Learning Robust Hierarchical fMRI across Sites

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-site fMRI studies face the challenge that the pooling introduces systematic non-biological site-specific variance due to hardware, software, and environment. In this paper, we propose to reduce site-specific variance in the estimation of hierarchical Sparsity Connectivity Patterns (hSCPs) in fMRI data via a simple yet effective matrix factorization while preserving biologically relevant variations. Our method leverages unsupervised adversarial learning to improve the reproducibility of the components. Experiments on simulated datasets display that the proposed method can estimate components with higher accuracy and reproducibility, while preserving age-related variation on a multi-center clinical data set.

1 Introduction

Multi-site fMRI studies have gained a lot of interest over the last decade [1, 2]. One reason for this is the necessity to evaluate a hypothesis in multiple settings/sites and make the hypothesis result generalizable to a diverse population. Also, the pooling of data is essential when studying rare disorders or neurological conditions where the aim is to generalize the results to diverse populations [3, 4]. However, the data pooling often results in the introduction of non-biological systematic variance due to differences in scanner hardware and imaging acquisition parameters [5]. This additional variability can lead to spurious results and a decrease in statistical power. The variability can also result in hindrance in the estimation of true biological changes or in inferring non-biological differences as biological because of the correlation between site effects and biological predictors. Many studies working with multi-site data fMRI have reported considerable variability due site or scanner effects [6, 7, 1].

The non-biological variability introduced due to inter-scanner and inter-protocol variations can affect the estimation of the common features derived from fMRI [8], such as functional connectivity [5] or sparse hierarchical factors (this paper). These features were used for the study of the brain's function during aging [9], of various neurological disorders [10, 11], and tasks [12]. The non-biological variability can considerably reduce these features' reproducibility across different datasets and hence their utility as biomarkers for diseases that disrupt functional connectivity. Thus the removal of non-biological variance introduced by pooling of the data is essential for many neuroimaging studies.

Independent Component Analysis [13], Non Negative Matrix Factorization [14], and Sparse Dictionary Learning [15] are some of the commonly used methods for estimating interpretable components that capture the complexity of the human brain function at rest. However, these methods do not account for the hierarchical organization of the human brain [16, 17]. Recently, deep learning based methods [18, 19] have been developed to estimate hierarchical networks. However, they fail to capture overlapping components and the existence of positively and negatively correlated nodes in a component that are shown to capture the underlying structure better [20, 21]. Hierarchical Sparse Connectivity Patterns (hSCPs) overcome these limitations by capturing interpretable hierarchical, sparse, and overlapping components using a linear deep matrix factorization approach. The method not only captures shared patterns but also the subject-specific weights of these patterns, thus capturing

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the heterogeneity of brain activity patterns across individuals. In this paper, we focus on robust estimation of hSCPs [21–23] in a multi-site regime.

Many existing methods to reduce site effects are based on an empirical Bayes method ComBat [24], which was developed to remove 'batch effects' in genetics and has been applied for harmonizing different measures derived from structural [25, 26] and functional MRI [8]. But these harmonization methods can not be directly applied to the hSCP method because of the loss in the structure of captured heterogeneity (more details in section 2.1). See Appendix A for extended discussion on related work.

Contribution. In this paper, we develop a new model that is robust to site-effects in the estimation of sparse hierarchical connectivity pattern components (rshSCP). For this, the method learns site-specific features and global space, storing the information about the scanner and site, and uses these features to reduced site effects in the components. We also use an existing adversarial learning approach [23] on top of our method to improve the reproducibility and generalizability of the components across components from the same site. We formulate the method as a non-convex optimization problem which is solved using stochastic gradient descent. Experiments on simulated and real datasets show that our method can improve the split-sample and leave one site reproducibility of the components while retaining age-related biological variability in the data, thus capturing informative heterogeneity.

2 Method

In this section, we first present a brief overview of the hierarchical Sparse Connectivity Pattern (hSCP) and its adversarial formulation. We then describe our main method and the joint formulation incorporating adversarial learning. We follow the notation in [22] and [23]. The set of symmetric positive definite matrices of size $P \times P$ is denoted by $\mathbb{S}_{++}^{P\times P}$. Matrix $\mathbf{A}$ with all the elements greater than or equal to 0 is denoted by $\mathbf{A} \geq 0$. $\mathbf{J}_P$ denotes $P \times P$ matrix with all elements equal to one. $P \times P$ identity matrix is denoted by $\mathbf{I}_P$ and element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \circ \mathbf{B}$.

2.1 Introduction to hierarchical Sparse Connectivity Patterns

Hierarchical Sparse Connectivity Patterns (hSCP) [22] is a hierarchical extension of Sparse Connectivity Patterns (SCPs), first defined by [21] to estimate sparse functional patterns in the human brain using fMRI data. Let there be N number of subjects or participants, and each subject's BOLD fMRI time series has T time points and P nodes representing regions of interest. The input to hSCP are correlation matrices $\mathbf{\Theta}^n \in \mathbb{S}_{++}^{P\times P}$ where ith and jth element of the matrix is the correlation between time series of ith and jth node. hSCP then outputs a set of shared hierarchical patterns following the below equations:

$$\mathbf{\Theta}^n \approx \mathbf{W}_1\mathbf{\Lambda}_1^n\mathbf{W}_1^\top, \quad \ldots \quad \mathbf{\Theta}^n \approx \mathbf{W}_1\mathbf{W}_2\ldots\mathbf{W}_K\mathbf{\Lambda}_K^n\mathbf{W}_K^\top\mathbf{W}_{K-1}^\top\ldots\mathbf{W}_1^\top,$$

where $\mathbf{\Lambda}_k^n$ is a diagonal matrix having positive elements storing relative contribution of the components for the nth subject at kth level, K is the depth of hierarchy and $P > k_1 > \ldots > k_K$ i.e. each successive level in the hierarchy has less number of components than the previous one. In the above formulation, $\mathbf{W}_1 \in \mathbb{R}^{P\times k_1}$ stores k_1 components at the bottom most level, and each successive multiplication by $\mathbf{W}_2, \mathbf{W}_3, \ldots, \mathbf{W}_K$ linearly transforms to a lower dimensional space of $k_2, k_3, \ldots, k_K$ dimension. Let $\mathcal{W} = \{\mathbf{W}_r \mid r = 1, \ldots, K\}$ be the set storing sparse components shared across all subjects and $\mathcal{D} = \{\mathbf{\Lambda}_r^n \mid r = 1, \ldots, K; n = 1 \ldots, N\}$ be set storing subject specific diagonal matrix with $\mathbf{\Lambda}_r^n \geq 0$. The hierarchical components are estimated by solving the below problem:

$$\begin{aligned} \min_{\mathcal{W},\mathcal{D}} \quad & H(\mathcal{W},\mathcal{D},\mathcal{C}) = \sum_{n=1}^{N}\sum_{r=1}^{K} \|\mathbf{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{\Lambda}_r^n(\prod_{j=1}^{r}\mathbf{W}_n)^\top\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{w}_l^r\|_1 < \lambda_r,\ \|\mathbf{w}_l^r\|_\infty \leq 1,\ \mathrm{trace}(\mathbf{\Lambda}_r^n) = 1, \\ & \mathbf{\Lambda}_r^n \geq 0,\ \forall n, r, l; \qquad \mathbf{W}_j \geq 0,\ j = 2, \ldots, K, \end{aligned} \tag{1}$$

where $\mathcal{C} = \{\mathbf{\Theta}^n \mid n = 1, \ldots, N\}$, $l = 1, \ldots, k_r,\ n = 1, \ldots, N$ and $r = 1, \ldots, K$. L_1, L_∞ and trace constraints help the problem to identify a decomposition which can provide reproducible components. More details can be found in the original paper [22]. We will be denoting above

constraint set as $\Omega_{\mathcal{W}} = \{\mathbf{W} \mid \|\mathbf{w}_l^r\|_1 < \lambda_r,\ \|\mathbf{w}_l^r\|_\infty \leq 1, \mathbf{W}_j \geq 0,\ j = 2, \ldots, K\}$ and $\Psi = \{\mathbf{\Lambda} \mid \mathrm{trace}(\mathbf{\Lambda}_r^n) = 1, \mathbf{\Lambda}_r^n \geq 0\}$. A detailed description of the method can be found in [22].

Can we use standard harmonization approaches? These methods reduce site effects by adjusting for additive and multiplicative effects for each feature in data separately and use emperical Bayes estimates the model parameters. These methods can be used in the case of hSCP in two ways. First, harmonization can be directly applied to each element of the correlation matrices, which is the input of hSCP. This will reduce site effects from each element of the correlation matrix, thus from the complete input, but the final matrix that does not necessarily follow the essential property of a correlation matrix i.e., positive definiteness. Second, harmonization can be directly applied to $\mathbf{\Lambda}$ to remove site effects. To understand this, we look at the hSCP formulation at one level:

$$\mathbf{\Theta}^n \approx \sum_{l=1}^{k} d_l^n \mathbf{w}_l \mathbf{w}_l^\top \approx \mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^\top,$$

where d_l^n are non-zero elements storing the subject-specific information, which can be affected by the variability introduced by the site. In this model, harmonizing each feature across different sites will change the relative contribution of the components in each subject's functional structure, which will remove the interpretability of the subject-wise weights, which is not desirable.

2.2 Adversarial Learning in hSCP

Adversarial learning has shown to achieve state-of-the-art performance of various matrix factorization approaches [27, 28]. Recently, Sahoo and Davatzikos [23] demonstrated that incorporating adversarial learning in the estimation problem of hSCP can improve the reproducibility of the hierarchical components. The method is based on perturbation of input data $\mathbf{\Theta}^n$ to learn stable components robust to adversaries. First the input data is perturbed to generate new data $\mathbf{\Gamma}^n = \mathbf{\Theta}^n + \sigma\mathbf{1}_P$ where σ is the standard deviation of the data. Then the perturbed set of components $\tilde{\mathbf{W}}_1$ are estimated using the new data and are ensured to be close to $\mathbf{W}_1$ by solving the below minimization problem:

$$A(\hat{\mathbf{W}}_1) = \alpha\|\tilde{\mathbf{W}}_1 - \mathbf{W}_1\|_F^2 + \sum_{n=1}^{N} \|\mathbf{\Gamma}^n - \tilde{\mathbf{W}}_1\mathbf{\Lambda}^n\tilde{\mathbf{W}}_1^\top\|_F^2. \tag{2}$$

In the above optimization problem, the first term constraints $\tilde{\mathbf{W}}_1$ to be close to $\mathbf{W}_1$ and second term is used for learning the components using the perturbed data. Aim of the learner is to estimate $\mathbf{\Lambda}^n$ and $\mathbf{W}_1$ by minimizing the below cost function:

$$D(\mathbf{W}, \mathbf{\Lambda}) = \sum_{n=1}^{N} \|\mathbf{\Theta}^n - \tilde{\mathbf{W}}_1\mathbf{\Lambda}^n\tilde{\mathbf{W}}_1^\top\|_F^2 + \beta\sum_{n=1}^{N} \|\mathbf{\Theta}^n - \mathbf{W}_1\mathbf{\Lambda}^n\mathbf{W}_1^\top\|_F^2, \tag{3}$$

for a fixed $\tilde{\mathbf{W}}_1$. Learner first estimates subject specific information $\mathbf{\Lambda}^n$ using perturbed weight matrix and then use it to learn $\mathbf{W}_1$. Below is the optimization problem using adversarial learning of hSCP:

$$\begin{aligned} \min_{\mathcal{W},\mathcal{D}} \quad & H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C}) + \beta H(\mathcal{W}, \mathcal{D}, \mathcal{C}) \\ \text{s.t.} \quad & \tilde{\mathbf{W}}_r = \arg\min_{\hat{\mathbf{W}}_r} \alpha\|\hat{\mathbf{W}}_r - \mathbf{W}_r\|_F^2 + H(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{P}) \quad r = 1, \ldots, K \\ & \tilde{\mathbf{W}}_r, \mathbf{W}_r \in \Omega, \qquad \mathcal{D} \in \Psi, \end{aligned} \tag{4}$$

where $\tilde{\mathcal{W}} = \{\tilde{\mathbf{W}}_r \mid r = 1, \ldots, K\}$ and $\mathcal{P} = \{\mathbf{\Gamma}^n \mid n = 1, \ldots, N\}$. We next discuss our main method which aims to reduce the site effects.

2.3 Robust to site hSCP

Estimating hSCPs in multi-site data can introduce non-biological variances in the components and the subject-specific information. One of the typical approaches would be to use harmonization methods mentioned previously, but it would lead to a loss in the structure of these features, which in turn will lose interpretability. Instead of removing site effects after estimating the components, we jointly

model the sparse components and the site effects, and estimate robust to site hSCP (rshSCP). We first look at the case when there is only one level in the hierarchy, which then can be extended to multiple levels. Let there be total S sites, $\mathcal{I}_s$ be the set storing subjects from site s and $\mathbf{y} \in \mathbb{R}^{N \times S}$ be the one-hot encoded site labels We hypothesize that there is a space $\mathbf{V}$ storing site and scanner information for all the possible available data, and for each site s, we have space $\mathbf{U}^s$ storing site-specific information for $s = 1, \ldots, S$. Based on the above hypothesis, we decompose the correlation matrix $\mathbf{\Theta}^n$ of $n \in \mathcal{I}_s$ to jointly estimate the hSCPs, $\mathbf{U}^s$ and $\mathbf{V}$ as:

$$\mathbf{\Theta}^n \approx \underbrace{\mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^\top}_{\substack{\text{decomposition of}\\ \text{subject components}}} + \underbrace{\mathbf{U}^s\mathbf{V}}_{\substack{\text{decomposition of}\\ \text{site components}}}, \tag{5}$$

where $\mathbf{U}^s$ is constrained to be a diagonal matrix and L_1 sparsity constraint is used for $\mathbf{V}$ to prevent overfitting. In addition to estimating the site effects, we train the model such that the predictive power of subject-specific information $\mathcal{D}$ for predicting site is reduced, which can assist in removing site information. For this, we train a differentiable classification model $F(\zeta, \mathcal{D})$ parameterized by ζ with input $\mathbf{\Lambda}^n$ that return site predictions $\hat{\mathbf{y}} \in \mathbb{R}^{N \times S}$. These predictions indicate the probabilities that each of N inputs belongs to each of S site labels. The classification model is trained by optimizing for ζ such that the cross-entropy loss $\mathcal{L}(\zeta, \mathcal{D}, \mathbf{y})$ between the predictions $\hat{\mathbf{y}}$ and the true site labels $\mathbf{y}$ is minimized:

$$\zeta^* = \arg\min_{\zeta} -\frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{S} y_{n,s} \log \hat{y}_{n,s}. \tag{6}$$

Using this model, we modify $\mathbf{\Lambda}^n$ such that its predictability power reduces. We achieve this by maximizing the above loss with respect to $\mathcal{D}$. This will result in a minimax game, where the classifier learns to minimize the cross-entropy or the surrogate classification loss, and $\mathcal{D}$ is adjusted to maximize the loss. The joint optimization problem can be written as:

$$\begin{aligned} \max_{\zeta} \min_{\mathbf{W},\mathcal{D},\mathcal{U},\mathbf{V}} \quad & \sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s} \|\mathbf{\Theta}^n - \mathbf{W}\mathbf{\Lambda}^n\mathbf{W}^\top - \mathbf{U}^s\mathbf{V}\|_F^2 + \gamma\mathcal{L}(\zeta, \mathcal{D}, \mathbf{y}) \\ s.t. \quad & \mathbf{W} \in \Omega, \quad \mathcal{D} \in \Psi, \quad \|\mathbf{v}_p\|_1 < \mu,\ p = 1, \ldots, P, \end{aligned} \tag{7}$$

where $\mathcal{U} = \{U_s | s = 1, \ldots, S\}$, $\mathbf{v}_p$ is the pth column of $\mathbf{V}$.

2.4 Complete Model

We can combine the above formulation (7) at multi level with the adversarial learning (4) to jointly model hSCPS and site effects. Let

$$G(\mathcal{W}, \mathcal{D}, \mathcal{C}) = \sum_{s=1}^{S}\sum_{n\in\mathcal{I}_s}\sum_{r=1}^{K} \|\mathbf{\Theta}^n - (\prod_{j=1}^{r}\mathbf{W}_j)\mathbf{\Lambda}_r^n(\prod_{n=1}^{r}\mathbf{W}_n)^\top - \mathbf{U}_r^s\mathbf{V}_r\|_F^2, \tag{8}$$

then the joint optimization problem can be written as:

$$\begin{aligned} \max_{\zeta} \min_{\mathcal{W},\mathcal{D},\mathcal{U},\mathbf{V}} \quad & J(\tilde{\mathcal{W}}, \mathcal{W}, \mathcal{D}, \mathcal{C}) = G(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{C}) + \beta G(\mathcal{W}, \mathcal{D}, \mathcal{C}) + \gamma\mathcal{L}(\zeta, \mathcal{D}, \mathbf{y}) \\ \text{s.t.} \quad & \tilde{\mathbf{W}}_r = \arg\min_{\hat{\mathbf{W}}_r} \alpha\|\hat{\mathbf{W}}_r - \mathbf{W}_r\|_F^2 + G(\tilde{\mathcal{W}}, \mathcal{D}, \mathcal{P}) \quad r = 1, \ldots, K \\ & \tilde{\mathbf{W}}_r, \mathbf{W}_r \in \Omega \quad \mathcal{D} \in \Psi, \quad \|\mathbf{v}_p\|_1 < \mu,\ p = 1, \ldots, P, \end{aligned} \tag{9}$$

The optimization problem defined above is a non-convex problem that we solve using alternating minimization. Complete algorithm and the details about the optimization are described in Appendix B. Note that the random initialization of the variables can result in a very different final solution that might be far from the ground truth. One such solution for $\mathbf{U}$ and $\mathbf{V}$ would be the identity matrix since all the correlation matrices have one as their diagonal element, which can drastically change the final components. It might also be possible that $\mathbf{V}$ might store highly reproducible components since they are present in most individuals, leading to a decrease in reproducibility of hSCPs. We prevent these cases by using $\mathrm{svd} - \mathrm{initialization}$ [22, Algorithm 2] for $\mathcal{W}$ and $\mathcal{D}$, where, in the starting, most of the variability associated with data is stored in $\mathcal{W}$ and $\mathcal{D}$. In this way, we can prevent $\mathbf{V}$ from storing

Table 1: Accuracy of the components on simulated dataset at one level.

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.789	0.787	0.745	0.736
Adv. hSCP	0.869	0.875	0.862	0.854
rshSCP	0.873	0.865	0.843	0.867
Adv. rshCP	$\mathbf{0.903}$	$\mathbf{0.910}$	$\mathbf{0.902}$	$\mathbf{0.908}$
rshSCP w/ rand.	0.856 ± 0.039	0.834 ± 0.055	0.824 ± 0.031	0.818 ± 0.036
Adv. rshSCP. w/ rand.	0.897 ± 0.030	0.895 ± 0.036	0.892 ± 0.023	0.886 ± 0.034

Table 2: Split sample reproducbility on simulated dataset at one level.

Method	$k_1 = 8$	$k_1 = 10$	$k_1 = 12$	$k_1 = 14$
hSCP	0.769 ± 0.052	0.798 ± 0.047	0.739 ± 0.053	0.734 ± 0.047
Adv. hSCP	0.781 ± 0.037	0.818 ± 0.031	0.780 ± 0.034	0.750 ± 0.031
rshSCP	0.825 ± 0.039	0.845 ± 0.030	$\mathbf{0.826 \pm 0.039}$	0.779 ± 0.036
Adv. rshSCP	$\mathbf{0.840 \pm 0.044}$	$\mathbf{0.869 \pm 0.034}$	0.815 ± 0.035	$\mathbf{0.802 \pm 0.030}$
rshSCP w/ rand.	0.804 ± 0.085	0.818 ± 0.086	0.780 ± 0.068	0.758 ± 0.071
Adv. rshSCP w/ rand.	0.826 ± 0.069	0.833 ± 0.081	0.801 ± 0.074	0.782 ± 0.077

highly reproducible components during initial iterations. We initialize $\mathbf{U}_r^s$ and $\mathbf{V}_r$ using the below equation:

$$\mathbf{U}_r^s = \left[\frac{1}{|\mathcal{I}_s|} \left(\sum_{n \in \mathcal{I}_s} \mathbf{\Theta}^n - (\prod_{j=1}^{r} \mathbf{W}_j) \mathbf{\Lambda}_r^n (\prod_{n=1}^{r} \mathbf{W}_n)^\top \right) \mathbf{J}_p \right] \circ \mathbf{I}_p, \qquad \mathbf{V}_r = \frac{1}{P} \mathbf{J}_P. \tag{10}$$

This complete initialization procedure ensures that the algorithm starts with the majority of variability in the data stored in $\mathcal{W}$ and $\mathcal{D}$, and $\mathbf{U}^s$ start from the residual variance left in site s after the $\mathrm{svd} - \mathrm{initialization}$ procedure. We show in the next sections that this simple strategy, though sub-optimal, can help estimate reproducible components with diminished site effects. All the code is implemented in MATLAB and will be released upon publication.

3 Experiment

3.1 Simulated Dataset

One level. We first generate simulated dataset at one level to evaluate the performance of our model against the vanilla hSCP. We simulate data with $p = 50$, $k_1 = 10$, $S = 4$ with 200, 300, 400 and 500 number of participants in each site. We generated sparse shared components $\mathbf{W}_1$ with percentage of non-zeros equal to 60% and each element sampled from $\mathcal{N}(0, 1)$. We then generate correlation matrix for nth subject belonging to sth site using:

$$\mathbf{\Theta}^n = (\mathbf{W}_1 + \mathbf{E}_1^n) \, \mathbf{\Lambda}^n \, (\mathbf{W}_1 + \mathbf{E}_2^n)^\top + \mathbf{U}^s \mathbf{V} + \mathbf{E}_2^n, \tag{11}$$

where $\mathbf{U}^s$ is a diagonal matrix with positive elements sampled from $\mathcal{N}(1, .1)$, $\mathbf{V}$ is a random matrix sampled from wishart distribution, each element of $\mathbf{\Lambda}^n$ is sampled from $\mathcal{N}(4, 1)$ and $\mathbf{E}_1^n$ is the noise matrix added to the components whose each element is sampled from $\mathcal{N}(0, .1)$ and $\mathbf{E}_2^n$ is added to ensure that the final matrix is positive definite. However the diagonal elements of $\mathbf{\Theta}^n$ are not equal to 1. To make them 1, we extract diagonal elements $\mathbf{D}$ of $\mathbf{\Theta}^n$ and get the new correlation matrix as $\mathbf{D}^{1/2} \mathbf{\Theta}^n \mathbf{D}^{1/2}$. We used a feed-forward neural network for the classification model with two hidden layers. The networks contain the following layers: a fully connected layer with 50 hidden unites, dropout layer with rate 0.2, ReLU, a fully-connected layer with 4 hidden units and a softmax layer. Optimal value of hyperparameters α, β and μ are selected from $[0.1, 1]$, $[1, 5]$ and $[0.1, 0.5, 1]$ respectively, and λ_1 is fixed to be 5 for estimating sparse components. The criterion for choosing the best hyperparameter is maximum split-sample reproducibility. The split sample reproducibility is the normalized inner product between the components estimated on two random equal splits of the data. Split sample reproducibility tries to answer the question of whether the components are

Table 3: Accuracy of the components on hierarchical simulated dataset.

	$k_2 = 4$				$k_2 = 6$			
Method \ k_1	8	10	12	14	8	10	12	14
hSCP	0.806	0.801	0.783	0.777	0.797	0.790	0.773	0.766
Adv. hSCP	0.875	0.872	0.870	0.864	0.863	0.859	0.849	0.851
rshSCP	0.881	0.876	0.860	0.862	0.874	0.871	0.852	0.859
Adv. rshSCP	**0.904**	**0.909**	**0.905**	**0.907**	**0.902**	**0.903**	**0.904**	**0.902**

generalizable across subjects from the same sites or not. Table 2 shows the reproducibility of the components generated from different methods. It is computed over 15 runs in all the experiments. We used accuracy of the estimated components as a performance measure. It is defined as the normalized inner product between ground truth components and estimated components. All the experiments were run on a four i7-6700HQ CPU cores single ubuntu machine.

Accuracy. Table 1 displays the accuracy of hSCP, hSCP with adversarial learning (Adv. hSCP), rshSCP, rshSCP with adversarial learning (Adv. rshSCP), and rshSCP and rshSCP with random initialization (rshSCP w/ rand. and Adv. rshSCP w/ rand.) on the simulated dataset. Here, accuracy is defined as the average correlation between estimated components and the ground truth components. From the results, we can see that the rshSCP with adversarial learning can significantly improve the components' accuracy and the reproducibility of the components. The results using random initialization instead of using the initialization strategy mentioned in the previous section indicates that random initialization brings significant variability to performance. On average, it performs worse than our strategy, but there might be instances where the random initialization can perform better, which might suggest that there might be some better strategy for initialization. Also, for $\mathbf{V}$, there is an optimal sparsity value, which achieves the best result. If $\mathbf{V}$ is dense, then it might remove essential information that might reduce reproducibility, and if it is too sparse, then we might not have desired effects to make the model robust. The results showing the variation in the accuracy and reproducibility with the sparsity of $\mathbf{V}$ are in Appendix D.1.

Site prediction. To check if the estimated subject information ($\mathbf{\Lambda}$) has reduced predictive power to predict the site to which the subject belonged, we performed a 5 fold cross-validation using SVM with RBF kernel. Since we already used feed forward network, we decided to use a different model for the prediction. Our model leads to a decrease in average cross-validation accuracy from 97.6% to 67%, which suggests that our model can reduce the prediction capability to predict site. More details about the result are in Appendix D.1.

Two level. Under the same settings as defined above, we generate correlation matrix from two level components with $k_2 = 4$ using:

$$\begin{aligned} \mathbf{\Theta}^n &= \tilde{\mathbf{W}}_1 \tilde{\mathbf{W}}_2 \mathbf{\Lambda}^n \tilde{\mathbf{W}}_2^\top \tilde{\mathbf{W}}_1^\top + \mathbf{U}^s \mathbf{V} + \mathbf{E}_3^n, \\ \tilde{\mathbf{W}}_1 &= \mathbf{W}_1 \mathbf{E}_1^n, \quad \tilde{\mathbf{W}}_2 = \mathbf{W}_2 + \mathbf{E}_2^n \end{aligned} \tag{12}$$

where each element of $\mathbf{W}_2$ is sampled from $\mathcal{N}(0, 1)$, the percentage of non-zeros equal to 40%, $\mathbf{E}_1^n$ and $\mathbf{E}_2^n$ is the noise added to the components whose each element is sampled from $\mathcal{N}(0, .1)$ and $\mathbf{E}_3^n$ is added to ensure that the final matrix is positive definite. Table 3 shows the accuracy for different values of k_1 and k_2. Selection of hyparameter is same as in the previous paragraph. We can see that the proposed method estimates most accurate ground truth components. Reproducbility results and random initialization results are availble in Appendix D.1.

3.2 Real Dataset

Data. We collected functional MRI data from 5 different multi-center imaging studies- 1) Baltimore Longitudinal Study of Aging (BLSA) [29, 30], the Coronary Artery Risk Development in Young Adults study (CARDIA) [31], UK BioBank (UKBB) [32], Open access series of imaging studies (OASIS) [33] and Aging Brain Cohort Study (ABC) from Penn Memory Center [34]. Although UK Biobank has more than 20000 scans, we only used 2023 randomly selected scans to avoid

Table 4: Summary characteristics of the real dataset.

Data Sites	Participants	% of Females	Age Range (Median)	Scanner
BLSA-3T	784	56.5	$[22, 95](68)$	3T Philips
CARDIA1	199	55.7	$[42, 61](52)$	3T Siemens Tim Trio
CARDIA2	321	51.4	$[43, 61](52)$	3T Philips Achieva
CARDIA3	278	55.3	$[43, 62](52)$	3T Philips Achieva
UKBB	2023	55.2	$[45, 79](63)$	3T Siemens Skyra
OASIS	847	56.0	$[42.6, 97](70)$	1.5T Siemens Vision
ABC	279	59.1	$[23, 95](70)$	3T Siemens Tim Trio

Table 5: Split-sample reproducbility on real dataset ($k_2 = 4$).

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.713 ± 0.039	0.707 ± 0.038	0.697 ± 0.035	0.683 ± 0.036
Adv. hSCP	0.737 ± 0.041	0.719 ± 0.033	0.715 ± 0.037	0.710 ± 0.043
rshSCP	0.806 ± 0.036	0.768 ± 0.032	0.742 ± 0.033	0.743 ± 0.044
Adv. rshSCP	$\mathbf{0.808 \pm 0.030}$	$\mathbf{0.772 \pm 0.036}$	$\mathbf{0.747 \pm 0.034}$	$\mathbf{0.746 \pm 0.036}$

Table 6: Leave one site out reproducbility on real dataset($k_2 = 4$).

Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	0.652 ± 0.038	0.618 ± 0.041	0.592 ± 0.033	0.571 ± 0.035
Adv. hSCP	0.656 ± 0.035	0.629 ± 0.039	0.601 ± 0.035	0.584 ± 0.034
rshSCP	0.712 ± 0.034	0.701 ± 0.036	0.676 ± 0.038	0.665 ± 0.034
Adv. rshSCP	$\mathbf{0.716 \pm 0.032}$	$\mathbf{0.709 \pm 0.031}$	$\mathbf{0.688 \pm 0.034}$	$\mathbf{0.671 \pm 0.033}$

estimating the results that would be heavily influenced by the UK Biobank. We projected the data into a lower-dimensional space such that the number of nodes in each subject's data was 100. Table 4 summarizes the number of participants in each site and age distribution. More details on the dataset and the preprocessing pipeline are given in Appendix C. CARDIA data is divided into three parts because of the acquisition at three different sites.

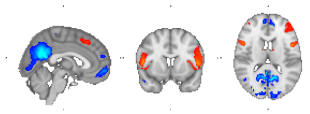
Reproducibility. Since we don't have access to ground truth here, we compare the methods based on the split sample and leave one site reproducibility. Leave one site out reproducibility is defined as the similarity between components derived from the site s and all sites except s. Split sample reproducibility tries to answer the question of whether the components are generalizable to other sites or not. For estimating rshSCP with only one site, we used $\mathbf{V}$ estimated from all sites except s since the idea behind $\mathbf{V}$ was to store information about the site/scanner from various sites. This would also help analyze the generalization power of $\mathbf{V}$. The optimum value of the hyperparameters is selected from the range defined in section 3.1. λ_1 and λ_2 are fixed to be 10 and 5 respectively for estimating sparse components. The criterion for choosing the best value is the maximum split sample reproducibility. Table 5 shows the split sample reproducibility for varied values of k_1 and $k_2 = 4$. Leave one site out reproducibility results are shown in Table 6. Table 13 and Table 14 in D.2 shows split sample reproducibility and leave one site out reproducibility respectively at two-level for $k_2 = 6$. The results demonstrate that the proposed method can significantly improve the split sample reproducibility and leave one site out reproducibility. For the remaining paper, we focus on the comparison between components learned using adversarial learning from hSCP and Adv. rshSCP.

Site prediction. We performed the same experiment under the same settings as mentioned in the previous section to check $\mathbf{\Lambda}$ has reduced predictive power to predict the site. Using SVM, our model leads to a decrease in average cross-validation accuracy from 51% to 32%, which suggests that our model can reduce the prediction capability to predict site.

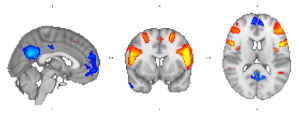
Age prediction We used subject specific information ($\mathbf{\Lambda}$) to predict age of each subject as the metric to check if our method is able to preserve age related biological variation. Table 7 shows 10 fold cross validation mean absolute error (MAE) using random forest . From the table, we can see

Table 7: Mean absolute error ($k_2 = 4$)

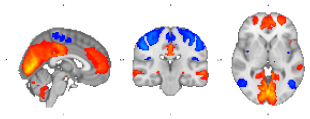
Method	$k_1 = 10$	$k_1 = 15$	$k_1 = 20$	$k_1 = 25$
hSCP	6.490 ± 1.485	6.468 ± 1.442	6.425 ± 1.412	6.414 ± 1.417
Adv. rshSCP	6.494 ± 1.501	6.467 ± 1.475	6.432 ± 1.483	6.409 ± 1.470



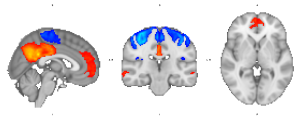
(a) Component 5 : $\rho = -0.02, -\ln(p) = 1.57$



(b) Component 5 : $\rho = -0.13, -\ln(p) = 28.6$



(c) Component 6 : $\rho = -0.07, -\ln(p) = 9.32$



(d) Component 6 : $\rho = -0.05, -\ln(p) = 5.27$

Figure 1: Left column ((a) & (c)) displays the components estimated using hSCP and right column ((b) & (d)) displays the components estimated using rshSCP. Red and blue regions are anti-correlated with each other but are correlated among themselves. The colors are not associated with negative or positive correlation since they can be swapped without affecting the final inference.

that the proposed method has comparable performance as the hSCP suggesting that it preserves age related biological variance. More details about the experiment are given in Appendix D.2.

3.3 Analysis of components

A robust method should be able to reduce non-biological variability caused by site and scanner while retaining biological variability. In this study, we look at brain aging-related associations and leave analysis with other variables for future work. We also discuss the difference between the components with and reduced site effects. We selected the subjects with age greater than 60 to find an association between brain aging and the components derived from hSCP and rshSCP. The total number of subjects having an age greater than 60 is 2746. We first computed Spearman correlation between subject-specific information ($\mathbf{\Lambda}$) and their respective age and used 0.05 as the significance level for the hypothesis test of no correlation against the alternative hypothesis of a nonzero correlation. We derived 10 fine-scale ($1 - 10$) and 4 coarse-scale components (I-IV) because of the high split sample reproducibility and easier interpretation of each component. The correlation and p-values are displayed in Table 15 and 16 in Appendix D.2 for the hSCP and rshSCP.

We first compare components from the two methods. Figure 1 shows the components derived from hSCP and the proposed method. The first row of the figure displays the components with anti-correlation between Default Mode Network (DMN) and Dorsal Attention Network (DAN). The component derived using hSCP has a part of the visual area positively associated with DMN, but the opposite is true, as shown by the previous sparse connectivity patterns [21]. On the other hand, the component with reduced site effects is cleaner since it does not include that relation. This component has a negative correlation with age which has been previously shown in resting-state fMRI and task-based fMRI [35]. The magnitude of anti-correlation has been connected to individual differences in task performances in healthy young adults [36]. However, in the case of older adults, the behavioral implications of reduced anti-correlation remain unclear. The second row of the Figure 1 displays another set of components for comparison. The components stores information about the anti-correlation between DMN and sensorimotor, which aligns with the previous literature [37]. But the addition of a positive correlation of DMN with visual areas will cause misleading inference since it contradicts the previous SCPs and studies. Hence making an inference without removing the site effects can be misleading. Discussion of the association between other relevant components and aging is in Appendix D.2 showing biological interpretability of the components in brain aging. From the results, we can see that there is an increase (or decrease in anti-correlation) in connectivity between different networks in the aging brain. This suggests that there is a reorganization of the aging brain aligning with the previous findings [38]. This can serve as a base to explore rshSCP as a biomarker of neurodegenerative diseases.

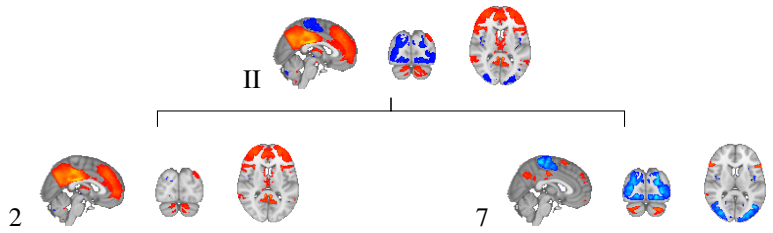


Figure 2: One of the hierarchical components derived from rshSCP comprising of component 2 and 7 at fine scale and component II at coarse scale.

Figure 2 displays one of the hierarchical components with coarse-scale component storing relation between different fine-scale components comprising DMN, sensorimotor, and visual areas, previously studied by [37]. These findings give evidence that even after removal site effects, the components can have a meaningful interpretation. The results indicate that our approach can extract robust informative patterns without using traditional seed-based methods that are dependent on the knowledge of the seed region of interest.

4 Conclusion

In this work, we have presented a method for estimating site effects in hSCP. We formulated the problem as a minimax non-convex optimization problem and solved it using AMSgrad. We also propose a simple initialization procedure to make the optimization procedure deterministic and improve the performance on an average on a simulated and real dataset. Experimentally, using a simulated dataset, we showed that our method accurately estimates the ground truth compared to the vanilla method with better reproducibility. On the real dataset, we show that the proposed method can capture components with a better split sample and leave one site out reproducibility without losing biological interpretability and information. We also show that without removal of site effects, we can have a noisy estimate of sparse components resulting in misleading downstream analysis.

There are several directions for future research. First, it would be interesting to consider the framework for the analysis of task-induced activity to investigate the extent of site effect and corrections on underlying networks activated by the task. Second, we can also look low-dimensional modeling of $\mathbf{V}$ along with sparse constraints which has been used several robust matrix factorization problems. Since we have only shown age related biological preservation, future studies will focus on whether the proposed method preserves components associated with other demographic, clinical phenotypes, and pathological biomarkers.

There are few weaknesses of our proposed model, which also adds directions for future work. First, our method only captures linear site effects, it would be interesting to see if explicitly capturing non-linear site effects can improve the performance of the model. Second, the result of the optimization algorithm depends on the initialization procedure, which has been shown to perform well on the simulated dataset and real dataset but can be sub-optimal.

5 Broader Impact

In this work, we provide a new method to diminish site effects in hSCPs and robustly estimate the components in multi-center studies. The formulation used in the method is not limited to hSCP. It can be easily extended to various matrix factorization approaches such as Independent Component Analysis, Non-negative Matrix Factorization, Dictionary Learning, etc., to improve the reproducibility of functional networks/components. Our work not only has broader applicability in terms of methods used for estimation of components but also to different types of neuroscience data, which includes EEG, MEG, etc. Meanwhile, it should be noted that any unsupervised machine learning model has risk associated with it in terms of accuracy of the model; our work is no departure from this. It can lead to a negative impact since the reproducibility of the components is not 100%, which can lead to misleading clinical and biological interpretability of some of the components. This can be prevented by analyzing the components having high reproducibility as shown in section 3.2 and 3.3.

References

[1] Stephanie Noble, Dustin Scheinost, Emily S Finn, Xilin Shen, Xenophon Papademetris, Sarah C McEwen, Carrie E Bearden, Jean Addington, Bradley Goodyear, Kristin S Cadenhead, et al. Multisite reliability of mr-based functional connectivity. *Neuroimage*, 146:959–970, 2017.

[2] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[3] Christian Dansereau, Yassine Benhajali, Celine Risterucci, Emilio Merlo Pich, Pierre Orban, Douglas Arnold, and Pierre Bellec. Statistical power and prediction accuracy in multisite resting-state fmri connectivity. *Neuroimage*, 149:220–232, 2017.

[4] Anisha Keshavan, Friedemann Paul, Mona K Beyer, Alyssa H Zhu, Nico Papinutto, Russell T Shinohara, William Stern, Michael Amann, Rohit Bakshi, Antje Bischof, et al. Power estimation for non-standardized multisite studies. *NeuroImage*, 134:281–294, 2016.

[5] Russell T Shinohara, Jiwon Oh, Govind Nair, Peter A Calabresi, Christos Davatzikos, Jimit Doshi, Roland G Henry, Gloria Kim, Kristin A Linn, Nico Papinutto, et al. Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *American Journal of Neuroradiology*, 38(8):1501–1509, 2017.

[6] Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

[7] Jorge Jovicich, Ludovico Minati, Moira Marizzoni, Rocco Marchitelli, Roser Sala-Llonch, David Bartrés-Faz, Jennifer Arnold, Jens Benninghoff, Ute Fiedler, Luca Roccatagliata, et al. Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fmri study. *Neuroimage*, 124:442–454, 2016.

[8] Meichen Yu, Kristin A Linn, Philip A Cook, Mary L Phillips, Melvin McInnis, Maurizio Fava, Madhukar H Trivedi, Myrna M Weissman, Russell T Shinohara, and Yvette I Sheline. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human brain mapping*, 39 (11):4213–4227, 2018.

[9] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

[10] Alex Fornito, Andrew Zalesky, and Edward Bullmore. *Fundamentals of brain network analysis*. Academic Press, 2016.

[11] Cornelis J Stam. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10): 683–695, 2014.

[12] Ian A Cook, Susan Y Bookheimer, Laura Mickes, Andrew F Leuchter, and Anand Kumar. Aging and brain activation with working memory tasks: an fmri study of connectivity. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 22(4):332–342, 2007.

[13] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, 106(31):13040–13045, 2009.

[14] Vamsi K Potluru and Vince D Calhoun. Group learning using contrast nmf: Application to functional and structural mri of schizophrenia. In *2008 IEEE International Symposium on Circuits and Systems*, pages 1336–1339. IEEE, 2008.

[15] Kangjoo Lee, Sungho Tak, and Jong Chul Ye. A data-driven sparse glm for fmri analysis using sparse dictionary learning with mdl criterion. *IEEE Transactions on Medical Imaging*, 30(5):1076–1089, 2010.

[16] Gaëlle Doucet, Mikaël Naveau, Laurent Petit, Nicolas Delcroix, Laure Zago, Fabrice Crivello, Gael Jobard, Nathalie Tzourio-Mazoyer, Bernard Mazoyer, Emmanuel Mellet, et al. Brain activity at rest: a multiscale hierarchical functional organization. *Journal of neurophysiology*, 105(6):2753–2763, 2011.

[17] Murat Demirtaş, Joshua B Burt, Markus Helmer, Jie Lisa Ji, Brendan D Adkinson, Matthew F Glasser, David C Van Essen, Stamatios N Sotiropoulos, Alan Anticevic, and John D Murray. Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron*, 101(6):1181–1194, 2019.

[18] Wei Zhang, Shijie Zhao, Xintao Hu, Qinglin Dong, Heng Huang, Shu Zhang, Yu Zhao, Haixing Dai, Fangfei Ge, Lei Guo, et al. Hierarchical organization of functional brain networks revealed by hybrid spatiotemporal deep learning. *Brain connectivity*, 10(2):72–82, 2020.

[19] Heng Huang, Xintao Hu, Yu Zhao, Milad Makkie, Qinglin Dong, Shijie Zhao, Lei Guo, and Tianming Liu. Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, 37 (7):1551–1561, 2017.

[20] Michael Andersen, Ole Winther, Lars Kai Hansen, Russell Poldrack, and Oluwasanmi Koyejo. Bayesian structure learning for dynamic brain connectivity. In *International Conference on Artificial Intelligence and Statistics*, pages 1436–1446. PMLR, 2018.

[21] Harini Eavani, Theodore D Satterthwaite, Roman Filipovych, Raquel E Gur, Ruben C Gur, and Christos Davatzikos. Identifying sparse connectivity patterns in the brain using resting-state fmri. *Neuroimage*, 105: 286–299, 2015.

[22] Dushyant Sahoo, Theodore D Satterthwaite, and Christos Davatzikos. Hierarchical extraction of functional connectivity components in human brain using resting-state fmri. *IEEE Transactions on Medical Imaging*, 2020.

[23] Dushyant Sahoo and Christos Davatzikos. Extraction of hierarchical functional connectivity components in human brain using adversarial learning. *arXiv preprint arXiv:2104.10255*, 2021.

[24] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[25] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M Nasrallah, Theodore D Satterthwaite, Yong Fan, et al. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450, 2020.

[26] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017.

[27] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 355–364, 2018.

[28] Lei Luo, Yanfu Zhang, and Heng Huang. Adversarial nonnegative matrix factorization. In *International Conference on Machine Learning*, pages 6479–6488. PMLR, 2020.

[29] Nicole M Armstrong, Yang An, Lori Beason-Held, Jimit Doshi, Guray Erus, Luigi Ferrucci, Christos Davatzikos, and Susan M Resnick. Predictors of neurodegeneration differ between cognitively normal and subsequently impaired older adults. *Neurobiology of aging*, 75:178–186, 2019.

[30] Susan M Resnick, Dzung L Pham, Michael A Kraut, Alan B Zonderman, and Christos Davatzikos. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *Journal of Neuroscience*, 23 (8):3295–3301, 2003.

[31] Gary D Friedman, Gary R Cutter, Richard P Donahue, Glenn H Hughes, Stephen B Hulley, David R Jacobs Jr, Kiang Liu, and Peter J Savage. Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology*, 41(11):1105–1116, 1988.

[32] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779, 2015.

[33] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[34] John Pluta, Paul Yushkevich, Sandhitsu Das, and David Wolk. In vivo analysis of hippocampal subfield atrophy in mild cognitive impairment via semi-automatic segmentation of t2-weighted mri. *Journal of Alzheimer's disease*, 31(1):85–99, 2012.

[35] R Nathan Spreng, W Dale Stevens, Joseph D Viviano, and Daniel L Schacter. Attenuated anticorrelation between the default and dorsal attention networks with aging: evidence from task and rest. *Neurobiology of aging*, 45:149–160, 2016.

[36] Joseph B Keller, Trey Hedden, Todd W Thompson, Sheeba A Anteraper, John DE Gabrieli, and Susan Whitfield-Gabrieli. Resting-state anticorrelations between medial and lateral prefrontal cortex: association with working memory, aging, and individual differences. *Cortex*, 64:271–280, 2015.

[37] Fikret Işik Karahanoğlu and Dimitri Van De Ville. Transient brain activity disentangles fmri resting-state dynamics in terms of spatially and temporally overlapping networks. *Nature communications*, 6(1):1–10, 2015.

[38] Jessica S Damoiseaux. Effects of aging on functional and structural brain connectivity. *Neuroimage*, 160: 32–40, 2017.

[39] Bhim M Adhikari, Neda Jahanshad, Dinesh Shukla, Jessica Turner, Dominik Grotegerd, Udo Dannlowski, Harald Kugel, Jennifer Engelen, Bruno Dietsche, Axel Krug, et al. A resting state fmri analysis pipeline for pooling inference across diverse cohorts: an enigma rs-fmri protocol. *Brain imaging and behavior*, 13 (5):1453–1467, 2019.

[40] Seyed Mostafa Kia, Hester Huijsdens, Richard Dinga, Thomas Wolfers, Maarten Mennes, Ole A Andreassen, Lars T Westlye, Christian F Beckmann, and Andre F Marquand. Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 699–709. Springer, 2020.

[41] Johanna MM Bayer, Richard Dinga, Seyed Mostafa Kia, Akhil R Kottaram, Thomas Wolfers, Jinglei Lv, Andrew Zalesky, Lianne Schmaal, and Andre Marquand. Accommodating site variation in neuroimaging data using hierarchical and bayesian models. *bioRxiv*, 2021.

[42] Haldun M Ozaktas. Paradigms of connectivity for computer circuits and networks. *Optical Engineering*, 31(7):1563–1567, 1992.

[43] Ian F. Akyildiz, Xudong Wang, and Weilin Wang. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, 2005. ISSN 1389-1286. doi: https://doi.org/10.1016/j.comnet.2004.12.001. URL `https://www.sciencedirect.com/science/article/pii/S1389128604003457`.

[44] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30:6338–6347, 2017.

[45] Osame Kinouchi and Mauro Copelli. Optimal dynamical range of excitable networks at criticality. *Nature physics*, 2(5):348–351, 2006.

[46] Danielle S Bassett, Daniel L Greenfield, Andreas Meyer-Lindenberg, Daniel R Weinberger, Simon W Moore, and Edward T Bullmore. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS comput biol*, 6(4):e1000748, 2010.

[47] Rishidev Chaudhuri, Alberto Bernacchia, and Xiao-Jing Wang. A diversity of localized timescales in network activity. *elife*, 3:e01239, 2014.

[48] Luca Ferrarini, Ilya M Veer, Evelinda Baerends, Marie-José van Tol, Remco J Renken, Nic JA van der Wee, Dirk J Veltman, André Aleman, Frans G Zitman, Brenda WJH Penninx, et al. Hierarchical functional modularity in the resting-state human brain. *Human brain mapping*, 30(7):2220–2231, 2009.

[49] Esraa Al-Sharoa, Mahmood Al-Khassaweneh, and Selin Aviyente. Tensor based temporal and multilayer community detection for studying brain dynamics during resting state fmri. *IEEE Transactions on Biomedical Engineering*, 66(3):695–709, 2018.

[50] Teddy J Akiki and Chadi G Abdallah. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9(1):1–15, 2019.

[51] Arian Ashourvan, Qawi K Telesford, Timothy Verstynen, Jean M Vettel, and Danielle S Bassett. Multi-scale detection of hierarchical community architecture in structural and functional brain networks. *PLoS One*, 14(5):e0215520, 2019.

[52] Annette L Fitzpatrick, Catherine K Buchanan, Richard L Nahin, Steven T DeKosky, Hal H Atkinson, Michelle C Carlson, and Jeff D Williamson. Associations of gait speed and other measures of physical function with cognition in a healthy cohort of elderly persons. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62(11):1244–1251, 2007.

[53] Liang Zhan, Lisanne M Jenkins, Ouri E Wolfson, Johnson Jonaris GadElkarim, Kevin Nocito, Paul M Thompson, Olusola A Ajilore, Moo K Chung, and Alex D Leow. The significance of negative correlations in brain connectivity. *Journal of Comparative Neurology*, 525(15):3251–3265, 2017.

[54] Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005.

[55] Xintao Hu, Heng Huang, Bo Peng, Junwei Han, Nian Liu, Jinglei Lv, Lei Guo, Christine Guo, and Tianming Liu. Latent source mining in fmri via restricted boltzmann machine. *Human brain mapping*, 39 (6):2368–2380, 2018.

[56] Qinglin Dong, Fangfei Ge, Qiang Ning, Yu Zhao, Jinglei Lv, Heng Huang, Jing Yuan, Xi Jiang, Dinggang Shen, and Tianming Liu. Modeling hierarchical brain networks via volumetric sparse deep belief network. *IEEE Transactions on Biomedical Engineering*, 67(6):1739–1748, 2019.

[57] Anastasia Podosinnikova, Matthias Hein, and Rainer Gemulla. *Robust Principal Component Analysis as a Nonlinear Eigenproblem*. PhD thesis, Saarland University, 2013.

[58] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[59] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[60] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

[61] Ludovica Griffanti, Gholamreza Salimi-Khorshidi, Christian F Beckmann, Edward J Auerbach, Gwenaëlle Douaud, Claire E Sexton, Enikő Zsoldos, Klaus P Ebmeier, Nicola Filippini, Clare E Mackay, et al. Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *Neuroimage*, 95:232–247, 2014.

[62] Stephen M Smith, Aapo Hyvärinen, Gaël Varoquaux, Karla L Miller, and Christian F Beckmann. Group-pca for very large fmri datasets. *Neuroimage*, 101:738–749, 2014.

[63] Yuhui Du and Yong Fan. Group information guided ica for fmri data analysis. *Neuroimage*, 69:157–197, 2013.

[64] Mark Jenkinson. Measuring transformation error by rms deviation. *Studholme, C., Hill, DLG, Hawkes, DJ*, 1999.

[65] Christina Triantafyllou, Richard D Hoge, Gunnar Krueger, Christopher J Wiggins, Andreas Potthast, Graham C Wiggins, and Lawrence L Wald. Comparison of physiological noise at 1.5 t, 3 t and 7 t and optimization of fmri acquisition parameters. *Neuroimage*, 26(1):243–250, 2005.

[66] Jonathan D Power, Kelly A Barnes, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *Neuroimage*, 59(3):2142–2154, 2012.

[67] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[68] Chao-Gan Yan, R Cameron Craddock, Yong He, and Michael P Milham. Addressing head motion dependencies for small-world topologies in functional connectomics. *Frontiers in human neuroscience*, 7: 910, 2013.

[69] Irene E Nagel, Claudia Preuschhof, Shu-Chen Li, Lars Nyberg, Lars Bäckman, Ulman Lindenberger, and Hauke R Heekeren. Load modulation of bold response and connectivity predicts working memory performance in younger and older adults. *Journal of Cognitive Neuroscience*, 23(8):2030–2045, 2011.

[70] Linda Geerligs, Natasha M Maurits, Remco J Renken, and Monicque M Lorist. Reduced specificity of functional connectivity in the aging brain during task performance. *Human brain mapping*, 35(1):319–330, 2014.

[71] Luiz Kobuti Ferreira, Ana Carolina Brocanello Regina, Natasa Kovacevic, Maria da Graça Morais Martin, Pedro Paim Santos, Camila de Godoi Carneiro, Daniel Shikanai Kerr, Edson Amaro Jr, Anthony Randal McIntosh, and Geraldo F Busatto. Aging effects on whole-brain functional connectivity in adults free of cognitive and psychiatric disorders. *Cerebral cortex*, 26(9):3851–3865, 2016.

[72] Linda Geerligs, Remco J Renken, Emi Saliasi, Natasha M Maurits, and Monicque M Lorist. A brain-wide study of age-related changes in functional connectivity. *Cerebral cortex*, 25(7):1987–1999, 2015.

[73] K Warner Schaie and Sherry Willis. *Handbook of the Psychology of Aging*. Academic Press, USA, 2021.

[74] Cheryl Grady, Saman Sarraf, Cristina Saverino, and Karen Campbell. Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiology of aging*, 41:159–172, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section 4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code is proprietary and data needs to be requested from the authors of the references. However, the complete algorithm and all hyperparameter values are stated for reproducing the experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See section 3.1 and 3.2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 3.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 3.2
   (b) Did you mention the license of the assets? [No] The code is proprietary.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No subjects are recognizable in the data
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]