmacology*, 2002, pp. 252–253.

[33] D. Veltman and C. Hutton, "SPM99," Wellcome Department of Imaging Neuroscience, University College London," Technical Report, 2000.

[34] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics and Computing*, vol. 10, no. 1, pp. 63–72, 2000.

[35] W. Ou and P. Golland, "From spatial regularization to anatomical priors in fMRI analysis," in *Proceedings of the 19th International Conference on Information Processing in Medical Imaging*, vol. 3565. Germany: Springer Verlag, 2005, pp. 88–100.