Monte Carlo Expectation Maximization with Hidden Markov Models to Detect Functional Networks in Resting-State fMRI

Wei Liu[1], Suyash P. Awate[1], Jeffrey S. Anderson[2], Deborah Yurgelun-Todd[3], and P. Thomas Fletcher[1]

[1] Scientific Computing and Imaging Institute, University of Utah, USA
[2] Department of Radiology, University of Utah, USA
[3] Department of Psychiatry, University of Utah, USA

Abstract. We propose a novel Bayesian framework for partitioning the cortex into distinct functional networks based on resting-state fMRI. Spatial coherence within the network clusters is modeled using a hidden Markov random field prior. The normalized time-series data, which lie on a high-dimensional sphere, are modeled with a mixture of von Mises-Fisher distributions. To estimate the parameters of this model, we maximize the posterior using a Monte Carlo expectation maximization (MCEM) algorithm in which the intractable expectation over all possible labelings is approximated using Monte Carlo integration. We show that MCEM solutions on synthetic data are superior to those computed using a mode approximation of the expectation step. Finally, we demonstrate on real fMRI data that our method is able to identify visual, motor, salience, and default mode networks with considerable consistency between subjects.

1 Introduction

Resting-state functional magnetic resonance imaging (fMRI) measures background fluctuations in the blood oxygen level-dependent (BOLD) signal of the brain at rest. The temporal correlations between these signals are used to estimate the functional connectivity of different brain regions. This technique has shown promise as a clinical research tool to describe functional abnormalities in Alzheimer's disease, schizophrenia, and autism [4]. In resting-state fMRI, a standard analysis procedure is to select a region of interest (ROI), or seed region, and find the correlation between the average signal of the ROI and other voxels in the brain. These correlations are thresholded so that only those voxels with significant correlations with the seed region are shown. Recent methods to find functional networks without seed regions include independent component analysis (ICA) [2], which often includes an ad hoc method to manually choose those components that are anatomically meaningful. Other approaches employ clustering techniques to automatically partition the brain into functional networks. A similarity metric is defined, e.g., correlation [5] or frequency coherence [9], and then a clustering method such as k-means or spectral clustering is used to group

K. Suzuki et al. (Eds.): MLMI 2011, LNCS 7009, pp. 59–66, 2011.

voxels with similar time series. A drawback of these approaches is that they disregard the spatial position of voxels, and thus ignore the fact that functional networks are organized into sets of spatially coherent regions.

We introduce a new data-driven method to partition the brain into networks of functionally-related regions from resting-state fMRI. The proposed algorithm does not require specification of a seed, and there is no ad hoc thresholding or parameter selection. We make a natural assumption that functionally homogeneous regions should be spatially coherent. Our method incorporates spatial information through a Markov random field (MRF) prior on voxel labels, which models the tendency of spatially-nearby voxels to be within the same functional network. Each time series is first normalized to zero mean and unit norm, which results in data lying on a high-dimensional unit sphere. We then model the normalized time-series data as a mixture of von Mises-Fisher (vMF) distributions [1]. Each component of the mixture model corresponds to the distribution of time series from one functional network. Solving for the parameters in this combinatorial model is intractable, and we therefore use a stochastic method called Monte Carlo Expectation Maximization (MCEM), which approximates the expectation step using Monte Carlo integration. The stochastic property of MCEM makes it possible to explore a large solution space, and it performs better than a standard mode approximation method using iterated conditional modes (ICM).

The proposed method in this paper is related to previous approaches using MRFs to model spatial relationships in fMRI data. Descombes et al. [3] use a spatio-temporal MRF to analyze task-activation fMRI data. Liu et al. [7] use an MRF model of resting state fMRI to estimate pairwise voxel connections. However, neither of these approaches tackle the problem of clustering resting-state fMRI into functional networks.

2 Hidden Markov Models of Functional Networks

We use a Bayesian statistical framework to identify functional networks of the gray matter in fMRI data. We formulate a generative model, which first generates a spatial configuration of functional networks in the brain, followed by an fMRI time series for each voxel based on its network membership. We employ an MRF prior to model network configurations, represented via unknown, or *hidden*, labels. Given a label, we assume that the fMRI time series, normalized to zero mean and unit norm, are drawn from a von Mises-Fisher likelihood.

Let $\mathcal{S} = \{1, \dots, N\}$ be the set of indices for all gray-matter voxels. We assume that the number of networks L is a known free parameter. Let $\mathcal{L} = \{1, 2, \cdots, L\}$ be the set of labels, one for each network. We denote a label map for functionally-connected networks as a vector $\mathbf{z} = (z_1, \dots, z_N), z_i \in \mathcal{L}$. Let $\mathcal{Z} = \mathcal{L}^N$ be the set of all possible $\mathbf{z}$ configurations.

2.1 Markov Prior Model

Functional networks should consist of few, reasonably-sized, possibly distant regions. We model such networks $\mathbf{z}$ using the *Potts* MRF [6]:

$$P(\mathbf{z}) = \frac{1}{C} \exp\left(-\beta \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j) \right),$$

where T is 1 if its argument is true and 0 otherwise; $\mathcal{N}_i$ is the set of neighbors of i as defined by the neighborhood system underlying the MRF; $\beta > 0$ is a model parameter controlling label-map smoothness; C is a normalization constant that is the sum of $P(\mathbf{z})$ over all possible configuration of $\mathbf{z}$. The Markov-Gibbs equivalence [6] implies that the conditional distribution of z_i at site i is:

$$P(z_i|\mathbf{z}_{-i}) = P(z_i|z_{\mathcal{N}_i}) = \frac{\exp\left(-\beta \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j)\right)}{\sum_{l \in \mathcal{L}} \exp\left(-\beta \sum_{j \in \mathcal{N}_i} T(l \neq z_j)\right)}, \tag{1}$$

where $\mathbf{z}_{-i}$ is the collection of all variables in $\mathbf{z}$ excluding site i. The neighborhood is the usual 6 adjacent voxels, which does not overly smooth across boundaries. Previous works [10,3] have demonstrated the advantages of MRF's over Gaussian smoothing in preserving segment boundaries.

2.2 Likelihood Model

To make the analysis robust to shifts or scalings of the data, we normalize the time series at each voxel to zero mean and unit length. This results in the data being projected onto a high-dimensional unit sphere. After normalization, the sample correlation between two time series is equal to their inner product, or equivalently, the cosine of the geodesic distance between these two points on the sphere. Thus, we re-formulate the problem of finding clusters of voxels with high correlations to the problem of finding clusters with small within-cluster distances on the sphere.

We use the notation $\mathbf{x} = \{(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) \,|\, \boldsymbol{x}_i \in S^{p-1}\}$ to denote the set of *normalized* time series. Observe that given $\mathbf{z} \in \mathcal{Z}$, the random vectors $\boldsymbol{x}_i$ are conditional independent. Thus, the likelihood $\log P(\mathbf{x}|\mathbf{z}) = \sum_{i \in \mathcal{S}} \log P(\boldsymbol{x}_i|z_i)$. We model the emission function $P(\boldsymbol{x}_i|z_i)$ using the von Mises-Fisher distribution

$$f(\boldsymbol{x}_i; \boldsymbol{\mu}_l, \kappa_l | z_i = l) = C_p(\kappa_l) \exp\left(\kappa_l \boldsymbol{\mu}_l^T \boldsymbol{x}_i\right), \quad \boldsymbol{x}_i \in S^{p-1}, \quad l \in \mathcal{L} \tag{2}$$

where, for the cluster labeled l, $\boldsymbol{\mu}_l$ is the mean direction, $\kappa_l \geq 0$ is the *concentration parameter*, and the normalization constant $C_p(\kappa) = \kappa^{\frac{p}{2}-1} / \{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)\}$, where I_ν denotes the modified Bessel function of the first kind with order ν. The larger the κ, the greater is the density concentrated around the mean direction. Since (2) depends on $\boldsymbol{x}$ only by $\boldsymbol{\mu}^T \boldsymbol{x}$, the vMF distribution is unimodal and rotationally symmetric around $\boldsymbol{\mu}$.

In the Bayesian framework, we also define distributions on parameters. We assume that $\forall l \in \mathcal{L}$, $\kappa_l \sim \mathcal{N}(\mu_\kappa, \sigma_\kappa^2)$ with hyperparameters μ_κ and σ_κ^2 that can be set empirically. This prior enforces constraints that the clusters should not have extremely high or low concentration parameters. We empirically tune the hyperparameters μ_κ and σ_κ^2 and have found the results to be robust to specific choices of the hyperparameters.

3 Monte Carlo EM

To estimate the model parameters and the hidden labels, we use a stochastic variant of expectation maximization (EM) called Monte Carlo EM (MCEM) [10]. The standard EM algorithm maximizes the expectation of the log-likelihood of joint pdf of $\mathbf{x}$ and the hidden variable $\mathbf{z}$ with respect to the posterior probability $P(\mathbf{z}|\mathbf{x})$, i.e. $\mathbb{E}_{P(\mathbf{z}|\mathbf{x})}[\log P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})]$. The combinatorial number of configurations for $\mathbf{z}$ makes this expectation intractable. Thus, we use Monte Carlo simulation to approximate this expectation as

$$\widetilde{Q}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \approx \frac{1}{M} \sum_{m=1}^{M} \log P(\mathbf{z}^m; \beta) + \log P(\mathbf{x}|\mathbf{z}^m; \boldsymbol{\theta}_L), \tag{3}$$

where $\mathbf{z}^m$ is a sample from $P(\mathbf{z}|\mathbf{x})$, $\boldsymbol{\theta}_L = \{\boldsymbol{\mu}_l, \kappa_l : l \in \mathcal{L}\}$ is the parameter vector of the likelihood, and $\boldsymbol{\theta} = \{\beta, \boldsymbol{\theta}_L\}$ is the full parameter vector of the model. Computing the MRF prior in (3) is still intractable due to the normalization constant, and we instead use a pseudo-likelihood approximation [6], which gives

$$\widetilde{Q} \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{i \in \mathcal{S}} \log P(z_i | z_{\mathcal{N}_i}; \beta) + \frac{1}{M} \sum_{m=1}^{M} \sum_{i \in \mathcal{S}} \log P(\boldsymbol{x}_i | z_i; \boldsymbol{\theta}_L) = \widetilde{Q}_P + \widetilde{Q}_L.$$

We use $\widetilde{Q}_P$ to denote the log-pseudo-likelihood of the prior distribution, and use $\widetilde{Q}_L$ to denote the log-likelihood distribution.

3.1 Sampling from the Posterior

Given the observed data $\mathbf{x}$ and parameter value $\boldsymbol{\theta} = \{\beta, \boldsymbol{\theta}_L\}$, we sample from the posterior distribution $P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$ using Metropolis sampling. We define the posterior energy, which is to be minimized, as the negative log of the posterior $P(z_i|\boldsymbol{x}_i)$. Thus, Bayes rule implies:

$$U(z_i = l|\mathbf{x}) = \beta \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j) - \log C_p(\kappa_l) - \kappa_l \boldsymbol{\mu}_l^T \boldsymbol{x}_i + \text{const}, \tag{4}$$

which is the sum of the prior energy, the conditional energy, and a parameter-independent quantity. Then, given a current configuration $\mathbf{z}^n$ Metropolis sampling generates a new candidate label map $\mathbf{w}$ as follows: (i) draw a new label l' at site i with uniform distribution; $\mathbf{w}$ has value l' at site i, with other sites remaining the same as $\mathbf{z}^n$; (ii) compute the change of energy $\Delta U(\mathbf{w}) = U(\mathbf{w}|\mathbf{x}) - U(\mathbf{z}^n|\mathbf{x}) = U(z_i = l'|\mathbf{x}) - U(z_i = l|\mathbf{x})$; (iii) accept candidate $\mathbf{w}$ as $\mathbf{z}^{n+1}$ with probability $\min(1, \exp\{-\Delta U(\mathbf{w})\})$; (iv) after a sufficiently long burn-in period, generate a sample of size M from the posterior distribution $P(\mathbf{z}|\mathbf{x})$.

3.2 Parameter Estimation

Estimating $\boldsymbol{\theta}_L$: By maximizing $\widetilde{Q}_L$ with the constraint $\|\boldsymbol{\mu}_l\| = 1$, we get

$$R_l = \sum_{m=1}^{M} \sum_{i \in \mathcal{S}_l} \boldsymbol{x}_i, \qquad \hat{\boldsymbol{\mu}}_l = \frac{R_l}{\|R_l\|}, \tag{5}$$

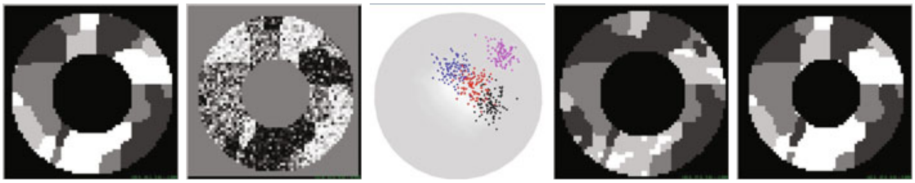

Fig. 1. Synthetic example. From left to right: true labels, first time point of observed time series, time series plot on sphere, label map estimated by mode-approximation, and label map estimated by MCEM.

where $S_l = \{i \in \mathcal{S} : z_i = l\}$ is the set of data points in cluster l. We have no *a priori* knowledge for $\boldsymbol{\mu}_l$, so a maximum likelihood estimation in (5) is the best we can do. For κ_l we maximize the posterior distribution $P(\kappa_l|\mathbf{x}, \mathbf{z}^1, \ldots, \mathbf{z}^M)$. Since $\widetilde{Q}_L$ is not dependent on κ, we maximize $\widetilde{Q}_L(\kappa_l) + \log P(\kappa_l; \mu_\kappa, \sigma_\kappa^2)$ and get

$$A_p(\hat{\kappa}_l) + \frac{\hat{\kappa}_l - \mu_\kappa}{N_l \sigma_\kappa^2} = R_l, \qquad (6)$$

where $A_p(\hat{\kappa}_l) = I_{\frac{p}{2}}(\hat{\kappa}_l)/I_{\frac{p}{2}-1}(\hat{\kappa}_l)$ and $N_l = |\mathcal{S}_l|$ is the number of data points in cluster l. Because (6) contains the ratio of two modified Bessel functions, an analytic solution is unavailable and we have to resort to a numerical solution. We use Newton's method for solving $g(\hat{\kappa}_l) = A_p(\hat{\kappa}_l) - (\hat{\kappa}_l - \mu_\kappa)/(N_l \sigma_\kappa^2) - R_l = 0$. The choice of initial value for Newton's algorithm depends on the strength of the prior on κ_l (i.e. the σ_κ value). For a noninformative prior, $\hat{\kappa}_l = (pR_l - R^3)/(1 - R^2)$ is a good a good initial value [1]. For a strong prior, a reasonable initial value is the current value of κ_l.

Estimating β: To estimate β, we again rely on Newton's method to find the solution numerically. The derivatives $\partial \widetilde{Q}_P/\partial\beta$ and $\partial^2 \widetilde{Q}_P/\partial\beta^2$ for the pseudo-likelihood approximation of the MRF prior are easily computed.

3.3 MCEM-Based Algorithm for Hidden-MRF Model Estimation

Given the methods for sampling and parameter estimation, we estimated the hidden-MRF model by iteratively using (i) MCEM to learn model parameters and (ii) using ICM to compute optimal network labels. In the expectation (E) step, we draw samples from the posterior $P(\mathbf{z}|\mathbf{x})$, given current estimates for parameters $\boldsymbol{\theta}$. In the maximization (M) step, we use these samples to update estimates for the parameters $\boldsymbol{\theta}$.

4 Results and Conclusion

Synthetic example: We first simulate low-dimensional time series (2D 64×64 image domain; 3 timepoints, for visualization on sphere S^2) to compare the (i) proposed method using MCEM with (ii) the mode-approximation approach that replaces the E step in EM with a mode approximation. We simulate a label map by sampling from a MRF having $\beta = 2$. Given the label map, we simulate vMF samples (on the sphere S^2). Figure 1 shows that the MCEM solution is close to the ground truth, while the mode-approximation solution is stuck in a local maximum.

Resting-State fMRI: We evaluated the proposed method on real data, from healthy control subjects, in a resting-state fMRI study. BOLD EPI images (TR = 2.0 s, TE = 28 ms, 40 slices at 3 mm slice thickness, 64 x 64 matrix, 240 volumes) were acquired on a Siemens 3 Tesla Trio scanner. The data was preprocessed in SPM, including motion correction, registration to T2 and T1 structural MR images, spatial smoothing by a Gaussian filter, and masked to include only the gray-matter voxels. We used the `conn` software [11] to regress out signals from the ventricles and white matter, which have a high degree of physiological artifacts. A bandpass filter was used to remove frequency components below 0.01 Hz and above 0.1 Hz. We then projected the data onto the unit sphere by subtracting the mean of each time series and dividing by the magnitude of the resulting time series. We then applied the proposed method to estimate the functional network labels with the number of clusters set to $L = 8$.

Figure 2 shows the optimal label maps, produced by the proposed method for 3 of all 16 subjects in the dataset. We note that among the 8 clusters, one cluster, with the largest κ value and largest number of voxels, corresponds to background regions with weakest connectivity and is not shown in the figure. Among the clusters shown, we can identify the visual, motor, dorsal attention, executive control, salience, and default mode networks (DMN) [8]. Four networks: the visual, motor, executive control, and DMN, were robustly found across all subjects. More variability was found in the dorsal attention network (notice that it is much larger in subject 3) and salience network (notice that it is missing in subject 2). We found that changing the number of clusters, although leading to different label maps, preserves the four robust networks. For instance, we also ran the analysis with the number of clusters set to 4 or 6 (results not shown) and were able to recover the same four robust networks.

The next experiment compares our results with ICA. A standard ICA toolbox (GIFT; `mialab.mrn.org`) was applied on the same preprocessed data of each subject independently, which we call "Individual ICA". We also applied standard Group ICA, using all data from the 16 subjects simultaneously. In both ICA experiments the number of components are set to 16. The component maps are converted to z score and thresholded at 1. For each method we computed an overlap map for each functional network by adding the corresponding binary

Table 1. The number of voxels with value greater than 8 in the overlapped label map

	DMN	Motor	Attention	Visual
MCEM	5043	7003	3731	5844
Individual ICA	114	167	228	134
Group ICA	3075	5314	3901	3509

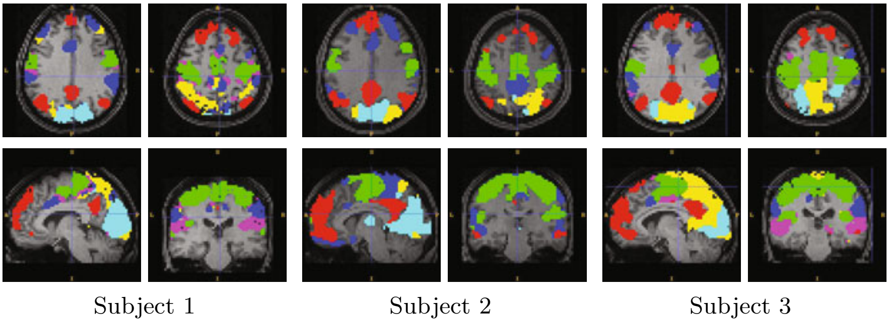

Fig. 2. Functional networks detected by the proposed method for 3 subjects overlaid on their T1 images. The clusters are the visual (cyan), motor (green), executive control (blue), salience (magenta), dorsal attention (yellow), and default mode (red) networks.

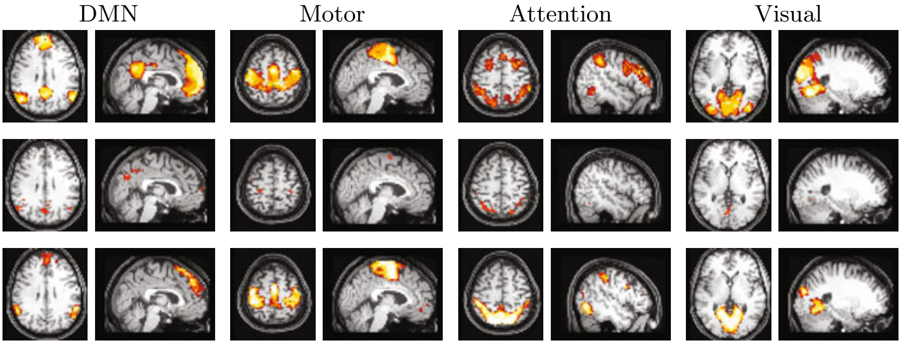

Fig. 3. Comparison of overlap label maps by our approach and ICA for 16 subjects. Top: our MCEM approach, middle: Individual ICA, bottom: Group ICA. Color ranges from 8 (red) to 16 (yellow).

label maps of all 16 subjects. The results in Figure 3 show our method can detect the motor, attention, and visual network with accuracy comparable with Group ICA. Besides, our method also detects DMN with posterior cingulate cortex (PCC) and medial prefrontal cortex (MPFC), while Group ICA split the DMN into two components, one with the MPFC and another with the PCC (not shown).

To see the consistency of the label map between subjects for all three methods, we look at each method's overlapped label map and count the number of voxels whose value are greater than 8. Table 1 shows that our method exhibits better consistency than both Individual and Group ICA.

Conclusion: We present a novel Bayesian approach to detect functional networks of the brain from resting-state fMRI that incorporates an MRF for spatial regularization, and we use MCEM to approximate the maximum posterior solution. Future work will include extending the model to group analysis, which can be achieved using a hierarchical Bayesian model. We also plan to investigate the use of non-parametric Bayesian methods to estimate the number of clusters.

Acknowledgments. This work was funded in part by NIH Grant R01 DA020269 (Yurgelun-Todd).

References

1. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions. J. Machine Learning Res. 6(2), 1345 (2006)
2. Beckmann, C., Smith, S.: Tensorial extensions of independent component analysis for multisubject fMRI analysis. Neuroimage 25(1), 294–311 (2005)
3. Descombes, X., Kruggel, F., Cramon, D.V.: Spatio-temporal fMRI analysis using Markov random fields. IEEE Trans. on Medical Imaging 17(6), 1028–1039 (1998)
4. Fox, M., Greicius, M.: Clinical Applications of Resting State Functional Connectivity. Frontiers in Systems Neuroscience 4 (2010)
5. Golland, P., Lashkari, D., Venkataraman, A.: Spatial patterns and functional profiles for discovering structure in fMRI data. In: 42nd Asilomar Conference on Signals, Systems and Computers, pp. 1402–1409 (2008)
6. Li, S.Z.: Markov random field modeling in computer vision. Springer, Heidelberg (1995)
7. Liu, W., Zhu, P., Anderson, J., Yurgelun-Todd, D., Fletcher, P.: Spatial regularization of functional connectivity using high-dimensional markov random fields. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 363–370. Springer, Heidelberg (2010)
8. Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L.: A default mode of brain function. PNAS 98(2), 676–682 (2001)
9. Thirion, B., Dodel, S., Poline, J.: Detection of signal synchronizations in resting-state fMRI datasets. Neuroimage 29(1), 321–327 (2006)
10. Wei, G., Tanner, M.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85(411), 699–704 (1990)
11. Whitfield-Gabrieli, S.: Conn Matlab toolbox (March 2011), `http://web.mit.edu/swg/software.htm`