A Nonparametric Bayesian Approach to Detecting Spatial Activation Patterns in fMRI Data

Seyoung Kim, Padhraic Smyth, and Hal Stern

Bren School of Information and Computer Sciences
University of California, Irvine, CA 92697-3425
{sykim, smyth}@ics.uci.edu, sternh@uci.edu

Abstract. Traditional techniques for statistical fMRI analysis are often based on thresholding of individual voxel values or averaging voxel values over a region of interest. In this paper we present a mixture-based response-surface technique for extracting and characterizing spatial clusters of activation patterns from fMRI data. Each mixture component models a local cluster of activated voxels with a parametric surface function. A novel aspect of our approach is the use of Bayesian nonparametric methods to automatically select the number of activation clusters in an image. We describe an MCMC sampling method to estimate both parameters for shape features and the number of local activations at the same time, and illustrate the application of the algorithm to a number of different fMRI brain images.

1 Introduction

Functional magnetic resonance imaging (fMRI) is a widely used technique to study activation patterns in the brain while a subject is performing a task. Voxel-level activations collected in an fMRI session can often be summarized as a β map, a collection of β coefficients estimated by fitting a linear regression model to the time-series of each voxel. Detection of activation areas in the brain using β maps is typically based on summary statistics such as the mean activation values of particular brain regions, or detection of significant voxels by thresholding based on statistics such as p-values computed at each voxel. These approaches do not directly model spatial patterns of activation—detection and characterization of such patterns can in principle provide richer and more subtle information about cognitive activity and its variation across individuals and machines.

In earlier work on spatial activation patterns, Hartvig [1] 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. This work focused on extracting activated voxels by thresholding after the model parameters were estimated, rather than characterizing activation patterns directly. Penny and Friston [2] proposed a mixture model similar to that described in this

paper, with each mixture component representing a local activation cluster. In prior work we proposed a response surface model that represents an activation pattern as a superposition of Gaussian shaped parametric surfaces [3].

While the approaches proposed in [2] and [3] provide richer information about spatial activation than voxel-based methods, they both have the drawback that users have to determine the number of activation clusters manually by looking at individual images. While this may be feasible for analyzing relatively small brain regions for a small number of images, in a large scale analysis of brain activation patterns automatic detection of the number of bumps becomes an important pragmatic issue.

In this paper we take a nonparametric Bayesian approach and use Dirichlet processes [4] to address the problem of automatically determining the number of activation clusters. Rasmussen [5] illustrated how the Dirichlet process could be used as a prior on mixing proportions in mixture of Gaussian models to automatically determine the number of mixture components from data in Bayesian manner. This approach is sometimes referred to as the *infinite mixture model* since it does not a priori specify a finite number of mixture components but instead allows the number to be determined by the data.

The primary novel contribution of this paper is the application of the general framework of infinite mixtures and Bayesian inference to the specific problem of characterizing spatial activation patterns in fMRI. We model spatial activation patterns in fMRI as a mixture of experts with a constant background component and one or more activation components. An expert assigned to each "activation cluster" models each local activation cluster with a parametric surface model (similar to the surface model we proposed in [3] and detailed in the next section) with free parameters for the heights and center locations of the "bumps." Combining this mixture of experts model with a Dirichlet process prior we can estimate the shape parameters and the number of bumps at the same time in a statistical manner, using the infinite mixture model framework.

The paper is organized as follows. In Sect. 2, we introduce a mixture of experts model for spatial activation patterns, describe inference procedures for this model and extend it using an infinite mixture model to find the number of activation clusters automatically from data. In Sect. 3, we demonstrate the performance of our model on fMRI data for two individuals collected at two different sites. Sect. 4 concludes with a brief discussion on future work.

2 Activation Surface Modeling

2.1 The mixture of experts 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. Under the assumption 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

model the activation y_i at voxel $\mathbf{x}_i$ with a mixture of experts model:

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

where $\mathscr{C} = \{c_{bg}, c_m, m = 1, \ldots, M-1\}$ is a set of M expert component labels for a background component c_{bg} and $M-1$ activation components (the c_m's). The first term on the right hand side of Equation (1) defines the expert for a given component. We model the expert for an activation component as a Gaussian-shaped surface centered at $\mathbf{b_m}$ with width Σ_m and height k_m as follows.

$$y_i = k_m \exp\left(-(\mathbf{x}_i - \mathbf{b}_m)'(\Sigma_m)^{-1}(\mathbf{x}_i - \mathbf{b}_m)\right) + \varepsilon, \tag{2}$$

where ε is zero-mean additive noise. The background component is modeled as having a constant activation level μ with additive noise. We use the same Gaussian distribution with mean 0 and variance σ^2 for the noise term ε for both types of experts.

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

$$P(c|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|c)\pi_c}{\sum_{c \in \mathscr{C}} p(\mathbf{x}_i|c)\pi_c}, \tag{3}$$

where π_c is a class prior probability $P(c)$. $p(\mathbf{x}_i|c)$ is defined as follows. For activation components, $p(\mathbf{x}_i|c_m)$ is a normal density with mean $\mathbf{b}_m$ and covariance Σ_m. $\mathbf{b}_m$ and Σ_m are shared with the Gaussian surface model for experts in Equation (2). This implies that the probability of activating the mth expert is highest at the center of the activation and gradually decays as $\mathbf{x}_i$ moves away from the center. $p(\mathbf{x}_i|c_{bg})$ for the background component is modeled as having a uniform distribution of $1/N$ for all positions in the brain. If $\mathbf{x}_i$ is not close to the center of any activations, the gate function selects the background expert for the voxel.

Putting this model in a Bayesian framework we define priors on all of the parameters. The center of activation $\mathbf{b}_m$ is *a priori* uniformly distributed over voxels that have positive β values within a predefined brain region of interest. We use an inverse Wishart(ν_0, S_0) prior for the width parameter Σ_m for activation, and a Gamma(a_{0k}, b_{0k}) prior for the height parameter k_m. A normal distribution $N(0, {\sigma_0}^2)$ is used as a prior on the background mean μ. σ^2 for noise is given a Gamma(a_0, b_0) prior. A Dirichlet distribution with a fixed concentration parameter α is used as a prior for class prior probabilities π_c's.

2.2 MCMC for mixture of experts model

Because of the nonlinearity of the model described in Sect. 2.1, we rely on MCMC simulation methods to obtain samples from the posterior probability

density of parameters given 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 $p(\boldsymbol{\theta}, \mathbf{c}|\mathbf{y}, \mathbf{X})$ where $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{c}$ represent a collection of y_i's, $\mathbf{x}_i$'s and c_i's for $i = 1, \ldots, N$ and $\boldsymbol{\theta} = \{\mu, \sigma^2, \{\mathbf{b}_m, \Sigma_m, k_m\}, m = 1, \ldots, M-1\}$. Notice that the mixing proportions π_c's are not included in $\boldsymbol{\theta}$ and dealt with separately later in this section. During each sampling iteration the parameters $\boldsymbol{\theta}$ and component labels $\mathbf{c}$ are sampled alternately.

To obtain samples for parameters $\boldsymbol{\theta}$, we consider each parameter θ_j in turn and sample from

$$p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{c}, \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{c}, \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\mathbf{c}, \boldsymbol{\theta}) p(\theta_j),$$

where the subscript $-j$ indicates all parameters except for θ_j. Gibbs sampling is used for the background mean μ. The width parameter Σ_m for activation can be sampled using the Metropolis-Hastings algorithm with an inverse Wishart distribution as a proposal distribution. For all other parameters the Metropolis algorithm with a Gaussian proposal distribution can be used.

Given a Dirichlet$(\alpha/M, \ldots, \alpha/M)$ prior for the π_c's we can integrate out the π_c's to obtain

$$p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j} + \alpha/M}{N - 1 + \alpha}, \tag{4}$$

where $n_{-i,j}$ indicates the number of observations excluding y_i that are associated with component j [5]. This is combined with the likelihood terms to obtain the conditional posterior for c_i:

$$p(c_i|\mathbf{c}_{-i}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}, \alpha) \propto p(y_i|c_i, \mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{x}_i|c_i, \boldsymbol{\theta}, \alpha) P(c_i|\mathbf{c}_{-i}, \alpha).$$

We can sample the component label c_i for observation y_i from this distribution.

2.3 The infinite mixture of experts model

In the previous sections, we assumed that the number of components M was fixed and known. For an infinite mixture model, with a Dirichlet process prior [5], in the limit as $M \to \infty$ the class conditional probabilities shown in Equation (4) become

$$\text{components where } n_{-i,j} > 0: \qquad P(c_i = j|c_{-i}, \alpha) = \frac{n_{-i,j}}{N - 1 + \alpha} \tag{5}$$

$$\text{all other components combined: } P(c_i \neq c_{i'} \text{ for all } i' \neq i|c_{-i}, \alpha) = \frac{\alpha}{N - 1 + \alpha} \tag{6}$$

When we sample component label c_i for observation y_i in a given iteration of MCMC sampling, if there are any observations associated with component j other than y_i, Equation (5) is used and this component has a non-zero prior probability of being selected as c_i. If y_i is the only observation associated with

label j, this component is considered as unrepresented and a new component and its parameters are generated based on Equation (6) and prior distributions for the parameters. Once all observations are associated with components, the parameters of these components can be sampled in the same way as in the finite mixture of experts model assuming M is the number of components represented in the current sample of $\mathbf{c}$.

The class conditional prior probabilities in the infinite mixture model above (Equations (5) and (6)) are not dependent on the input positions $\mathbf{x}_i$, whereas the gate function for $P(c|\mathbf{x}_i)$ is a function of $\mathbf{x}_i$. To allow for dependence on the gate function we separate the input independent term from the gate function by writing it as in Equation (3), and apply the infinite mixture model to the second term of Equation (3).

To sample from the class conditional posterior, we use Algorithm 7 from Neal [6] that combines a Metropolis-Hastings algorithm with partial Gibbs sampling. In the Metropolis-Hastings step, for each c_i if it is not a singleton, we propose a new component with parameters drawn from the prior and decide whether to accept it or not based on $p(y_i|c_i, \mathbf{x}_i, \boldsymbol{\theta})p(\mathbf{x}_i|c_i, \boldsymbol{\theta})$. If it is a singleton, we consider changing it to other classes represented in the data. This Metropolis-Hastings algorithm is followed by partial Gibbs sampling for labels to improve efficiency.

At the start of the MCMC sampling the model is initialized to one background component and one or more activation components. Because we are interested in deciding the number of activation components, whenever a decision to generate a new component is made, we assume the new component is an activation component and sample appropriate parameters from their priors. If the number of voxels associated with a component at a given iteration is 0, the component is removed. It is reasonable to assume that not all regions of the brain are activated during a task and, thus, that there will always be some voxels in the background class. To prevent the sampling algorithm from removing the background component, the algorithm assigns the n_{labeled} voxels with the lowest β values in the image to the background component, and does partially supervised learning with these labeled voxels. In this case we are interested in sampling from the joint posterior of the parameters and the labels for unlabeled data $p(\mathbf{c}_{\text{unlabeled}}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{c}_{\text{labeled}})$.

In general the number of components depends on the concentration parameter α of the Dirichlet process prior. It is possible to sample α from the posterior using a gamma distribution as a prior on α [7]. In our experiments, we found fixing α to a small value ($\alpha < 5$) worked well.

3 Experimental Results

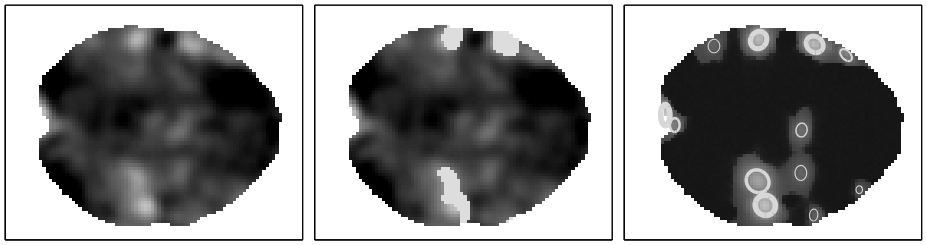
We demonstrate the performance of our algorithm using fMRI data collected from two subjects (referred to as Subjects 1 and 2) performing the same sensorimotor task. fMRI data was collected for each subject at multiple sites as part of a large multi-site fMRI study[1] (here we look at data from the Stanford (3T)

[1] http://www.nbirn.net

and Duke (4T) sites). Each run of the sensorimotor task produces a series of 85 scans that can be thought of as a large time-series of voxel images. The set of scans for each run is preprocessed in a standard manner using SPM99 with the default settings. The preprocessing steps include correction of head motion, normalization to a common brain shape (SPM EPI canonical template), and spatial smoothing 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 to produce β maps. 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.

The hyperparameters for prior distributions are set based on prior knowledge on local activation clusters. σ_0^2, for the prior on the background mean, is set to 0.1 and the noise parameter σ^2 is given a prior distribution Gamma(1.01, 1) so that the mode is at 0.01 and the variance is 1. The prior for height k_m of an activation component is set to Gamma(2,2) based on the fact that maximum β values in this data are usually around 1. The width of an activation component Σ_m is given an inverse-Wishart prior with degree of freedom 4. The 2×2 scale matrix is set to variance 8 and covariance 1. The concentration parameter α of the Dirichlet process prior is set to 1.

To initialize the model for posterior sampling, we use a heuristic algorithm to find candidate voxels for local activation centers and assign a mixture component to each of the candidates. To find candidate voxels, we take all of the positive voxels of a cross section, and repeatedly select the largest voxel among the voxels that have not been chosen and that are at least four 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. The width parameters are set to the mean of the prior on width. As mentioned earlier we fix the labels of $n_{\text{labeled}} = 10$ voxels with the lowest β values as



(a) (b) (c)

Fig. 1. Results for subject 2, data from Stanford MRI machine: (a) raw data (β maps) for a cross section ($z = 48$) (b) SPM showing active voxels colored white ($p \leq 0.05$) (c) Predictive values for activation given the estimated mixture components with height $k_m > 0.1$. The width of the ellipse for each bump is 1 standard deviation of the width parameter for that component. The thickness of ellipses indicates the estimated height k_m of the bump.

background, and perform partially supervised learning with these labeled voxels.

The MCMC sampler is run for 3000 iterations. The heights of the estimated bumps range between 0.003 and 1.2. Since low heights are likely to correspond to weak background fluctuations we display only those activation components above a certain threshold (height $k_m \geq 0.1$).

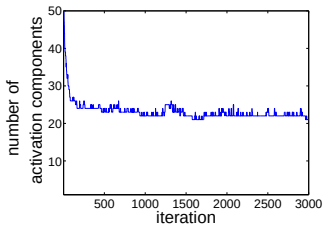


Fig. 2. The number of activation components as a function of MCMC iterations for the data in Fig. 1.

We fit the infinite mixture of experts model to a β map cross section (at $z = 48$) for subject 2 on the Stanford MRI machine, and summarize the results in Fig. 1. After 3000 iterations of the sampling algorithm using the β map shown in Fig. 1(a), we sample posterior predictive values for β values at each voxel using the last 1000 samples of the parameters. The medians of these samples are shown in Fig. 1(c) as predicted β values. Parameters from a single sample are overlaid as ellipses centered around the estimated activation centers $\mathbf{b}_m$. The model was able to find all of the significant clusters of activation in the raw data even though the number of activation components was unknown a priori to the learning algorithm. Most of the larger activation clusters are in the auditory area in the upper and in the lower middle part of the images, consistent with the activation pattern of sensorimotor tasks. For comparison, in Fig. 1(b) we show the activated voxels found by thresholding the z map (normalized beta-map) with $p \leq 0.05$. We can see that the thresholding method cannot separate the two bumps in the lower middle part of Fig. 1(a), and it completely misses the activations in the center of the image.

In Fig. 2 we assess the convergence of the Markov chain based on the number of active components at each iteration. The initially large number of components (from the heuristic initialization algorithm) quickly decreases over the first 300 iterations and after around 900 iterations stabilizes at around 22 to 25 components.

Fig. 3 shows results for a cross section (at $z = 53$) of the right precentral gyrus area for subject 1 collected over two visits. The β maps are shown on

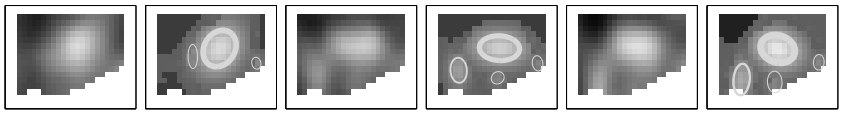


Fig. 3. Results from subject 1 at Duke. (a) Visit 1, run 1, (b) visit 2, run 1, and (c) visit 2, run 2. β maps for a cross section ($z = 53$) of right precentral gyrus and surrounding area are shown on the left. On the right are shown predictive values for activation given the mixture components estimated from the images on the left. The width of the ellipse for each bump is 1 standard deviation of the width parameter for that component. The thickness of ellipses indicates the estimated height k_m of the bump.

the left (Fig. 3(a) for visit 1, and Fig. 3(b)(c) for two runs in visit 2) and the estimated components on the right given the images on the left. Even though the β maps in Fig. 3(a)-(c) were collected from the same subject using the same fMRI machine there is variability in activation across visits such as the bump on the lower left of the β map for visit 2 in addition to the bigger bump at the center of β maps common to both visits. This information is successfully captured in the estimated activation components.

4 Conclusions

We have shown that infinite mixtures of experts can be used to locate local clusters of activated voxels in fMRI data and to model the spatial shape of each cluster, without assuming a priori how many local clusters of activation are present. Once the clusters are identified the characteristics of spatial activation patterns (shape, intensity, relative location) can be extracted directly and automatically. This can in turn provide a basis for systematic quantitative comparison of activation patterns in images collected from the same subject over time, from multiple subjects, and from multiple sites.

Acknowledgments The authors acknowledge the support of the following grants: the Functional Imaging Research in Schizophrenia Testbed, Biomedical Informatics Research Network (FIRST BIRN; 1 U24 RR021992, www.nbirn.net); 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. Author PS was also supported in part by the National Science Foundation under awards number IIS-0431085 and number SCI-0225642.

References

1. Hartvig, N. (1999) A stochastic geometry model for fMRI data. Research Report 410, Department of Theoretical Statistics, University of Aarhus.
2. Penny, W. & Friston, K. (2003) Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging,* **22**(4):504-514.
3. Kim, S., Smyth, P., Stern, H. & Turner, J. (2005) Parametric Response Surface Models for Analysis of Multi-Site fMRI data Mixtures of general linear. In *Proceedings of the 8th International Conference on Medical Image Computing and Computer Assisted Intervention.*
4. Furguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.
5. Rasmussen, C.E. (2000) The infinite Gaussian mixture model. In S.A. Solla, T.K. Leen and K.-R. Muller (eds.), *Advances in Neural Information Processing Systems 12*, pp. 554-560. Cambridge, MA: MIT Press.
6. Neal, R.M. (1998) Markov chain sampling methods for Dirichlet process mixture models. Technical Report 4915, Department of Statistics, University of Toronto.
7. Escobar, M. & West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association,* 90:577–588.