Modeling Covariate Effects in Group Independent Component Analysis with Applications to Functional Magnetic Resonance Imaging

Ran Shi and Ying Guo

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, US

Abstract

Independent component analysis (ICA) is a powerful computational tool for separating independent source signals from their linear mixtures. ICA has been widely applied in neuroimaging studies to identify and characterize underlying brain functional networks. An important goal in such studies is to assess the effects of subjects’ clinical and demographic covariates on the spatial distributions of the functional networks. Currently, covariate effects are not incorporated in existing group ICA decomposition methods. Hence, they can only be evaluated through ad-hoc approaches which may not be accurate in many cases. In this paper, we propose a hierarchical covariate ICA model that provides a formal statistical framework for estimating and testing covariate effects in ICA decomposition. A maximum likelihood method is proposed for estimating the covariate ICA model. We develop two expectation-maximization (EM) algorithms to obtain maximum likelihood estimates. The first is an exact EM algorithm, which has analytically tractable E-step and M-step. Additionally, we propose a subspace-based approximate EM, which can significantly reduce computational time while still retain high model-fitting accuracy. Furthermore, to test covariate effects on the functional networks, we develop a voxel-wise approximate inference procedure which eliminates the needs of computationally expensive covariance estimation. The performance of the proposed methods is evaluated via simulation studies. The application is illustrated through an fMRI study of Zen meditation.

Keywords:
Independent component analysis; Multi-subject imaging data; Covariate effects; EM algorithm; Sparsity; Functional magnetic resonance imaging (fMRI)
1 INTRODUCTION

Functional magnetic resonance imaging (fMRI) is one of the most commonly used imaging technologies to assess neural activities in the brain. In fMRI studies, the observed data represents combined signals generated from various brain functional networks. One major goal in fMRI analysis is to characterize the spatial patterns as well as the temporal dynamics of these underlying networks. Independent component analysis (ICA) is an effective tool to achieve this goal. As a special case of blind source separation, ICA separates observed data into linear combinations of latent components that are statistically independent. As a fully data-driven algorithm, ICA does not require any prior information about the underlying source signals. ICA was initially applied to analyze single-subject fMRI data (McKeown et al., 1998; Biswal and Ulmer, 1999; Calhoun et al., 2001; Beckmann and Smith, 2005; Lee, 2011). Denote by $Y$ the $T \times V$ fMRI data matrix for one subject, where $T$ is the number of MR scans and $V$ is the number of voxels in the 3D brain image acquired during each scan. That is, each row of $Y$ represents a concatenated 3D image. Classical noise-free ICA model can be applied to decompose the observed data as follows,

$$Y_{T \times V} = A_{T \times q}S_{q \times V},$$

where $q$ is the total number of source signals. Each row of $S$ concatenates the 3D map of a spatial source signal. $A$ is the temporal mixing matrix which linearly combines the $q$ spatial sources to generate the observed series of fMRI images. The $q$ source signals are assumed to be independent in the spatial domain and hence are called independent components (ICs).

To decompose multi-subject fMRI data, ICA is extended for group analysis, which is referred to as group ICA (Calhoun et al., 2001). One commonly used group ICA method in fMRI analysis is the temporal concatenation group ICA (TC-GICA). In TC-GICA, the $T \times V$ fMRI data matrices for $N$ subjects are stacked on the temporal domain to form a tall $TN \times V$ matrix. The concatenated matrix is then decomposed into the product of a $TN \times q$ group mixing matrix and a $q \times V$ spatial source matrix with independent rows. A notable restriction of the TC-GICA framework is the assumption on homogeneous spatial source signals across subjects. To overcome this restriction, Guo and Tang (2013) proposed a hierarchical group ICA (H-GICA) model to accommodate between-subject variability in spatial source signals by incorporating subject-specific random effects in ICA.

One major limitation of the existing group ICA methods is that they do not incorporate subject’s covariate information in the ICA decomposition. The need to consider covariate effects in ICA arises from the fact that the spatial distributed patterns of brain functional networks vary considerably with subjects’ clinical, biological and demographic characteristics. By assessing and testing the
covariate effects on the ICs, we can address a series of important questions in mental health research such as how the functional networks are associated with mental disorders and how they respond to treatments (Greicius, et al., 2007; Chen et al., 2007; Sheline, et al., 2009). Currently, investigators take two ad-hoc approaches to investigate covariate effects in ICA. The first approach is to first conduct single-subject ICA separately on each subject’s data, select matching components from each subject and then perform group analyses on the selected single subject IC maps (Greicius, et al., 2007). A major problem with this approach is that it is often challenging to identify matching ICs across subjects since ICA output is invariant under permutations. Furthermore, since most ICA algorithms are stochastic, ICs extracted in separate ICA runs for different subjects are often not comparable to each other. The second approach is heuristic two-step regression strategies based on TC-GICA models such as the back-construction method (Calhoun et al., 2001) and the dual regression method (Beckman et al., 2009). Basically, these two-step methods first perform TC-GICA to extract group common IC maps and then reconstruct subject-specific maps by performing post-ICA analysis such as adjusting the mixing matrices with respect to subject-specific observed fMRI data. In the second step, covariate effects are evaluated via secondary tests, such as two-sample t-tests or MANCOVA, on the subject-specific maps. These two-step procedures are developed based on heuristic reconstructions and comparisons without supporting underlying probabilistic models. More importantly, the common group IC maps extracted in the TC-GICA step can be severely inaccurate when there are strong covariate effects or the sample data are imbalanced in terms of the covariates patterns. Since all subsequent analyses in the two-step procedures are based on the TC-GICA maps, the accuracy of the results become questionable.

In this paper, we propose a new hierarchical covariate ICA regression model (hc-ICA) that directly model covariate effects in group ICA. In the first-level model, we decompose the fMRI data for each subject as linear mixtures of subject-specific spatial source signals. In the second-level model, we model the subject-specific spatial source signals in terms of baseline population-level spatial source signals, covariate effects and between-subject random variabilities. To the best of our knowledge, hc-ICA is the first model-based statistical method for estimating and testing the covariate effects on brain functional networks. Comparing to the existing methods, hc-ICA has the following advantages: first, it can provide model-based estimates of brain functional network spatial maps for clinical and demographic subpopulations such as mentally disordered vs. normal subjects. secondly, it can potentially offer more accurate estimation and inference for the primary effects of interest, e.g. disease or treatment, on functional networks by appropriately accounting for other confounding factors in
We propose a maximum likelihood (ML) approach to estimate the parameters in our model via the EM algorithm. We first develop an exact EM which provides explicit forms for both the E-step and M-step. To provide a faster estimation method, we further develop a subspace-based approximate EM algorithm which focuses on a subspace of the latent states of the source signals. We provide theoretical justification as well as empirical evidence through simulations that the subspace-based EM can provide highly accurate approximation to the exact EM. We show that the computation load of the approximate EM scales linearly with the number of ICs, which is significantly more efficient as compared to the exponential growth of computing time in the exact EM. To test covariate effects in hc-ICA, we propose a voxel-wise approximate inference procedure which eliminates the needs of computationally expensive covariance estimation. We illustrate the performance of our estimation and inference methods through simulation studies. The proposed methods are applied to an fMRI study on Zen meditation.

2 METHODS

2.1 Preprocessing prior to ICA

Suppose that our original fMRI dataset consists of \( N \) subjects. For each subject, the fMRI signals are acquired at \( T \) time points across \( V \) voxels. Let \( \tilde{y}_i(v) \in \mathbb{R}^T \) be the time series recorded for subject \( i \) at voxel \( v \). Then \( \tilde{Y}_i = [\tilde{y}_i(1), ..., \tilde{y}_i(V)] \) is the \( T \times V \) fMRI data matrix for subject \( i \). Prior to ICA, several preprocessing steps, which include centering, dimension reduction and whitening, are generally performed to reduce the complexity of data and guarantee the identifiability of models (Hyvärinen, Karhunen and Oja, 2001). Under the paradigm of group ICA, we perform a unified dimension reduction and whitening procedure on the original fMRI data as: for \( i = 1, ..., N \),

\[
Y_i = (\Lambda_{i,q} - \bar{\sigma}_{i,q}^2 I_q)^{-\frac{1}{2}} U_{i,q}^\prime \tilde{Y}_i, \tag{2}
\]

where \( \Lambda_{i,q} \) contains the largest \( q \) eigenvalues of \( Y_i Y_i^\prime \) and \( U_{i,q} \) contains the \( q \) corresponding eigenvectors as its columns. The residual variance, \( \bar{\sigma}_{i,q}^2 \), is the average of the smallest \( T - q \) eigenvalues that do not appear in \( \Lambda_{i,q} \). The value of \( q \), which is the number of ICs, is determined using the Lapalce approximation (Minka, 2000). Throughout the rest of our paper, all the models and methodologies will be developed based on the preprocessed data \( Y_i = [y_i(1), ..., y_i(V)] \) for \( i = 1, ..., N \), which are \( q \times V \) matrices.
2.2 A hierarchical covariate ICA regression model (hc-ICA)

In this part, we present a hierarchical covariate ICA regression model to model the covariate effects on the ICs. We provide detailed introduction on our model in the following.

2.2.1 The first-level model

As in the single-subject ICA, our first-level model decomposes each subject’s preprocessed signals as the linear combinations of $q$ independent spatial source signals that are mixed through a $q \times q$ mixing matrix. Both the spatial source signals and the mixing matrices are subject-specific. Additionally, we include noise terms to capture residual variabilities that are not explained by the $q$ independent components. The noise terms can also help avoid overfitting and enable formal statistical inferences (Beckmann and Smith, 2004). Then our first-level model is specified as follows:

$$y_i(v) = A_is_i(v) + e_i(v),$$  \hspace{1cm} (3)

where $s_i(v) = [s_{i1}(v),...,s_{iq}(v)]'$ is a $q \times 1$ vector with $s_{il}(v)$ representing the spatial source signal of the $l$th IC at voxel $v$. The $q$ elements of $s_i(v)$ are assumed to be independent and non-Gaussian (or Gaussian for at most one element). For the pre-whitened data, it can be shown that the mixing matrix, $A_i$, is orthogonal (Hyvärinen and Oja, 2000). The noise term $e_i(v)$ is a $q \times 1$ vector and $e_i(v) \sim N(0, E_v)$ for $v = 1,...,V$. Since the spatial variabilities and correlations among $y_i(v)$ across all voxels have been explained by the ICs, we assume $e_i(v)$ are independent with spatial stationarity in their variance, i.e., $E_v = E$ for $v = 1,...,V$. Recall that after the pre-whitening steps, the elements of $y_i(v)$ are uncorrelated and their variances are standardized. Based on these facts, we further assume that $E = \nu_0^2 I_q$.

2.2.2 The second-level model

In the second stage of the hierarchical model, we model subject-specific spatial source signals $s_i(v)$ in term of population-level source signals modulated by covariate effects, that is,

$$s_i(v) = s_0(v) + \beta(v)'x_i + \gamma_i(v),$$  \hspace{1cm} (4)

where $s_0(v) = [s_{01}(v),...,s_{0q}(v)]'$ is the $q$ population-level spatial source signals of the $q$ statistically independent and non-Gaussian ICs, $x_i = [x_{i1},...,x_{ip}]'$ is the $p \times 1$ vector of covariates containing subject-specific characteristics such as the treatment or disease group, demographical variables and biological traits, $\beta(v)$ is a $p \times q$ matrix where the element $\beta_{kl}(v)$ ($k = 1,...p, \ell = 1,...,q$) in $\beta(v)$ captures the effect of the $k$th covariate on $\ell$ th IC at voxel $v$, $\gamma_i(v)$ is a $q \times 1$ vector reflecting the random variabilities among subjects after adjusting the covariate effects. We assume $\gamma_i(v) \overset{i.i.d.}{\sim}$
N(0, D), i = 1, ..., N, v = 1, ..., V where D = diag(ν²₁, ..., ν²ₗ). By specifying IC-specific variances in D, we allow the between-subject variabilities to vary across ICs. Our proposed covariate ICA model can estimate subpopulation-level spatial source signals for subjects sharing the same covariate measures. It can also allow us to assess the adjusted effects of primary covariates, such as treatment or disease group, while controlling for other potential confounding factors. These strengths demonstrate the importance to build regression models under the ICA paradigm.

2.2.3 Source distribution model

We propose to model the unobserved population-level spatial source signals \( s₀(v) \) in (4) through probability distribution functions (pdf), which help provide a statistical framework for estimation and inference within our model. Following previous work (Guo and Tang, 2013), we choose a mixture of Gaussian distributions (MoG) as our source distribution model. Mathematically, MoG is dense in the functional space of probability distributions (Kostantinos, 2000). In our model for fMRI data, MoG reflects the true spatial distribution patterns of the functional networks in brain (Biswal and Ulmer, 1999), captures various types of non-Gaussian fMRI signals (Xu et al., 1997) and offers tractable likelihood-based estimations (McLachlan and Peel, 2000). Specifically, for a given \( ℓ \), we assume that

\[
s₀ℓ(v) \iid \text{MoG}(\pi ℓ, \mu ℓ, \sigma² ℓ), \quad v = 1, ..., V, \tag{5}
\]

where \( \pi ℓ = [\pi ℓ,1, ..., \pi ℓ,m]’ \) with \( \sum_{j=1}^{m} \pi ℓ,j = 1 \), \( \mu ℓ = [\mu ℓ,1, ..., \mu ℓ,m]’ \) and \( \sigma² ℓ = [\sigma² ℓ,1, ..., \sigma² ℓ,m]’ \) and \( m \) is the number of Gaussian components to construct the mixtures. The pdf of MoG(\( \pi ℓ, \mu ℓ, \sigma² ℓ \)) is \( \sum_{j=1}^{m} \pi ℓ,j g(x; \mu ℓ,j, \sigma² ℓ,j) \) where \( g(\cdot) \) is the pdf of the multivariate Gaussian distribution. In real applications, mixtures of two to three Gaussian components are sufficient to capture the distribution of fMRI spatial signals, with the different Gaussian components corresponding to the background fluctuation and the negative or positive BOLD effects (Beckmann and Smith, 2004; Guo and Pagnoni, 2008; Guo and Pagnoni, 2011).

To facilitate derivations in models involving MoG, it is helpful to define latent states variables (McLachlan and Peel, 2000). Here we define \( z(v) = [z₁(v), ..., zₗ(v)]’ \) at voxel \( v \) as follows. \( zℓ(v), ℓ = 1, ..., q, \) takes value in \( \{1, ..., m\} \) with probability \( p[zℓ(v) = j] = \pi ℓ,j \) for \( j = 1, ..., m \). All elements of \( z(v) \) across all voxels are assumed to be independent. Conditional on \( zℓ(v) \), we have that \( p[s₀ℓ(v)|zℓ(v) = j] = g(s₀ℓ(v); \mu ℓ,j, \sigma² ℓ,j) \). We can generalize this result to all ICs as follows: conditional on \( z(v) \),

\[
s₀(v) = \mu z(v) + ψ z(v), \tag{6}
\]
where \( \mu_{z(v)} = (\mu_{1,z_1(v)}, \ldots, \mu_{q,z_q(v)})', \ \psi_{z(v)} = (\psi_{1,z_1(v)}, \ldots, \psi_{q,z_q(v)})' \) and \( \psi_{z(v)} \sim N(0, \Sigma_{z(v)}) \) with \( \Sigma_{z(v)} = \text{diag}(\sigma_{1,z_1(v)}^2, \ldots, \sigma_{q,z_q(v)}^2) \). We view (6) as the third-level of our hc-ICA model.

### 2.3 Maximum likelihood estimation

We propose to estimate parameters in the hc-ICA model through a maximum likelihood (ML) approach using the EM algorithm. Based on models in (3), (4) and (6), the complete data log-likelihood for our model is

\[
l(\Theta; \mathcal{Y}, \mathcal{X}, S, Z) = \sum_{v=1}^{V} l_v(\Theta; \mathcal{Y}, \mathcal{X}, S, Z),
\]

where \( \mathcal{Y} = \{Y_i : i = 1, \ldots, N\}, \ \mathcal{X} = \{x_i : i = 1, \ldots, N\}, \ S = \{s_i(v) : i = 0, \ldots, N, v = 1, \ldots, V\} \) and \( Z = \{z(v) : v = 1, \ldots, V\}. \) All the parameters in the likelihood function are represented by \( \Theta = \{\beta(v), \{A_i\}, E, D, \{\pi_l\}, \{\mu_i\}, \{\sigma_i^2\}\} \). Specifically,

\[
l_v(\Theta; \mathcal{Y}, \mathcal{X}, S, Z) = \sum_{i=1}^{N} \left[ \log g(y_i(v); A_i, s_i(v), E) + \log g(s_i(v); s_0(v) + \beta(v)' x_i, D) \right] + \log g(s_0(v); \mu_{z(v)}, \Sigma_{z(v)}) + \sum_{l=1}^{q} \log \pi_{l,z_i(v)}. \tag{8}\n\]

#### 2.3.1 The exact EM algorithm

Through this part, we develop an EM algorithm to find the the ML estimates (MLEs) of our model. The EM algorithm iterates between the E-step and the M-step to update the parameter estimates until convergence. We briefly introduce our algorithm in this section. More details about the EM algorithm are presented in Appendix.

**E-step:** In the E-step, given the parameter estimates \( \hat{\Theta}^{(k)} \) from the last step, we derive the conditional expectation of the complete data log-likelihood as follows:

\[
Q(\Theta|\hat{\Theta}^{(k)}) = \sum_{v=1}^{V} E_{{s(v), z(v)|y(v)}} [l_v(\Theta; \mathcal{Y}, \mathcal{X}, S, Z)], \tag{9}\n\]

where \( y(v) = [y_{1(v)}', \ldots, y_{N(v)}']' \) is the vector concatenating all the observed data at voxel \( v, \ s(v) = [s_1(v)', \ldots, s_N(v)', s_0(v)']' \) is the vector containing latent source signals for each subject as well as the whole population. The detailed expression of \( Q(\Theta|\hat{\Theta}^{(k)}) \) is available in Appendix. To evaluate \( Q(\Theta|\hat{\Theta}^{(k)}) \), the joint conditional distributions of \( s(v) \) and \( z(v) \) given our observed data, \( y(v) \), need to be obtained at each voxel. The direct derivation of the conditional distributions through the standard integration approach is cumbersome due to the hierarchical model structure. We develop a fast and convenient approach by collapsing the three-level hierarchical model into a nonhierarchical model and then rewriting it in the form of a classical linear model (please see Appendix A.2 for details). The specific steps to obtain the conditional distributions are summarized in the following. First, we find
p \left[ s(v)|z(v), y(v); \hat{\Theta}^{(k)} \right]$, which is exactly a multivariate Gaussian density (Equation (25) in Appendix). Secondly, we evaluate the probability mass functions, $p \left[ z(v)|y(v); \hat{\Theta}^{(k)} \right]$ through the Bayes’s Theorem (Equation (26) in Appendix). Thirdly, we obtain the joint distribution $p \left[ s(v), z(v)|y(v); \hat{\Theta}^{(k)} \right]$ by convolving the two distributions derived in the previous two steps. Given this joint conditional distribution, all the conditional expectations in (9) are available and they all have analytical forms. Specifically, we only need to evaluate the conditional expectations in the Q-functions such as $E[s_i(v)|y(v); \Theta]$, $E[s_0(v)|y(v); \Theta]$, $E[s_i(v)s_i(v)'|y(v); \Theta]$, $E[s_i(v)s_0(v)'|y(v); \Theta]$, $E[s_0(v)s_0(v)'|y(v); \Theta]$ for $i = 1, ..., N$. These terms are all available in $E[s(v)|y(v); \Theta]$ and $E[s(v)s(v)'|y(v); \Theta]$, which can be given as

$$
E[s(v)|y(v); \Theta] = \sum_{z(v) \in \mathcal{R}} p(z(v)|y(v); \Theta) E[s(v)|y(v), z(v); \Theta],
$$

$$
E[s(v)s(v)'|y(v); \Theta] = \sum_{z(v) \in \mathcal{R}} p(z(v)|y(v); \Theta) E[s(v)|y(v), z(v); \Theta] E[s(v)|y(v), z(v); \Theta]' + \sum_{z(v) \in \mathcal{R}} p(z(v)|y(v); \Theta) \text{Var}[s(v)|y(v), z(v); \Theta],
$$

where $\mathcal{R}$ represents the set of all possible values of $z(v)$ which has a cardinality of $m^2$. Based on what we have presented above, our E-step is fully analytically tractable. No iterative numerical integration routines are needed.

**M-step:** In the M-step, we update the current parameters estimates $\hat{\Theta}^{(k)}$ to

$$
\hat{\Theta}^{(k+1)} = \arg\max_{\Theta} Q(\Theta|\hat{\Theta}^{(k)}).
$$

(10)

We have derived explicit solutions for all parameter updates. The updating rules are as follows:

- **Update $\beta(v)$:** for $v = 1, ..., V$,

$$
\hat{\beta}^{(k+1)} = \left( \sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} \left\{ x_i \left( E[s_i(v)|y(v); \hat{\Theta}^{(k)}] - E[s_0(v)|y(v); \hat{\Theta}^{(k)}] \right) \right\}.
$$

(11)

- **Update $A_i$:** for $i = 1, ..., N$, we let

$$
\tilde{A}_i^{(k+1)} = \left\{ \sum_{v=1}^{V} y_i(v) E[s_i(v)|y(v); \hat{\Theta}^{(k)}] \right\} \left\{ \sum_{v=1}^{V} E[s_i(v)s_i(v)'|y(v); \hat{\Theta}^{(k)}] \right\}^{-1},
$$

and then update $\hat{A}_i^{(k+1)} = \mathcal{H}(\tilde{A}_i^{(k+1)})$ where $\mathcal{H}(\cdot)$ is the orthogonalization transformation.

- **Update $E = I_q \nu_0^2$:**

$$
\nu_0^{2(k+1)} = \frac{1}{TNV} \sum_{v=1}^{V} \sum_{i=1}^{N} \left\{ y_i(v)'y_i(v) - 2y_i(v)'\tilde{A}_i^{(k+1)}E[s_i(v)|y(v); \hat{\Theta}^{(k)}] + \text{tr} \left[ \tilde{A}_i^{(k+1)} \tilde{A}_i^{(k+1)'E[s_i(v)s_i(v)'|y(v); \hat{\Theta}^{(k)}] \right] \right\}.
$$

(13)
Here, $E$ is defined as $E[\{s_{\ell}(v) \mid s_{\ell}(v) = j, y(v) \}; \hat{\Theta}(k)]$.

In fMRI analysis, researchers are often interested in thresholded IC maps to identify activated voxels in each functional network. Following Guo and Tang (2013), we propose a thresholding method based on the mixture distributions for this purpose.

After obtaining $\hat{\Theta}$, we can estimate the population-level source signal maps and their variances based on the mean and variance of the conditional distribution $p[s_0(v) \mid y(v), \hat{\Theta}]$. In addition, we can provide model-based estimates of subject-specific source signal maps and their variances based on $p[s_i(v) \mid y(v), \hat{\Theta}]$. These two distributions are readily available from the marginal of $p[s(v), z(v) \mid y(v), \hat{\Theta}]$. In fMRI analysis, researchers are often interested in thresholded IC maps to identify activated voxels in each functional network. Following Guo and Tang (2013), we propose a thresholding method based on the mixture distributions for this purpose.

One major limitation for the standard EM algorithm is that it requires $O(m^q)$ flops to complete. This is because that, in the E-step, we need to evaluate $p[z(v) \mid y(v), \hat{\Theta}(k)]$ for all the $m^q$ possible values of $z(v)$. The evaluations of certain conditional moments also scale exponentially with regard
Algorithm 1 the exact EM algorithm

**Initial values:** Start with initial values $\Theta^{(0)}$ which can be obtained based on estimates from existing group ICA software.

repeat

E-step:

1. Determine $p[s(v) \mid y(v); \Theta^{(k)}]$:
   1.a Evaluate $p[s(v) \mid y(v), z(v); \Theta^{(k)}]$ according to (25)
   1.b Evaluate $p[z(v) \mid y(v); \Theta^{(k)}]$ according to (26)
   1.c Evaluate $p[s(v) \mid y(v), \Theta^{(k)}]$ by convolving the distributions derived in 1.a and 1.b.

2. Evaluate conditional expectations in $Q(\Theta | \hat{\Theta}^{(k)})$

M-step:

Update $\beta(v), A_i, \pi_{\ell,j}, \mu_{\ell,j}, \sigma_{\ell,j}^2$ according to (11), (12), (15), (16), (17)

Update the variance parameters $D, E$ according to (13), (14)

until $\frac{\|\Theta^{(k+1)} - \Theta^{(k)}\|}{\|\Theta^{(k)}\|} < \epsilon$

To $q$. Consequently, as the number of spatial source signals increases, the algorithm will become extremely slow. Normally the exact EM is not feasible when $q$ is larger than 10 for real data analysis. To solve this problem, in the next part, we propose an approximate EM which runs fast even when the number of spatial sources is large.

2.3.2 The approximate EM algorithm based on subset contraction

In this section, we propose an approximate EM algorithm that requires only $O(mq)$ flops to complete. Like the exact EM, the approximate version is analytically tractable. Denote by $R$ the set containing all possible values of $z(v) = [z_1(v), ..., z_q(v)]'$, which has a cardinality of $m^q$. We enumerate the elements in $R$ as $z^r = [z_1^r, ..., z_q^r]'$ where $z_{\ell}^r \in \{1, ..., m\}$, $r = 1, ..., m^q$ for all $\ell = 1, ..., q$. The key of the approximate algorithm is that instead of considering all the $m^q$ elements in $R$ during the E-step, we only need to focus on a small subset of $R$ for neuro imaging data. The following proposition provides the theoretical justification for the proposed subset contraction approximation.

**Proposition 1.** Suppose that the $q$ random elements in $z(v) = [z_1(v), ..., z_q(v)]'$ are independent with $p[z_{\ell}(v) = j] = \pi_{\ell,j}$. Divide $R$ as follows:

- $R_0 = \{[1, ..., 1]'\}$;
- $R_1 = \{z^r : \exists$ a unique integer $\ell$, s.t., $z_{\ell}^r \in \{2, ..., m\}\}$;
- $R_2 = R \setminus (R_0 \cup R_1)$.

Then for all $0 < \epsilon < 1$, there exists $0 < \delta < 1$ such that when $\pi_{\ell,1} > 1 - \delta$ for all $\ell = 1, ..., q$, we have $p[z(v) \in R_0 \cup R_1] > 1 - \epsilon$. 


The proof of this proposition is relegated to Appendix. This result illustrates the concentration of measures to the subset $\mathcal{R}_0 \cup \mathcal{R}_1$. The two assumptions in the proposition are reasonable in ICA for brain fMRI data. The independence among elements in $\mathbf{z}(v) = [z_1(v), ..., z_q(v)]'$ is a fundamental assumption in the ICA model when the ICs are modeled through the MoG distributions. Furthermore, in the MoG distribution, we can treat that $z_\ell(v) = 1$ as the fact that the $\ell$th latent source signals at voxel $v$ exhibits background fluctuation. Then the condition that $\pi_{\ell,1} > 1 - \delta$ with a small positive value $\delta$ is in fact characterizing the sparse nature of the spatial source signals (McKeown, et al., 1998; Daubechies, et al., 2009, Lee, et al., 2011). That is, within a specific brain functional network (IC), most of the voxels exhibit background fluctuations while only a small number of voxels are activated (or deactivated).

This result implies that when the spatial source signals are independent and sparse, their probability masses will concentrate to a small subspace. Based on this fact, we modify our exact EM algorithm into an approximate version specifically for fMRI data. In the E-step, when evaluating the probability masses, $p[\mathbf{z}(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}]$, we will only consider elements in $\mathcal{R}_0 \cup \mathcal{R}_1$ as nonzero and normalize the relevant probability masses by enforcing $p[\mathbf{z}(v) \in \mathcal{R}_0 \cup \mathcal{R}_1] = 1$. Since $\mathcal{R}_0 \cup \mathcal{R}_1$ only contains $(m-1)q + 1 = \mathcal{O}(mq)$ elements, the approximate EM requires $\mathcal{O}(mq)$ flops to complete.

Comparing with the exact EM, the major change in the approximate EM arises from evaluating the conditional distribution $\mathbf{z}(v) | \mathbf{y}(v)$. We use a sparse vector of probability masses to approximate the exact conditional distributions of $\mathbf{z}(v)$ given $\mathbf{y}(v)$ as defined in (31). The concentration of measures to the subset $\bar{\mathcal{R}} = \mathcal{R}_0 \cup \mathcal{R}_1$ leads to the simplification in evaluating the conditional expectations. For example, in the E-step, under the approximate EM algorithm,

$$E[\mathbf{s}(v) | \mathbf{y}(v); \Theta] = \sum_{\mathbf{z}(v) \in \bar{\mathcal{R}}} \tilde{p}[\mathbf{z}(v) | \mathbf{y}(v); \Theta] E[\mathbf{s}(v) | \mathbf{y}(v), \mathbf{z}(v); \Theta],$$

$$E[s_{0\ell}(v) | z_\ell(v) = j, \mathbf{y}(v); \Theta] = \sum_{\mathbf{z}(v) \in \bar{\mathcal{R}}^{(\ell,j)}} \tilde{p}[\mathbf{z}(v) | \mathbf{y}(v); \Theta] E[s_{0\ell}(v) | \mathbf{y}(v), \mathbf{z}(v); \Theta],$$

(18)

where $\mathcal{R}^{(\ell,j)}$ is defined correspondingly as $\bar{\mathcal{R}}^{(\ell,j)} = \{z^r \in \bar{\mathcal{R}} : z^r_\ell = j\}$. This equations implies that, instead of summing over all possible realizations of the $m^q$ latent states in $\mathcal{R}$ and the $m^{q-1}$ latent states in $\mathcal{R}^{(\ell,j)}$, we only need to perform less than $mq$ summations (since both the cardinality of $\bar{\mathcal{R}}$ and the cardinality of $\bar{\mathcal{R}}^{(\ell,j)}, \forall \ell, j$, are bounded by $mq$) by imposing the approximated conditional distribution $\tilde{p}[\mathbf{z}(v) | \mathbf{y}(v); \Theta]$ as defined in (31) in Appendix. We summarize the approximate EM algorithm as Algorithm 2.
Algorithm 2 the approximate EM algorithm

**Initial values:** Start with initial values \( \Theta^{(0)} \).

repeat

**E-step:**
1. Determine \( \tilde{p}[s(v) \mid y(v); \Theta^{(k)}] \):
   1.a Evaluate \( p[s(v) \mid y(v), z(v); \Theta^{(k)}] \) according to (25)
   1.b Evaluate \( \tilde{p}[z(v) \mid y(v); \Theta^{(k)}] \) according to (31)
   1.c Evaluate \( \tilde{p}[s(v) \mid y(v); \Theta^{(k)}] \) by convolving the distributions derived in 1.a and 1.b.
2. Evaluate conditional expectations in \( Q(\Theta \mid \hat{\Theta}^{(k)}) \) with regard to \( \tilde{p}[s(v) \mid y(v); \Theta^{(k)}] \).

**M-step:**
Update \( \beta(v), A_i, \pi_{\ell,j}, \mu_{\ell,j}, \sigma_{\ell,j}^2 \) according to (11), (12), (15), (16), (17)
Update \( D, E \) according to (13), (14)

until \( \frac{\|\Theta^{(k+1)} - \Theta^{(k)}\|}{\|\Theta^{(k)}\|} < \epsilon \)

2.4 Inference for the covariate effect estimates

For inferential purposes in a fully parameterized model, the inverse of the information matrix is usually used for estimating the asymptotic variance-covariance matrix of the MLEs. Standard EM algorithms only provide parameter estimates without considering the information matrices. Many extensions to the EM algorithm have been developed to estimate the information matrices (Louis, 1982; Meilijson, 1989; Meng and Rubin, 1991). However, for our model, these information matrix based methods can be computationally expensive. First, the dimension of the information matrix for our model is huge because of the large number of parameters, which scales to the number of voxels. Second, the ML estimates, \( \hat{\beta}(v), v = 1, ..., V \), are not independent across voxels because they depend on the same set of parameters such as the mixing matrices. As a result, the information matrix of our model is ultrahigh dimensional and is not sparse, which makes it extremely challenging to invert.

In this part, we propose an efficient method to approximately estimate the asymptotic standard errors of \( \hat{\beta}(v) \) by directly using the output from our EM algorithms. Our method circumvent the computationally expensive matrix inversions in those information matrix based methods. Considering our original model given in (3) and (4), we can collapse the two equations and then left multiplying the orthogonal matrix, \( A_i \), on both sides of the collapsed equation, then we have

\[
A_i'y_i(v) = s_0(v) + X_i \text{vec } [\beta(v)'] + \gamma_i(v) + A_i' e_i(v). \tag{19}
\]

In (19), \( X_i = x_i' \otimes I_q \) and \( \gamma_i(v) + A_i' e_i(v) \sim N(0, W) \) with \( W = D + E \) being an arbitrary \( q \times q \) diagonal matrix. We can rewrite (19) as

\[
y_i^*(v) = X_i \text{vec } [\beta(v)'] + \zeta_i(v), \tag{20}
\]

where \( y_i^*(v) = A_i'y_i(v) - s_0(v) \), \( \zeta_i(v) = \gamma_i(v) + A_i' e_i(v) \). We propose voxel-wise approximation to the
original model by imposing voxel dependent covariance structure \( \xi_i(v) \sim N(0, W(v)) \) in (20). Model (20) can be viewed as a voxel wise multivariate linear regression model with parameter \( \text{vec} [\beta(v)'] \) and \( W(v) \). Our standard error estimates for \( \hat{\beta}(v) \) is then derived from this voxel-wise approximate model, which can be given as follows:

\[
\hat{\text{Var}} \left\{ \text{vec} [\hat{\beta}(v)'] \right\} = \frac{1}{N} \left( \sum_{i=1}^{N} X'_i \hat{W}(v)^{-1} X_i \right)^{-1},
\]

where \( \hat{\beta}(v) \) is the ML estimates based on our EM algorithms while the term \( \hat{W}(v) \) in (21) is given as

\[
\hat{W}(v) = \frac{1}{N} \left[ \sum_{i=1}^{N} \left( y_i^*(v) - X_i \text{vec} [\hat{\beta}(v)'] \right) \left( y_i^*(v) - X_i \text{vec} [\hat{\beta}(v)'] \right)' \right] \odot I_q.
\]

The operator “\( \odot \)" in (22) represents the Hadamard product of matrices. We finally plugging \( \hat{y}_i^*(v) = \hat{A}_i y_i(v) - \hat{s}_0(v) \) into (22) to replace \( y_i^*(v) \). Here, \( \hat{A}_i \) is the ML estimates for the mixing matrices. The estimator \( \hat{s}_0(v) \) is given as the posterior mean, \( E \left[ s_0(v)|y(v), \hat{\Theta} \right] \), which is exactly our estimates for the population-level spatial sources.

Given the variance estimators, hypothesis testings on the covariate effects are performance using the asymptotic Wald tests. Our method can provide voxel wise Z-statistics and p-values for testing whether a specific covariate has significant influences on a brain functional network at each voxel. Various techniques handling multiple comparison issues such as the Bonferroni correction or FDR can be directly applied to our voxelwise p-value maps or Z-statistic maps.

## 3 SIMULATION STUDIES

### 3.1 Simulation study I: performance validation of our model

Within the group ICA framework, dual-regression (Beckmann, et al., 2009) is comparable to ours. The output from the dual-regression can be readily used to estimate the covariate effects. We compare our method with the dual-regression and report the results in this part. We simulate fMRI data with \( q = 3 \) spatial sources for \( N = 10, 20, 40 \) subjects. We generate \( p = 2 \) covariates for each subject with one being categorical \( (x_{i1} \overset{iid}{\sim} \text{Bernoulli}(0.5)) \) and the other being continuous \( (x_{i2} \overset{iid}{\sim} \text{Uniform}(-1, 1)) \). The spatial source signals across subjects consist of 4 slices of images and each slide has 25 \( \times \) 25 voxels, resulting in a total number of \( V = 2500 \) voxels. These subject-specific IC maps are created following (4) where \( s_0(v), \beta(v) \) and \( \gamma_i(v) \) are simulated respectively as follows: generate 3 distinct population-level IC maps of size 25 \( \times \) 25 \( \times \) 4 with the value of each voxel being 4 (activated) or 0 (not activated); for each of the two covariates, generate 3 distinct covariate effect maps of the same size with the value of each voxel being 0, 1.5, 1.8, 2.5 or 3.0; sample the between-subject random variabilities from one of the three noise levels: low \( (D = \text{diag}(0.1, 0.3, 0.5)) \), medium \( (D = \text{diag}(1.0, 1.2, 1.4)) \) and high
\( \mathbf{D} = \text{diag}(1.8, 2.0, 2.5) \).

After the subject-specific IC maps are created, their temporal responses are generated by randomly permutating the phase of three time series of length \( T = 200 \), which are estimated from real fMRI data analysis. Then Gaussian background noise with a variance of 1 is linearly added to the mixed spatial sources to generate one simulated fMRI data matrix of size \( 200 \times 2500 \) for each subject.

Both our method and the dual-regression are implemented for the simulated fMRI data over 100 runs. The performance criteria to compare these two methods are primarily the correlations between the true and estimated signals in both temporal and spatial domains. To compare the estimates for the covariate effects, we report the mean square errors (MSEs) of \( \hat{\beta}(v) \), which is defined as

\[
\text{MSE} \left[ \hat{\beta}(v) \right] = \frac{1}{100V} \sum_{s=1}^{100} \sum_{v=1}^{V} \left\| \hat{\beta}_{(s)}(v) - \beta(v) \right\|_F^2,
\]

(23)

where \( \hat{\beta}_{(s)}(v) \) denote the estimator from the simulated dataset from the \( s \)th run and \( \| \cdot \|_F \) is the Frobenius norm for matrix. We present the results of the correlations and the mean square errors in Table 1. The results demonstrate that hc-ICA provides more accurate estimates compared with the dual-regression for signals from both the spatial and time domains. The covariate effect estimates from hc-ICA have smaller mean square errors. These advantages become more dramatic as the sample size decreases or the the between-subject random variabilities increases. This implies that our method can better adapt to noisy fMRI datasets with smaller sample sizes. To further illustrate the mechanism of the outperformance of hc-ICA, we plots estimated population-level IC maps and the covariate effect maps from both methods in Figure 1. We can see clearly in Figure 1 that the estimates of the population-level IC maps from the dual-regression are heavily interfered by the covariate effects, especially the one for the categorical covariate. The estimated covariate effect maps from the dual-regression also show mismatches across the three spatial sources. These are in sharp contrast to the accurate estimates from hc-ICA.

[Table 1 about here.]

[Figure 1 about here.]

### 3.2 Simulation study II: performance of the approximate EM

In this part, we compare the performance of the exact EM algorithm with the approximate version. We simulate fMRI data with \( q = 3, 6, 10 \) spatial sources for ten subjects with two covariates. The simulated datasets are generated similarly to 3.1. It is worthwhile to note that the between-subject random variabilities now have a variance-covariance matrix \( \mathbf{D} = \mathbf{I}_q \) for all scenarios \( q = 3, 6, 10 \). We
fit the hc-ICA model to the simulated fMRI datasets using both the exact EM and its approximate version over 50 runs. Table 2 presents the results of the simulation study. We evaluate the accuracy of the two algorithms based on the correlations between the estimated population-level IC maps and the truth as well as the mean square errors of $\hat{\beta}(v)$ as defined in [23]. We also report the average computation time in minute for each run and the proportion of converged runs in Table 2. Based on the results in Table 2, we claim that the accuracy of the approximate EM is comparable to that of the exact EM. However, as the number of ICs become larger, the approximate EM runs significantly faster than the exact one.

3.3 Simulation study III: performance of our inference procedures

We examine the performance of our inference procedures for $\hat{\beta}(v)$ through this part. We simulate simple fMRI datasets with $q = 2$ spatial sources and $p = 2$ covariates for $N = 20, 40, 80$ subjects. The covariates are specified the same as in 3.1. To facilitate computation, we only consider $V = 400$ voxels and thus simulate $20 \times 20$ 2D images with similar patterns of the slices in 3.1. The variance of between-subject random variabilities is fixed as 0.5 for both of the two spatial sources. The variance of the noise terms added to the observed signals is 0.8. We implement our method and the dual-regression for the simulated datasets over 1,000 runs. The standard errors for $\hat{\beta}(v)$ from fitting our model are generated using the procedures described in 2.4. For the dual-regression, these standard errors are the direct output of the post-hoc OLS regressions, where the dual-regression estimates for $s_i(v)$ are regressed on the vectors of covariates. For those $\beta_{k\ell}(v) = 0$, we can estimate the Type-I error rate for testing $H_0: \beta_{k\ell}(v) = 0$ versus $H_0: \beta_{k\ell}(v) \neq 0$ at a given significance level, $\alpha$. We report the average of these Type-I error rates in Table 3. For those $\beta_{k\ell}(v) = 1.5, 1.8, 2.5$ and 3.0, we can estimate the test powers when $\alpha$ is fixed as 0.05. The powers for nonzero $\beta_{k\ell}(v)$s with the same values are averaged and the averages are reported in Table 3. We can see from Table 3 that our method can provide more accurate Type-I error rates compared with the dual-regression as the sample size increases. Our method is also generally more powerful than the dual-regression when the covariate effects are larger than 1.5.

4 REAL FMRI DATA APPLICATION

We apply our methods to an fMRI study comparing the neural activity patterns between Zen meditators and controls. In the study, twelve Zen meditators with more than 3 years of daily practice
were recruited along with twelve control subjects who have never practiced meditation. Demographic factors such as age, gender and education level were matched for the two groups. In the study, 50 words and 50 phonologically and orthographically matched nonword items were displayed on a screen in a pseudo-random temporal order. Subjects were asked to respond whether the presented item was a real English word or not via a button-box with their left hand (index finger = “yes”, middle finger = “no”). Subjects were instructed to return attention to their breathing once they would catch themselves mind-wandering. A total number of 520 MR scans were acquired on a 3T Siemens Magnetom Trio system. Each scan contained $53 \times 63 \times 46$ voxels. The images were corrected for slice acquisition timing differences and subject movements, registered and spatially normalized to the MNI standard brain space. We analyzed the Zen meditation data by first performing the preprocessing steps as described in 2.1. The number of ICs was selected to be 14 by Laplace approximation (Minka, 2000). Therefore, we fit our model with 14 spatial sources for 24 subjects using the approximate EM. The covariates of the model consisted of the intercept and the group indicator (1 for the meditators, 0 for the controls). The sign of the output signals are all calibrated by guaranteeing positive correlations between the estimated time courses and the rhythm of experimental stimulus.

Based on the extracted spatial sources, we identify two functional networks of particular interest. Figure 2 and 3 depicts spatial maps for these networks among meditators and controls. The activated brain region in each component include voxels with an estimated conditional probability of activation exceeding 0.95. Specifically, the first network includes supplementary motor area, the hand region of the right sensorimotor cortex (contralateral to the left hand pressing the button) and the visual cortex. We call this network the task-related network (TRN) because the functions of the activated regions in this network are clearly associated with the tasks in this study. Figure 2 also include the $Z$-statistic and $p$-value (both thresholded at $p < 0.01$, the same as in Figure 3) maps for comparing the TRN between the two groups. Results show that meditations demonstrate stronger activation in visual cortex and hand region of the right sensorimotor cortex, but not in the supplementary motor area. Since our task involves responding to visual stimuli via button clicking with left hands, these two brain areas represent the key functional subregions in the TRN. The comparison between the two groups implies more coherent synergy within TRN for the meditators.

Figure 3 shows the second network which includes the posterior cingulate cortex, the medial prefrontal cortex, the lateral parietal cortex and the hippocampus. This network is called the “default mode network (DMN)” (Raichle et al., 2001) which features high metabolism at resting states and decreased activities during tasks. The $Z$-statistic and $p$-value maps for group comparisons show
that meditators display significantly stronger activation in the medial prefrontal cortex and posterior cingulate cortex (PCC) than the controls. These two regions, particularly the PCC, play a central role in the DMN (Greicius et al., 2003; Leech et al., 2011). Our results implies these two regions have stronger functional connectivity among meditators. Figure 3 also shows that meditators have stronger activation in the left lateral parietal cortex (LLPC). The LLPC is known to be associated with language processing and often becomes a more prominent subregion in the DMN, i.e. demonstrating stronger connectivity with the other regions within DMN, when the experimental stimuli are language-based or semantic as is the case with the Zen meditation study. Our findings suggest that compared to control subjects, this language-related subregion in DMN is more functionally connected with the central regions of the DMN such as PCC, again indicating meditators may have stronger functional connectivity within DMN than controls.

[Figure 3 about here.]

Figure 4 demonstrated the multiple comparison results. We applied the procedure introduced by Storey (2002) to conduct FDR corrections. The p-values (thresholded at 0.01) after FDR control at the identified brain regions were plotted for each voxel in Figure 4. The results indicate that in the task-related network, the spatial source signals of the meditators are still significantly different from those of the control, especially at the visual cortex. The results also validate our conclusion that in the default mode network, the meditators have significantly stronger source signals at the medial prefrontal cortex and the posterior cingulate cortex.

[Figure 4 about here.]

5 DISCUSSION

We proposed a hierarchical covariate ICA regression model to systematically study the covariate effects on brain function networks. Numerical analysis justified the reliability and efficiency of the related estimating and inferential procedures for our model. A real fMRI data example illustrated how our methods can be applied to stimulate new findings in neurobiology, psychology and related biomedical fields.

In the approximated EM algorithm, we decrease the computational burdens by concentrating on a subset of the latent class variables. Concentration of measures to this subset can be interpreted as the belief that activation regions within a certain brain functional network do not overlap. This belief is usually true for most fMRI studies. However, we still evaluated the performance of the approximate
EM when the non-overlapping assumption was broken through simulation studies. The criterion we adopt was the correlations of the estimated population-level spatial maps and the truth. We found that the approximate EM could always provide accurate estimates in the non-overlapping regions. For the overlapping regions, the estimates from the approximate EM were slightly biased. However, the overall estimates are acceptable (correlation larger than 0.95) even when the overlapping rate reaches 20%. For our model, the inference procedures are derived through voxel wise approximations to the original model. It involves a few ad-hoc adjustments. By plugging in the estimates for the mixing matrices and population-level spatial signals, we fail to account for the variabilities in their estimates. Thus we may underestimate the standard errors. The slightly inflated Type-I errors in our simulation studies agree with this argument. Some potential directions of future works might be: establishing rigorous asymptotic theories about our model, providing sparse solutions for the covariate effect estimates and extending our model to clustered or longitudinal settings.

References

Beckmann, C. and Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, **23**, 137-152.

Beckmann, C. and Smith, S. (2005). Tensorial extensions of independent component analysis for multi-subject fMRI analysis. *NeuroImage*, textbf{25}, 294-311.

Beckmann, C., Mackay, C., Filippini, N., and Smith, S. (2009). Group comparison of resting-state fMRI using multi-subject ICA and dual regression. *Proc. HBM. San Fransico*.

Biswal, B. and Ulmer, J. (1999). Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of Computer Assisted Tomography* **23**, 265-271.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping* **14**, 140-1151.

Chen, C., Ridler, K., Suckling, J., Williams, S., Fu, C., Merlo-Pich, E., Bullmore, E. (2007). Brain imaging correlates of depressive symptom severity and predictors of symptom improvement after antidepressant treatment. *Biological Psychiatry* **62**, 407-14.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., Ardenne, K. D., Richter, W., Cohen, J. D., and Haxby, J. (2009). Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences* **106**, 10415C10422.
Greicius, M. D., Flores, B., Menon, V., Glover, G., Solvason, H., Kenna, H., Reiss, A. and Schatzberg, A. (2007). Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry* **62**, 429-37.

Greicius, M. D, Krasnow, B, Reiss A. L. and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Science* **100**, 253-258.

Guo, Y. (2011). A general probabilistic model for group independent component analysis and its estimation methods. *Biometrics* **67**, 1531-1542.

Guo, Y. and Pagnoni, G. (2008). A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage* **42**, 1078-1093.

Guo, Y. and Tang, L. (2013). A hierarchical model for probabilistic independent componenet analysis of multi-subject fMRI studies. *Biometrics* (upcoming issues).

Kostantinos, N. Gaussian mixtures and their applications to signal processing. (2000). *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*. Taylor&Francis Group: CRC Press.

Lee, K., Tak, S. and Ye, J. C. (2011), A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion. *IEEE Transactions on Medical Imaging*, **30**, 1076 - 1089.

Lee, S., Shen, H., Truong, Y., Lewis, M. and Huang, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging. *Journal of the American Statistical Association* **106**, 1009-1024.

Leech, R., Kamourieh, S., Beckmann, C ans Sharp, D. (2011). Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *The Journal of Neuroscience* **31**, 3217C3224.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Ser. B* **44**, 226-233.

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping* **6**, 160-188.
McLachlan, G. and Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Ser. B* **51**, 127-138.

Meng, X., Rubin D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899-908.

Sheline, Y., Barch, D., Price, J., Rundle, M., Vaishnavi, S., Snyder, A., Mintun, M., Wang, S., Coalson, R., Raichle, M. (2009). The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences* **106**, 1942-7.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Ser. B* **64**, 479-498.

Xu, L., Cheung, C., Yang, H. and Amari, S. (1997). Maximum equalization by entropy maximization and mixture of cumulative distribution functions. *Proc. ICCN’97* 1821-1826.
Appendix

A 1: The expression of Q-functions

In this section we present details about our EM algorithm.

E-step: The Q-function given in (9) can be expressed as

$$Q(\Theta | \hat{\Theta}^{(k)}) = Q_1(\Theta | \hat{\Theta}^{(k)}) + Q_2(\Theta | \hat{\Theta}^{(k)}) + Q_3(\Theta | \hat{\Theta}^{(k)}) + Q_4(\Theta | \hat{\Theta}^{(k)}),$$

where

$$Q_1(\Theta | \hat{\Theta}^{(k)}) = -\frac{NV}{2} \log |E| - \frac{1}{2} \sum_{v=1}^{V} \sum_{i=1}^{N} \text{tr} \left( E^{-1} \left[ y_i(v) y_i(v)' - 2 A_i E[s_i(v) | y(v); \hat{\Theta}^{(k)}] y_i(v)' \right] \right) + A_i E[s_i(v) s_i(v)' | y(v); \hat{\Theta}^{(k)}] A_i',$$

$$Q_2(\Theta | \hat{\Theta}^{(k)}) = -\frac{NV}{2} \log |D| - \frac{1}{2} \sum_{v=1}^{V} \sum_{i=1}^{N} \text{tr} \left( D^{-1} \left[ E[s_i(v) s_i(v)' | y(v); \hat{\Theta}^{(k)}] + E[s_0(v) s_0(v)' | y(v); \hat{\Theta}^{(k)}] + \beta(v)' x_i x_i' \beta(v) - 2 E[s_i(v) s_0(v)' | y(v); \hat{\Theta}^{(k)}] + 2 E[s_0(v) | y(v); \hat{\Theta}^{(k)}] x_i' \beta(v) \right] \right),$$

$$Q_3(\Theta | \hat{\Theta}^{(k)}) = -\frac{1}{2} \sum_{v=1}^{V} \sum_{\ell=1}^{q} \sum_{j=1}^{m} p[z_{t}(v) = j | y(v); \hat{\Theta}^{(k)}] \left\{ \log \sigma_{\ell,j}^2 + \frac{1}{\sigma_{\ell,j}^2} \left[ \mu_{\ell,j}^2 \right] \right\} + E[s_0(v)^2 | z_{t}(v) = j; y(v), \hat{\Theta}^{(k)}] - 2 \mu_{\ell,j} E[s_0(v) | z_{t}(v) = j, y(v); \hat{\Theta}^{(k)}],$$

$$Q_4(\Theta | \hat{\Theta}^{(k)}) = \sum_{v=1}^{V} \sum_{\ell=1}^{q} \sum_{j=1}^{m} p[z_{t}(v) = j | y(v); \hat{\Theta}^{(k)}] \log \pi_{\ell,j},$$

and $y(v) = [y_1(v)', ..., y_N(v)']'$ contains all the observed data at voxel $v$ (for all the $N$ subjects). To evaluate the Q-functions, we need the joint conditional distribution, $p[s(v), z(v) | y(v); \Theta]$ where $s(v) = [s_1(v)', ..., s_N(v)', s_0(v)']'$.

A 2: The derivation of $p[s(v) | y(v), z(v); \Theta]$ and $p[z(v) | y(v); \Theta]$

To derive the joint distribution $p[s(v), z(v) | y(v); \Theta]$, we need the expression of $p[s(v) | y(v), z(v); \Theta]$ and $p[z(v) | y(v); \Theta]$. To achieve this goal, we first collapse our model defined in (3), (4) and (6) across the $N$ subjects as, for $v = 1, ..., V$,

$$A' y(v) = B x + U \mu_{z(v)} + R r_{z(v)} + e(v),$$

where $r_{z(v)} = [\gamma_1(v)', ..., \gamma_N(v)', \psi_{z(v)}']'$ concatenates error terms in the second and third level models, $e(v) = [e_1(v)', ..., e_N(v)']'$ contains random errors for the first level model across all subjects, $x = [x_1', ..., x_N']'$ represents all the covariate measurements, $B = I_N \otimes \beta(v)'$, $U = 1_N \otimes I_q$, $R = [I_{Nq}, 1_N \otimes I_q]$ and $A = \text{blockdiag}(A_1, ..., A_N)$ is a combined mixing matrix with $A_i$s as its block diagonal.
elements \((A)\) is also orthogonal). It is trivial to have that in \((24), e(v) \sim N(0, \Sigma)\) and \(r_{z(v)} \sim N(0, \Gamma_{z(v)})\) where \(\Sigma = I_N \otimes E\) and \(\Gamma_{z(v)} = \text{blockdiag}(I_N \otimes D, \Sigma_{z(v)})\). Thus \((24)\) can be represent as

\[
y_0(v) \sim N(Rr_{z(v)}, \Sigma), \quad r_{z(v)} \sim N(0, \Gamma_{z(v)})
\]

where \(y_0(v) = A'y(v) - Bx - U'\mu_{z(v)}\). This representation is a canonical Bayesian general linear model given \(z(v)\). Then given \(z(v)\) and conditional on \(y(v), p[r_{z(v)} | y(v), z(v); \Theta] = g(\mu_{r(v)|y(v)}, \Sigma_{r(v)|y(v)})\) where

\[
\mu_{r(v)|y(v)} = \Sigma_{r(v)|y(v)} R' \Sigma^{-1}[A'y(v) - Bx - U'\mu_{z(v)}],
\]
\[
\Sigma_{r(v)|y(v)} = (R' \Sigma^{-1} R + \Gamma_{z(v)}^{-1})^{-1}.
\]

Note that from equation \((4)\) and \((6)\), \(s(v) = Pr_{z(v)} + Q_{z(v)},\) where

\[
P = \begin{pmatrix} I_{Nq} & U \\ 0 & I_q \end{pmatrix}, \quad Q_{z(v)} = \begin{pmatrix} Bx + U'\mu_{z(v)} \\ \mu_{z(v)} \end{pmatrix},
\]

we can easily have that:

\[
p[s(v) | y(v), z(v); \Theta] = g(P \mu_{r(v)|y(v)} + Q_{z(v)}, P \Sigma_{r(v)|y(v)} P') \quad (25).
\]

Next we need to find \(p[z(v) | y(v); \Theta].\) From \((24),\) we have that \(p[A'y(v) | z(v)] = g(Bx + U'\mu_{z(v)}, R_{z(v)} R' + \Sigma).\) Notice that \(p[z(v)] = \prod_{\ell=1}^q \pi_{\ell,z(v)}\) for all \(v,\) by simply applying the Bayes' theorem,

\[
p[z(v) | y(v); \Theta] = \frac{\left[\prod_{\ell=1}^q \pi_{\ell,z(v)}\right] g(A'y(v); Bx + U'\mu_{z(v)}, R_{z(v)} R' + \Sigma)}{\sum_{z(v) \in R} \left[\prod_{\ell=1}^q \pi_{\ell,z(v)}\right] g(A'y(v); Bx + U'\mu_{z(v)}, R_{z(v)} R' + \Sigma)}, \quad (26)
\]

where \(R\) is the range of \(z(v) = [z_1(v), ..., z_q(v)]', z_t(v) = 1, ..., m,\) which contains \(m^q\) distinct vectors in \(\mathbb{R}^q.\)

**A 3: Proof of Proposition 1**

We prove Proposition 1 by introducing a lemma.

**Lemma 1.** If the elements of \(z(v) = [z_1(v), ..., z_q(v)]'\) are independent with \(p[z_\ell(v) = j] = \pi_{\ell,j}\) for \(j = 2, ..., m, \ell = 1, ..., q,\) then

\[
p[z(v) \in R_0 \cup R_1] = \frac{1 + \sum_{\ell=1}^q \kappa_{\ell}}{\prod_{\ell=1}^q (1 + \kappa_{\ell})}, \quad (27)
\]

where \(\kappa_{\ell} = \sum_{j=2}^m \tau_{\ell,j}\) and \(\tau_{\ell,j} = \pi_{\ell,j}/\pi_{\ell,1}.\)

**Proof.** By definition \(R_1 \cap R_0 = \emptyset\) and \(p[z(v) \in R_0] = \prod_{\ell=1}^q p[z_\ell(v) = 1] = \prod_{\ell=1}^q \pi_{\ell,1}.\) For a given \(z(v) \in R_1,\) suppose \(z_t(v) = j > 1\) for \(t = 1, ..., q\) and \(z_{t\neq t}(v) = 1,\) then \(p[z(v)] = \tau_{t,j} \prod_{\ell=1}^q \pi_{\ell,1}.\) This implies that

\[
p[z(v) \in R_1] = \left(\sum_{t=1}^q \sum_{j=2}^m \tau_{t,j}\right) \prod_{\ell=1}^q \pi_{\ell,1} = \left(\sum_{\ell=1}^q \kappa_{\ell}\right) \prod_{\ell=1}^q \pi_{\ell,1}.
\]
Also we have that \( \sum_{j=1}^{m} \pi_{\ell,j} = 1 \) for all \( \ell = 1, \ldots, q \), then \( \pi_{\ell,1} + \pi_{\ell,1} \sum_{j=2}^{m} \tau_{\ell,j} = (1 + \kappa_{\ell})\pi_{\ell,1} = 1 \), which gives \( \pi_{\ell,1} = 1/(1 + \kappa_{\ell}) \). Thus
\[
p[z(v) \in R_0 \cup R_1] = p[z(v) \in R_0] + p[z(v) \in R_1] \\
= \left(1 + \sum_{\ell=1}^{q} \kappa_{\ell}\right) \prod_{\ell=1}^{q} \pi_{\ell,1} \\
= \frac{1 + \sum_{\ell=1}^{q} \kappa_{\ell}}{\prod_{\ell=1}^{q}(1 + \kappa_{\ell})} \tag{28}
\]

Now we prove the Proposition 1.

\textbf{Proof.} Based on the definition of \( \tau_{\ell,j} \) in Lemma 1,
\[
\kappa_{\ell} = \sum_{j=2}^{m} \tau_{\ell,j} = \frac{p[z_{\ell}(v) \neq 1]}{p[z_{\ell}(v) = 1]} = \frac{1 - \pi_{\ell,1}}{\pi_{\ell,1}}. \tag{29}
\]
For all \( 0 < \epsilon < 1 \), let \( \delta = \frac{\sqrt{\epsilon}}{\sqrt{\epsilon} + q} \in (0, 1) \). Then if \( \pi_{\ell,1} > 1 - \delta \), we have that \( 0 < \kappa_{\ell} < \frac{\delta}{1-\delta} \) for all \( \ell = 1, \ldots, q \). Then based on the Taylor expansion for \( p[z(v) \in R_0 \cup R_1] = F(\kappa_1, \ldots, \kappa_q) = \frac{1+\sum_{\ell=1}^{q} \kappa_{\ell}}{\prod_{\ell=1}^{q}(1 + \kappa_{\ell})} \)
at 0, \( \exists 0 < \kappa_0 < \kappa_{\ell} \) for all \( \ell = 1, \ldots, q \), such that
\[
p[z(v) \in R_0 \cup R_1] = 1 + \sum_{\ell=1}^{q} \left. \frac{\partial F}{\partial \kappa_{\ell}} \right|_{\kappa_{\ell} = \kappa_0} \kappa_{\ell} \\
= 1 - \sum_{\ell=1}^{q} \sum_{j \neq \ell} \frac{\kappa_0}{\prod_{j \neq \ell}(1 + \kappa_0)} \frac{1}{(1 + \kappa_0)^2} \kappa_{\ell} \\
> 1 - \sum_{\ell=1}^{q} \sum_{j \neq \ell} \frac{\kappa_0^2}{\prod_{j \neq \ell}(1 + \kappa_0)} \\
> 1 - \left( \frac{q \delta}{1-\delta} \right)^2 = 1 - \epsilon \tag{30}
\]
\( \square \)

\textbf{A 4: Remarks on the approximate EM}

In the approximate EM, the conditional distribution \( z(v) \mid y(v) \) is determined by the probability masses
\[
\tilde{p}[z(v) \mid y(v), \Theta] = \begin{cases} 
\frac{[\prod_{\ell=1}^{q} \pi_{\ell,z_{\ell}(v)}]g(A'y(v); \textbf{B}_x + \textbf{U} \mu_{z_{\ell}(v)}) \textbf{R}' + \textbf{Y}}{\sum_{z(v) \in \tilde{R}} [\prod_{\ell=1}^{q} \pi_{\ell,z_{\ell}(v)}]g(A'y(v); \textbf{B}_x + \textbf{U} \mu_{z_{\ell}(v)}) \textbf{R}' + \textbf{Y}}, & z(v) \in \tilde{R} \\
0, & z(v) \in \textbf{R} \setminus \tilde{R}
\end{cases}
\tag{31}
\]
where \( \tilde{R} = R_0 \cup R_1 \). Thus we use a sparse vector of probability masses, with concentration of measures on the subset \( \tilde{R} = R_0 \cup R_1 \), to approximate the exact conditional distribution of \( z(v) \) given
$y(v)$. The follow-up evaluations of the conditional moments in the E-step only involves $z(v) \in \tilde{R}$. And the corresponding definition of $R^{(\ell,j)}$ is adapted to $\tilde{R}^{(\ell,j)} = \{z^r \in \tilde{R} : z^r_\ell = j\}$. 

23
Table 1: Simulation results for comparing our hc-ICA method against the dual-regression. Values presented are mean and standard deviation of correlations between the true and estimated: Subject-specific spatial maps, Population-level spatial maps and Subject-specific time courses. The mean and standard deviation of the MSE of the covariate estimates are also provided.

| Btw-subj Var | Population-level spatial maps | Subject-specific spatial maps |
|--------------|--------------------------------|-----------------------------|
|              | Corr.(SD)                      | Corr.(SD)                   |
|              | hc-ICA                        | Dual.Reg.                   | hc-ICA                        | Dual.Reg.                   |
| Low          |                                |                             |                               |
| N=10         | 0.982 (0.003)                  | 0.955 (0.017)               | 0.984 (0.004)                 | 0.945 (0.023)               |
| N=20         | 0.990 (0.002)                  | 0.956 (0.002)               | 0.996 (0.002)                 | 0.949 (0.008)               |
| N=40         | 0.992 (0.002)                  | 0.955 (0.003)               | 0.996 (0.001)                 | 0.956 (0.002)               |
| Medium       |                                |                             |                               |
| N=10         | 0.942 (0.017)                  | 0.911 (0.107)               | 0.943 (0.011)                 | 0.882 (0.030)               |
| N=20         | 0.954 (0.002)                  | 0.916 (0.059)               | 0.959 (0.004)                 | 0.890 (0.016)               |
| N=40         | 0.961 (0.002)                  | 0.917 (0.024)               | 0.968 (0.003)                 | 0.893 (0.009)               |
| High         |                                |                             |                               |
| N=10         | 0.833 (0.146)                  | 0.726 (0.365)               | 0.894 (0.108)                 | 0.689 (0.303)               |
| N=20         | 0.850 (0.129)                  | 0.745 (0.273)               | 0.909 (0.084)                 | 0.695 (0.281)               |
| N=40         | 0.871 (0.055)                  | 0.747 (0.206)               | 0.928 (0.035)                 | 0.705 (0.259)               |
| Btw-subj Var | Subject-specific time courses  | Covariate Effects           |
|              | Corr.(SD)                      | MSE(SD)                     |
|              | hc-ICA                        | Dual.Reg.                   | hc-ICA                        | Dual.Reg.                   |
| Low          |                                |                             |                               |
| N=10         | 0.998 (0.001)                  | 0.987 (0.010)               | 0.048 (0.019)                 | 0.154 (0.055)               |
| N=20         | 0.998 (0.001)                  | 0.995 (0.004)               | 0.021 (0.003)                 | 0.127 (0.044)               |
| N=40         | 0.998 (0.001)                  | 0.994 (0.004)               | 0.012 (0.001)                 | 0.111 (0.030)               |
| Medium       |                                |                             |                               |
| N=10         | 0.993 (0.010)                  | 0.970 (0.028)               | 0.273 (0.088)                 | 0.485 (0.151)               |
| N=20         | 0.998 (0.003)                  | 0.976 (0.016)               | 0.117 (0.015)                 | 0.285 (0.076)               |
| N=40         | 0.998 (0.002)                  | 0.991 (0.008)               | 0.064 (0.005)                 | 0.187 (0.041)               |
| High         |                                |                             |                               |
| N=10         | 0.948 (0.021)                  | 0.903 (0.045)               | 0.387 (0.157)                 | 0.783 (0.325)               |
| N=20         | 0.978 (0.018)                  | 0.925 (0.029)               | 0.224 (0.075)                 | 0.532 (0.271)               |
| N=40         | 0.990 (0.015)                  | 0.934 (0.022)               | 0.131 (0.056)                 | 0.389 (0.198)               |
Table 2: Simulation results for comparing the exact EM and its approximated version based on 50 runs. Mean and standard deviation of correlations between the true and estimated population-level spatial maps are presented. The mean and standard deviation of the MSE of the covariate estimates are also provided. Both algorithms failed to converge in 2 common cases.

| # of IC | Population-level spatial maps | Covariate Effects |
|---------|-------------------------------|-------------------|
|         | Corr(SD)                      | MSE(SD)           |
| q=3     | Exact EM                     | 0.981(0.003)      | 0.048(0.020) |
|         | Approx. EM                   | 0.981(0.001)      | 0.048(0.019) |
| q=6     | Exact EM                     | 0.980(0.006)      | 0.069(0.024) |
|         | Approx. EM                   | 0.980(0.006)      | 0.070(0.022) |
| q=10    | Exact EM                     | 0.969(0.022)      | 0.105(0.033) |
|         | Approx. EM                   | 0.963(0.020)      | 0.112(0.028) |

| # of IC | Time in minute | Proportions of Convergence |
|---------|----------------|---------------------------|
|         | Exact EM       | Approx. EM                |
| q=3     | 9.91           | 5.22                      | 100% | 100% |
| q=6     | 71.05          | 9.09                      | 100% | 100% |
| q=10    | 860.10         | 19.02                     | 96%  | 96%  |

Table 3: Simulation results for the inference of $\beta(v)$ based on 1000 runs. Type-I errors and powers are averaged across voxels having the same values of $\beta_{k\ell}(v)$.

|                  | N=20 | N=40 | N=80 |
|------------------|------|------|------|
|                  | hci-ICA | Dual.Reg. | hci-ICA | Dual.Reg. | hci-ICA | Dual.Reg |
| Type-I error analysis: |
| Test size $\alpha$ |      |      |      |
| $\alpha = 0.01$   | 0.016 | 0.029 | 0.012 | 0.025 | 0.012 | 0.018 |
| $\alpha = 0.05$   | 0.067 | 0.084 | 0.060 | 0.076 | 0.056 | 0.062 |
| $\alpha = 0.10$   | 0.140 | 0.205 | 0.125 | 0.190 | 0.117 | 0.149 |
| $\alpha = 0.50$   | 0.547 | 0.580 | 0.536 | 0.565 | 0.521 | 0.557 |
| $\alpha = 0.80$   | 0.851 | 0.872 | 0.842 | 0.856 | 0.825 | 0.840 |

Power analysis (test size $\alpha = 0.05$):

| Covariate effect | N=20 | N=40 | N=80 |
|------------------|------|------|------|
|                  | hci-ICA | Dual.Reg. | hci-ICA | Dual.Reg. | hci-ICA | Dual.Reg |
| $\beta_{k\ell}(v) = 1.5$ | 0.142 | 0.130 | 0.257 | 0.203 | 0.370 | 0.284 |
| $\beta_{k\ell}(v) = 1.8$ | 0.269 | 0.224 | 0.460 | 0.390 | 0.751 | 0.548 |
| $\beta_{k\ell}(v) = 2.5$ | 0.581 | 0.475 | 0.836 | 0.705 | 0.930 | 0.839 |
| $\beta_{k\ell}(v) = 3.0$ | 0.892 | 0.845 | 1.000 | 0.922 | 1.000 | 1.000 |
Figure 1: Comparison between our method and the dual-regression: truth, estimates from our model, estimates from the dual regression (N=10, between-subject variabilities are high) based on 100 runs. All the images displayed are averaged across the 100 Monte Carlo data sets. Population-level spatial maps are shown in Figure 1(A). The results of the dual-regression are interfered by the covariate effects for the categorical covariate $X_1$. The results from our method are accurate. Covariate effect estimates are shown in Figure 1(B1) and Figure 1(B2) respectively. The results of the dual-regression show clear mismatching while our method provide accurate estimates.
(A) Images for the supplementary motor area and the visual cortex

(B) Images for the supplementary motor area and the right sensorimotor cortex (contralateral to the left hand)

Figure 2: Results for the task-related network. The subpopulation-level maps for the controls and meditators as well as the parametric Z-statistic map and p-value map for testing the differences are presented. Z-statistic and p-value maps thresholded at $p < 0.01$.

Figure 3: Results for the default mode network. The subpopulation-level maps for the controls and meditators as well as the parametric Z-statistic map and p-value map for testing the differences are presented. Z-statistic and p-value maps thresholded at $p < 0.01$. 
Figure 4: FDR controlled $p$-values at voxels for the identified brain regions. FDR controled $p$-value maps thresholded at $p < 0.01$. 

(A) Task-related network: the visual cortex

(B) Task-related network: the right sensorimotor cortex

(C) The default mode network