Published in final edited form as:

Neuroimage. 2014 October 15; 100: 520–534. doi:10.1016/j.neuroimage.2014.06.001.

A Functional Networks Estimation Method of Resting-State fMRI Using a Hierarchical Markov Random Field

Wei Liu^a, Suyash P. Awate^a, Jeffrey S. Anderson^b, and P. Thomas Fletcher^a

Wei Liu: weiliu@sci.utah.edu; Suyash P. Awate: suyash@sci.utah.edu; Jeffrey S. Anderson: andersonjeffs@gmail.com; P. Thomas Fletcher: fletcher@sci.utah.edu

^aScientific Computing and Imaging Institute, University of Utah, USA

^bDepartment of Radiology, University of Utah, USA

Abstract

We propose a hierarchical Markov random field model that estimates both group and subject functional networks simultaneously. The model takes into account the within-subject spatial coherence as well as the between-subject consistency of the network label maps. The statistical dependency between group and subject networks acts as a regularization, which helps the network estimation on both layers. We use Gibbs sampling to approximate the posterior density of the network labels and Monte Carlo expectation maximization to estimate the model parameters. We compare our method with two alternative segmentation methods based on K-Means and normalized cuts, using synthetic and real fMRI data. The experimental results show our proposed model is able to identify both group and subject functional networks with higher accuracy, more robustness, and inter-session consistency.

Keywords

resting-state functional MRI; segmentation; functional connectivity; hierarchical Markov random field; Bayesian

1. Introduction

To study the brain's intrinsic activity with resting-state fMRI (rs-fMRI) data, one models either the data of a single subject or a group of subjects. A single subject's data are often contaminated with the noise of various sources, and its results are typically unreliable for an inference of the whole population. On the other hand, combining data from multiple subjects and jointly estimating the common functional networks is more robust. Group analysis of rs-fMRI assumes that all subjects in the group share common functional connectivity patterns and that these group networks can be estimated more accurately because the noise from each subject is canceled by averaging. In practice, it is a major challenge to summarize the consistent patterns across subjects, as each subject's network structure appears similar but has slight variations.

Recent years have seen substantial interest in estimating functional networks of individual subjects during group analysis. An accurate estimate of an individual's network is an important step towards the understanding of brain-behavior relationships on a per-subject basis. Therefore, the intersubject variation must be modeled explicitly for a good estimation of single subjects as well as the group networks. Current methods (Yeo et al., 2011; Damoiseaux et al., 2006) either do not estimate individual functional network maps, or do not have an explicit statistical model on the intersubject variations (Calhoun et al., 2001a,b). Among the methods that do estimate subject functional networks, some have one or more drawbacks.

First, some methods map a common group functional network by concatenating the blood-oxygen-level-dependent (BOLD) signal from all subjects. To do this, they have to assume that the voxels across subjects map to the same anatomical structure after coregistration and share the same functional connectivity patterns. These assumptions are often violated due to the anatomical inhomogeneity between subjects and to the imperfect alignment of the existing coregistration routine. In addition, the simple concatenation does not take into account the possible difference between subjects. In particular, some participants may experience spontaneous but active cognition during the scan even in the resting-state. These activities modulate each subject's functional network in a different way and to a different extent, and hence interfere the estimation of the group's functional network. Such subject-specific confounding factors are less likely to be negligible by simple averaging compared to other sources of noise such as scanner noise, subject motion and coregistration.

Second, group analyses are often conducted in a *one way* procedure. In some scenarios (Van Den Heuvel et al., 2008; Craddock et al., 2012; Greicius et al., 2004, 2007; Seeley et al., 2009; Mohammadi et al., 2009), each subject's functional network is estimated independently, and a group map is simply summarized by averaging the subjects' connectivity maps. The estimates of the subject's map by these procedures do not use other subjects' information and may not be robust to noise. The group summary map extracted from these subject maps is also suboptimal. In other scenarios (Calhoun et al., 2001a), a group map is estimated first from the concatenated data and then is back-reconstructed to obtain the subject network maps. More recently, the subject network maps have been estimated from the averaged group map using a dual regression approach (Filippini et al., 2009; Beckmann et al., 2009). Such methods treat voxel intensity from all subjects the same way as for group map estimation, ignoring that they may have subject-specific variances. Both classes of approaches do not iteratively refine the initial group or subject estimates, and the estimation of one subject's connectivity does not benefit from the information of other subjects. Figure 1 gives an illustration of the various methods and their order of estimation.

Last, spatial smoothing is often used during preprocessing in order to address the issue of imperfect intersubject alignment. While spatially blurring the time series raises the signal-to-noise ratio, the choice of the smoothing kernel size has a big impact on the estimated functional maps. Over-smoothing inevitably results in the loss of fine structures of the functional maps. One need a model that makes use of the spatial dependency of the rs-fMRI signals, without losing the finer details of the functional patterns.

A data-driven, unified probabilistic framework will help solve the above issues. The model should integrate both group and subject connectivity variables. One can make inferences from the posterior distribution of the variables in both subject and group levels given the observed BOLD signal.

In this work we propose a Bayesian hierarchical model for estimating the functional networks by using the rs-fMRI data from a group of subjects. The hierarchy comes from an additional level of *group* map defined on top of the conventional subject functional network maps. The group effect goes into the subject network label's probabilistic distribution as a parameter. Both group and subject networks are jointly estimated in an iterative way. We give a natural interpretation of the regularization with a Bayesian perspective. Once the group's network map is known, it can help the individual subject's estimation as a prior probability. Because the group map combines the information from all subject maps, this prior distribution is equivalent to using other subjects' data for the current subject's inference. Besides, a subject's network estimates also help to iteratively refine the group map estimates. We model the intersubject variability by balancing between a common map for all subjects (no variability, maximal shared information) and a separate map for each subject (no shared information, maximal variability). We achieve the optimal balance in the Bayesian sense by computing the subject network label's posterior density. This posterior density combines the prior information from the group map and the data likelihood from the subject-specific BOLD signal. We further model the within-subject spatial coherence by a Markov random field (MRF). In the remaining part of the paper, we refer to our model a *hierarchical Markov random field* (HMRF).

A classical occurrence of hierarchical modeling in fMRI is the inclusion of random effects in a general linear model (GLM) (Beckmann et al., 2003), which is later extended to a full Bayesian framework (Woolrich et al., 2004a). The multilevel model has richer structures and can capture the structures in multiplegroup, multiple-session data, and distinguish between the influence of the fixed effect and that of the random factors. In our model, the hierarchy is defined on a latent variable mixture representation.

A Markov random field is a multivariate distribution defined on an undirected graph to represent the soft constraints between the variables. In fMRI analysis, it is a principal regularization method of obtaining a spatially coherent solution. Depending on the context, previous works have defined MRF on different variables. They have been used for the regularization priors on the coefficients of the general linear model (GLM) (Penny et al., 2005), on the parameters of a spatio-temporal auto-regression model (Woolrich et al., 2004b), an on the hidden activation variables in task-based experiments (Hartvig and Jensen, 2000). In this article, we define MRF on the latent network label variables of hidden Markov model (HMM), to represent our prior knowledge of the spatial coherence of the network patterns within a subject. There is a key difference between our model and conventional HMMs, though. We generalize the conventional concept of spatial regularization by defining a *joint* graph that includes the network variables of both the group and subject levels. In our model, the neighbors of each node on the graph include the corresponding nodes at another level, as well as the spatially adjacent voxels in the same level. The new graph introduces our additional assumption that one subject's functional networks should

share similar patterns with another subject's, implicitly represented by the group. With this definition, we map all the variables in a hierarchical model on to a single graph, and formulate a problem conceptually appealing and feasible in practice.

The exact inference of MRF is a combinatorial optimization of discrete variables, hence it is computationally infeasible except in special cases (Greig et al., 1989; Ng et al., 2012). The iterated conditional mode (ICM) is typically used to obtain a local optimum of the mode of the posterior density (Besag, 1986). In this work we are interested in the posterior variance of the network label variables as well as the mode, and we use a Monte Carlo expectation maximization (MCEM) sampling algorithm for the inference of both group and subject label maps. MCEM is data-driven in that the model parameters are estimated together with the network label variables. The only parameter that needs special treatment is the link strength between the group and subjects. MCEM integrates the Markov chain Monte Carlo sampling in the expectation-maximization loop. The price to pay is the a longer computation time than other approximate inference methods such as variational Bayes.

We show our HMRF model is able to recover both group and subject functional networks in simulated group fMRI data. While HMRF's group estimates are comparable or more accurate than the two other methods under comparison, we are especially interested in the higher accuracy of the individual subjects' estimates. We further show the strength of the model by a real multiple-session dataset, where we achieve significantly higher intersession consistency by using our joint-estimation model. The method also proves to be more stable under the data perturbation in a bootstrap experiment. This paper is based on our earlier work Liu et al. (2012), and we extend previous work to redefine the model in an integrated graphical model context. The new simulated data experiments explore the performance of the algorithm under various levels of spatial smoothing. In the real data experiments, we added a new intersession consistency test and the algorithm stability test with bootstrapping. We also improved the parameter estimation by using the Bayesian posterior predictive distribution of the test subjects in a cross-validation framework.

In the remainder of the paper, we give some background information and related works in section 2, define the model in section 3, and give the approximate inference procedure in section 4. We compare the accuracy and consistency of our method, together with other methods by testing on synthetic and real data experiments in section 5 and section 6 and discuss the model and algorithm performance in section 7.

2. Related Works

Independent component analysis (ICA) is one powerful tool for identifying functional networks by using rs-fMRI of a single subject (Calhoun et al., 2001b). ICA is used to recover the statistically independent functional components without *a priori* knowledge of the regions of interest. Group ICA is used as an extension of single-subject ICA in order to seek a set of independent components shared across all subjects (Calhoun et al., 2001b). In a typical group ICA study all subjects are registered to a common atlas and assumed to share a common spatial component map but have distinct time courses. The BOLD signals from all subjects are concatenated temporally followed by a single-subject ICA analysis. The subject

component maps are then obtained by a back-reconstruction procedure. Alternatively single-subject ICA is applied on each subject first, and a self-organized clustering algorithm applies to all subjects' components such that similar components are assigned into one cluster. The group ICA components are represented by the centers of the clusters (Esposito et al., 2005). Neither of the above approaches iteratively refine group (or subject) maps once the subject (or group) maps are estimated.

ICA as a Signal decomposition method obtains overlapped spatial components and needs ad-hoc thresholding. Such ambiguity makes interpreting the results difficult. On the other hand, functional networks estimation can also be defined as an image segmentation problem. The region-of-interest (ROI), or even the whole brain voxels can be partitioned into disjoint spatial patches. The patches with the same network labels, even when spatially remote from each other, belong to the same functional networks. To extend the segmentation method to a group of subjects, segmentations are performed first on individual subjects. The connectivity maps are averaged to obtain a group affinity matrix. A second level segmentation is performed on this affinity matrix (Bellec et al., 2010; Van Den Heuvel et al., 2008). Again, subject network maps are not refined from the estimates of the group network.

It is worth noting that Ng et al. (2012) also uses MRF for group study. The spatial neighborhood is extended to cross-subject voxels, in order to address the imperfect anatomical alignment and functional correspondence. Our model is different from Ng et al.'s group MRF model in that 1) a group level is defined in our model, whereas in Ng et al.'s work, a combined structure including all subjects is defined without a group level. In such a flat model, a voxel directly uses the information of the corresponding voxels of other subjects. Instead, we expect a second level enriches the model and better decomposes the fixed and random effects in the subject network map. 2) Ng et al. defines the MRF prior on the GLM coefficients in task-based experiments so the posterior inference is a two-class problem (active versus inactive), and an exact solution can be obtained by a graph-cuts algorithm. Our model applies to the network labels in a rs-fMRI study, and hence is a multiclass segmentation problem that is significantly more difficult to solve. 3) The unary potential function in Ng et al.'s model is defined via the posterior probability of the label variable given the GLM's coefficients, and there is no explicit splitting between the MRF prior and the likelihood. This modeling method is essentially a conditional random field (Lafferty, 2001). In our model, the unary potential in the MRF prior is defined as the links between group and subject level, and thus is separated from the data likelihood, so our model is a variant of the hidden Markov model. Thanks to the Bayesian perspective of our hierarchical model, the trade-off between prior and likelihood is accounted for automatically. 4) Ng et al.'s MRF is built on certain regions of the brain, while ours is built on the whole brain's gray matter voxels.

Another class of methods identifies the functional spatial patterns by decomposing the BOLD signal, which can be seen as generalizations of ICA (Varoquaux et al., 2010, 2011) methods. The authors of both works introduce generative models that include a population level and subject level. Each subject's mixing weight coefficients are regarded as being generated from population level latent factors, and the population and subject level's mixing matrices are solved jointly as a convex optimization problem (Varoquaux et al., 2011). With

regard to the joint estimation of both levels of the hierarchy, we can see such methods as a counterpart of our model in the class of signal decomposition methods.

3. Hierarchical MRF For Modeling Group fMRI

We begin by defining each subject's network label map as a Markov random field (MRF) with the neighborhood structure given by a regular lattice. The statistical dependency between adjacent voxels acts as a prior model favoring spatial coherence of estimated functional regions. To generalize the MRF to a hierarchical setting, an additional group label map is defined in addition to all subject label maps. The group label map has the same number of voxels and the same Markov structure as the individuals' maps, again to encourage spatial coherence of the functional regions in the group level. In addition, each voxel in the group map is connected to the corresponding voxel in each subject map. These connections model the relationships between the group and the individuals. The subjects' functional network labels are regarded as generated from the group labels, and the rs-fMRI time courses are regarded to be generated from a mixture of high-dimensional distributions given the subject network labels. All voxels of subjects and group label map are jointly connected into a single MRF. The functional network estimation is the inverse problem of the above data generation process, as the labels are inferred from their posterior distribution given the data. See Figure 2 for an illustration.

More specifically, we define an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. The set of vertices $\mathbf{V} = (\mathbf{V}_G, \mathbf{V}_1, \dots, \mathbf{V}_J)$ is the union of the gray matter voxels $\mathbf{V}_j$ for all J subjects as well as those in the group volume $\mathbf{V}_G$. An edge $(s, t) \in \mathbf{E}$ is defined in one of three types: (1) $s \in \mathbf{V}_G, t \in \mathbf{V}_j$ and s, t have the same physical coordinates, (2) $s, t \in \mathbf{V}_G$, and s, t are spatial neighbors, or (3) $s, t \in \mathbf{V}_j$, and s, t are spatial neighbors. In our model we use a 26-neighbor system in a 3D volume image (a voxel at the boundary of the gray matter may have < 26 neighbors). We will refer to the first type of links as *between-level links*, and the second and third types of links as *within-subject links*. On each node $s \in \mathbf{V}$, a discrete random variable $y_s \in \mathbf{L} = \{1, \dots, L\}$ is defined to represent the functional network label. We use $-s$ for the set of nodes excluding node s , and $N(s)$ for the set of neighboring nodes of s . Last we define clique c as a complete subgraph of $\mathbf{G}$, such that every pair of nodes in c has a link between them, and define $\mathbf{C}$ the set of all cliques in $\mathbf{G}$.

3.1. MRF Prior

MRF is a principal regularization method for modeling spatial context information. In a Bayesian setting, we use it as a prior distribution of the network label variables $Y = \{y_s \in \mathbf{L} | s \in \mathbf{V}\}$. Formally, with the definition of the graph $\mathbf{G}$ and neighbor system $N(s), \forall s \in \mathbf{V}$ above, Y is said to be a MRF on $\mathbf{G}$ if $P(y_s | y_{-s}) = P(y_s | y_{N(s)})$, i.e. a variable is conditional independent of the variables on the remaining nodes of the graph given its neighbors (Li, 1995). This local *conditional independence* property is difficult to apply to the inference of the joint distribution. Thanks to the equivalence of MRF and Gibbs fields (Besag, 1974), one can transform the local property into a global property. A Gibbs random field or Gibbs distribution takes the form of $P(Y) = (1/Z) \exp\{-U(Y)\}$, where Z is a normalization constant called partition function in order to guarantee the function integrates to one, and $U(Y) =$

$\sum_{c \in C} V_c(Y)$ is called the energy function. Each clique potential function V_c only depends on the variables in the corresponding clique c . The Hammersley-Clifford theorem (Hammersley and Clifford, 1968) states that Y is a MRF if and only if it obeys a Gibbs distribution. In this specific problem, the energy function takes the following form:

$$U(Y) = \sum_{s,r \in \mathbf{V}_G} \beta \psi(y_s, y_r) + \sum_{j=1}^J \left(\sum_{s \in \mathbf{V}_G, \tilde{s} \in \mathbf{V}_j} \alpha \psi(y_s, y_{\tilde{s}}) + \sum_{s,r \in \mathbf{V}_j} \beta \psi(y_s, y_r) \right). \quad (1)$$

The binary function ψ takes zero if the two inputs are equal and takes one otherwise. Parameters α and β determine the strength of the links. The pair of voxels (s, r) is spatially adjacent within the subject volume or the group volume (the type 2 and type 3 links), and $(s, \tilde{s})$ is a pair of neighboring voxels at a different level in the hierarchy, but sharing the same physical coordinates (type 1 link).

This regularization encodes two physiologically meaningful *a priori* assumptions on the functional networks under investigation: (1) The networks are spatially coherent within the single subject map and within the group map. This spatial coherency is modeled by the β potential term. (2) The subject's intrinsic functional activity must share similar patterns, regardless of the possible confounding of the noise and subject-specific effect. This between-subject constraint is modeled by the α potential term. The proposed energy function represents both priors without introducing blurring artifacts. As for the inference, although appearing different in the image domain, the three types of links are no different when looking from the abstract graph layer, and can be treated equivalently in the inference procedure. Our MRF prior is essentially a Potts model with different weights defined on three types of edges (Potts, 1952). However, we extend the Potts model such that the cliques in a graph include both within-subject links and between-level links, so the model favors not only spatial coherence but also the intersubject coherence.

3.2. Likelihood Model

In the generative model, for any individual subject, the observed time course at each voxel is assumed to be generated from a distribution conditioned on the network label at that voxel. In fMRI analysis the BOLD signal is typically normalized to be zero mean and unit norm, so the analysis is invariant of shifting or scalings of the data (Golland et al., 2008). The normalization results in the data being projected onto a high-dimensional unit sphere, and the sample correlation between the two time series is equal to their inner product. The rs-fMRI segmentation aims at a clustering such that within-cluster voxels have a high correlation, and between-cluster voxels have a low correlation. The equivalence of the correlation and inner product makes it possible to reformulate the original problem into a new one. Now we can find a clustering where voxels with a larger inner product are put into one cluster. And the new problem can be modeled and solved using a mixture of the von Mises-Fisher (vMF) distribution.

We use $X = \{(x_1, \dots, x_N) \mid x_s \in S^{p-1}\}$ to denote the set of *normalized* time series on the p -sphere, where p is the number of time points in the original BOLD signal, and N is the total

number of gray matter voxels of all subjects. Given Y , the random vectors x_s are conditionally independent, hence $\log P(X|Y) = \sum_j \sum_{s \in \mathbf{V}_j} \log P(x_s | y_s)$. The likelihood function $P(x_s | y_s)$ is naturally modeled by a vMF distribution

$$f(x_s | y_s = l; \mu_l, \kappa_l) = C_p(\kappa_l) \exp(\kappa_l \mu_l^\top x_s), \quad x_s \in S^{p-1}, \quad l \in \mathbf{L}, \quad (2)$$

where for the network cluster label l , μ_l is the mean time course, $\kappa_l \geq 0$ is the *concentration parameter*, and C_p is the normalization constant. The larger the κ_l , the greater the density concentrated around the mean. Since eq. (2) depends on x only through $\mu^\top x$, the vMF distribution is unimodal and rotationally symmetric around μ .

4. Bayesian Inference

The exact inference of $P(Y|X)$ is computationally intractable due to the pairwise interaction of MRF prior distribution. Various approximate solutions exist for such types of undirected graphical model inference problems, including Gibbs and Metropolis sampling, expectation-propagation, and some variation inference methods such as mean field approximation and message-passing methods. In this work, we choose Gibbs sampling because of its simple formulation and straightforward implementation in a multi-processor system. In addition, compared to a point estimate such as *maximum a posteriori* (MAP) framework, the samples of the label map can be used to approximate the full posterior density, and to help understand the confidence of the point estimates such as posterior mean or modes.

4.1. Gibbs Sampling

The Gibbs sampler, as a special case of the Metropolis-Hastings sampler, solves a multivariate sampling problem using iterative univariate sampling. When all the random variables but one are fixed, the transition probabilities depend only on the local conditional distributions. The resulting Markov chain's equilibrium density is exactly the target density $P(Y|X)$. In general MCMC sampling, the variables are visited either at random, or according to a predefined order. As a way of incorporating domain-specific information in the design of our Gibbs sampler, we schedule the sampling order also in a multilevel fashion. At the image level, we draw the m th sample of the group label map Y_G^m given all the previous subject label maps $\{Y_j^{m-1}, j=1 \dots J\}$. Next, we draw a sample of subject j 's label map Y_j^m given the current group map sample Y_G^m . At the voxel level, we sample and update y_s given the rest of the nodes fixed. We call it a *scan* when each $y_s, \forall s \in V$ is updated once. The conditional distribution used to generate samples at the group and subject voxels can be derived from eq. (1) and are given as

$$P(y_s | y_{-s}, X) = \frac{1}{Z_s} \exp\{-U_p(y_s | y_{N(s)}, x_s)\}, \quad \text{where,} \\ \forall s \in \mathbf{V}_G, U_p = \alpha \sum_{j=1}^J \Psi(y_s, y_s^j) + \beta \sum_{r \in N(s)} \Psi(y_s, y_r), \quad (3)$$

$$\forall s \in \mathbf{V}_j, U_p = \alpha \Psi(y_s, y_{\bar{s}}) + \beta \sum_{r \in N(s)} \Psi(y_s, y_r) - \kappa_l \mu_l^\top x_s - \log C_p, \quad (4)$$

where $-s$ is the set of all nodes excluding s , Z_s is the partition function of y_s , U_p is the posterior energy, and $N(s)$ is the set of neighbors of s . The y_s^j in (3) is the network label of subject j 's voxel with the same physical coordinates with s , and the y_s in (4) is the label of the group map's voxel with the same physical coordinates as s . Note the evaluation of Z_s is easy since it is in a univariate distribution and is the sum of only L terms. Because of the dependency on previous samples, the sequence of label map samples $\{Y^m, m = 1 \dots, M\}$ is indeed a Markov chain; hence our method falls into Markov chain Monte Carlo (MCMC) sampling. After a sufficient burn-in period, a series of samples $\{Y^m, m = 1, \dots, M\}$ is saved. The samples have all the information of $P(Y|X)$ and can be used for approximating the expectation $\mathbb{E}_{P(Y|X)}[\log P(X, Y; \theta)]$ as well as estimating the posterior variance.

4.2. Parameter Estimation

The parameters $\{\beta, \kappa, \mu\}$ in our model are data-dependent, and manual assignment can easily result in over-fitting. For example, β 's optimal value depends on the number of neighbors of a voxel and also on the number of subjects in the group. In this data-driven model, we propose to estimate the parameters θ from the data using an expectation maximization (EM) algorithm, with the network labels Y as the hidden variable. However, the high-dimensionality and dependency between spatially adjacent voxels in MRF make it infeasible to obtain a closed form solution of the expectation of $\log P(X, Y; \theta)$ with respect to $P(Y|X)$. Here we propose to approximate the expectation using Monte Carlo EM (MCEM) algorithm. The set of samples, $(Y^1, \dots, Y^M)$ generated from density $P(Y|X)$ is used to approximate the expectation by the empirical average $(1/M) \sum_{m=1}^M \log P(X, Y^m; \theta)$. Furthermore, in order to evaluate $\log P(X, Y^m; \theta) = \log P(Y^m; \theta) + \log P(X|Y^m; \theta)$ as a function of θ , we face the difficulty of evaluating the partition function Z in $P(Y^m)$. In practice the likelihood function $P(Y; \theta)$ is approximated by pseudo-likelihood (Besag, 1974), which is defined as the product of the conditional likelihoods $P(y_s|y_{-s}; \theta), \forall s \in \mathbf{V}$. Therefore the label map's log-likelihood can be written as

$$\begin{aligned} \log P(Y; \theta) &\approx \sum_{s \in \mathbf{V}} -U(y_s|y_{-s}; \theta) - \log Z_s, \\ \forall s \in \mathbf{V}_G, U(y_s|y_{-s}) &= \alpha \sum_{j=1}^J \Psi(y_s, y_s^j) + \beta \sum_{r \in N(s)} \Psi(y_s, y_r); \quad (5) \\ \forall s \in \mathbf{V}_j, U(y_s|y_{-s}) &= \alpha \Psi(y_s, y_{\bar{s}}) + \beta \sum_{r \in N(s)} \Psi(y_s, y_r); \end{aligned}$$

where y_s^j and y_s have the same definition as in equation (3) and (4). With the pseudo-likelihood approximation, there is no need to compute the original Z . Instead we compute Z_s for each voxel s , just like what we do in the Gibbs sampling.

4.3. HMRF Algorithm Using MCEM

With all the preparation above, parameter estimation can be done by maximizing $(1/M) \sum_{m=1}^M \log P(X, Y^m)$. More specifically, β exists only in the prior, and can be estimated by maximizing $\frac{1}{M} \sum_{m=1}^M \log p(Y^m)$ with the Newton-Raphson method. Since $\{\mu, \kappa\}$ exist only in the data likelihood, the normalization constant Z in the prior is not a problem, hence $\{\mu, \kappa\}$ are estimated by maximizing $(1/M) \sum_{m=1}^M \log P(X|Y^m)$. The α parameter is treated differently and will be discussed in section 4.4. In order for MCMC sampling to converge quickly to the posterior, we need a reasonably good initial network label map. Here the K-Means clustering on a concatenated group dataset is used for the initial maps of both the group and subjects. After the EM iteration converges, we save M Monte Carlo samples as output. The Monte Carlo samples have all the information of the posterior distribution of network labels, and will be used in postprocessing for inference. Putting this all together, the HMRF method to estimate the group and individual label maps is given in Algorithm 1.

Algorithm 1: HMRF: Monte Carlo EM algorithm for network label inference and parameter estimation

Data: Normalized rs-fMRI, initial group label map

Result: MC samples of label maps $\{Y^m, m = 1, \dots, M\}$, parameters $\{\beta, \mu, \alpha\}$

while $\mathbb{E}_{p(Y|X)}[\log P(X, Y; \theta)]$ not converged **do**

repeat

foreach $s \in V_G$ **do** Draw sample of y_s from $P(y_s|y_{-s}, x_s; \theta)$ using (3);

foreach $j = 1 \dots J$ **do**

foreach $s \in V_j$ **do** Draw sample of y_s from $P(y_s|y_{-s}, x_s; \theta)$ using (4);

 Save sample Y^m after B burn-ins;

until $B + M$ times;

foreach $l = 1 \dots L$ **do**

 Estimate $\{\mu_l, \kappa_l\}$ by maximizing $(1/M) \sum_{m=1}^M \log P(X|Y^m; \theta)$;

 Estimate β by maximizing (5);

4.4. Estimating α Parameter By Cross-Validation

The parameter α in our model represents the strength of the links between the group and subject network label maps. The parameter implicitly represents the extent to which the functional patterns are shared among the subjects. Unfortunately, this parameter cannot be estimated in a MCEM framework by a Newton-Raphson method, as such a direct optimization will result in a collapsed solution. A solution of $\alpha = 0$ would minimize the energy associated with the between-level links, and the group map V_G would degenerate into a constant label map because such a map would minimize the energy associated with the links within the group map. We instead use the posterior predictive distribution (Gelman et al., 2003) of a test subject's BOLD signal X_r , defined as

$$P(X_t|X;\alpha,\theta_t)=\int P(X_t|X_t;\theta_t)P(Y_t|X;\alpha)dY_t, \quad (6)$$

where $\theta_t = \{\mu_t, \kappa_t, \beta_t\}$ is the parameter set of the test subject. With a leave-one-out procedure the same as that in the standard cross-validation, we pick one subject as the test subject X_t , and the remaining $J - 1$ subjects as the training data. We then compute the average $P(X_t|X;\alpha, \theta_t)$ across all test subjects given a list of prespecified α values, and choose α with the highest average predictive distribution. The detailed procedure to compute (6) is given in Appendix A.

5. Experiments On Simulated Data

Given the lack of ground truth of the functional network of the *in vivo* rs-fMRI data, we begin the experiments with a simulated dataset. We focus primarily on the estimation accuracy on the simulated dataset, and on the estimation consistency on the *in vivo* data.

We compare our method with two other clustering methods – K-Means and normalized-cuts (N-Cuts) – as well as two degenerated versions of the HMRF algorithm: HMRF-A and HMRF-B. The K-Means algorithm, as a simple and fast clustering method, is applied to the paradigm fMRI study in Baumgartner et al. (1998), and is later used by Bellec et al. (2010) for bootstrap analysis of the rs-fMRI group study. In our experiment, the distance metric of K-Means is defined as $1 - x_s^\top x_r$. To estimate an individual subject's network, we apply K-Means on each subject's BOLD signal 20 times, and choose the segmentation map with the minimal ratio of the sum of the intercluster distance and the sum of the intracluster distance. For the group study, we construct a *group* dataset by concatenating all subjects' time courses and run K-Means 20 times also on this group's dataset to estimate a group network label map. The initial cluster centers for both subject and group clustering are chosen randomly while at the same time maximizing the between-center distance (Arthur and Vassilvitskii, 2007).

N-Cuts formulates the fMRI image segmentation as a graph partitioning problem. A global criterion is used to find a subset of edges to remove from a full-connected graph, and the voxels are partitioned into multiple disjoint sets (Shi and Malik, 2000). N-Cuts is used by Van Den Heuvel et al. (2008) and Craddock et al. (2012) for the group rs-fMRI study. Following Van Den Heuvel et al., we also apply N-Cuts in two stages. First N-Cuts is run on each subject's affinity matrix, as computed from the pairwise correlation between time courses. A second N-Cuts is applied on a group affinity matrix, computed by summing all subjects' binarized affinity matrices derived from their segmentation maps. We use the same toolbox `Ncutclustering_9` (Shi and Malik, 2000) as in Van Den Heuvel et al., as well as the same parameter setting.

Both HMRF-A and HMRF-B, as simplified versions of HMRF, serve to check whether a reduced model would be able to achieve the same or better performance compared to the proposed full model. Both models are the same as HMRF except $\alpha = 0$ for HMRF-A, and $\beta = 0$ for HMRF-B. The model HMRF-B indeed amounts to defining a MRF on each single subject and estimating each subject's networks independent of other subjects. Such a strategy is equivalent to the hidden Markov model we proposed in Liu et al. (2011).

For HMRF, we skip the first 500 burn-in samples before saving 100 samples of the label map at each EM iteration. The convergence testing of MCMC sampling, especially in high-dimensional space is an open question and there is no widely accepted method to address this issue. We empirically choose the number of burn-in and MC samples by observing that the posterior probability estimated from samples has no significant change. The β parameter is estimated by the M step, as well as the μ and κ for each vMF component. As an optional postprocessing step, the discrete label map is obtained by running a iterated conditional mode (Besag, 1986) algorithm based on the last MCMC sample map.

Before a discussion of synthetic data generation, we briefly discuss how to measure the data quality of rs-fMRI. The separability of a data set for the purpose of clustering depends on both the within-cluster variance and between-cluster variance. In this specific rs-fMRI dataset, the signal-to-noise ratio (SNR) is represented by the ratio of the average between-cluster distance (defined as $1 - \mu_i^\top \mu_j$, where μ_i and μ_j are the cluster's mean time series), and the average within-cluster variance (defined by $1/\kappa$).

We generated synthetic rs-fMRI data in two steps. First, a group network map with five network labels is generated by drawing samples from a Potts model with $\beta = 2.0$ and 500 scans. Given the group map, a subject map is generated according to equation (1) with $\alpha = 0.5$ and $\beta = 2.0$. The subject map generation procedure is repeated 25 times to obtain a group of 25 subjects. To simulate the BOLD signal given the network label map, we first generate mean time courses μ_l , $l = \{1, \dots, 5\}$ from a first-order auto-regressive process $x_t = \phi x_{t-1} + \varepsilon$, with $\phi = 0.8$ and $\varepsilon = 0.1$. The sample correlations between the mean time series are in the range of $(-0.15, 0.3)$. Then, we add independent Gaussian white noise on each cluster's mean time course. The variance of the white noise is chosen such that the simulated BOLD signals have SNR=24, which is close or slightly lower than that of the real rs-fMRI data used in our experiments. Once the time series are generated, they are spatially smoothed with a Gaussian filter. Because the size of the smoothing filter may have interactions with our HMRF model and hence have an impact on the estimation accuracy, we spatially smoothed the generated BOLD signals with three levels of scale: FWHM = 0, FWHM = 1.88 mm, and FWHM = 4.7 mm. Furthermore, the synthetic data are generated randomly, so the experimental results from the data may also vary. To take account of the variability of the results, we repeated the above data generation process 100 times. For each generated data set, we run the five methods on the BOLD signals preprocessed by three levels of Gaussian filters respectively and compare the Monte Carlo average of the estimated label maps with the ground truth.

5.1. Synthetic Data Results

Among the 100 Monte Carlo runs of the data generation and estimation procedure, we choose one dataset smoothed at FWHM = 1.88 mm. The corresponding estimates are shown in Figure 3. We use the Rand index (Rand, 1971) to measure the similarity between simulated ground truth subject maps and the true group map. Rand index (RI) ranges in $[0, 1]$, and takes 1 if the two maps under comparison are the same. The RI value for this particular simulated dataset is 0.88 (similar values for other generated datasets), which we find is empirically close to the real data. From the figure, all methods appear to estimate the

group map well (except HMRF-B, which does not allow a group map estimate), but perform differently on the subjects. K-Means tries to identify the finer details of the individual subject's spatial patterns but fails due to the high noise level. N-Cuts and HMRF-A can detect the large patterns but lose some detail; HMRF-B does estimate the smooth subject map thanks to the within-subject smoothness links but the maps do not match the ground truth well. Finally, the HMRF is able to recover subjects' network maps with good matching to the ground truth.

To quantitatively evaluate the accuracy of the segmentation map from various methods, we calculate the RI values between the true map and the estimated map. The boxplot in Figure 4 shows the RI across all Monte Carlo runs and subjects. In all three settings of smoothing kernel size, HMRF achieves higher accuracies compared to other methods. In addition, for individual subjects' estimation, our model performs best at a moderate smoothing size of FWHM = 1.88 mm, which is smaller than the typical 5-8 mm smoothing size. This is because the HMRF model benefits from the reduced noise variance resulting from the moderate smoothing, but avoids losing finer details due to excessive smoothing. In practice, this means when applying HMRF, the BOLD signal should be smoothed by a small-kernel Gaussian filter, and we choose FWHM = 1.5 mm in the following real data experiments. We also note that the K-Means optimal smoothing kernel size is larger than that of HMRF, because it lacks the spatial coherence regularization and hence needs more smoothing in preprocessing stage. Last, we found that the two reduced models HMRF-A and HMRF-B do not perform as well as the full model, indicating that the hierarchy in the full model is indeed necessary. For all possible smoothing sizes, HMRF's estimation accuracy is comparable or moderately better than the other four methods on the group label map, and significantly higher on subject maps.

6. Real Data Experiments

In this work we test our methods on the publicly available NYU test-retest (TRT) dataset that has been used previously by Shehzad et al. (2009) and Zuo et al. (2010). While the original goal of the above works was to verify the voxelwise intra- and intersession TRT reliability, our goal is to verify whether the methods under consideration are able to estimate consistent functional network maps across sessions, given the fair amount of intersession consistency in the data set (Damoiseaux et al., 2006; Chen et al., 2008; Meindl et al., 2010; Franco et al., 2009). We present two experiments with the NYU-TRT datasets. The first experiment aims at demonstrating the intersession consistency of the estimated subject network maps, and the second one evaluates how the algorithms behave under the perturbation of the data by using bootstrap sampling. We compare three methods, HMRF, K-Means, and N-Cuts, in both experiments. The other two methods, HMRF-A and HMRF-B, are not taken into account in this section since they are a simplified version of HMRF and have been shown to be suboptimal compared to the full model.

6.1. Preprocessing

Twenty-six healthy control participants (11 males, mean age 20.5 ± 4.8 years) were scanned three times. The participants had no history of psychiatric or neurological illness. BOLD EPI images (TR = 2 s, TE = 25 ms, flip angle = 90, 39 slices at 3 mm slice thickness, 64×64

matrix, field of view = 192 mm, 197 volumes) were acquired on a Siemens Allegra 3.0 Tesla scanner. Scans 2 and 3 were conducted in a single session, 45 minutes apart, and were 5-16 months after the first scan. The subjects were asked to relax and remain still with their eyes open during the scan. A high resolution T1-weighted image was also obtained (MPRAGE with TR = 2.5 s, TE = 4.35 ms, TI = 900 ms, flip angle = 8°, 176 slices, FOV = 256 mm).

The fMRI data was preprocessed using the scripts of the 1000 functional connectomes projects ¹, as well as FMRIB's FSL toolset and the Medical College of Wisconsin's AFNI tool ². The volumes are motion corrected by aligning to the mean volume with a six-parameter rigid body transformation. The BOLD signals were bandpass filtered to 0.01 to 0.1 Hz, and were regressed out nuisance variables including white matter, CSF mean time courses and six motion parameters. The signal is then filtered by a FWHM = 1.5 mm Gaussian filter for spatial smoothness. The small kernel size of spatial smoothing guarantees that noise is canceled by averaging neighboring voxels while keeping the finer functional network patterns (see the simulated test and Figure 4). The functional images are first registered to the corresponding T1 images, and both functional and T1 images are registered to MNI152 (Montreal Neurological Institute) space with a 12-parameter affine transformation. Finally, after masking out white matter and CSF voxels, we have 39080 gray matter voxels remained in each subject. We construct a joint graph with over one million nodes including all subjects and the group map.

6.2. Choosing Parameters

In this work we do not address the problem of how many functional networks exist in the human brain. Instead, we use existing reports (Yeo et al., 2011) and choose seven functional networks for segmentation throughout the real data experiments. With this setting, we expect to identify the following typical functional networks: visual and primary motor (Damoiseaux et al., 2006), attention (Fox et al., 2006), default mode network (DMN) (Greicius et al., 2004), saliency, and executive control system (Seeley et al., 2007), regardless of the segmentation methods used. The K-Means is repeated 20 times with random initialization (Arthur and Vassilvitskii, 2007) for segmentation of both the subject and group maps. For N-Cuts, we threshold each subject's correlation matrix at 0.4 before applying N-Cuts on single subject. After the individual segmentation, we average all subjects' binary segmentation matrices, and threshold the averaged matrix at 0.3. The result represents the group correlation matrix. Both cut-off thresholds are suggested by Van Den Heuvel et al. (2008). Our implementation is different with from Van Den Heuvel et al. only in that we partition the subject map into seven clusters instead of 20. This is because we need to compare the subject maps estimated by N-Cuts with those estimated by the HMRF method at the same number of networks. We also run N-Cuts with 20 clusters on subject maps to compare with our seven-cluster configuration (results now shown), and find the group level segmentation has not been impacted by our lack of over-segmentation at the subject level. For HMRF, we initialize both the group and subject label maps with the group label map estimated from K-Means. The sampling routine (E-step of MCEM algorithm)

¹www.nitrc.org/plugins/mwiki/index.php/fcon_1000:ScriptUse

²<http://afni.nimh.nih.gov>

skips 500 burn-in samples before saving 100 MC samples. The parameters $\{\beta, \mu, \kappa\}$ are estimated from the data. With α estimated from the posterior predictive distribution (see section 4.4), we found the similarity between estimated group and subject maps measured is around 0.85 measured by RI value.

6.3. Interession Consistency

Since the TRT dataset and the general rs-fMRI data have been shown to share consistent functional networks across all sessions (Damoiseaux et al., 2006; Chen et al., 2008; Meindl et al., 2010; Franco et al., 2009), we verify the consistency of the HMRF algorithm by applying it to all three sessions of data. A method is said to be consistent if it is able to derive similar network estimates across sessions. We compare three pairs of sessions' consistency: S1 vs S2, S1 vs S3 and S2 vs S3. For each subject in each pair of sessions, we compute the consistency score between this subject's network map estimates in two sessions. The similarity is again represented by the RI values. We expect the proposed HMRF algorithm has higher average similarity compared with other methods. The consistency scores of all subjects are summarized in a boxplot as in Figure 5. For comparison, the same boxplots are also drawn for K-Means and N-Cuts. From the figure, the subject network label maps estimated from HMRF have significant higher interession consistency scores compared to the other two methods. This indicates that our algorithm is able to capture the common functional patterns across sessions. In addition, both K-Means and HMRF have higher interession consistency scores between session two and session three, compared to the other two interession comparisons. This is consistent with the fact that sessions two and three have a smaller interval (45 minutes apart), compared to session one and two (5-16 months). K-Means has slightly better between-session consistency than N-Cuts, probably because we have run K-Means multiple times and have chosen the best solutions.

The RI values in Figure 5 only gives a single number of similarity between two network label maps, rather than a voxel-wise consistency map. To visualize the consistency at the voxel level, we first match session two and session three's segmentation maps to session one's by permuting the cluster labels (this is not needed for the between-session RI, which is invariant to label permutation). Then we define a variance map as follows: the variance at certain voxels takes the value 0 if the estimates of all three sessions have the same labels. The variance takes 1 if two of the three sessions have the same labels, and takes 2 if none of the estimates are the same. We then average the variance map across all subjects and obtain a mean variance map. This map shows how the algorithm performs in term of consistency at the voxel level across all subjects. The results are shown in Figure 6. Image visualizaiton is done by using nipy, a python package for neuroimaging data.³ We note that although K-Means and N-Cuts have low variance at the visual cortex, they have larger variance in most voxels of dorsal attention and the DMN. These findings confirm the different level of consistency between the functional networks, as has been shown in the original work of Zuo et al. (2010). Overall, the HMRF method's estimates have the lowest level of variance and hence the highest level of consistency.

³<http://nipy.org/nipy/stable/index.html>

6.4. Bootstrapping

In these experiments we aim to evaluate the performance of the three algorithms with bootstrapping. In the bootstrapping method, one covers the whole distribution of the estimator with the independent samples drawn from the original dataset with replacements, and estimates the stability of an algorithm (Efron and Tibshirani, 1993). An approximate solution of an algorithm is stable if the solution is not highly sensitive to the input data. It is unstable if a slight change in the data can cause the predicted values to change significantly. In this experiment, the bootstrapped samples can be seen as a small perturbation of the input data and will be used to test the algorithm stability.

There are various approaches for resampling the available data. One may resample the subjects from the original dataset (Damoiseaux et al., 2006). Here for each voxel of each subject in session 1, we sample with replacement from the 197 time points of preprocessed data, and obtain a bootstrap sample volume with the same BOLD signal length and number of subjects with the original dataset. The sampling is similar to the circular block bootstrap in Bellec et al. (2010), except that we do not model the temporal correlation between time points. Since all methods under comparison here do not model temporal correlation, the shuffling of the time points has no effect on the segmentation results. After repeating the sampling 100 times, we obtain a set of 100 bootstrap samples, each of which includes all subjects' time series data. Then, all three segmentation methods are applied on each of the bootstrap datasets. We estimate group and subject level maps from each bootstrap dataset by using the three methods. All the estimated label maps are postprocessed by a label permutation routine to guarantee that the same networks have the same labels.

Figure 7 shows seven average group-level functional network maps across all bootstrap sampled data. For each network, we extract a binary map with voxel intensity taking 1 in that network and 0 outside. This binary map is then averaged over all bootstrap samples. We also show the variance of this binary label map over all samples in Figure 8. Small variance indicates more stability under bootstrap sampling. All three methods have moderate to high stability across bootstrap samples. For visual, motor, and DMN networks, K-Means and N-Cuts have reasonably high stability, although some voxels at the boundary of the network regions are labeled differently across bootstrap samples. For the attention, salience and executive control networks estimated by K-Means and N-Cuts, the ambiguity not only happens on the boundary of the network regions, but also on some bigger regions inside the networks. For example, in some bootstrap runs, K-Means incorrectly assigns the posterior cingulate cortex (PCC) to the attentive network (see the red regions in dorsal attentive in Figure 7), whereas PCC has been shown to be part of the DMN (Greicius et al., 2003). K-Means also miss part of the primary motor network. K-Means even in a few runs merges the brain stem into the DMN. For N-Cuts, the dorsal attentive, salience, and executive control networks have larger variance under this data perturbation. Compared to the other two methods, HMRF has the smallest variance, and hence the highest stability in all seven networks including the brain stem. A small number of voxels in motor and DMN still shows unstable assignments.

To demonstrate the stability of the estimates on each of the subject functional networks, we first pick 3 subjects from the 25 subjects in the dataset. For each subject, we show the

average network patterns over all bootstrap samples. See Figure 9. We show only six physiologically meaningful networks, excluding the one corresponding to the brain stem. For each network, one representative slice is shown. From the figure, all three subjects' mean network maps have lower stability compared to their corresponding group networks. Certain subjects' networks are significantly less stable than other subjects, due to the various degree of perturbation by the random sampling even using the same bootstrapping procedure. Among the six networks, attentive networks exhibit the most dramatic change under bootstrap sampling. Some voxels of salience and executive control networks are absorbed into attentive networks. This misclassification happens most on subject 2, and also happens a moderate amount on subjects 1 and 3. Compared to the other two methods, HMRF is able to estimate reasonably stable functional networks even with data resampling. Attentive networks and executive control networks tend to change more than other networks, but still less than K-Means and N-Cuts.

Another way to show the stability of the subject label maps is the variance map. Since we are interested in comparing among three methods the variance of the networks across all subjects, we show the variance not for each single subject, but an average variance over all subjects. See Figure 10. Because of the averaging over all subjects, the variance is more spread over the voxels. Again, HMRF shows significantly smaller variance than the other two methods, indicating its subject map estimates are more stable under bootstrap sampling.

6.5. Between-Level Links estimation

We also run the cross-validation and use the posterior predictive distribution in equation (6) for estimating the optimal α parameter. Figure 11 gives a plot of the average predictive density with alpha ranging in [0.15, 0.5], with interval 0.05. We found that with too small α , the model has low prediction values on the test data, and too large α values improve the prediction but still not the optimal. The best α value is around 0.3 to 0.35.

7. Discussion

We proposed a new hierarchical model for identifying functional networks of the human brain from a group of rs-fMRI dataset. The model assumes a group functional network map is shared among all subjects in the population, and individual subjects' functional patterns are generated as variations from this group level network. If we see the functional network pattern as a clustering of the fMRI data, we actually assume the subject maps are samples from an unknown distribution of the clusterings, with its mean given by the group map. We reformulate the distribution of clusterings as a distribution of network labels, where a subject's labels at each voxel are seen as generated from the group's network labels. While the intersubject statistical dependency is defined by the links between group and subject labels, the spatial coherence of the functional networks is guaranteed by the within-subject MRF. All the network label variables at both levels with their links, and the parameters are defined in an integrated graph, and the general techniques of graphical models can be used for (approximate) inference. This multilevel view is typically used in general statistical analysis when the individuals are grouped into units, and the variance of the variables is decomposed into the group-specific and subjectspecific terms. We borrow this idea and apply it in a clustering problem where the intensities of voxels at each time point are

grouped into units (subjects), and the vMF's κ parameter represents the individual subject's variance. The α parameter is equivalent to the *pooling factor* in the standard hierarchical linear model (Gelman and Pardoe, 2006), and controls the degree to which the estimates of subject label maps are pooled together.

We use the MCMC sampling for the inference because of its good approximation of the posterior density. An alternative class of methods is variational inference, including mean field approximation and expectation propagation. Both variational methods and MCMC are the approximation of the true posterior distribution. The former approximate the target distribution by the product of factorized distributions, and the latter achieve the approximation by Monte Carlo averaging. Both classes of methods depend on initial conditions. However, with Gibbs sampling, we obtain a full posterior density estimates of the network variables, while variational methods usually have point estimates. Besides, the derivation of the conditional expectation used for updating of variational methods would be cumbersome in our multilevel model. On the other hand, the Gibbs sampling is straightforward as the conditional probability is easy to compute in our Bayesian setting. Therefore, we choose Gibbs sampling due to its simplicity, as well as the fact that the application does not require real time computation. An additional critical property of the MCMC sampling is that its convergence does not depend on the dimension of the variables (Robert and Casella, 2005); thus we can achieve reasonable compute time even in this million-dimensional problem. The whole Monte Carlo expectation maximization procedure uses 45-50 cores on a multiprocessor machine, and takes about 2 hours for a group of 25 subjects.

As a practical guide for applying HMRF, the introduction of within-subject MRF is not meant to replace the spatial smoothing in the preprocessing steps. This is one step further from what we found in our previous work (Liu et al., 2012), where no spatial smoothing is conducted when the HMRF model is used. In the simulated experiments, we found a moderate spatial smoothing, plus our HMRF model can achieve the best estimation accuracy. The best accuracy of the combined model is because moderate smoothing does help to decrease the noise, without overly impacting the signal at finer spatial scales. The MRF regularization further favors spatial coherence and intersubject coherence.

8. Conclusion

The main contribution of our work is a hierarchical model for estimating population functional network maps as well as individual subjects' maps from rs-fMRI data. The relationship of the network labels in both group and subjects is represented probabilistically. We solve the multivariate inference problem by MCEM sampling. The results for a synthetic group of fMRI data show that our method achieves higher accuracy on both group and subject network map estimation than the regular *one-way* approach, such as K-Means and normalized cuts. The experiments on a multiple-session rs-fMRI dataset show that our HMRF algorithm is able to estimate the network maps for each session with higher between-session consistency. A further bootstrapping experiment also proves the proposed algorithm has robust solutions under data perturbations.

Appendix A: Predictive Distribution

The test subject's predictive distribution in (6) for a chosen α can be evaluated through a Monte-Carlo approximation

$$P(Y_t|X; \alpha; \theta_t) \approx \frac{1}{M} \sum_m P(X_t|Y_t^m; \alpha; \theta_t), Y_t^m \sim P(Y_t|X; \alpha). \quad (\text{A.1})$$

One economical way of generating sample $\{Y_t^m, m=1 \dots, M\}$ can be done within the MCEM loop of algorithm 1. After the current group map is generated in E step, one sample Y_t^m can be generated from $P(Y_t|X; \alpha, \theta)$. The corresponding posterior energy function at voxel s is $U_p(y_s|y_{N(s)}) = \alpha\psi(y_s, y_s) + \beta\sum_{r \in N(s)} \psi(y_s, y_r)$. This energy is same to equation (4), except that there is no time series data term $\kappa_l \mu_l^\top x_s - \log C_p$ since the test subject data X_t are not given in this distribution. For one sample map Y_t^m the test subject parameter set θ_t is obtained by optimizing $P(X_t|Y_t^m)$. As a simple reasoning of why we can use the equation (6) for estimating α , when α is too small, most of the Y_t^m will depend less on the group map Y_G and tend to be random clusterings, which will have low data likelihoods in (A.1). When α is too big, Y_t^m will be almost the same as Y_G , again resulting in a suboptimal value for (A.1). Only with an appropriate α , could Y_t^m sufficiently explores the sampling space including the regions where the predictive distribution is maximized. In practice, we evaluate eq. (A.1) for a fixed set of α values, and choose α with the largest predictive density value $P(Y_t|X; \alpha)$.

References

- Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, Society for Industrial and Applied Mathematics; 2007. p. 1027-1035.
- Baumgartner R, Windischberger C, Moser E. Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging*. 1998; 16:115–125. [PubMed: 9508268]
- Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in fMRI. *Neuroimage*. 2003; 20:1052–1063. [PubMed: 14568475]
- Beckmann CF, Mackay CE, Filippini N, Smith SM. Group comparison of resting-state fMRI data using multi-subject ica and dual regression. *Neuroimage*. 2009; 47:S148.
- Bellec P, Rosa-Neto P, Lyttelton O, Benali H, Evans A. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage*. 2010; 51:1126–1139. [PubMed: 20226257]
- Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B (Methodological)*. 1974; 36:192–236.
- Besag J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B (Methodological)*. 1986:259–302.
- Calhoun V, Adali T, Pearlson G, Pekar J. A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapping*. 2001a; 14:140–151. [PubMed: 11559959]
- Calhoun V, Adali T, Pearlson G, Pekar J. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*. 2001b; 13:43–53. [PubMed: 11284046]
- Chen S, Ross TJ, Zhan W, Myers CS, Chuang KS, Heishman SJ, Stein EA, Yang Y. Group independent component analysis reveals consistent resting-state networks across multiple sessions. *Brain Research*. 2008; 1239:141. [PubMed: 18789314]

- Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*. 2012; 33:1914–1928. [PubMed: 21769991]
- Damoiseaux J, Rombouts S, Barkhof F, Scheltens P, Stam C, Smith SM, Beckmann C. Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences*. 2006; 103:13848–13853.
- Efron, B.; Tibshirani, R. *An introduction to the bootstrap*. Chapman & Hall; New York: 1993.
- Esposito F, Scarabino T, Hyvarinen A, Himberg J, Formisano E, Comani S, Tedeschi G, Goebel R, Seifritz E, Di Salle F, et al. Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage*. 2005; 25:193–205. [PubMed: 15734355]
- Filippini N, MacIntosh BJ, Hough MG, Goodwin GM, Frisoni GB, Smith SM, Matthews PM, Beckmann CF, Mackay CE. Distinct patterns of brain activity in young carriers of the apoe- $\epsilon 4$ allele. *Proceedings of the National Academy of Sciences*. 2009; 106:7209–7214.
- Fox MD, Corbetta M, Snyder AZ, Vincent JL, Raichle ME. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences*. 2006; 103:10046–10051.
- Franco AR, Pritchard A, Calhoun VD, Mayer AR. Interrater and intermethod reliability of default mode network selection. *Human Brain Mapping*. 2009; 30:2293–2303. [PubMed: 19206103]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian data analysis*. CRC press; 2003.
- Gelman A, Pardoe I. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*. 2006; 48:241–251.
- Golland, P.; Lashkari, D.; Venkataraman, A. Spatial patterns and functional profiles for discovering structure in fMRI data. *Signals, Systems and Computers, 2008 42nd Asilomar Conference on, IEEE*; 2008. p. 1402-1409.
- Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H, Reiss AL, Schatzberg AF. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological psychiatry*. 2007; 62:429–437. [PubMed: 17210143]
- Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*. 2003; 100:253–258.
- Greicius MD, Srivastava G, Reiss AL, Menon V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences*. 2004; 101:4637–4642.
- Greig D, Porteous B, Seheult AH. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society Series B (Methodological)*. 1989:271–279.
- Hammersley JM, Clifford P. *Markov fields on finite graphs and lattices*. 1968
- Hartvig NV, Jensen JL. Spatial mixture modeling of fMRI data. *Human Brain Mapping*. 2000; 11:233–248. [PubMed: 11144753]
- Lafferty, J. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*; Morgan Kaufmann; 2001.
- Li, SZ. *Markov random field modeling in computer vision*. Springer-Verlag; New York, Inc.: 1995.
- Liu W, Awate S, Anderson J, Yurgelun-Todd D, Fletcher P. Monte Carlo expectation maximization with hidden Markov models to detect functional networks in resting-state fMRI. *Machine Learning in Medical Imaging*. 2011:59–66.
- Liu, W.; Awate, SP.; Fletcher, PT. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012*. Springer; 2012. Group analysis of resting-state fMRI by hierarchical Markov random fields; p. 189-196.
- Meindl T, Teipel S, Elmouden R, Mueller S, Koch W, Dietrich O, Coates U, Reiser M, Glaser C. Test-retest reproducibility of the default-mode network in healthy individuals. *Human Brain Mapping*. 2010; 31:237–246.
- Mohammadi B, Kollewé K, Samii A, Krampfl K, Dengler R, Münte T. Changes of resting state brain networks in amyotrophic lateral sclerosis. *Klinische Neurophysiologie*. 2009; 40:V148.

- Ng B, Hamarneh G, Abugharbieh R. Modeling brain activation in fMRI using group MRF. *Medical Imaging, IEEE Transactions on* 31. 2012:1113–1123.
- Penny WD, Trujillo-Barreto NJ, Friston KJ. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*. 2005; 24:350–362. [PubMed: 15627578]
- Potts, R. *Proceedings of the Cambridge Philosophical Society*. Cambridge Univ Press; 1952. Some generalized order-disorder transformations; p. 106–109.
- Rand WM. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*. 1971; 66:846–850.
- Robert CP, Casella G. *Monte carlo statistical methods* (springer texts in statistics). 2005
- Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD. Neurodegenerative diseases target large-scale human brain networks. *Neuron*. 2009; 62:42. [PubMed: 19376066]
- Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, Reiss AL, Greicius MD. Dissociable intrinsic connectivity networks for salience processing and executive control. *The Journal of Neuroscience*. 2007; 27:2349–2356. [PubMed: 17329432]
- Shehzad Z, Kelly AC, Reiss PT, Gee DG, Gotimer K, Uddin LQ, Lee SH, Margulies DS, Roy AK, Biswal BB, et al. The resting brain: unconstrained yet reliable. *Cerebral Cortex*. 2009; 19:2209–2229. [PubMed: 19221144]
- Shi J, Malik J. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22. 2000:888–905.
- Van Den Heuvel M, Mandl R, Pol H. Normalized cut group clustering of resting-state fMRI data. *PLoS One*. 2008; 3:e2001. [PubMed: 18431486]
- Varoquaux, G.; Gramfort, A.; Pedregosa, F.; Michel, V.; Thirion, B. *Information Processing in Medical Imaging*. Springer; 2011. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity; p. 562–573.
- Varoquaux G, Sadaghiani S, Piel P, Kleinschmidt A, Poline JB, Thirion B. A group model for stable multi-subject ica on fmri datasets. *Neuroimage*. 2010; 51:288–299. [PubMed: 20153834]
- Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM, et al. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*. 2004a; 21:1732–1747. [PubMed: 15050594]
- Woolrich MW, Jenkinson M, Brady JM, Smith SM. Fully Bayesian spatio-temporal modeling of fMRI data. *Medical Imaging, IEEE Transactions on* 23. 2004b:213–231.
- Yeo B, Krienen F, Sepulcre J, Sabuncu M, Lashkari D, Hollinshead M, Roffman J, Smoller J, Zöllei L, Polimeni J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*. 2011; 106:1125–1165. [PubMed: 21653723]
- Zuo X, Kelly C, Adelstein J, Klein D, Castellanos F, Milham M. Reliable intrinsic connectivity networks: test-retest evaluation using ICA and dual regression approach. *Neuroimage*. 2010; 49:2163. [PubMed: 19896537]

Figure 1.

Comparison of segmentation methods for group study of fc-fMRI. Most methods use a one-way approach, either in a subject-group order (left) or a group-subject order (middle). Our method (right) aims at a joint estimation of both levels of network maps, where group and subject maps help each other in a bidirectional flow

Figure 2.

We define a MRF on a graph that includes the voxels of all subject maps as well as the group map. The set of edges includes between-level links with weight α , and withinsubject links with weight β . The square box on the subject level and time courses repeats J times the nodes in the square, representing all the subjects. Only the central voxels connection is shown for the between-level links, whereas in practice the links exist on all other voxels. The BOLD signal variables are shaded, meaning they are set to the observed value.

Figure 3.

The estimated group and subject functional network label maps from various methods, as well as the ground truth maps. Only two are shown among the 25 subjects.

Figure 4.

Box-and-whiskers plots of the estimation accuracies of all methods for three levels of spatial smoothing. The top row is the accuracies of subject labels across all subjects and MC samples, and the bottom is group map accuracies across all MC samples. The upper and lower ‘hinges’ correspond to the 25th and 75th percentiles. The asterisk on top of each box indicates the p-value of the standard two-tailed T test between HMRF and the corresponding method. No asterisk: significant $p > 0.05$; *: significant at $p < 0.05$; **: significant at $p < 0.01$; ***: significant at $p < 0.001$. The group map is not applicable to *HMRF-B* due to its lack of between-level links.

Figure 5.

Box-and-whiskers plots of the RI value between each pair of sessions over the all subjects' label map. The bottom and top of the boxes are the 25th and 75th percentile, and the whiskers extend to the whole range of the data except the outliers.

Figure 6.

The inter-session variance maps for three segmentation methods. The variance maps are obtained for each subject, averaged across subjects, and finally normalized to [0, 1]. A few voxels with intensity above 0.8 are rendered the same as those with intensity 0.8. This single map covers all seven functional networks, and we selectively show the slices corresponding to the three major networks. The images left are the subjects' left, and we use the same convention in the following figures.

Figure 7.

The group level's mean functional networks estimated from all bootstrapped data by three segmentation methods. The binary map of each network is averaged over all bootstrap samples. The average intensity ranges from 0 to 1.

Figure 8.

The group variance map estimated from all bootstrap data by the three segmentation methods. The variance ranges from 0 to 0.25.

Figure 9.

The three subjects' average network label maps estimated from all bootstrap samples. One representative slice is shown for each of the seven networks for each subject (row) and each method (column), excluding brain stem component. The average values range from 0 to 1.

Figure 10.

The subjects' variance maps estimated from all bootstrap samples. The maps are averaged across all subjects, and their values range from 0 to 0.25. The color map is in $[0, 0.15]$ since most of the variance values fall into this range.

Figure 11.

Estimation of parameter α with the average predictive distributions using the leave-one-out cross-validation. We use the data from only the first session of the NYU-TRT dataset but find similar patterns in the other two sessions, α are sampled between 0.15 and 0.5, with interval 0.05.