

COMMON AND INDIVIDUAL STRUCTURE OF BRAIN NETWORKS¹

BY LU WANG, ZHENGWU ZHANG AND DAVID DUNSON

Central South University, University of Rochester and Duke University

This article focuses on the problem of studying shared- and individual-specific structure in replicated networks or graph-valued data. In particular, the observed data consist of n graphs, $G_i, i = 1, \dots, n$, with each graph consisting of a collection of edges between V nodes. In brain connectomics, the graph for an individual corresponds to a set of interconnections among brain regions. Such data can be organized as a $V \times V$ binary adjacency matrix A_i for each i , with ones indicating an edge between a pair of nodes and zeros indicating no edge. When nodes have a shared meaning across replicates $i = 1, \dots, n$, it becomes of substantial interest to study similarities and differences in the adjacency matrices. To address this problem, we propose a method to estimate a common structure and low-dimensional individual-specific deviations from replicated networks. The proposed Multiple GRaph Factorization (M-GRaF) model relies on a logistic regression mapping combined with a hierarchical eigenvalue decomposition. We develop an efficient algorithm for estimation and study basic properties of our approach. Simulation studies show excellent operating characteristics and we apply the method to human brain connectomics data.

1. Introduction. Binary undirected networks, encoding the presence or absence of connections between pairs of nodes, have wide applications in biology and social science [Girvan and Newman (2002)]. While most available procedures focus on modeling a single network, we consider the case where a network over a common set of nodes is measured for each individual under study, leading to multiple network observations. One particular example is structural or functional brain networks, with the brain parcellated into a fixed number of regions. Multimodal magnetic resonance imaging (MRI) scans, together with advanced image processing tools, can give us a connectivity pattern of the brain represented by an undirected binary network [Zhang et al. (2018a)]. Such networks from multiple subjects typically share a common structure while exhibiting their own features.

In this context, it becomes of particular interest to study similarities and differences in human brain networks. Shared connectivity patterns provide important insights into evolutionarily-conserved structures in the human brain. For example,

Received July 2017; revised April 2018.

¹Supported in part by United States Office of Naval Research (ONR) Grant N00014-14-1-0245 and Army Research Institute (ARI) Grant W911NF-16-1-0544.

Key words and phrases. Binary networks, multiple graphs, penalized logistic regression, random effects, spectral embedding.

pairs of brain regions that have very high or very low probabilities of connection for essentially all individuals. Individual-specific structure of the brain network may help to predict and add to mechanistic understanding of causes of variation in human cognitive traits and behaviors. [Lock et al. \(2013\)](#) proposed a useful tool to separate joint and individual variation for multiple datasets associated with a common set of objects. However, their method was not designed for network-valued data. There is a strong need for new statistical methods that identify and separate the common and individual structure for multiple replicated networks.

The focus of this article is on extracting common and low-dimensional individual-specific structure from replicated binary networks. In structural brain connectivity applications, providing the main motivation of this article, the individual-specific components reflect distinct characteristics of that individual's brain structure which may relate to her traits. We focus on data from the Human Connectome Project (HCP) [[Van Essen et al. \(2012\)](#)] (www.humanconnectomeproject.org/), which contains rich brain imaging data along with a range of cognitive, motor, sensory and emotional traits [[Barch et al. \(2013\)](#)]. [Figure 1](#) displays two binary structural brain networks we extracted from two HCP subjects and the difference of their adjacency matrices. The left panel shows the network for an individual with a low visuospatial processing score, while the middle panel shows one with a high score. Potentially, the individual difference, for example, the cross-hemisphere connectivity, may predict a range of traits, such as cognitive, motor, and sensory abilities.

There is a large literature on statistical modeling of binary networks [[Goldenberg et al. \(2010\)](#)]. For example, exponential family random graph models (ERGMs) assume the probability of observing a graph is determined by a vector of graph statistics, such as the total number of edges, the degrees of the vertices and so on. However, since the mappings from graphs to features are often many to one mappings, one drawback of the ERGM is that simply relying on some summary features of a graph can not represent complex structures of networks. Latent space models [[Hoff, Raftery and Handcock \(2002\)](#)], however, are more flexible at

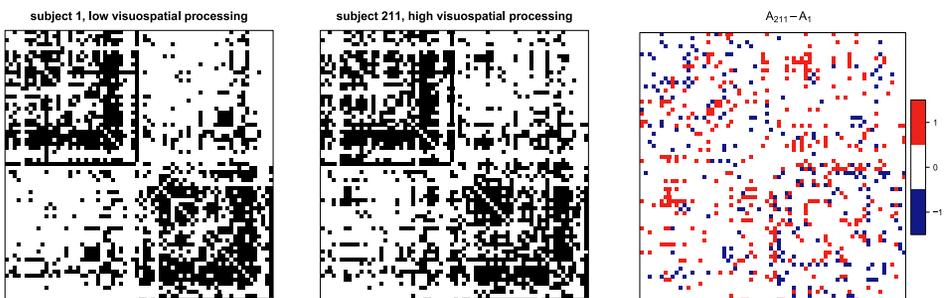


FIG. 1. Adjacency matrices of two structural brain networks in the HCP data (left and middle) and a heatmap of their differences (right).

characterizing the distribution of graphs because they can effectively model each edge probability, while still maintaining rich types of dependence structure in the graph.

A variety of latent space models have been developed [Durante, Dunson and Vogelstein (2017), Hoff (2008), Tang et al. (2015)], which are appealing in defining rich types of network structure while achieving dimensionality reduction by embedding each node in a low-dimensional latent space. Edges are typically assumed to be conditionally independent given the latent positions of nodes. The edge probabilities are described as functions of distance or (weighted) inner products of node-specific latent vectors with a *logit* or *probit* link. Bayesian inference is often employed, but substantial computational problems can arise for large multiple-network data.

Considering a single network, efficient algorithms have been developed for estimating its low-dimensional latent structure. Sussman et al. (2012) estimate nodes' latent positions from a low rank approximation to the adjacency matrix for a random dot product graph (RDPG) with identity link. Though Sussman et al. (2012) proved the consistency of assigning nodes to blocks by clustering over their latent vectors, the dot product of the estimated latent positions may not be valid probabilities. O'Connor, Médard and Feizi (2015) proposed to do node clustering on an RDPG with a logistic link to address this problem. They provided an efficient algorithm for maximum likelihood inference of nodes' latent positions which contains a spectral decomposition on the mean-centered adjacency matrix and a logistic regression with positive constraint on the coefficients.

There is a literature on analysis methods for data consisting of a set of networks that share a common vertex set. For multi-layer or multi-view graphs, vertices correspond to entities and different graph layers capture different types of relationships among the entities [Dong et al. (2014)]. Linked Matrix Factorization (LMF) [Tang, Lu and Dhillon (2009)] approximates each graph by a graph-specific factor and a factor matrix common to all graphs. The goal is to cluster vertices into communities, and LMF focuses on merging the information from multiple graphs instead of characterizing unique structure for each graph. Other relevant methods include principal component analysis (PCA) and tensor decomposition [Kolda and Bader (2009), Tucker (1966)]. Usual PCA requires *flattening* of the data, which destroys the network structure, while tensor methods that concatenate the adjacency matrices together might be more appropriate [Zhang et al. (2018b)]. None of these approaches directly addresses our problem of interest.

We develop a promising framework for studying the common and low-dimensional individual structure of multiple binary networks with similar patterns. Our approach provides a data generating process for each graph, which shows how the low-dimensional structure drives the high-dimensional networks. Specifically, the logit of the edge-probability matrix for each network is decomposed into the sum of a common term and a low-rank individual-specific deviation. Based on the

idea of an unrestricted eigen-decomposition (no positive constraints on eigenvalues), our model is able to capture complex network patterns, such as hubs [Hoff (2008)], better than latent distance or latent inner-product models. A novel algorithm inspired by O'Connor, Médard and Feizi (2015) is proposed for efficiently estimating the model.

The rest of the paper is organized as follows. The model and algorithm together with two variants are proposed in Section 2. Section 3 contains simulation studies demonstrating the computational performance of our algorithm and basic properties of parameter estimates. Applications to scan-rescan brain network data and the HCP data are reported in Section 4 and Section 5 concludes.

2. Methodology. We focus on undirected binary networks with a common node set and no self-loops. Let A_1, \dots, A_n be the corresponding adjacency matrices of these networks. Each A_i is a $V \times V$ symmetric matrix with $A_{i[uv]} = 1$ if node u and v are connected in network i and $A_{i[vu]} = 0$ otherwise.

2.1. *M-GRAF model.* We take the conditional independence approach of latent space models by assuming for each pair of nodes (u, v) in network A_i , an edge is drawn independently from a Bernoulli distribution given the corresponding edge probability:

$$(2.1) \quad A_{i[uv]} \mid \Pi_{i[uv]} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(\Pi_{i[uv]}), \quad u > v; u, v \in \{1, 2, \dots, V\},$$

where Π_i denotes the $V \times V$ symmetric edge probability matrix corresponding to network i , $i = 1, \dots, n$.

In our exploratory analyses of brain network data, we observe that brain structural networks generally share some common connectivity patterns such as hemisphere modularity, as shown in Figure 2. In addition, the deviation of individual networks from the average tends to be much sparser, with many entries in the deviation matrix $|A_i - \bar{A}|$ of small magnitude (shown in the right most panel of Figure 2). We expect that these deviations can be accurately approximated as low rank.

Therefore, we assume the logit of each probability matrix Π_i can be decomposed into two parts: a symmetric matrix $Z \in \mathbb{R}^{V \times V}$ shared by all networks representing the baseline log odds for each connection and a symmetric low rank matrix $D_i \in \mathbb{R}^{V \times V}$ representing the deviation of unit i from the baseline:

$$(2.2) \quad \text{logit}(\Pi_i) = Z + D_i, \quad i = 1, \dots, n.$$

Suppose D_i has rank K , typically with $K \ll V$. Taking an eigenvalue decomposition of D_i ,

$$(2.3) \quad D_i = Q_i \Lambda_i Q_i^\top,$$

where $Q_i \in \mathbb{R}^{V \times K}$ satisfies $Q_i^\top Q_i = I_K$ and $\Lambda_i = \text{diag}(\lambda_{i1}, \dots, \lambda_{iK})$ is a $K \times K$ diagonal matrix.

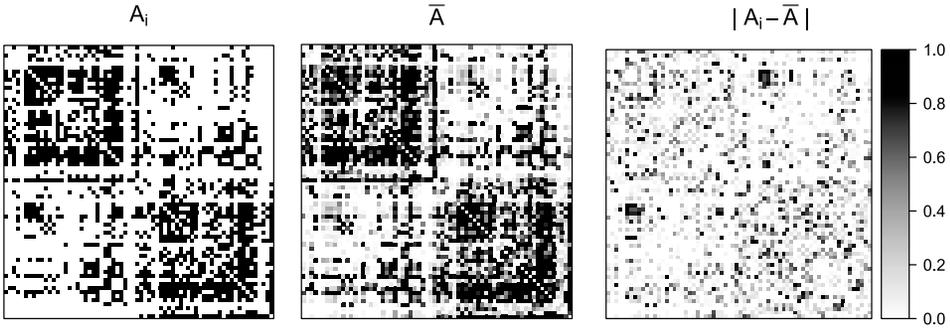


FIG. 2. Left: the adjacency matrix A_i of a subject's structural brain network in the HCP data. Middle: average of the adjacency matrices $\bar{A} = \sum_{i=1}^n A_i/n$. Right: absolute value of $(A_i - \bar{A})$.

Equations (2.2)–(2.3) imply that the individual elements of Π_i can be expressed as

$$(2.4) \quad \text{logit}(\Pi_{i[uv]}) = Z_{uv} + \sum_{k=1}^K \lambda_{ik} Q_{i[uk]} Q_{i[vk]},$$

for $u \neq v, u, v \in \{1, \dots, V\}, i = 1, \dots, n$.

Z_{uv} in (2.4) represents the baseline log odds for the node pair u, v across all networks. Interpretation of the rest of (2.4) is similar to that of the eigenmodel in Hoff (2008) where the relationship between two nodes is represented as a weighted dot product of node-specific latent vectors. For each network i , $Q_{i[uk]}$ can be interpreted as node u 's value of some unobserved characteristic k or latent coordinate along axis k . λ_{ik} is the scaling parameter of latent axis k . The magnitude of λ_{ik} controls the impact of axis k in determining the edge probabilities of network i , while the sign of λ_{ik} determines whether similar values of $Q_{i[uk]}$ and $Q_{i[vk]}$ would contribute positively or negatively to the connection probability between node u and v .

This model based on the idea of eigenvalue decomposition is flexible at characterizing a wide array of patterns in network data [Hoff (2008)], including transitivity and hubs. Transitivity describes the phenomenon that a friend of a friend is a friend, which is well represented by a latent distance model or RDPG but is poorly characterized by a stochastic block model. A hub refers to a center node that connects to many other nodes while these nodes do not connect to each other. Such structure could be described by a stochastic block model with a large number of groups. However, latent distance models or RDPGs often perform poorly or require high dimension of latent characteristics to capture the pattern of hubs. As Hoff (2008) pointed out, the flexibility of the eigenmodel is “due to the fact that it provides an unrestricted low rank approximation to the adjacency matrix” and is therefore able to represent more complicated patterns than the other three latent space models.

2.2. *Inference of Q_1, \dots, Q_n .* For estimation of the model (2.1)–(2.3), we first simplify the joint log-likelihood of the n network observations A_1, \dots, A_n as

$$\begin{aligned}
 & \log L(A_1, \dots, A_n \mid Z, D_1, \dots, D_n) \\
 &= \sum_{i=1}^n \sum_{u=1}^V \sum_{v < u} [A_{i[uv]} \log(\Pi_{i[uv]}) + (1 - A_{i[uv]}) \log(1 - \Pi_{i[uv]})] \\
 (2.5) \quad &= \sum_{i=1}^n \sum_{u=1}^V \sum_{v < u} \left[A_{i[uv]} \log\left(\frac{\Pi_{i[uv]}}{1 - \Pi_{i[uv]}}\right) + \log(1 - \Pi_{i[uv]}) \right] \\
 &= \sum_{i=1}^n \sum_{u=1}^V \sum_{v < u} [A_{i[uv]}(Z_{uv} + D_{i[uv]}) + \log(1 - \Pi_{i[uv]})].
 \end{aligned}$$

PROPOSITION 2.1. *Assuming that the common structure Z is given,*

$$\begin{aligned}
 & \arg \max_{D_1, \dots, D_n} \log L(A_1, \dots, A_n \mid Z, D_1, \dots, D_n) \\
 &= \arg \max_{D_1, \dots, D_n} \sum_{i=1}^n \frac{1}{2} \operatorname{tr}([A_i - \pi(Z)]D_i),
 \end{aligned}$$

where $\operatorname{tr}(\cdot)$ is the matrix trace, $\pi(\cdot)$ is the logistic function and $\pi(Z)$ denotes applying $\pi(\cdot)$ to each entry in matrix Z . The diagonal elements of $\pi(Z)$ and A_i are set to 0.

The proof of Proposition 2.1 can be found in Appendix A.1. From the form of the joint log-likelihood (2.5), it is clear the D_i s can be estimated independently for $i = 1, \dots, n$ conditionally on Z . According to Proposition 2.1, $\operatorname{tr}([A_i - \pi(Z)]D_i)$ is a good surrogate function of the log-likelihood $\log L(A_i \mid Z, D_i)$, which is easier to maximize since it is linear in D_i . Hence given Z and recalling the low rank assumption on D_i , we solve the following optimization (2.6) to estimate D_i :

$$\begin{aligned}
 (2.6) \quad & \max_{D_i} \operatorname{tr}([A_i - \pi(Z)]D_i) \\
 & \text{s.t. } \operatorname{rank}(D_i) = K.
 \end{aligned}$$

Plugging in the eigen-decomposition (2.3) of D_i into the target function of (2.6), we have

$$\begin{aligned}
 \operatorname{tr}([A_i - \pi(Z)]D_i) &= \operatorname{tr}([A_i - \pi(Z)]Q_i \Lambda_i Q_i^\top) \\
 &= \operatorname{tr}(Q_i^\top [A_i - \pi(Z)]Q_i \Lambda_i) \\
 &= \sum_{k=1}^K \lambda_{ik} Q_{i[\cdot, k]}^\top [A_i - \pi(Z)]Q_{i[\cdot, k]},
 \end{aligned}$$

where $Q_{i[\cdot k]}$ denotes the k th column of Q_i . Then we obtain the following equivalent optimization to (2.6):

$$(2.7) \quad \begin{aligned} \max_{Q_i, \Lambda_i} \quad & \sum_{k=1}^K \lambda_{ik} Q_{i[\cdot k]}^\top [A_i - \pi(Z)] Q_{i[\cdot k]} \\ \text{s.t.} \quad & Q_i^\top Q_i = I_K, \quad Q_i \in \mathbb{R}^{V \times K}. \end{aligned}$$

Suppose the diagonal entries of Λ_i are sorted decreasingly so that $\lambda_{i1} \geq \dots \geq \lambda_{ik} > 0 > \lambda_{i,k+1} \geq \dots \geq \lambda_{iK}$. Then the optimal Q_i in (2.7) can be solved according to the following Proposition 2.2.

PROPOSITION 2.2. *Let B be a $V \times V$ symmetric real matrix. Suppose the eigenvalues of B are $\sigma_1(B) \geq \dots \geq \sigma_V(B)$ and the corresponding orthonormal eigenvectors are $\mathbf{q}_1, \dots, \mathbf{q}_V$. For any $k \in \{1, \dots, V\}$, given k positive real numbers $c_1 \geq \dots \geq c_k > 0$, and for any orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ in \mathbb{R}^V , one has*

$$(2.8) \quad \max_{\mathbf{u}_1, \dots, \mathbf{u}_k} \sum_{j=1}^k c_j \mathbf{u}_j^\top B \mathbf{u}_j = c_1 \sigma_1(B) + \dots + c_k \sigma_k(B)$$

and

$$(2.9) \quad \min_{\mathbf{u}_1, \dots, \mathbf{u}_k} \sum_{j=1}^k c_j \mathbf{u}_j^\top B \mathbf{u}_j = c_1 \sigma_V(B) + \dots + c_k \sigma_{V-k+1}(B).$$

Therefore an optimal solution to (2.8) is $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ and an optimal solution to (2.9) is $\{\mathbf{q}_V, \dots, \mathbf{q}_{V-k+1}\}$.

The proof is in Appendix A.2. Let $\mathbf{q}_1^{(i)}, \dots, \mathbf{q}_k^{(i)}$ be the first k eigenvectors of $A_i - \pi(Z)$ corresponding to the largest eigenvalues, and $\mathbf{q}_{V-K+k+1}^{(i)}, \dots, \mathbf{q}_V^{(i)}$ the last $(K - k)$ eigenvectors of $A_i - \pi(Z)$ corresponding to the smallest eigenvalues. Then according to Proposition 2.2, an optimal solution Q_i to (2.7) is $Q_i = (\mathbf{q}_1^{(i)}, \dots, \mathbf{q}_k^{(i)}, \mathbf{q}_{V-K+k+1}^{(i)}, \dots, \mathbf{q}_V^{(i)})$.

2.3. Logistic regression for Z and $\{\lambda_{ik}\}$. Once $\{Q_i : i = 1, \dots, n\}$ is estimated, it remains only to estimate the parameters $\{\lambda_{ik} : k = 1, \dots, K; i = 1, \dots, n\}$ and Z . Note that λ_{ik} 's and entries of Z are linear in the logistic link function (2.4). Therefore the MLE of $\{\lambda_{ik}\}$ and Z given $\{Q_i\}$ can be solved by logistic regression of the lower triangular entries of $\{A_i : i = 1, \dots, n\}$ on the corresponding entries of $\{Q_{i[\cdot k]} Q_{i[\cdot k]}^\top : k = 1, \dots, K; i = 1, \dots, n\}$. Let $\mathcal{L}(\cdot)$ be a function mapping the lower triangular entries of a $V \times V$ matrix into a $V(V - 1)/2 \times 1$ long vector, let $\boldsymbol{\pi}_i = \mathcal{L}(\Pi^{(i)}) = (\pi_{i1}, \dots, \pi_{iL})^\top$, where $L = V(V - 1)/2$, let

$\mathbf{z} = \mathcal{L}(Z) = (z_1, \dots, z_L)^\top$, and let M_i be a $L \times K$ matrix with each column being $M_{i[\cdot k]} = \mathcal{L}(Q_{i[\cdot k]} Q_{i[\cdot k]}^\top)$ for $k = 1, \dots, K$. Then (2.4) can be written as

$$(2.10) \quad \text{logit}(\pi_{il}) = z_l + \sum_{k=1}^K \lambda_{ik} M_{i[lk]}, \quad l = 1, \dots, L; i = 1, \dots, n.$$

However, as K increases, overfitting could cause a serious separation issue in the logistic regression (2.10), where the binary outcomes can be almost perfectly predicted by a linear combination of predictors. The separation issue is well known to cause nonidentifiability of logistic regression coefficients with the MLE being $\pm\infty$. A solution to this problem is to place a penalty or prior on the coefficients. Penalized likelihood estimation proposed by Firth (1993) is equivalent to the use of Jeffreys invariant prior. The Newton–Raphson algorithm by Heinze (2006) based on Firth’s method was very slow even for a small synthetic dataset in our simulation. Gelman et al. (2008) propose independent Cauchy priors with center 0 and scale 2.5 for each of the logistic regression coefficients as a weakly informative default. However, such Cauchy priors have very heavy tails and often do not have good performance in sparse data settings with separation issues in our experience. Hence, we instead recommend the following weakly informative Gaussian prior distributions:

$$(2.11) \quad Z_{uv} \sim N(0, 10^2/\gamma), \quad u > v, u, v \in \{1, \dots, V\},$$

$$(2.12) \quad \lambda_{ik} \sim N\left(0, \frac{2.5^2}{\gamma \cdot (2 \cdot \text{sd}_{ik})^2}\right), \quad k = 1, \dots, K; i = 1, \dots, n,$$

where γ is a prior precision factor, sd_{ik} is the standard deviation (sd) of $M_{i[\cdot k]}$, and the factor $1/(2 \cdot \text{sd}_{ik})^2$ in (2.12) is equivalent to standardizing the predictors to have sd of 0.5 as suggested by Gelman et al. (2008).

The Gaussian prior is equivalent to L_2 regularization or a ridge penalty for generalized linear models. Hence, we could compute maximum-a-posteriori (MAP) estimates for Z and $\{\lambda_{ik}\}$ with the `glmnet` function in R. The algorithm implemented in `glmnet` uses cyclical coordinate descent [Hastie and Tibshirani (1990)] and can handle large problems efficiently in our experience. γ is selected through cross validation.

2.4. CISE algorithm. Based on the derivations above, we develop a CISE (common and individual structure explained) algorithm for estimating the M-GRAF model (2.1)–(2.3). CISE is essentially a block coordinate descent algorithm and Algorithm 1 presents the details.

2.5. Distance-based classification of networks. In many applications, in addition to the network variable A_i there may be a class label $l_i \in \{1, 2, \dots, m\}$ associated with each subject i in the dataset, such as high IQ or low IQ, healthy or

Algorithm 1: Common and individual structure explained (CISE) for multiple binary networks

Input: Adjacency matrices A_1, \dots, A_n of size $V \times V$, low rank K , tolerance $\varepsilon \in \mathcal{R}_+$.

Output: Estimates of $\{Q_i : i = 1, \dots, n\}$, Z , and $\{\lambda_{ik} : k = 1, \dots, K; i = 1, \dots, n\}$.

- 1 Initialize $\hat{\pi}(Z) = \sum_{i=1}^n A_i/n$;
 - 2 Initialize each \hat{Q}_i to be the K eigenvectors of $A_i - \hat{\pi}(Z)$ corresponding to the largest eigenvalues in magnitude.
 - 3 **do**
 - 4 (I) Perform L_2 -penalized logistic regression (2.10)–(2.12) to obtain the MAP estimates of $\{Z_{uv}\}$ and $\{\lambda_{ik}\}$. Time complexity of this step is $O(nV^2K)$, according to Minka (2003).
 - 5 (II) For each i , let k_i be the number of positive values in $\lambda_{i,1:K}$;
 - 6 Compute the first k_i eigenvectors of $A_i - \pi(Z)$, $\mathbf{q}_1^{(i)}, \dots, \mathbf{q}_{k_i}^{(i)}$, and the last
 - 7 $(K - k_i)$ eigenvectors, $\mathbf{q}_{V-K+k_i+1}^{(i)}, \dots, \mathbf{q}_V^{(i)}$ (with sorted eigenvalues).
 - 8 Let $\hat{Q}_i = (\mathbf{q}_1^{(i)}, \dots, \mathbf{q}_{k_i}^{(i)}, \mathbf{q}_{V-K+k_i+1}^{(i)}, \dots, \mathbf{q}_V^{(i)})$. Time complexity of this
 - 9 partial eigen-decomposition is $O(V^2K)$ or less [Woolfe et al. (2008)].
 - 10 **while** percent change of joint log-likelihood (2.5) $\geq \varepsilon$;
-

with Alzheimer’s disease. People may want to predict the class membership for a new unlabeled subject based on her brain connectivity. After estimating the low-rank components $\{Q_i, \Lambda_i\}$, representing individual-specific features of a subject’s network data, classification can proceed via a simple distance-based procedure. We define the following distance measure between subject i and j , which avoids misalignment and rotation issues of eigenvectors across subjects:

$$\begin{aligned} d(i, j) &:= \|D_i - D_j\|_F \\ &= \|Q_i \Lambda_i Q_i^\top - Q_j \Lambda_j Q_j^\top\|_F, \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Since Q_i and Q_j lie on the Stiefel manifold $\mathcal{S}_{K,V} = \{X \in \mathbb{R}^{V \times K} : X^\top X = I_K\}$, we can further simplify this distance metric as

$$\begin{aligned} d^2(i, j) &= \text{tr}[(Q_i \Lambda_i Q_i^\top - Q_j \Lambda_j Q_j^\top)^\top (Q_i \Lambda_i Q_i^\top - Q_j \Lambda_j Q_j^\top)] \\ &= \text{tr}(\Lambda_i^2) + \text{tr}(\Lambda_j^2) - 2 \text{tr}(\Lambda_i Q_i^\top Q_j \Lambda_j Q_j^\top Q_i) \end{aligned}$$

so that we only need to compute traces of several small $K \times K$ matrices instead of the large $V \times V$ matrices. For a new unlabeled subject i^* , the proximity measure between i^* and a class c is defined as the average distance from i^* to all the subjects in the class c . Subject i^* is then allocated to the class with the minimum proximity.

2.6. *Variants.* The model described in Section 2.1 is very flexible, since for each subject, we have $\{Q_i, \Lambda_i\}$ to represent its individual structure. This model can be modified to further reduce the number of parameters in two different settings so as to accommodate different degrees of heterogeneity in the data.

2.6.1. *Variant 1:* $D_i = Q_i \Lambda Q_i^\top$. In this case, Λ_i 's are assumed to be the same over all networks so that the number of unknown coefficients in $\{\Lambda_i\}$ declines from nK to K . This model implies that the scaling parameters controlling the impacts of the latent axes are equal for all networks (as discussed in Section 2.1). In this case, the estimation of Q_{i^*} for a new network i^* becomes quite efficient once Z and Λ have been estimated from the training set of networks. Suppose the diagonal entries of $\hat{\Lambda}$ are sorted decreasingly: $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_k > 0 > \hat{\lambda}_{k+1} \geq \dots \geq \hat{\lambda}_K$, \hat{Q}_{i^*} therefore consists of the first k and the last $(K - k)$ eigenvectors of $A_{i^*} - \pi(\hat{Z})$. This variant provides competitive goodness-of-fit to the brain network data compared with the more flexible model $D_i = Q_i \Lambda_i Q_i^\top$ as shown in the applications.

Only a small modification to Algorithm 1 is needed for estimation of Λ . Again we choose a weakly informative prior for λ_k 's:

$$(2.13) \quad \lambda_k \sim N\left(0, \frac{2.5^2}{\gamma \cdot (2 \cdot \text{sd}_k)^2}\right), \quad k = 1, \dots, K,$$

where sd_k is the standard deviation of $(M_{i[\cdot k]}^\top, \dots, M_{n[\cdot k]}^\top)^\top$. As in (2.12), the factor $1/(2 \cdot \text{sd}_k)^2$ in (2.13) is equivalent to standardizing the predictor to have sd of 0.5; the numerator 2.5 is the suggested scale of the Cauchy prior by Gelman et al. (2008); γ adds flexibility to the shrinkage of this prior, which is often tuned by cross validation in practice. Then the MAP estimates for Z and Λ can be obtained via a L_2 -penalized logistic regression.

2.6.2. *Variant 2:* $D_i = Q \Lambda_i Q^\top$. Alternatively, we might do a joint embedding by restricting Q_i 's to be the same. Then the individual structure of each network is represented by a linear combination of K common rank-one matrices and a $K \times 1$ loading vector $\lambda_{i,1:K}$, which greatly reduces dimensionality. In this joint embedding setting, we could still follow an iterative algorithm to do inference on the parameters. Z and $\{\Lambda_i : i = 1, \dots, n\}$ can be estimated from a logistic regression with ridge penalty as discussed in Section 2.3 with M_i replaced by $M = \mathcal{L}(Q Q^\top)$. The challenge lies in estimating Q given Z and $\{\Lambda_i\}$.

Similar to the previous cases, given Z and $\{\Lambda_i\}$, Q can be estimated from the following optimization

$$(2.14) \quad \begin{aligned} \max_{Q \in \mathbb{R}^{V \times K}} \quad & \sum_{i=1}^n \text{tr}([A_i - \pi(Z)]D_i) \\ \text{s.t.} \quad & D_i = Q\Lambda_i Q^\top, \\ & Q^\top Q = I_K. \end{aligned}$$

Plugging in $D_i = Q\Lambda_i Q^\top$ into the target function of (2.14), we have

$$\begin{aligned} & \sum_{i=1}^n \text{tr}([A_i - \pi(Z)]D_i) \\ &= \sum_{i=1}^n \text{tr}([A_i - \pi(Z)]Q\Lambda_i Q^\top) \\ &= \sum_{i=1}^n \text{tr}(Q^\top [A_i - \pi(Z)]Q\Lambda_i) \\ &= \sum_{i=1}^n \sum_{k=1}^K \lambda_{ik} \mathbf{q}_k^\top [A_i - \pi(Z)] \mathbf{q}_k \\ &= \sum_{k=1}^K \mathbf{q}_k^\top \left\{ \sum_{i=1}^n \lambda_{ik} [A_i - \pi(Z)] \right\} \mathbf{q}_k, \end{aligned}$$

where \mathbf{q}_k is the k th column of Q . Define $W_k := \sum_{i=1}^n \lambda_{ik} [A_i - \pi(Z)]$, $k = 1, \dots, K$. Then the optimization (2.14) can be written as

$$(2.15) \quad \begin{aligned} \max_{\mathbf{q}_1, \dots, \mathbf{q}_K} \quad & \sum_{k=1}^K \mathbf{q}_k^\top W_k \mathbf{q}_k \\ \text{s.t.} \quad & \mathbf{q}_k^\top \mathbf{q}_k = 1, \quad \mathbf{q}_k^\top \mathbf{q}_j = 0 \quad (k \neq j). \end{aligned}$$

Let $\text{vec}_1(W)$ denote the first eigenvector (unit length) of W corresponding to the largest eigenvalue. If $\text{vec}_1(W_1), \dots, \text{vec}_1(W_K)$ are close to K orthonormal vectors, we will obtain a global maxima for (2.15), otherwise, we can only get a local maxima due to the fact that the optimization is non-convex and there is no closed form solution available. A greedy algorithm is developed to solve (2.15), and the details are presented in Appendix A.3.

3. Simulation studies. In this section, we conduct a number of simulation experiments to evaluate the efficiency of CISE algorithm. We also assess the performance of M-GRAF model in inference on the common and individual-specific

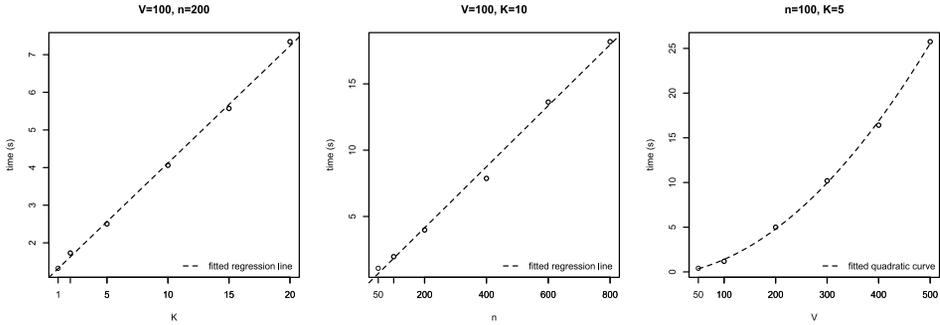


FIG. 3. Average computation time (in seconds) per iteration of CISE (Algorithm 1) for 30 runs versus latent dimension K (left), number of networks n (middle), and number of nodes V (right). All the numerical experiments are conducted in R (version 3.3.1).

components of variability in synthetic networks. CISE algorithm is implemented in both R and Matlab and all the numerical experiments are conducted in a machine with 8 Intel Core i7 3.4 GHz processor and 16 GB of RAM. The Matlab and R codes are available in the Supplementary Material [Wang, Zhang and Dunson (2019)]. The algorithm is also implemented in the R package CISE available on CRAN.

3.1. Computational performance. Each iteration of CISE includes two steps: (1) L_2 -penalized logistic regression and (2) n partial eigenvalue decompositions of $V \times V$ matrices. We simulated a sequence of Erdős–Rényi graphs (each edge is present with probability 0.5) for different numbers of nodes and then assess how the execution time increases with the problem size. Figure 3 displays the average computation time per iteration of CISE algorithm (in R) as a function of the latent dimension K , the number of networks n , and the number of nodes V . We can see that for large problem size with $n = 800$, $V = 100$, and $K = 10$, each iteration of CISE on average takes less than 20 seconds; with $V = 500$, $n = 100$, and $K = 5$, the average running time is around 25 seconds. The runtime of each CISE iteration in Matlab is similar to that in R though a bit longer for small problem size. From Figure 3, it is clearly seen that CISE exhibits a linear order with K and n , and a quadratic order with V , that is, $O(V^2nK)$, which is the same as our theoretical analysis in Algorithm 1.

CISE is a block coordinate descent algorithm, and is guaranteed to converge to a (local) mode. In our experience with simulated and real data, CISE generally converges very fast with a good initialization as specified in Algorithm 1: it usually takes less than five steps before the relative change in the joint log-likelihood becomes less than 1% even for very large problem size. Figure 4 shows how the joint log-likelihood (2.5) evolves over iterations under different problem sizes. CISE is much more efficient than the Gibbs sampler in Durante, Dunson and Vogelstein (2017) which conducts Bayesian inference on a related model to M-GRAF but

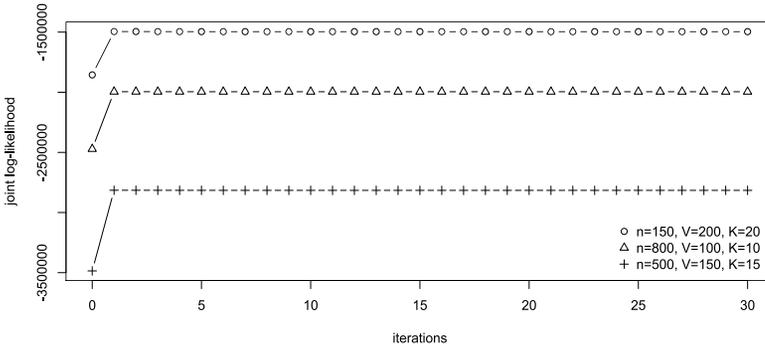


FIG. 4. CISE algorithm: joint log-likelihood over iterations under different values of n , V , and K .

could take hours or days to run for the same problem size. In practice when dealing with real brain network data, we suggest setting $\varepsilon = 0.01$ in Algorithm 1 based on our experiments.

3.2. Inference on common and individual structure. The goal in this section is to assess the performance of our proposed method in terms of inference on the shared versus individual-specific components of variability in replicated networks. To mimic the real brain network data, we first estimate Z and $\{D_i\}$ under $K = 3$ from about 800 68×68 structural brain networks extracted from HCP data. Then the networks are simulated from the M-GRAF model based on the estimated \hat{Z} and $\{\hat{D}_i\}$.

We conduct a sequence of numerical experiments to demonstrate properties of the estimated parameters in the M-GRAF model as the number of networks grows. The true values of Z and $\{D_i\}$ are denoted as Z_0 and $\{D_{i0}\}$ where each D_{i0} has rank $K = 3$. We generate different numbers n ($n = 50, 100, 200, 400, 800$) of 68×68 adjacency matrices from the M-GRAF model based on Z_0 and randomly selected D_{i0} 's. At each value of n , we run CISE algorithm with $K = 3$ to obtain the estimated parameters \hat{Z} and $\{\hat{D}_i\}$. Element-wise differences between the lower triangular entries of \hat{Z} and Z_0 and the counterpart between \hat{D}_i and D_{i0} for 20 randomly selected networks are recorded. The procedure described above is repeated 50 times where each time we randomly permute 10% of the entries in Z_0 . Figure 5 displays boxplots of the pooled differences between estimated parameters and their true values under each n across 50 simulations. Based on the plot, the differences between \hat{Z} and Z_0 seem to converge to 0 as n increases. We also notice that the differences between \hat{D}_i and D_{i0} are centered around 0 and stable across n , which is as expected since the number of parameters in $\{D_i\}$ increases with n . Figure 6 displays the estimated \hat{Z} and \hat{D}_i 's versus their corresponding true values from one experiment under $n = 800$.

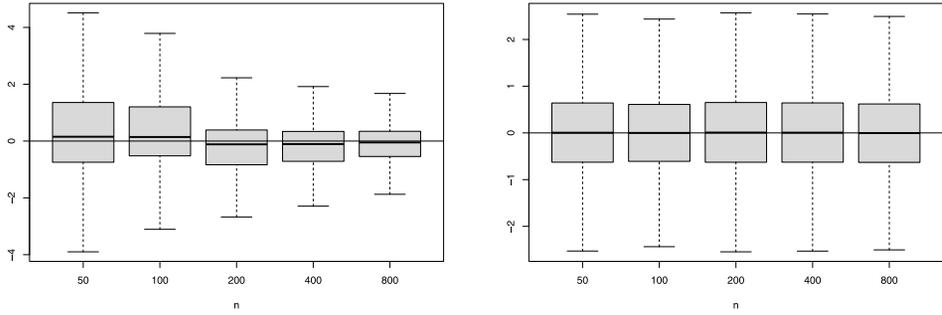


FIG. 5. Box plots of pooled differences across 50 simulations between $\mathcal{L}(\hat{Z})$ and $\mathcal{L}(Z_0)$ (left) and $\mathcal{L}(\hat{D}_i) - \mathcal{L}(D_{i0})$ for 20 randomly selected networks (right) under each number of networks n . The networks are generated from M-GRAF model with $K = 3$.

3.3. Selection of the dimensionality K . In the above simulation experiments of this section, we assume the dimensionality K is known and simply set K equal to its true value. But in practice, we face a model selection problem.

In the scenario that we have some extra categorical variable in the dataset and the goal is to do prediction, we can use cross validation to choose K as illustrated in Section 4. Otherwise we recommend the classical “Elbow” method to determine K , which requires first running CISE algorithm for a sequence of K ’s and plotting the joint log-likelihood (2.5) at convergence versus dimension K . Then the optimal K is determined to be the bend point where the objective function starts to increase slowly as shown in Figure 7. The plot implies that the bend point is at $K = 3$ for different numbers n of networks, which coincides with the true dimension in our data generating process. Based on our study, this approach outperforms AIC or BIC particularly when n is large.

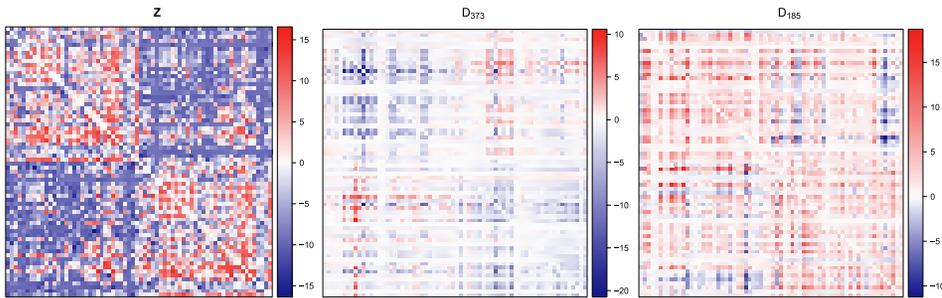


FIG. 6. Level plots for estimated parameters (lower triangular) versus their true values (upper triangular) with $n = 800$ and $K = 3$. Left: \hat{Z} versus Z_0 ; Middle and Right: \hat{D}_i versus the true D_{i0} for two networks, where the 373th network (middle) has the lowest network density and the 185th network (right) has the highest network density in the synthetic data.

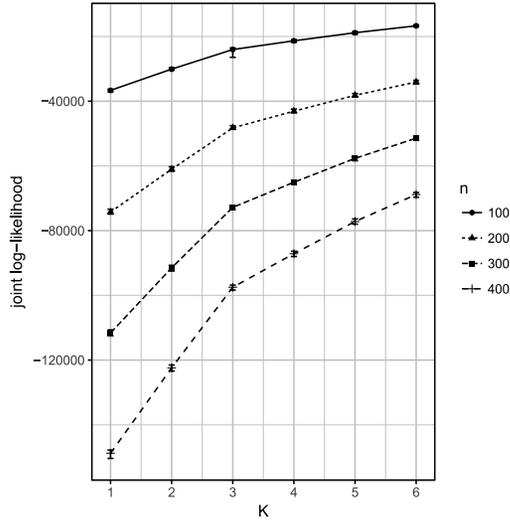


FIG. 7. Mean joint log-likelihood at convergence of CISE algorithm with their 95% confidence intervals across 50 simulations versus dimension K under different numbers of replicated networks. The networks are simulated from M-GRAF model with $K = 3$ as described in Section 3.2.

4. Applications to structural brain networks. In this section, we apply M-GRAF to two real datasets involving 256 HCP subjects: (1) HCP scan-rescan dataset and (2) a subset of HCP 1, 200 subjects dataset. Each subject is preprocessed using a state-of-the-art dMRI preprocessing pipeline [Zhang et al. (2018a)] to extract 68×68 binary structural networks based on the Desikan parcellation atlas [Desikan et al. (2006)]. Certainly, even state-of-the-art tractography is subject to measurement errors, but ground truth measurements on actual neurofibers are unavailable given current technology. Hence, there will be two components of variability in the measured brain networks, one attributed to systematic variability across subjects in their brain connection structure, and one due to measurement errors. Our model can accommodate these two components of variability, with the low-rank assumption on individual deviation not only capturing the main variation of each graph but also serving as a denoising procedure.

In addition to the network data, we also extract a cognitive trait, measuring the subject's visuospatial processing ability, to study the relationship between brain connectivity and this cognitive score.

4.1. Scan-rescan brain network data. In this application, we compare the performance of CISE (Algorithm 1) with several other low-rank approximation methods on the scan-rescan brain network data. The data were collected for 44 healthy subjects under a scan-rescan session, and therefore two 68×68 binary adjacency matrices are available for each subject for a total of $n = 88$ brain networks. Two

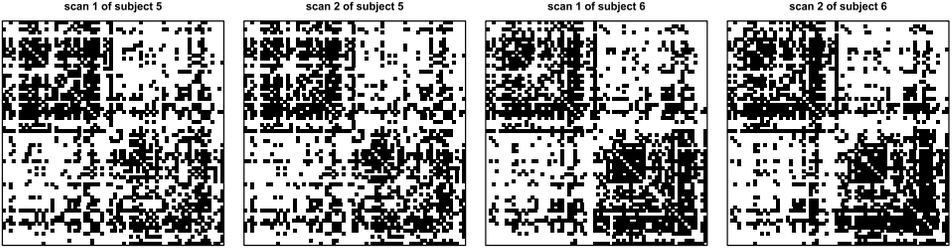


FIG. 8. Paired adjacency matrices for two subjects in the HCP scan-rescan data.

examples of the scan-rescan networks extracted for two subjects are shown in Figure 8. It is easy to observe that differences between scan-rescan adjacency matrices for the same subject are much smaller than those between the adjacency matrices for different subjects

These scan-rescan data provide an appealing setting for studying how discriminative the latent structure can be in identification of subjects. The idea is to first learn a low-rank representation for each brain network and then check whether the pairs of networks having the closest low-rank representations correspond to the same subjects. Specifically, we use the distance measure $d(i, j)$ between scan i and j as introduced in Section 2.5 and then conduct leave-one-out cross validation (LOOCV): for a test subject i^* , find $j^* = \arg \min_{j \neq i^*} d(i^*, j)$ and check if i^* and j^* correspond to the same person. Similarly, for the model where $D_i = Q_i \Lambda Q_i^\top$, the pairwise distance is defined as

$$\begin{aligned} d^2(i, j) &= \|Q_i \Lambda Q_i^\top - Q_j \Lambda Q_j^\top\|_F^2 \\ &= 2 \operatorname{tr}(\Lambda^2) - 2 \operatorname{tr}(\Lambda Q_i^\top Q_j \Lambda Q_j^\top Q_i). \end{aligned}$$

Another variant of our model $D_i = Q \Lambda_i Q^\top$ does not provide good fit to the data and thus we do not display the results below.

We compare the performance of CISE with some popular matrix and tensor decompositions on multiple network data as below. For a fair comparison, we apply these low-rank approximation methods to the demeaned adjacency matrices $\{A_i - \bar{A} : i = 1, \dots, n\}$ where $\bar{A} = \sum A_i / n$, so as to better capture the deviation of each network from their common structure.

- Separate factorization. We apply the spectral embedding method [Sussman et al. (2012)] separately to each network in the dataset where each probability matrix Π_i is estimated by the sum of \bar{A} and a low rank approximation to $(A_i - \bar{A})$ via SVD.
- CP decomposition. Let \mathcal{A}_d denote the $V \times V \times n$ tensor of demeaned adjacency matrices. The CP decomposition seeks to model \mathcal{A}_d as a sum of rank-one tensors: $\mathcal{A}_d \approx \sum_{k=1}^K d_k \mathbf{u}_k \circ \mathbf{v}_k \circ \mathbf{w}_k$, where $\mathbf{u}_k \in \mathbb{R}^V$, $\mathbf{v}_k \in \mathbb{R}^V$, $\mathbf{w}_k \in \mathbb{R}^n$, $d_k \geq 0$ and \circ denotes the outer product [Kolda and Bader (2009)]. Unlike the singular

TABLE 1

LOOCV identification accuracy on scan-rescan data for different methods

	M-GRAF1 $D_i = Q_i \Lambda_i Q_i^\top$	M-GRAF2 $D_i = Q_i \Lambda Q_i^\top$	Separate factorization	CP decomposition	Tucker decomposition
$K = 2$	0.705	0.761	0.761	0.114	0.136
$K = 5$	0.886	0.932	0.932	0.477	0.670
$K = 7$	0.966	0.989	0.943	0.591	0.761
$K = 8$	0.977	1.000	0.966	0.625	0.841

value decomposition (SVD) for a matrix, CP decomposition does not uniquely decompose the data [Kolda and Bader (2009)], which may complicate the analysis. Similar to Sussman et al. (2012), each probability matrix Π_i is estimated by $\hat{\Pi}_i = \bar{A} + \sum_{k=1}^K d_k w_{ki} \mathbf{u}_k \mathbf{v}_k^\top$, where w_{ki} is the i th entry of \mathbf{w}_k .

- Tucker decomposition. Tucker decomposition seeks to model \mathcal{A}_d as $\mathcal{A}_d \approx \mathcal{D} \times_1 U_1 \times_2 U_2 \times_3 W$ where \mathcal{D} is a $K_1 \times K_2 \times K_3$ core tensor and the factors $U_1 \in \mathbb{R}^{V \times K_1}$, $U_2 \in \mathbb{R}^{V \times K_2}$ and $W \in \mathbb{R}^{n \times K_3}$ are orthonormal matrices [Kolda and Bader (2009)]. We set $K_1 = K_2 = K$ and $K_3 = n$ in this case, and again we consider each matrix along the third dimension of the low rank tensor plus \bar{A} as the estimated probability matrix $\hat{\Pi}_i$.

We use R package `rTensor` to compute the components in tensor decompositions. The distance measure in these methods is defined as $d(i, j) := \|\hat{\Pi}_i - \hat{\Pi}_j\|_F$. We report the LOOCV accuracy of subject identification on the scan-rescan data in Table 1. The results show that the accuracy from the variant of our model $D_i = Q_i \Lambda Q_i^\top$ is always the highest under the same rank K and reaches 1 at $K = 8$. Although separate factorization has the same accuracy as our model at rank 2 and 5, its accuracy increases more slowly with K . The two tensor decomposition methods have poor classification performance here, implying that their low rank approximations are not discriminative enough in this scenario.

After obtaining a discriminative latent structure, we want to further check how well edges in the networks can be predicted. We compute the area under the ROC curve (AUC) in predicting $\mathcal{L}(A_i)$ with estimated probability matrix $\hat{\Pi}_i$ and the residual sum of squares (RSS), that is, the L_2 -norm of the difference between $\mathcal{L}(A_i)$ and $\mathcal{L}(\hat{\Pi}_i)$. The mean and standard deviation of AUC and RSS across all the subjects are reported in Table 2, which shows that CISE has higher AUC and lower RSS than other methods with the same rank K . The results from the two variants of our model are quite similar, though allowing Λ_i to vary across individuals performs slightly better due to more flexibility.

We assess goodness-of-fit by comparing some key topological features of networks observed in the data to those estimated from different methods. The selected topological measures include network density, average shortest path length, transitivity, and mean of node degrees (degree mean) [Newman (2010)]. Specifically,

TABLE 2
Mean and standard deviation of AUC and RSS across subjects under different K 's

	M-GRAF1 $D_i = Q_i \Lambda_i Q_i^\top$	M-GRAF2 $D_i = Q_i \Lambda Q_i^\top$	Separate factorization	CP decomposition	Tucker decomposition
	AUC				
$K = 2$	0.9880 \pm 0.0024	0.9877 \pm 0.0024	0.9846 \pm 0.0031	0.9758 \pm 0.0043	0.9758 \pm 0.0043
$K = 5$	0.9948 \pm 0.0014	0.9945 \pm 0.0014	0.9928 \pm 0.0017	0.9768 \pm 0.0042	0.9777 \pm 0.0040
$K = 7$	0.9969 \pm 0.0009	0.9968 \pm 0.0009	0.9959 \pm 0.0011	0.9774 \pm 0.0037	0.9791 \pm 0.0037
$K = 8$	0.9976 \pm 0.0008	0.9974 \pm 0.0007	0.9970 \pm 0.0008	0.9779 \pm 0.0036	0.9800 \pm 0.0037
	$\ \mathcal{L}(A_i) - \mathcal{L}(\hat{\Pi}_i)\ _2$				
$K = 2$	9.63 \pm 0.51	9.68 \pm 0.52	10.75 \pm 0.46	11.62 \pm 0.51	11.61 \pm 0.51
$K = 5$	7.65 \pm 0.53	7.75 \pm 0.53	9.63 \pm 0.40	11.50 \pm 0.50	11.41 \pm 0.49
$K = 7$	6.64 \pm 0.48	6.72 \pm 0.52	8.99 \pm 0.37	11.42 \pm 0.47	11.27 \pm 0.48
$K = 8$	6.18 \pm 0.50	6.34 \pm 0.50	8.69 \pm 0.36	11.38 \pm 0.47	11.19 \pm 0.48

we first obtain the predictive distributions of these topological measures for each subject by simulating 100 networks from the estimated $\hat{\Pi}_i$ under different models, and then compare the predictive means to the empirical topological features via scatterplots along with 95% confidence intervals as shown in Figure 9. Each dot in these scatterplots corresponds to a subject with x -coordinate being her empirical topological measure and y -coordinate the predictive mean. The closer the points are to the dashed diagonal line, the better fit of the model. For a fair comparison, we choose $K = 17$ for Separate factorization and $K = 36$ for Tucker decomposition in Figure 9 since these choices of K provide an accuracy of 1 for the two methods in the scan-rescan classification task. We set $K = 100$ for CP decomposition, which provides an accuracy of around 0.989. Figure 9 shows that the two variants of our model provide much better characterization of network topological features than the other methods. In addition, the variant $D_i = Q_i \Lambda Q_i^\top$ provides almost indistinguishable predictive results from those under the M-GRAF model with $D_i = Q_i \Lambda_i Q_i^\top$. Therefore, restricting Λ_i to be the same across subjects seems to be a reasonable assumption for brain network data.

4.2. *Brain networks and cognitive traits.* The HCP collects measurements on a range of motor, sensory, cognitive, and emotional processes for each participant, with an overarching goal being improved understanding of the relationship between brain connectivity and human traits [Barch et al. (2013)]. For sake of clarity and brevity, we focus here on studying relationships between brain structural connectivity and one particular trait—visuospatial processing.

Visuospatial processing is commonly assessed using the Variable Short Penn Line Orientation Test (VSPLLOT), where two line segments are presented on the screen and participants are asked to rotate a movable line so that it is parallel to

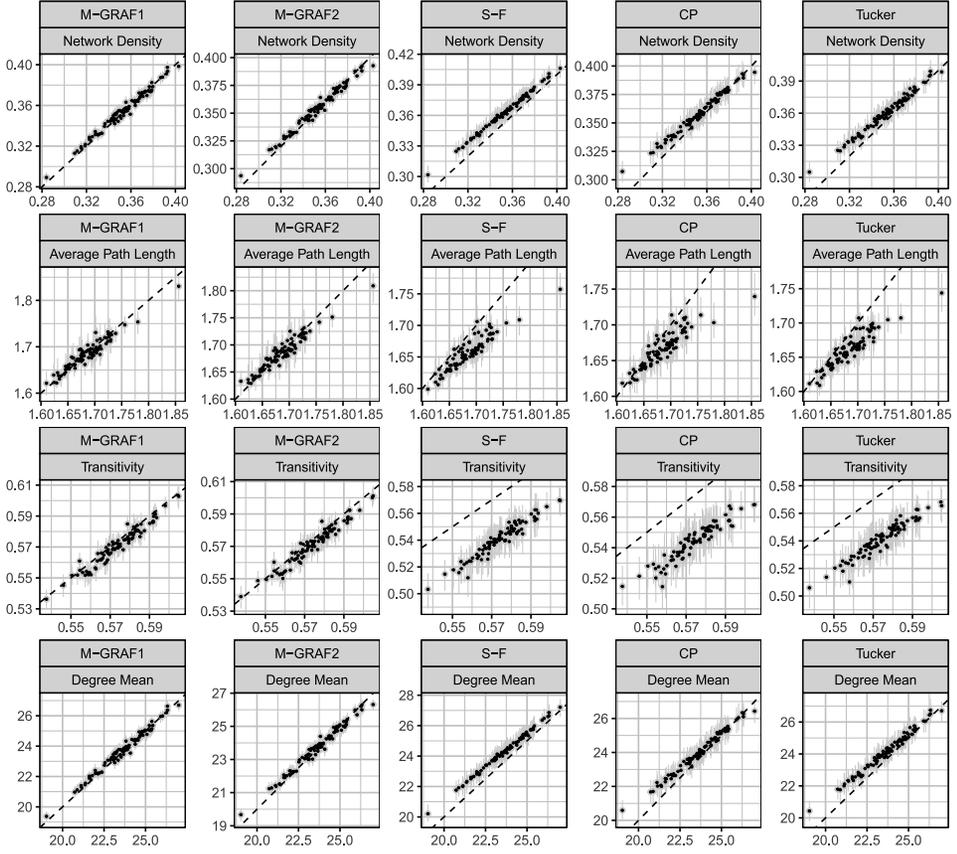


FIG. 9. Goodness-of-fit assessment for selected network topological features under different methods. The methods from left to right: M-GRAF1 is M-GRAF with $D_i = Q_i \Lambda_i Q_i^\top$ under $K = 8$; M-GRAF2 is M-GRAF with $D_i = Q_i \Lambda Q_i^\top$ under $K = 8$; S-F is separate factorization with $K = 17$; CP decomposition with $K = 100$ and Tucker decomposition with $K = 36$. The topological features from top to bottom are network density, average shortest path length, transitivity, and degree mean. Each dot of the scatterplot corresponds to a subject, where x-coordinate denotes her observed topological feature, y-coordinate denotes the corresponding predictive mean, and the grey segment denotes the 95% predictive confidence interval. The dashed line in each scatterplot denotes the $y = x$ line.

the fixed line; for more details, we refer the readers to Moore et al. (2015). The latest released HCP data contain VSPLLOT scores of about 1200 healthy adults. We preselected subjects having high (top 10%) and low (bottom 10%) VSPLLOT scores with 106 subjects in each group. Hence the resulting dataset contains an indicator of high/low visuospatial processing score $l_i \in \{0, 1\}$ and an adjacency matrix A_i representing the structural connectivity among 68 brain regions for 212 individuals.

TABLE 3
*Mean and standard deviation of
 prediction accuracies in repeated 10-fold
 cross validation*

K	Accuracy
1	0.561 \pm 0.101
2	0.621 \pm 0.107
3	0.622 \pm 0.104
4	0.623 \pm 0.105
5	0.643 \pm 0.102
6	0.641 \pm 0.104
7	0.629 \pm 0.105

We followed the same goodness-of-fit assessment procedure as described in Section 4.1 and observed very similar performance between the models $D_i = Q_i \Lambda Q_i^\top$ and $D_i = Q_i \Lambda_i Q_i^\top$. Therefore, we choose the variant $D_i = Q_i \Lambda Q_i^\top$ to further reduce the number of parameters. We use the distance described in Section 2.5 to classify subjects with high and low visuospatial processing score using their estimated low-rank components $\{Q_i\}$ and Λ . The prediction accuracy is measured by repeating 10-fold cross validation (CV) 30 times. We report the mean and standard deviation of the CV accuracies under different choices of K in Table 3. It seems that $K = 5$ is enough to provide a good prediction accuracy of 0.643 on average, implying that individual-specific components of brain connectivity are related to visuospatial processing. The estimated common structure \hat{Z} of brain connectivity underlying all subjects is displayed via a heatmap in Figure 10. The chord diagram in Figure 10 shows the selected 277 edges with $\pi(\hat{Z}_{uv}) > 0.999$, where $\pi(\cdot)$ is the logistic function. Hence we expect these connections to be present with probability almost 1 for an average brain of 212 HCP subjects.

Table 3 also shows that $K = 2$ leads to a jump in performance relative to $K = 1$. Since $\hat{\lambda} = 83.6$ under $K = 1$ and $(\hat{\lambda}_1, \hat{\lambda}_2) = (77.6, -71.7)$ under $K = 2$, we display the second column of Q_i corresponding to λ_2 via a heatmap across the 68 brain regions for two subjects in Figure 11 (their adjacency matrices are shown in Figure 1). According to Moore et al. (2015), visuospatial processing is linked to posterior cortical function and thus we focus on the regions in the occipital lobe, which is located in the posterior portion of the human cerebral cortex and is the visual processing center of the brain containing most of the anatomical regions of the visual cortex [Zeki et al. (1991)]. Subject 1 in the left plot of Figure 11 has the lowest score in VSPLLOT and we can see that her brain regions located in the occipital lobe (bottom of the plot) all have similar positive coordinates. Since $\hat{\lambda}_2 < 0$, this indicates that Subject 1 tends to have few connections within the occipital lobe. Subject 211 in the right plot of Figure 11 has the highest score in VSPLLOT and the

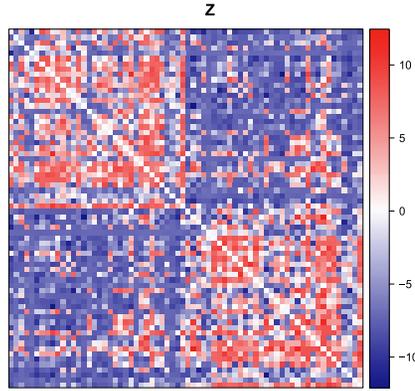


FIG. 10. Heatmap of the estimated \hat{Z} under $K = 5$ (left) and a chord diagram of the connections uv 's (277 in total) with $\pi(\hat{Z}_{uv}) > 0.999$ (right), where $\pi(\cdot)$ is the logistic function.

coordinates of her brain regions in occipital lobe are not similar, indicating more connections within this lobe.

To identify a subnetwork that might relate to visuospatial processing, we test for differences in the log odds of each connection between the two groups. Specifically, for each connection uv in the brain network, we applied a t -test on the $D_{i[uv]}$'s in high and low visuospatial functioning groups under $K = 5$. We adjusted for multiple comparisons by rejecting all local nulls having a p -value below the Benjamini and Hochberg (1995) threshold to maintain a false discovery rate $FDR \leq 0.15$. The significant connections are displayed via a chord diagram in Figure 12. Figure 12 shows that many connections in the selected subnetwork relate to

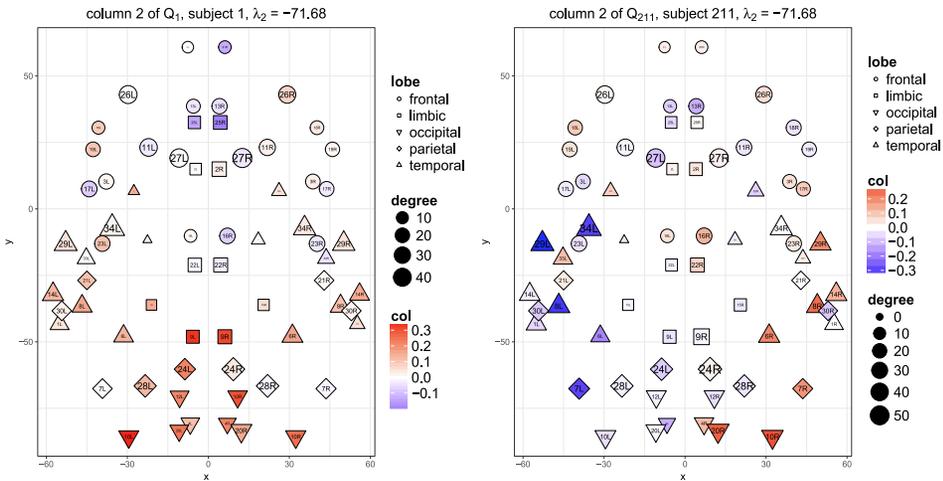


FIG. 11. Heatmap of the second column of Q_i for subject $i = 1$ (left) and $i = 211$ (right) under $K = 2$. Their adjacency matrices are presented in Figure 1.

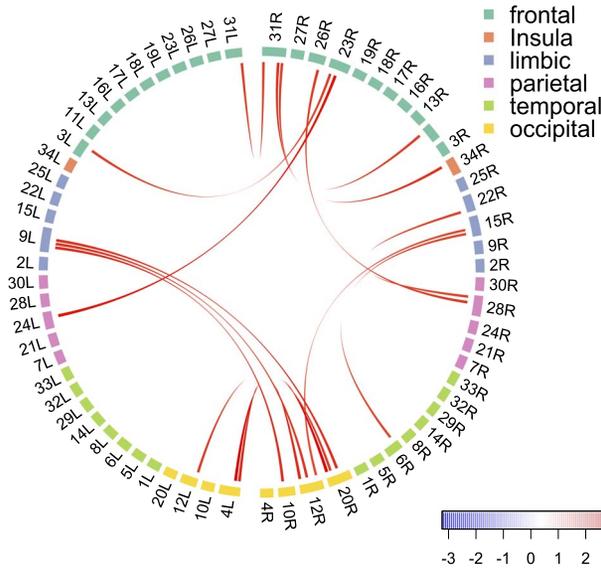


FIG. 12. Selected subnetwork that might be related to the visuospatial processing: significant connections (15 in total) in the t -test of $D_{i[uv]}$'s between high and low visuospatial processing group for each edge uv under $FDR \leq 0.15$. The color of the chords represents the corresponding t statistic, which goes from blue to red as t statistic goes from -3.20 (minimum among all connections) to 3.88 (maximum).

regions in the occipital lobe, especially the right occipital lobe. This seems consistent with neuroimaging and lesion studies which provide evidence of dysfunction in right posterior regions of the brain for deficits in visuospatial processing [Moore et al. (2015)]. In particular in the occipital lobe, Region 12R (right lingual) and 20R (right peri calcarine) in Figure 12 seem to be the most affected regions related to visuospatial processing since they have more connections with differences between the two groups. This agrees with the findings that damage to the lingual gyrus leads to a form of topographic disorientation [Kravitz et al. (2011)] and abnormalities in calcarine sulcus, which is a key node of the ventral visual pathway, are related to impaired visual information processing [Wu et al. (2015)].

5. Conclusion. In this paper, we develop a framework for studying common and individual structure of multiple binary undirected networks with similar patterns. Two variants of the model have been proposed to account for different degrees of heterogeneity in the data so as to avoid overfitting. We have developed an efficient algorithm—CISE—for estimating the model based on spectral decomposition. Simulation studies have illustrated the fast computation of CISE algorithm on large datasets and good properties in inference. We also demonstrated how accounting for common structure can lead to a much lower dimensional individual latent structure, which is highly discriminative in scan-rescan identification. Our

approach also provides better prediction and goodness-of-fit (in terms of topological properties) to brain network data than some popular dimension-reduction methods.

Although CISE algorithm has good performance when the latent dimension K is small, it can get trapped at some local modes when K is large due to high dimensionality of the parameter space. A multi-resolution approach might be a solution to this issue, where we apply a coarse to fine factorization of Z and the estimates of the parent entries in the previous layer provide prior information for the daughter entries in the next layer. This technique may prevent some parameters from getting trapped in local modes leading to a better optima.

APPENDIX: PROOFS OF PROPOSITIONS

This appendix contains proofs of Proposition 2.1 and Proposition 2.2 in Section 2.2 as well as an algorithm for inference of Q in the variant $D_i = Q\Lambda_i Q^\top$.

A.1. Proof of Proposition 2.1.

PROOF. Note that the expression inside the brackets of (2.5) is a univariate function of $D_{i[uv]}$ given Z_{uv} . Let $x = D_{i[uv]}$, $a = A_{i[uv]}$, $\mu = Z_{uv}$. Then $\Pi_{i[uv]} = \pi(\mu + x)$, where $\pi(x) := 1/[1 + \exp(-x)] \in (0, 1)$. Let $h(x) := \log[1 - \pi(x)]$. Then

$$(A.1) \quad h'(x) = \frac{-\pi'(x)}{1 - \pi(x)} = \frac{-\pi(x)(1 - \pi(x))}{1 - \pi(x)} = -\pi(x).$$

Consider μ as known and the expression inside the brackets of (2.5) is defined as

$$f(x) := a\mu + ax + h(\mu + x).$$

Next we will show that given μ ,

$$(A.2) \quad \arg \max_x f(x) = \arg \max_x [a - \pi(\mu)]x.$$

According to (A.1), the first derivative of $f(x)$ becomes

$$f'(x) = a - \pi(\mu + x).$$

Note that $a \in \{0, 1\}$ since a is the realization of a binary random variable.

(i) $a = 1$. $f'(x) = 1 - \pi(\mu + x) > 0$ indicating that $f(x)$ is maximized at $x = +\infty$ which also maximizes $[a - \pi(\mu)]x$.

(ii) $a = 0$. $f'(x) = -\pi(\mu + x) < 0$ indicating that $f(x)$ is maximized at $x = -\infty$ which also maximizes $[a - \pi(\mu)]x$.

Then (A.2) is verified based on (i) and (ii) and the optimal $\{D_i : i = 1, \dots, n\}$ maximizing (2.5) given Z can be written as

$$\begin{aligned} & \arg \max_{\{D_i\}} \sum_{i=1}^n \sum_{u=1}^V \sum_{v<u} [A_{i[uv]}(Z_{uv} + D_{i[uv]}) + \log(1 - \Pi_{i[uv]})] \\ &= \arg \max_{\{D_i\}} \sum_{i=1}^n \sum_{u=1}^V \sum_{v<u} [A_{i[uv]} - \pi(Z_{uv})] D_{i[uv]} \\ &= \arg \max_{\{D_i\}} \sum_{i=1}^n \frac{1}{2} \operatorname{tr}([A_i - \pi(Z)] D_i). \end{aligned}$$

The last line follows because $\pi(Z)$ and each A_i are symmetric matrices and their diagonal elements are set at 0. \square

A.2. Proof of Proposition 2.2.

PROOF. It suffices to prove (2.8) as (2.9) follows by replacing B with $-B$ noting that $\sigma_j(-B) = -\sigma_{V-j+1}(B)$, $j = 1, \dots, V$.

We do induction on dimension k and first verify the $k = 1$ case. By Rayleigh–Ritz Theorem [Parlett (1998)], for any unit vector $\mathbf{u} \in \mathbb{R}^V$ we have

$$\max_{\mathbf{u}^\top \mathbf{u} = 1} \mathbf{u}^\top B \mathbf{u} = \sigma_1(B).$$

Since $c_1 > 0$, then $\max_{\mathbf{u}^\top \mathbf{u} = 1} c_1 \mathbf{u}^\top B \mathbf{u} = c_1 \sigma_1(B)$. So (2.8) holds for $k = 1$.

Assume (2.8) holds for $k = j - 1$. We now show that (2.8) also holds for $k = j$.

Let $\mathbf{u}_1, \dots, \mathbf{u}_j$ be an orthonormal basis of a j -dimensional subspace U in \mathbb{R}^V . We define a scaled partial trace to represent the objective function in (2.8) for notational simplicity in the rest proof:

$$\operatorname{ptr}(B \mid U, c_{1:j}) := \sum_{i=1}^j c_i \mathbf{u}_i^\top B \mathbf{u}_i.$$

Then

$$\max_{\dim(U)=j} \operatorname{ptr}(B \mid U, c_{1:j}) = \max_{\mathbf{u}_1, \dots, \mathbf{u}_j} \sum_{i=1}^j c_i \mathbf{u}_i^\top B \mathbf{u}_i$$

for any orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_j\}$ in \mathbb{R}^V .

According to Courant–Fischer Theorem [Parlett (1998)],

$$(A.3) \quad \sigma_j(B) = \max_{\dim(U)=j} \min_{\mathbf{u} \in U : \mathbf{u}^\top \mathbf{u} = 1} \mathbf{u}^\top B \mathbf{u}.$$

Then for every j -dimensional subspace U of \mathbb{R}^V and any orthonormal basis of U , $\mathbf{u}_1, \dots, \mathbf{u}_j$, there is some \mathbf{u}_m ($m \in \{1, \dots, j\}$) such that $\mathbf{u}_m^\top \mathbf{B} \mathbf{u}_m \leq \sigma_j(B)$. Since $c_j > 0$, then

$$(A.4) \quad c_j \mathbf{u}_m^\top \mathbf{B} \mathbf{u}_m \leq c_j \sigma_j(B).$$

The remaining vectors $\{\mathbf{u}_i : i \neq m\}$ is also an orthonormal basis of a $(j-1)$ -dimensional subspace \tilde{U} . By induction,

$$(A.5) \quad \text{ptr}(B \mid \tilde{U}, c_{1:(j-1)}) \leq \sum_{i=1}^{j-1} c_i \sigma_i(B).$$

Adding the two inequalities (A.4) and (A.5), we have

$$\text{ptr}(B \mid U, c_{1:j}) \leq c_1 \sigma_1(B) + \dots + c_j \sigma_j(B)$$

for any j -dimensional subspace U . Therefore

$$\max_{\dim(U)=j} \text{ptr}(B \mid U, c_{1:j}) \leq c_1 \sigma_1(B) + \dots + c_j \sigma_j(B).$$

On the other hand, by selecting U to be the span of the first j orthonormal eigenvectors of B , we obtain the reverse inequality

$$\max_{\dim(U)=j} \text{ptr}(B \mid U, c_{1:j}) \geq c_1 \sigma_1(B) + \dots + c_j \sigma_j(B). \quad \square$$

A.3. Inference of Q in the joint embedding model $D_i = Q \Lambda_i Q^\top$. We are going to solve $\mathbf{q}_1, \dots, \mathbf{q}_K$ sequentially, where \mathbf{q}_k is the k th column of Q . Let $\text{eval}_1(W)$ denote the largest eigenvalue of W . Suppose $s_1 = \arg \max_k \text{eval}_1(W_k)$. Then set $\mathbf{q}_{s_1} = \text{evec}_1(W_{s_1})$. To decide the next \mathbf{q}_k to update, let U be a $V \times (V-1)$ matrix comprising of a set of orthonormal basis of the space orthogonal to \mathbf{q}_{s_1} . For $k \neq s_1$, we know $\mathbf{q}_k \in \text{span}(U)$ and hence assume $\mathbf{q}_k = U \mathbf{a}_k$ for some vector $\mathbf{a}_k \in \mathbb{R}^{V-1}$. $\mathbf{q}_k^\top \mathbf{q}_k = 1$ implies that $\mathbf{a}_k^\top U^\top U \mathbf{a}_k = \mathbf{a}_k^\top \mathbf{a}_k = 1$. So \mathbf{a}_k is of unit length. Then the optimization problem $\max_{\mathbf{q}_k} \{\mathbf{q}_k^\top W_k \mathbf{q}_k : \mathbf{q}_k^\top \mathbf{q}_k = 1; \mathbf{q}_k^\top \mathbf{q}_{s_1} = 0\}$ transforms to the optimization

$$(A.6) \quad \begin{aligned} \max_{\mathbf{a}_k \in \mathbb{R}^{V-1}} \mathbf{a}_k^\top U^\top W_k U \mathbf{a}_k \\ \text{s.t. } \mathbf{a}_k^\top \mathbf{a}_k = 1. \end{aligned}$$

By Rayleigh–Ritz Theorem, the solution of \mathbf{a}_k is $\text{evec}_1(U^\top W_k U)$ and

$$\max\{\mathbf{q}_k^\top W_k \mathbf{q}_k : \mathbf{q}_k^\top \mathbf{q}_k = 1, \mathbf{q}_k^\top \mathbf{q}_{s_1} = 0\} = \text{eval}_1(U^\top W_k U).$$

Note that the eigenvalues of a symmetric matrix are invariant to orthogonal transformation, that is, $\text{eval}_1(U^\top W_k U) = \text{eval}_1(R^\top U^\top W_k U R)$ for any orthogonal matrix R . Hence U in (A.6) can be an arbitrary orthonormal basis of the subspace \mathbf{q}_k^\perp . Suppose $s_2 = \arg \max_{k \neq s_1} \text{eval}_1(U^\top W_k U)$. Then set $\mathbf{q}_{s_2} = U \mathbf{a}_{s_2}$ where $\mathbf{a}_{s_2} = \text{evec}_1(U^\top W_{s_2} U)$. Repeat the above process and we can obtain the other \mathbf{q}_k 's. We summarize the procedure to solve Q given Z and $\{\lambda_{ik}\}$ in Algorithm 2.

Algorithm 2: Inference of Q in the joint embedding model $D_i = Q \Lambda_i Q^\top$

1 Let $W_k = \sum_{i=1}^n \lambda_{ik} [A_i - \pi(Z)]$, $k = 1, \dots, K$.
 2 Find $s_1 = \arg \max_k \text{eval}_1(W_k)$ and set $\mathbf{q}_{s_1} = \text{evec}_1(W_{s_1})$.
 3 **for** $k = 2 : K$ **do**
 4 find a set of orthogonal basis U of the subspace $\text{span}(\mathbf{q}_{s_1}, \dots, \mathbf{q}_{s_{(k-1)}})^\perp$;
 5 find $s_k = \arg \max_{k \in \{1, \dots, K\} \setminus \{s_1, \dots, s_{(k-1)}\}} \text{eval}_1(U^\top W_k U)$;
 6 set $\mathbf{q}_{s_k} = U \mathbf{a}_{s_k}$ where $\mathbf{a}_{s_k} = \text{evec}_1(U^\top W_{s_k} U)$;
 7 **end**
Output: $Q = (\mathbf{q}_1, \dots, \mathbf{q}_K)$.

Acknowledgements. We would like to thank David Choi for insightful comments, and Peter Hoff for useful comments on method comparisons.

SUPPLEMENTARY MATERIAL

Code and Data (DOI: [10.1214/18-AOAS1193SUPP](https://doi.org/10.1214/18-AOAS1193SUPP); .zip). The R and Matlab codes for CISE algorithm and the HCP data can be found in the file above.

REFERENCES

- BARCH, D. M., BURGESS, G. C., HARMS, M. P., PETERSEN, S. E., SCHLAGGAR, B. L., CORBETTA, M., GLASSER, M. F., CURTISS, S., DIXIT, S., FELDT, C. et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80** 169–189.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](https://doi.org/10.1111/j.1467-9868.1995.tb01704.x)
- DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980. DOI:[10.1016/j.neuroimage.2006.01.021](https://doi.org/10.1016/j.neuroimage.2006.01.021).
- DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2014). Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Trans. Signal Process.* **62** 905–918. [MR3160322](https://doi.org/10.1109/TSP.2013.2282222)
- DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. [MR3750873](https://doi.org/10.1080/01621459.2017.1358888)
- FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. [MR1225212](https://doi.org/10.1093/biomet/80.1.27)
- GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. [MR2655663](https://doi.org/10.1214/08-ANAS367)
- GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. [MR1908073](https://doi.org/10.1073/pnas.020610999)
- GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E., AIROLDI, E. M. et al. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.
- FRIEDMAN, J. HASTIE, T. J. and TIBSHIRANI, R. J. (1990). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

- HEINZE, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25** 4216–4226. [MR2307586](#)
- HOFF, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 657–664. Curran Associates, Inc., Red Hook, NY.
- HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. [MR1951262](#)
- KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. [MR2535056](#)
- KRAVITZ, D. J., SALEEM, K. S., BAKER, C. I. and MISHKIN, M. (2011). A new neural framework for visuospatial processing. *Nat. Rev., Neurosci.* **12** 217–230.
- LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. [MR3086429](#)
- MINKA, T. P. (2003). A comparison of numerical optimizers for logistic regression. Unpublished manuscript.
- MOORE, T. M., SCOTT, J. C., REISE, S. P., PORT, A. M., JACKSON, C. T., RUPAREL, K., SAVITT, A. P., GUR, R. E. and GUR, R. C. (2015). Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. *Psychol. Assess.* **27** 955–964.
- NEWMAN, M. E. J. (2010). *Networks: An Introduction*. Oxford Univ. Press, Oxford. [MR2676073](#)
- O’CONNOR, L., MÉDARD, M. and FEIZI, S. (2015). Clustering over logistic random dot product graphs. *Stat* **1050** 3.
- PARLETT, B. N. (1998). *The Symmetric Eigenvalue Problem. Classics in Applied Mathematics* **20**. SIAM, Philadelphia, PA. Corrected reprint of the 1980 original. [MR1490034](#)
- SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. [MR3010899](#)
- TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In *IEEE International Conference on Data Mining* 1016–1021.
- TANG, R., KETCHA, M., VOGELSTEIN, J. T., PRIEBE, C. E. and SUSSMAN, D. L. (2016). Law of large graphs. ArXiv preprint. Available at [arXiv:1609.01672](#).
- TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. [MR0205395](#)
- VAN ESSEN, D. C., UGURBIL, K., AUERBACH, E., BARCH, D., BEHRENS, T., BUCHOLZ, R., CHANG, A., CHEN, L., CORBETTA, M., CURTISS, S. W. et al. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage* **62** 2222–2231.
- WANG, L., ZHANG, Z. and DUNSON, D. (2019). Supplement to “Common and individual structure of brain networks.” DOI:[10.1214/18-AOAS1193SUPP](#).
- WOOLFE, F., LIBERTY, E., ROKHLIN, V. and TYGERT, M. (2008). A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.* **25** 335–366. [MR2455599](#)
- WU, X., LV, X.-F., ZHANG, Y.-L., WU, H.-W., CAI, P.-Q., QIU, Y.-W., ZHANG, X.-L. and JIANG, G.-H. (2015). Cortical signature of patients with HBV-related cirrhosis without overt hepatic encephalopathy: A morphometric analysis. *Front. Neuroanat.* **9** 82.
- ZEKI, S., WATSON, J., LUECK, C., FRISTON, K. J., KENNARD, C. and FRACKOWIAK, R. (1991). A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11** 641–649.
- ZHANG, Z., DESCOTEAUX, M., ZHANG, J., GIRARD, G., CHAMBERLAND, M., DUNSON, D., SRIVASTAVA, A. and ZHU, H. (2018a). Mapping population-based structural connectomes. *NeuroImage* **172** 130–145.

ZHANG, Z., ALLEN, G., ZHU, H. and DUNSON, D. (2018b). Relationships between human brain structural connectomes and traits. *BioRxiv* 256933.

L. WANG
DEPARTMENT OF STATISTICS
CENTRAL SOUTH UNIVERSITY
CHANGSHA HUNAN 410083
P.R. CHINA
E-MAIL: wangronglu22@gmail.com

Z. ZHANG
DEPARTMENT OF BIostatISTICS
AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF ROCHESTER
ROCHESTER, NEW YORK 14642
USA
E-MAIL: Zhengwu_Zhang@URMC.Rochester.edu

D. DUNSON
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: dunson@duke.edu