The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2095–2117
https://doi.org/10.1214/22-AOAS1709

GRAPH-AWARE MODELING OF BRAIN CONNECTIVITY NETWORKS

BY YURA KIM[a], DANIEL KESSLER[b] AND ELIZAVETA LEVINA[c]

Department of Statistics, University of Michigan, [a]*kimyr@umich.edu,* [b]*kesslerd@umich.edu,* [c]*elevina@umich.edu*

Functional connections in the brain are frequently represented by weighted networks, with nodes representing locations in the brain and edges representing the strength of connectivity between these locations. One challenge in analyzing such data is that inference at the individual edge level is not particularly biologically meaningful; interpretation is more useful at the level of so-called functional systems or groups of nodes and connections between them; this is often called "graph-aware" inference in the neuroimaging literature. However, pooling over functional regions leads to significant loss of information and lower accuracy. Another challenge is correlation among edge weights within a subject which makes inference based on independence assumptions unreliable. We address both of these challenges with a linear mixed effects model, which accounts for functional systems and for edge dependence, while still modeling individual edge weights to avoid loss of information. The model allows for comparing two populations, such as patients and healthy controls, both at the functional regions level and at individual edge level, leading to biologically meaningful interpretations. We fit this model to resting state fMRI data on schizophrenic patients and healthy controls, obtaining interpretable results consistent with the schizophrenia literature.

1. Introduction. Networks have been frequently used as a model for the brain's structural or functional connectome. The types of nodes and edges depend on the data collection modality; we focus on data collected from functional magnetic resonance imaging (fMRI), although the statistical models we propose are applicable to other forms of brain imaging and potentially to other network data settings, particularly those involving multiplex networks, that is, multiple networks observed on a common node set. In brief, fMRI is obtained by recording blood oxygenation level dependent (BOLD) signals from subjects over time and at multiple locations in the brain; the raw data for each subject is thus a four-dimensional array (BOLD signal indexed by three spatial coordinates and time). When extracting a network from fMRI data, a node is typically taken to be either a single location in the brain (a voxel) or a spatially contiguous group of voxels, otherwise known as a region of interest, or ROI (Smith (2012), Zalesky, Fornito and Bullmore (2010)). Edges in brain networks capture connections between nodes, which can reflect either structural or functional connections, depending on the type of data collected (Bullmore and Bassett (2011), Bullmore and Sporns (2009)). Structural connectivity has anatomical origins and can be inferred from fiber tracking methods such as diffusion MRI (Craddock et al. (2013), Zalesky, Fornito and Bullmore (2010), Zalesky et al. (2012)). Since we are working with fMRI data, we focus on functional connectivity which represents temporal correlations between different parts of the brain (Friston (1994), van den Heuvel and Pol (2010)). However, these two types of connectivity are frequently linked (van den Heuvel et al. (2009)), and our methods are equally applicable to both types. To take advantage of all the information available, we work with signed, weighted, dense networks where each pair of nodes is associated with a distinct real number; in contrast to some previous methods, for example, Simpson and Laurienti (2015), we do not apply thresholding to convert this matrix of weights into a binary network.

Received June 2019; revised September 2022.
Key words and phrases. Neuroimaging, functional MRI, network analysis.

Multiple studies (Power et al. (2011), Yeo et al. (2011)) have produced brain atlases suggesting a robust parcellation of brain ROIs into functional systems, though details vary (Craddock et al. (2013), van den Heuvel and Pol (2010)). Many neurological and psychiatric disorders have been associated with changes in functional connectivity between such systems (Bullmore (2012), Bullmore and Bassett (2011), Craddock et al. (2009), Craddock et al. (2013)). For instance, in schizophrenia a decrease in connectivity between the frontal and temporal cortices has been reported (Friston and Frith (1995)). The statistical challenge here is that while hypotheses and interpretation are framed at the level of connections between and within functional systems, the data are collected and modeled at a finer resolution of edges between ROIs. The scientific questions are often amenable to two-sample inference, comparing patients to healthy controls while identifying and locating specific changes in functional connectivity associated with a given disorder.

While the regression-based framework we propose is appreciably more general and can accommodate continuous covariates, the data we work with in this paper fall into the two-sample setting, containing resting state fMRI data from 54 schizophrenic patients and 70 healthy controls. A more thorough description of the study and imaging parameters is available in Aine et al. (2017). "Resting state" fMRI involves imaging participants who are told to just relax in the scanner without being given any specific task to perform. As a result, the per-voxel fMRI time series cannot be meaningfully temporally aligned between participants because there is no synchronized task or stimulus, and it is only meaningful to consider information averaged over time (after appropriate preprocessing, see Arroyo Relión et al. (2019) for details). Resting state data are especially well suited for network-based analysis, since raw time-series data cannot be meaningfully compared between patients but connectivity networks can.

Much of the work on the problem of two-sample inference for brain connectivity networks falls into one of several categories, as reviewed in Chung et al. (2021), which differ chiefly in which features of the networks are to be compared. The "bag of (graph-theoretic) features" framework uses a few global network summary measures, such as modularity or the clustering coefficient, and compares them across samples (Bullmore and Sporns (2009)). In this vein some recent work by Fujita et al. (2017) uses the spectral radius of each (thresholded) network to compare samples, considering, for example, whether the spectral radii of two subnetworks of a larger network are empirically correlated in a sample of networks. Another common approach, sometimes referred to as "bag of edges" or "massively univariate," arranges all the network edge weights into one vector, ignoring the network structure but allowing for usual multivariate inference at the edge level. This approach requires correcting for massive multiple testing, which tends to be overly conservative when test statistics are correlated, as is the case here (Craddock et al. (2013)). In addition, some approaches focus on a more global comparison in the so-called "bag of networks" framework; for example, Tang et al. (2017) proposes a test to evaluate whether two networks were drawn from the same generative model. In contrast to these approaches, our method proposed below offers multiresolution inference that can operate on both local edge-level features as well as more intermediate system-level features.

To improve power, Zalesky, Fornito and Bullmore (2010) proposed a network-based statistic (NBS) approach which reduces the number of multiple comparisons by focusing on large connected subnetworks. Specifically, they consider the size of the largest connected component of the graph, obtained by retaining the edges with two-sample test statistic exceeding a given threshold, and compare to a null permutation-based distribution. However, the result depends on the threshold for the test statistic (Kim et al. (2014)). Belilovsky, Varoquaux and Blaschko (2016) also aim to exploit structure by assuming sparsity in the group differences: this method has the advantage of using the time courses directly but does not account for

multiple subjects in each group. Another set of methods uses logistic regression with the binary group indicator as response and the individual's edge weights as explanatory variables to obtain various likelihood-based scores further used as a test statistics with permutation-based p-values. These include the sum of powered scores test (Pan et al. (2014)) and weighted and adaptive variants; for a detailed review and comparisons, see Kim et al. (2014).

A more recent algorithm, proposed by Narayan, Allen and Tomson (2015), directly uses fMRI time series to test the hypothesis that the probability of each edge is the same for the two populations. Narayan and Allen (2016) extended this method, called R^3 (resampling, random penalization, and random effects), to testing any (discrete or continuous) covariate effects, using random effects to account for between- and within-subject variability. Mixed effects models generally have been gaining popularity for brain networks, because they allow for individual heterogeneity and provide a framework for testing covariate effects. For example, Sobel and Lindquist (2014) use a linear mixed model for causal inference, using fMRI time series data, though this work is not designed for two-sample inference.

None of the methods discussed so far incorporates brain system structure in the ROIs, but some recent work leverages network community structure in modeling the association between brain connectivity and various phenotypes. Xia et al. (2020) models each participant's brain connectivity as the sum of two low-rank matrices, one capturing the population mean and the other reflecting the contribution of covariates and their interactions at the level of functional systems. We will compare this method to ours in Section 3.

Our approach to modeling brain networks inherits all the advantages of a linear mixed effects model (binary or discrete covariates, individual effects, flexible variance structure) while accounting for network structure and enabling inference at both the system and the edge levels. This allows for more accurate inference than what one can obtain by treating edges as a bag of features. We incorporate system structure through a brain parcellation into functional systems. ROIs within the same system of a meaningful parcellation tend to have similar connectivity patterns (Smith et al. (2013)), and we leverage this property to parameterize the model in a more interpretable and concise fashion. We also allow for some edge dependence, induced by the parcellation, which leads to a more accurate assessment of uncertainty.

Of course, implicit in our approach are various assumptions, for example, on the form of the variance structure, as discussed in Section 2.4. While these assumptions are unlikely to be perfectly satisfied, in line with the classic wisdom of Box (1976) we believe that our model may be wrong yet useful. For example, as we see in Section 3, our approach offers far more accurate uncertainty quantification, and thus more valid inference, than a more naive approach. In addition, conducting inference at the level of the brain system better aligns with prevailing scientific thinking and enables the confirmation of existing scientific insights as well as the identification of potentially novel effects.

A particularly related line of work on mixed effects models for brain networks was initiated by Simpson and Laurienti (2015) and continued in Bahrami, Laurienti and Simpson (2019), Bahrami et al. (2017), Simpson, Bahrami and Laurienti (2019); we will collectively refer to this body of work as the S–L (Simpson–Laurienti) approach. There are several important differences between the S–L approach and ours. First, S–L estimates a relatively small number of global coefficients, whereas we have both system-level and edge-level parameters, allowing for greater model flexibility. Second, we directly model edge weights without thresholding, while S–L fits a two-stage model, with the first stage determining a subset of important edges and the second stage only modeling those. Finally, the covariance structure of our model is appreciably more general, including nonzero covariance between both random effects and residuals. Of course, if the simpler assumptions of the S–L approach hold, their model may have both computational and power advantages, but such assumptions are unlikely to be verifiable in practice.

Another related paper by Fiecas et al. (2017) proposes a mixed effects model somewhat similar to ours for two-sample testing at either the level of the entire network or for individual edges. In contrast to our approach, they conduct two-sample inference by fitting two separate models and then comparing resulting statistics instead of testing parameters corresponding to group differences within a single model. They model within-subject covariance using a nonparametric approach, based on sampling distributions of correlations under autocorrelation, combined with somewhat restrictive assumptions about remaining error structure. This works well for their setting of 11 ROIs, but they note the approach does not scale well, whereas our covariance model is easily applicable to our dataset with 235 ROIs.

Next, we present the graph-aware linear mixed model for brain networks and the fitting algorithm in Section 2. Section 3 presents empirical results including analysis of the COBRE dataset, and Section 4 concludes with discussion and possible directions for future work.

2. Statistical methods: A network-aware mixed effects model.

2.1. *Setup and notation.* We assume that we are given a sample of N networks on n nodes; each network is represented by its weighted $n \times n$ adjacency matrix $A_m, m = 1, \ldots, N$. The nodes are aligned across all networks, corresponding to a common ROI atlas in the brain application. The entry $A_{m,ij}$ represents the connectivity between nodes i and j for network m, and we focus on the undirected setting $A_{m,ij} = A_{m,ji}$ with no selfloops appropriate for fMRI data. In the brain application, the weights are Fisher-transformed Pearson correlations between time series at different ROIs which are standard in the neuroimaging literature. Alternative measure of connectivity, such as partial correlations or thresholded correlations, can also be used; see Craddock et al. (2013), Liang et al. (2012), Smith et al. (2011), Zhen et al. (2007) for discussions on various ways of measuring functional connectivity. There are strong local spatial correlations between the time series at neighboring ROIs which leads to highly dependent edge weights.

We assume that the network nodes are divided into groups corresponding to network communities; in the application this corresponds to ROIs grouped into functional systems. In network analysis communities are typically viewed as groups of nodes with similar connectivity patterns; in many cases, including typical brain networks, this means that there are stronger connections within communities than between them. In this paper we use an existing known parcellation of the brain into functional systems; alternatively, one could apply one of the many community detection techniques first to estimate such a parcellation.

Let c_i be the community label of node i, common across all networks and taking values in $\{1, \ldots, K\}$. We refer to an unordered pair of communities (a, b), where $a, b \in \{1, \ldots, K\}$ as a network cell; there are a total of $K(K+1)/2$ cells corresponding to K communities. We will use these cells as a target of inference when characterizing effects at the cell level, and the cells will also inform our covariance structure, as discussed in Section 2.4.

Let $n^{(a)} = |\{i : c_i = a\}|$ denote the number of nodes in community a, and let $n^{(a,b)}$ be the number of edges in cell (a, b), where

$$n^{(a,b)} = \begin{cases} n^{(a)} n^{(b)}, & \text{if } a \neq b, \\ n^{(a)}(n^{(a)} - 1)/2 & \text{if } a = b, \end{cases}$$

and let $n^{(\cdot,\cdot)} = n(n-1)/2$ be the total number of distinct edge weights.

Next, let $y_m^{(a,b)} \in \mathbb{R}^{n^{(a,b)}}$ be the vector of edge weights in cell (a, b) for subject m. We use $y_{m,i}^{(a,b)}$ to refer to an element of these vectorized edge weights, with i ranging from 1 to $n^{(a,b)}$ for cell (a, b). Finally, let $y_m \in \mathbb{R}^{n^{(\cdot,\cdot)}}$ be the vector of *all* edges for subject m, that is, the concatenation of $y_m^{(a,b)}$ across all cells and $y \in \mathbb{R}^{Nn^{(\cdot,\cdot)}}$ the vector of all edges for all subjects.

2.2. *A linear mixed effects model.* Brain connectivity naturally varies across subjects, even if they have the same disease status and other covariates. A mixed effects model is a natural tool to incorporate this individual variation while also modeling effects of other covariates. A linear model allows for a straightforward interpretation of these effects and can account for the high correlations among edge weights by including a general covariance error structure.

For simplicity, we first write out the model for a given network cell (a, b). Let $x_m \in \mathbb{R}^p$ be a vector of subject-level covariates (generally including a "1" term for the intercept), such as disease status, age, gender, and so on. For each cell (a, b), we partition covariate effects into cell-level effects, denoted by coefficients $\alpha^{(a,b)} \in \mathbb{R}^p$, and additional edge-level effects, denoted by coefficients $\eta_i^{(a,b)} \in \mathbb{R}^p$, for each edge i in cell (a, b). We then model the expected edge weight as

$$\mathbb{E}\big(y_{m,i}^{(a,b)}\big) = x_m^T\big(\alpha^{(a,b)} + \eta_i^{(a,b)}\big),$$

where $m = 1, \ldots, N$ is the subject index, $i = 1, \ldots, n^{(a,b)}$ is the edge index within the cell, and $a, b \in \{1, \ldots K\}$ are community labels. For identifiability we require that $\sum_i \eta_{i,j}^{(a,b)} = 0$ for all j.

Adding a cell-specific subject random effect term $\gamma_m^{(a,b)}$ and an error term $\epsilon_{m,i}^{(a,b)}$, we get the proposed linear mixed effects model for edge weights,

$$y_{m,i}^{(a,b)} = x_m^T\alpha^{(a,b)} + x_m^T\eta_i^{(a,b)} + \gamma_m^{(a,b)} + \epsilon_{m,i}^{(a,b)}.$$

The noise variables $\epsilon_{m,i}^{(a,b)}$, specific to each subject and each edge, have mean 0 and are independent of the random effects.

As an example, consider the two-sample setting where we have a single subject covariate, say d_m, which is an indicator of, for example, disease status of subject m, set to 1 if subject m has the disease and 0 otherwise. The intercept then represents the global cell mean of subjects who do not have the disease. In this case we have $x_m = \begin{pmatrix}1 & d_m\end{pmatrix}^T$, and the model becomes

$$y_{m,i}^{(a,b)} = \big(\alpha_0^{(a,b)} + d_m\alpha_1^{(a,b)}\big) + \big(\eta_{i,0}^{(a,b)} + d_m\eta_{i,1}^{(a,b)}\big) + \gamma_m^{(a,b)} + \epsilon_{m,i}^{(a,b)}.$$

For every cell (a, b), the term α_0 represents the cell-level mean for patients with no disease, α_1 the cell-level shift due to disease, $\eta_{i,0}$ the edge-specific intercept for patients with no disease, and $\eta_{i,1}$ the edge-specific disease effect. These are all fixed effects, whereas $\gamma_m^{(a,b)}$ is the subject-specific random effect for the given network cell representing individual heterogeneity with mean 0 over the population of subjects.

2.3. *Modeling edge dependence.* While we have now set up a mean model for each network cell (a, b), there are correlations among edge weights across the whole brain. Not modeling these correlations is inaccurate and will result in overly optimistic estimates of uncertainty (Li (2015)); modeling all of them will result in an unmanageable number of parameters, so a compromise is needed.

First, we rewrite the model collecting the terms for all edges together. Recall that y_m is the vector of all edges for subject m, and let $\gamma_m \in \mathbb{R}^{K(K+1)/2}$ be the vector collecting all cell-level random effects for subject m. We assume that each random effects vector γ_m has mean 0 and covariance matrix U, and the edge weights satisfy

$$\mathbb{E}(y_m \mid \gamma_m) = X_m\beta + Z\gamma_m, \qquad \mathbb{V}\mathrm{ar}(y_m \mid \gamma_m) = V, \tag{2.1}$$

where β is a vector that captures the contribution of the coefficients in α and the η_i's for all cells (one can easily move between unconstrained β and the original constrained parameterization). The random effect design matrix is responsible for "broadcasting" the cell-level

effects $\gamma_m^{(a,b)}$ across edges and has the block-diagonal form

$$Z = \begin{pmatrix} 1_{n^{(1,1)}} & 0_{n^{(1,1)}} & 0_{n^{(1,1)}} & \dots & 0_{n^{(1,1)}} \\ 0_{n^{(2,1)}} & 1_{n^{(2,1)}} & 0_{n^{(2,1)}} & \dots & 0_{n^{(2,1)}} \\ 0_{n^{(2,2)}} & 0_{n^{(2,2)}} & 1_{n^{(2,2)}} & \dots & 0_{n^{(2,2)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{n^{(K,K)}} & 0_{n^{(K,K)}} & 0_{n^{(K,K)}} & \dots & 1_{n^{(K,K)}} \end{pmatrix},$$

where $1_{n^{(a,b)}}$ and $0_{n^{(a,b)}}$ are a vector of either all ones or all zeroes with length $n^{(a,b)}$. The fixed effects design matrix for cell (a, b) is given by

$$X_m^{(a,b)} = \begin{pmatrix} x_m^T & x_m^T & \dots & 0 \\ x_m^T & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_m^T & 0 & \dots & x_m^T \\ x_m^T & -x_m^T & \dots & -x_m^T \end{pmatrix},$$

where the initial columns hold the cell-level effects $\alpha^{(a,b)}$ and subsequent columns capture the contribution of the $\eta_i^{(a,b)}$'s. The full design matrix for subject m is then given by "tiling" cell-level design matrices into a block-diagonal $X_m = \operatorname{diag}(X_m^{(a,b)})$. Integrating out the random effect gives

$$\mathbb{E}(y_m) = X_m\beta, \qquad \mathbb{V}\mathrm{ar}(y_m) = V + ZUZ^T \equiv \Sigma. \tag{2.2}$$

We assume that subjects are not correlated in which case the overall covariance of y is given by a block diagonal matrix with repeated blocks of Σ on the diagonal and 0 elsewhere, and the overall design matrix $X = (X_1^T\ X_2^T\ \dots\ X_N^T)^T$. This block structure in the covariance allows us to avoid direct inversion of a very large matrix: assuming all inverses are well defined, the best linear unbiased estimator of β is given by the generalized least squares (GLS) estimator,

$$\hat{\beta} = \left(\sum_m X_m^T \Sigma^{-1} X_m\right)^{-1} \left(\sum_m X_m^T \Sigma^{-1} y_m\right). \tag{2.3}$$

In fact, all estimators of this form, even when the assumed covariance structure $\check{\Sigma} \neq \mathbb{V}\mathrm{ar}(y_m)$, are unbiased estimators of β under model (2.2), since

$$\mathbb{E}(\check{\beta}) = \left(\sum_m X_m^T \check{\Sigma}^{-1} X_m\right)^{-1} \left(\sum_m X_m^T \check{\Sigma}^{-1} \mathbb{E}(y_m)\right) = \beta.$$

2.4. *Graph-aware variance structure.* In addition to computational savings, there are multiple reasons to impose structure on the variance. With no additional assumptions on V, the decomposition $\Sigma = V + ZUZ^T$ is not unique, and V and U are not identifiable. There are multiple options for solving this identifiability problem, and incorporating network cell structure into variance assumptions, as we have done for the mean, has the additional benefit of being faithful to the data structure in the application.

Recall that we impose network structure through using cells (a, b). In the decomposition $\Sigma = V + ZUZ^T$, the second term already has a block structure corresponding to network cells, and it would be natural to impose some structure on V that creates a structure corresponding to network cells in Σ. Figure 1 shows two examples that achieve this goal: a diagonal V and a block-diagonal V. Diagonal V allows for heteroscedastic noise at each edge (which we can estimate because we have multiple subjects), and block-diagonal V further allows for dependence between edge noise variables belonging to the same network cell. In both cases the resulting covariance matrix $\Sigma = V + ZUZ^T$ is dense, allowing for dependence between all edge weights.

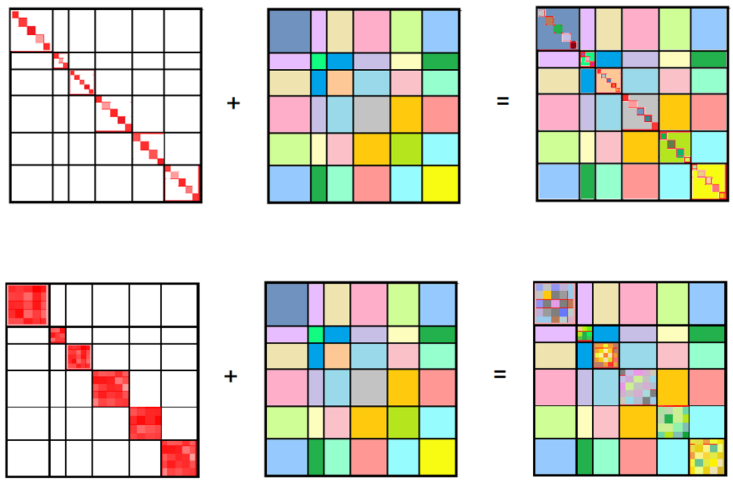


FIG. 1. *Decomposition of the intrasubject covariance matrix of edge weights,* $V + ZUZ^T = \Sigma$ *with diagonal* V *(top) and block-diagonal* V *(bottom).*

2.5. *Model fitting with the EM algorithm.* In practice, the GLS estimator (2.3) is not computable, since V and U are unknown. We take the approach of jointly estimating V, U, and β by maximum likelihood under the normal assumption on the random effects and the errors. Specifically, we assume that $\gamma_m \sim N(0, U)$ and

$$y_m = X_m\beta + Z\gamma_m + \epsilon_m,$$

where $\epsilon_m = \{\{\epsilon_{m,i}^{(a,b)}, 1 \le i \le n^{(a,b)}\} : 1 \le a \le b \le K\} \sim N(0, V)$ is the concatenated error term independent of γ_m.

The normal assumption is reasonable for our edge weights measured by Fisher's z-transformation of the Pearson correlation coefficient r which is designed to make the correlations close to normally distributed. The transformation is defined as

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

and is commonly used in neuroimaging (Varoquaux and Craddock (2013)). Since normality is difficult to verify in practice, this can also be viewed as a generic M-estimation approach with a loss function corresponding to the normal likelihood.

We use the EM algorithm to find maximum likelihood estimates (MLE) of V, U, and β. The derivation under the normal assumption is straightforward and is omitted here. The algorithm consists of the following two steps, iterated until convergence once initialized. For conciseness we write $< f >$ to denote the conditional expectation of f, given the observed data y.

E-step. Calculate posterior means of γ_m and $\gamma_m\gamma_m^T$, given the data, as

$$\langle\gamma_m\rangle = UZ^T\left(V + ZUZ^T\right)^{-1}(y_m - X_m\beta),$$

$$\langle\gamma_m\gamma_m^T\rangle = \left(U - UZ^T\left(V + ZUZ^T\right)^{-1}ZU\right) + \langle\gamma_m\rangle\langle\gamma_m\rangle^T.$$

M-step. Update the estimate of U to

$$\hat{U} = \frac{1}{N}\sum_m \langle\gamma_m\gamma_m^T\rangle,$$

and update the estimates of V and β by repeating the following steps until convergence:

$$\hat{V}_0 = \frac{1}{N}\sum_m \langle (y_m - X_m\hat{\beta} - Z\gamma_m)(y_m - X_m\hat{\beta} - Z\gamma_m)^T \rangle,$$

$$\hat{V} = \begin{cases} \operatorname{diag}(\hat{V}_0), & \text{if } V \text{ is modeled as diagonal,} \\ \text{block-diag}(\hat{V}_0), & \text{if } V \text{ is modeled as block-diagonal,} \end{cases}$$

$$\hat{\beta} = \Big(\sum_m X_m^T \hat{V}^{-1} X_m\Big)^{-1} \sum_m X_m^T \hat{V}^{-1}\big(y_m - Z\langle\gamma_m\rangle\big).$$

Initialization. The coefficients $\hat{\beta}$ can be initialized by ordinary least squares, taking

$$\hat{\beta}^{(a,b)} = \Big(\sum_m (X_m^{(a,b)})^T X_m^{(a,b)}\Big)^{-1} \sum_m (X_m^{(a,b)})^T y_m^{(a,b)}, \qquad \hat{\beta} = \{\hat{\beta}^{(a,b)} : 1 \le a \le b \le K\}.$$

To initialize V, we first calculate, for each pair of cells (a, b), (c, d), the empirical covariance

$$\hat{\Sigma}^{(a,b),(c,d)} = \frac{1}{N-1}\sum_m \big(y_m^{(a,b)} - X_m^{(a,b)}\hat{\beta}^{(a,b)}\big)\big(y_m^{(c,d)} - X_m^{(c,d)}\hat{\beta}^{(c,d)}\big)^T.$$

Then we initialize $U = (U^{(a,b),(c,d)})$ by taking

$$\hat{U}^{(a,b),(c,d)} = \frac{1}{n^{(a,b)}n^{(c,d)}} \sum_{i=1}^{n^{(a,b)}} \sum_{j=1}^{n^{(c,d)}} \hat{\Sigma}_{ij}^{(a,b),(c,d)}$$

and initialize V with a diagonal matrix, regardless of what assumptions we later make about it, as

$$\hat{V}^{(a,b)} = \left(\frac{1}{n^{(a,b)}} \sum_{i=1}^{n^{(a,b)}} \hat{\Sigma}_{ii}^{(a,b),(c,d)} - \hat{U}^{(a,b),(c,d)}\right) I_{n^{(a,b)}}, \qquad \hat{V} = \operatorname{diag}\big(\hat{V}^{(a,b)}\big),$$

where I_n is the $n \times n$ identity matrix.

Implementation. To increase the stability of the algorithm and to speed up convergence, we implement the EM algorithm for the equivalent model,

$$y_{m,i}^{(a,b)} = \zeta_m^{(a,b)} + x_m^T \eta_i^{(a,b)} + \epsilon_{m,i}^{(a,b)}.$$

Writing $\zeta_m = \{\zeta_m^{(a,b)}\}_{1\le a\le b\le K}$, $\zeta_m^{(a,b)} = x_m^T\alpha^{(a,b)} + \gamma_m^{(a,b)}$ and $\mu_m = \{x_m^T\alpha^{(a,b)}\}_{1\le a\le b\le K}$, we have

$$\zeta_m \sim N(\mu_m, U).$$

This simply combines the terms $x_m^T\alpha^{(a,b)}$ and $\gamma_m^{(a,b)}$, both depending only on m and (a, b), to get "mean-shifted" random effects terms, $\{\zeta_m^{(a,b)}\}$. This model is mathematically equivalent to the previous one but empirically converges faster and is less dependent on the initial value of β, due to centering.

The most time-consuming part of the algorithm is updating β in the M-step which involves inverting the large matrix

$$\sum_m (X_m^T V^{-1} X_m).$$

In a typical neuroimaging application, the size of this matrix will be in the tens of thousands; for the COBRE dataset analyzed in Section 3, it is approximately $28{,}000 \times 28{,}000$. To avoid inverting this matrix, we instead solve for β using a block coordinate descent algorithm. This implementation takes around five minutes to fit the model to the COBRE data with a diagonal V and around 15 minutes with a block-diagonal V on a machine with 10 2.8 GHz Intel Xeon E5-2680v2 processors, each with eight GB of memory. In contrast, the OLS estimator that we compare against is almost trivially fast to compute (less than one minute) and indeed is essentially solved during the initialization of our approach.

3. Empirical results. The COBRE schizophrenia dataset, introduced in Section 1, contains resting state fMRI connectivity brain networks of 54 patients with schizophrenia and 70 healthy controls. Data were downloaded via NITRC (http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html) and processed by Professor Stephan Taylor's lab in the Department of Psychiatry at the University of Michigan. This dataset is also available via the COINS platform (Landis et al. (2016), Wood et al. (2014)).

After preprocessing, individual voxels are combined through spatial smoothing into 264 ROIs from the functional parcellation by Power et al. (2011), and Fisher-transformed pairwise Pearson correlations between the time series at each of the ROIs are used as edge weights. Empirically, these weights are approximately normal which is expected from Fisher-transformed correlations. The parcellation of Power et al. (2011) divides the 264 ROIs into 14 functional systems which play the role of communities; these systems are shown in Table 1 and Figure 2. We used only systems 1 through 13 and excluded the 28 nodes of the "Uncertain" system from the analysis, since we have no a priori reason to believe that nodes that could not be clearly assigned to any system have a homogeneous connectivity pattern. Also, the data for node 75 are missing from the COBRE dataset which leaves a total of $264 - 28 - 1 = 235$ ROIs for the subsequent analysis. In addition, as part of our collaborator's preprocessing pipeline, nuisance covariates, including age, gender, motion (summarized as mean framewise displacement and its square), and handedness, were removed before fitting our model. This processed data was also used in Arroyo Relión et al. (2019) and is available in the graphclass package.

We use this dataset as a basis for two experiments where we can control ground truth: in Section 3.1 we assess the performance of our methods on synthetic data drawn from a model based on a fit to the COBRE data, while in Section 3.2 we assess the validity of our methods under the global null by using only randomly labeled healthy controls. We next fit our method to the full COBRE dataset and present the estimated parameters in Section 3.3, conduct inference in Section 3.4, and then apply a related, competing method in Section 3.5.

3.1. *Assessing the effect of variance modeling in synthetic data.* Before we proceed to analyze the COBRE data, we perform a comparison of several versions of our method on synthetic data simulated based on the COBRE dataset, but in a way that allows us to vary parameters of interest. The goal is to understand the effect of variance modeling on discovering individual effects of interest in a realistic setting where we nonetheless know the truth. To generate synthetic data, we first fitted the linear mixed effects model (2.1) to the COBRE dataset using a diagonal matrix V. Then, for each network cell (a, b), we calculated the p-value for the z-test of the hypothesis $H_0 : \alpha_1^{(a,b)} = 0$, where $\alpha_1^{(a,b)}$ is the coefficient of the binary disease indicator x_m in cell (a, b) ($x_m = 0$ for healthy controls and 1 for schizophrenic patients). Standard errors were computed from the standard GLS variance estimator, taking the square root of the diagonal elements of

$$\left(\sum_m X_m^T (V + ZUZ^T)^{-1} X_m\right)^{-1}.$$

TABLE 1
Functional systems from Power et al. (2011)

	System	Number of nodes
1	Sensory/somatomotor Hand	30
2	Sensory/somatomotor Mouth	5
3	Cingulo-opercular Task Control	14
4	Auditory	13
5	Default mode	58
6	Memory retrieval	5
7	Visual	31
8	Fronto-parietal Task Control	25
9	Salience	18
10	Subcortical	13
11	Ventral attention	9
12	Dorsal attention	11
13	Cerebellar	4
−1	Uncertain	28

We chose as "true positives" the 13 cells with $p < 0.05$ and set $\alpha_1^{(a,b)} = 0$ for the remaining 78 cells for the purpose of simulating synthetic data. Using this information, we refitted the GLS model with EM to obtain a diagonal $\hat{V}$, $\hat{U}$, and $\hat{\beta}$ (which comprises both cell- and edge-specific effects) with the 78 true negative α_1's set to 0. Then we generated a new synthetic dataset by drawing

$$y_m \sim N\big(X_m\hat{\beta}, \hat{V} + Z\hat{U}Z^T\big), \quad m = 1, \ldots, 100$$

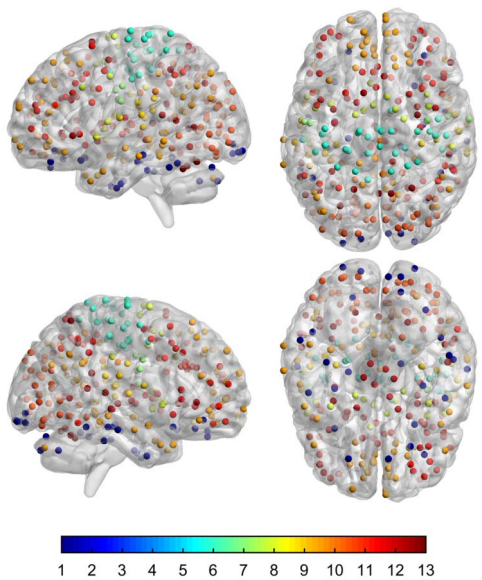


FIG. 2. *The* 13 *functional systems from Power et al. (2011) represented by color. Left column*: *Sagittal view from the left* (*top*) *and the right* (*bottom*). *Right column*: *Axial view from above* (*top*) *and from below* (*bottom*). *Figure generated using BrainNet* (*Xia, Wang and He* (2013)).

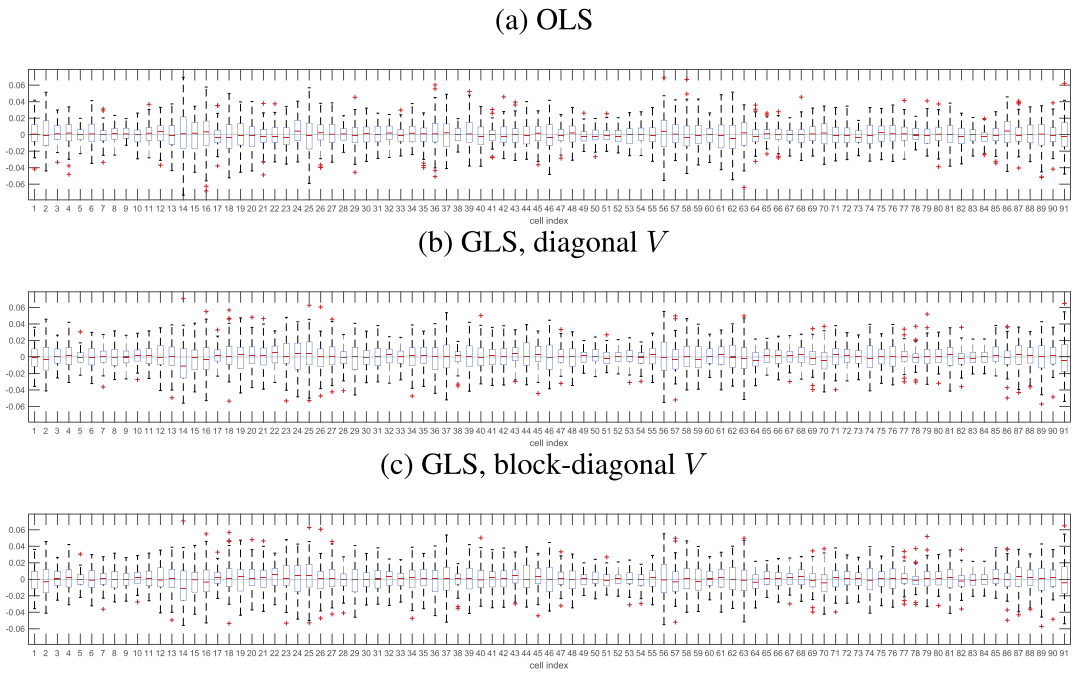


FIG. 3. *Boxplots of the errors,* $\hat{\alpha}_1 - \alpha_1$, *with* 100 *replications. Outliers are indicated by* $+$*'s, and a horizontal line marks* 0 (*no error*).

with 50 subjects each from healthy and schizophrenia populations. Our goal was to compare the usual OLS estimator,

$$\hat{\beta}^{\mathrm{OLS}} = \left(\sum_m X_m^T X_m\right)^{-1} \left(\sum_m X_m^T y_m\right),$$

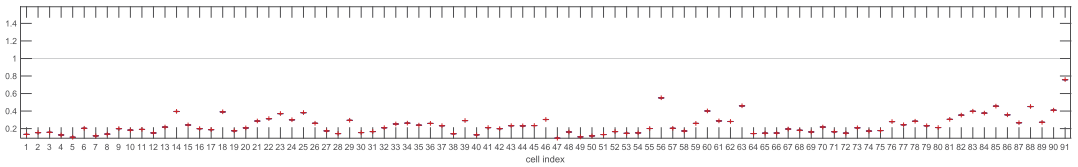
to the proposed GLS estimator. For each generated dataset, we fitted our model with all 91 covariates in two ways, with either diagonal or block-diagonal V, and we also fitted OLS for comparison. For each estimator we computed standard errors for GLS and OLS according to their respective standard formulas. The entire simulation was repeated 100 times.

While we fitted GLS using both a diagonal V and block-diagonal V covariance structure, the model from which we drew data *perfectly* coincides with the former. While the block-diagonal V covariance structure contains diagonal V as a special case, it is appreciably more flexible, and we anticipated that this additional flexibility may result in some overfitting in this case, although, as we shall see, the effect is quite modest.

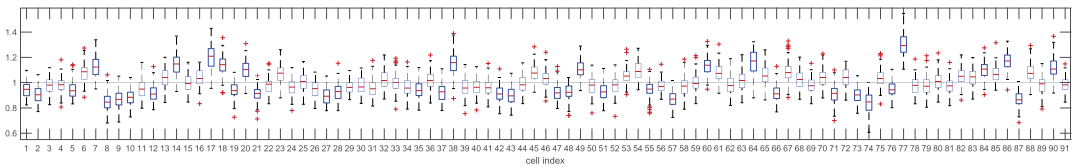
Figure 3 shows the boxplots of errors for the main parameter of interest, cell-level difference between populations $\hat{\alpha}_1 - \alpha_1$, for the three types of estimators and the 91 network cells. All three estimators look similar, and all the boxplots are centered around 0, as they should be since all three estimators are unbiased. A perhaps surprising result is that the point estimates for both diagonal and block-diagonal V methods are precisely the same. This is a consequence of looking at only effects at the level of network cells and of the particular covariance structures assumed.

For valid inference we need not just an accurate point estimate but also an accurate standard error. A comparison of the standard error of $\hat{\alpha}_1$, estimated using OLS and GLS to the empirical standard deviation of the estimated parameter $\mathrm{sd}(\hat{\alpha}_1)$, is shown in Figure 4, with boxplots of the ratio estimated $\mathrm{s.e.}/\,\mathrm{sd}(\hat{\alpha}_1)$ for the three estimators. The three rows in Figure 4 show the results from OLS, GLS with a diagonal V, and GLS with a block-diagonal V, from (a) to (c). Ideally, these boxplots should be centered around 1, but the OLS ratios

(a) OLS



(b) GLS, diagonal V



(c) GLS, block-diagonal V

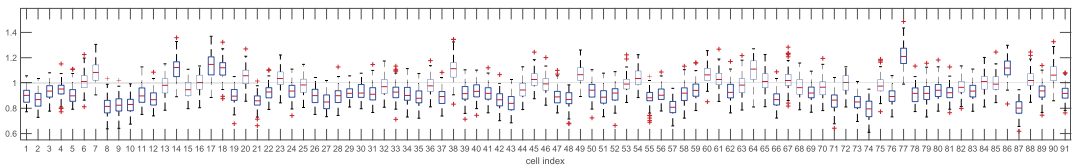


FIG. 4. *Boxplots of the ratios of the standard errors computed from the corresponding formula to the empirical standard deviation of* $\hat{\alpha}_1$ *with* 100 *replications. Outliers are indicated by* $+$*'s, and a horizontal line marks* 1 (*estimated standard deviation coincides with empirical standard deviation*). *Note that the vertical scale of panel* (*a*) *is different from* (*b*) *and* (*c*), *since OLS severely underestimates standard errors. Also, for OLS the variability for a given cell across simulations is dramatically smaller than variability between cells which makes outliers the only easily visible part of the boxplots.*

are much smaller than 1, though also the most stable; the GLS standard errors, on the other hand, are much closer to the truth and also more variable themselves. This shows that OLS severely underestimates standard errors by assuming independence, while GLS leads to honest inference. Careful scrutiny of Figure 4 reveals that the block-diagonal V method *slightly* underestimates standard errors, perhaps due to the slight overfitting resulting from its unneeded (in this setting) additional flexibility: the mean ratio of the diagonal V method is 1.001, whereas the mean ratio of the block-diagonal V method is 0.951.

We also compared coverage rates of 95% confidence intervals for α_1, defined by $[\hat{\alpha}_1 - 1.96 * s.e.(\hat{\alpha}_1), \hat{\alpha}_1 + 1.96 * s.e.(\hat{\alpha}_1)]$ for the three estimators, as shown in Figure 5. As we can expect from Figure 4, OLS confidence intervals have poor coverage, but both GLS methods give coverage close to 95%: averaging across network cells, coverage is approximately 94.7% for diagonal V and 93.2% for block-diagonal V. Because coverage for the GLS method is reasonably close to nominal, for these two methods we further computed the false positive rate (FPR) and the true positive rate (TPR) to assess the size and power of the procedure, respectively. The diagonal V method had an FPR of approximately 0.054 and a TPR of approximately 0.369, while the block-diagonal V method had an FPR of approximately 0.070 and a TPR of approximately 0.336.

While we do not claim that this data-based simulation gives us a good estimate of just how far off the standard errors are in real data, we argue that it does show the errors will be unrealistically small if the model is fitted with OLS. This is generally what we expect when the observations are positively correlated, and with many aspects of the simulated data

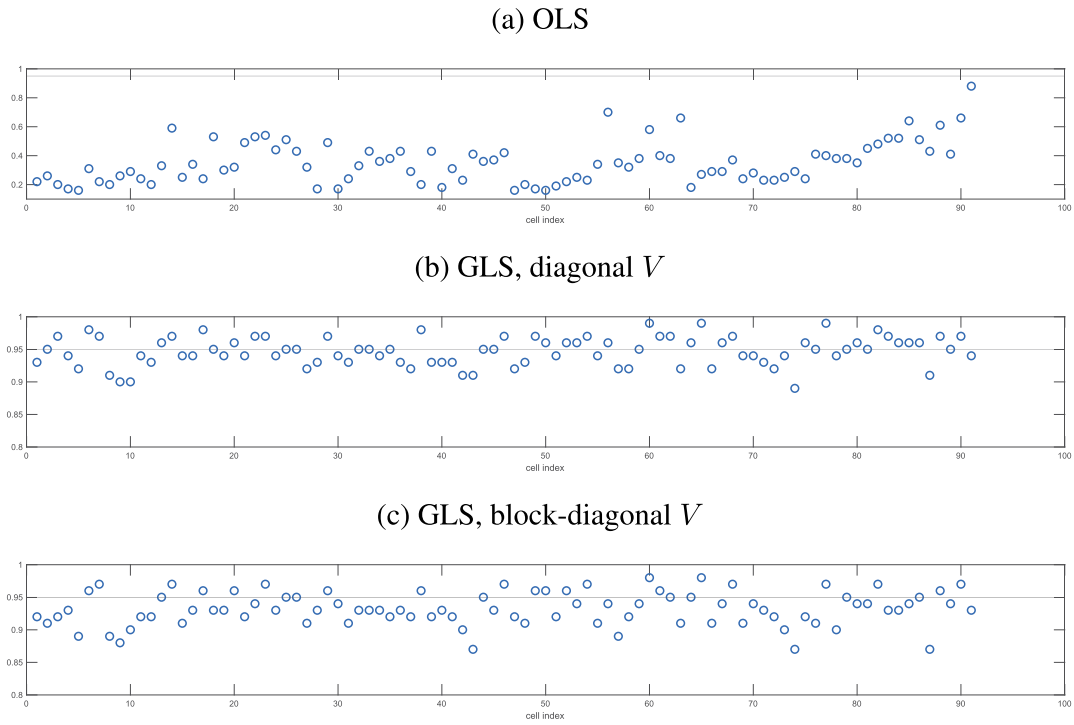


FIG. 5. *Coverage of the confidence intervals,* $[\hat{\alpha}_1 - 1.96 * s.e.(\hat{\alpha}_1), \hat{\alpha}_1 + 1.96 * s.e.(\hat{\alpha}_1)]$, *with* 100 *simulations. The horizontal line corresponds to* 95% *coverage. Note that the vertical scale in panel* (*a*) *is different from panels* (*b*) *and* (*c*) *due to poor coverage by OLS.*

matching the real data, this simulation gives us some idea of the differences between OLS and GLS performance that can be expected to arise in the application.

3.2. *Validity under global null.* Next, we evaluate the distribution of p-values obtained by our method under the null hypothesis of no difference between two populations. We use only the 70 healthy controls and split them randomly into two groups of 35. In this case p-values obtained by fitting the model should be uniformly distributed if the model is performing valid inference.

We repeat the random splits into two groups 100 times, fitting our model and computing p-values for the hypotheses $\alpha_1^{(a,b)} = 0$ every time, resulting in a total 9100 p-values (91 cells $\times$ 100 repetitions). Figure 6 shows the histogram of these 9100 p-values from OLS and GLS. The GLS histograms look close to the uniform, whereas the OLS histogram has a large number of small p-values, and approximately 2/3 are less than 0.05. After applying the Benjamini–Hochberg multiple testing correction (Benjamini and Hochberg (1995)) to control FDR at 5% (as we do in our application), the OLS method rejects the null hypothesis, on average, for 56.7 cells (out of 91) in each replication of the simulation. In contrast, the models with diagonal and block-diagonal V reject, respectively, 0.23 and 0.37 out of 91 hypotheses on average. Even with the more conservative Bonferroni correction, OLS still rejects 41 cells on average, whereas the rates for the two GLS methods are 0.14 and 0.28. To quantify the comparison of the distribution of the p-values to the uniform, we apply a Kolmogorov–Smirnov test and compare the p-values for a given cell across the 100 replications to the uniform distribution. This gives us 91 resulting p-values for each method shown in Figure 6, which we again compare to the uniform distribution by a Kolmogorov–Smirnov test (note that because there may be some dependence across cells, the p-value obtained from this procedure

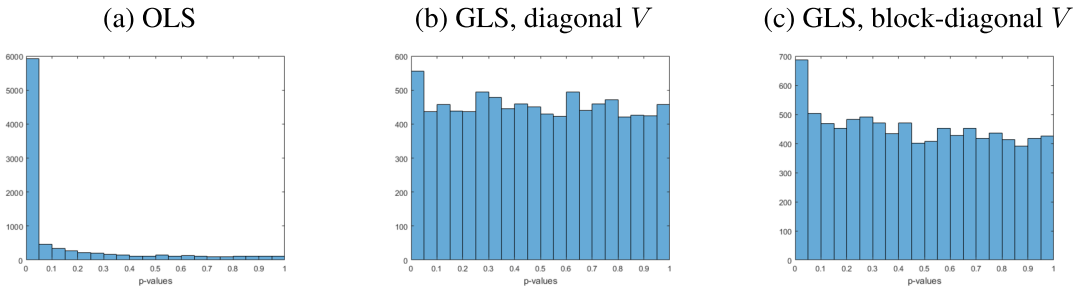


FIG. 6. *Distribution of p-values under the null hypothesis.*

is not strictly valid but still serves as a concise data summary). As may be anticipated from Figure 6, applying this procedure gives a p-value of 0 (to machine precision) for OLS, 0.6098 for GLS with a diagonal V, and 0.01653 for GLS with a block-diagonal V.

3.3. *Parameter estimation in full dataset.* We now fit model (2.1) to the entire COBRE dataset with disease status as a binary covariate. Figure 7 shows the estimated mean networks for the two populations, that is, $\{\hat{\alpha}_0 + \hat{\eta}_{i,0}\}$ and $\{\hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\eta}_{i,0} + \hat{\eta}_{i,1}\}$, with either diagonal

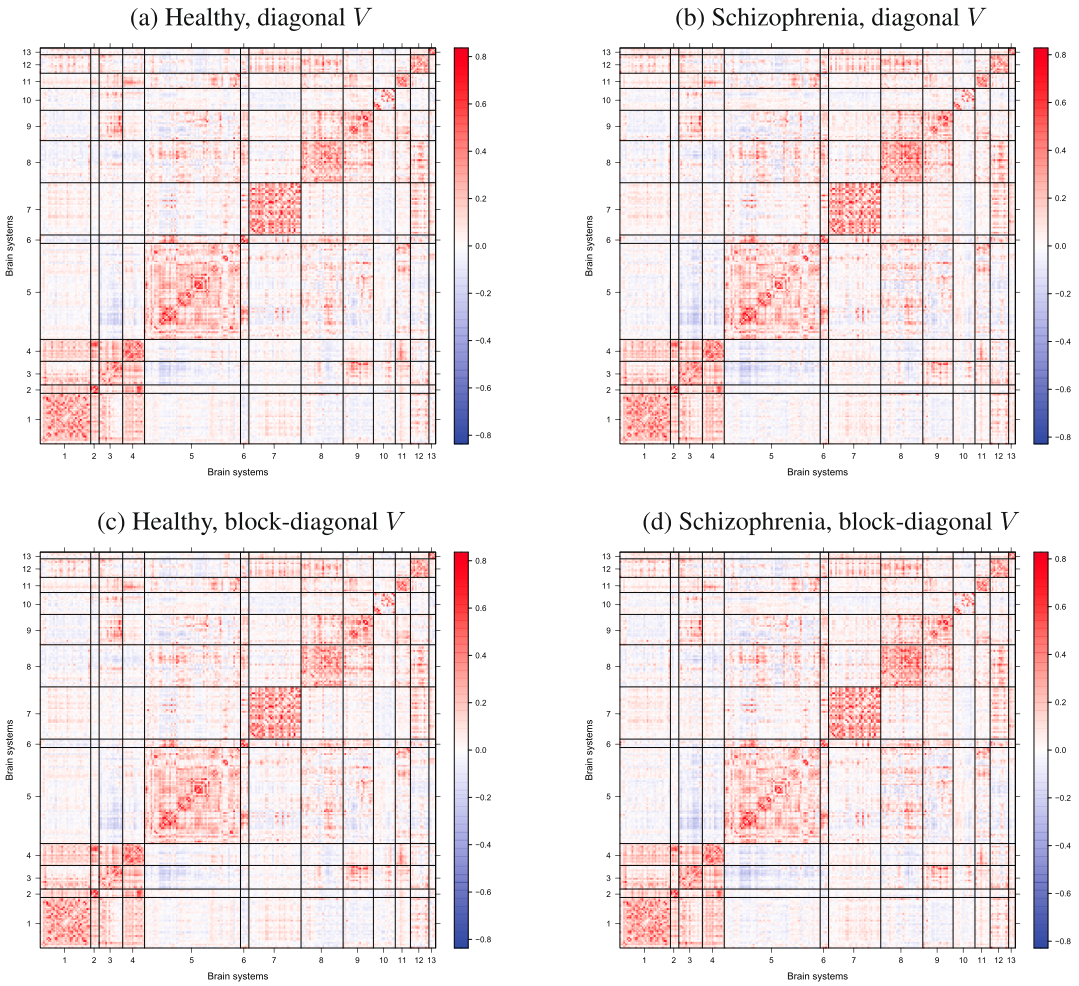


FIG. 7. *Estimates of the mean networks for healthy and schizophrenia patients.*

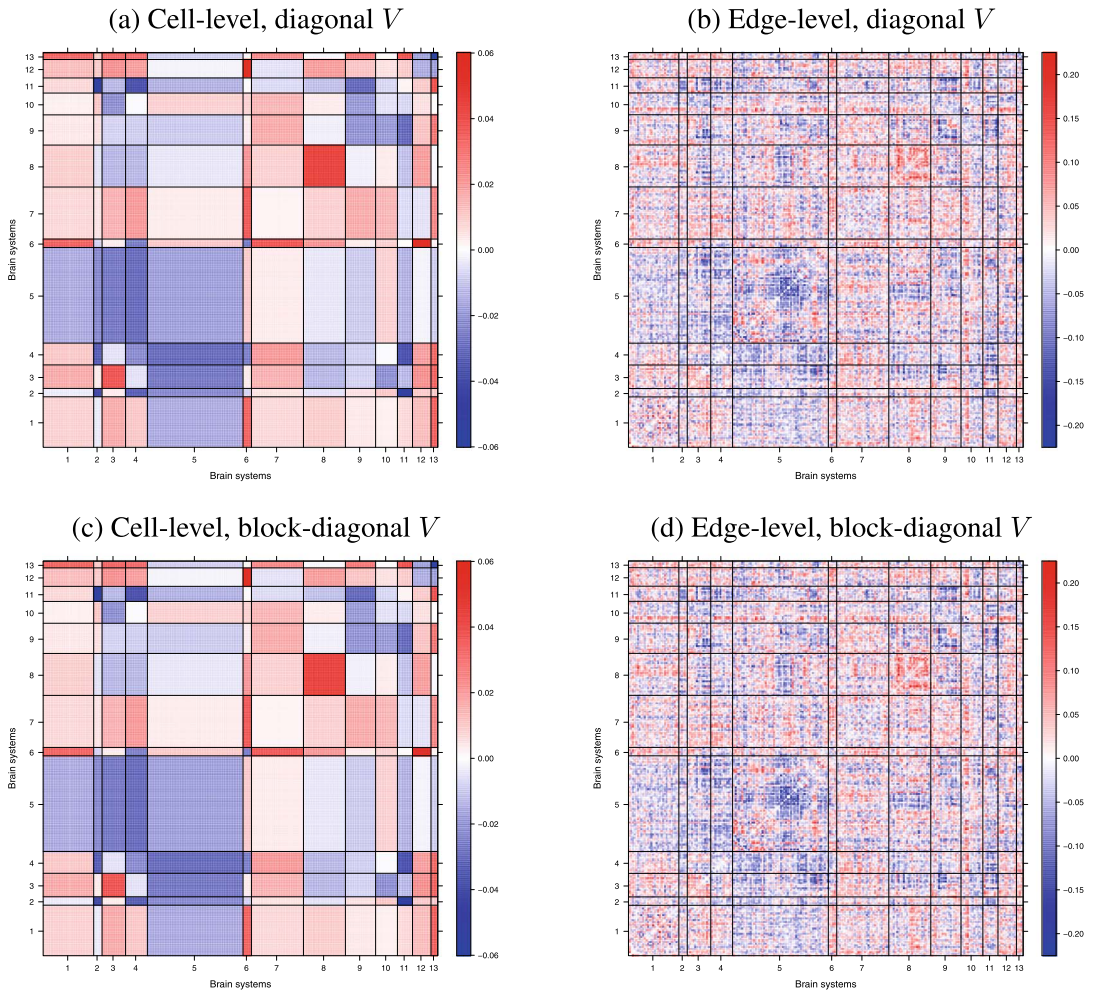


FIG. 8. *Estimates of differences between the two populations, healthy and schizophrenic, with diagonal and block-diagonal V. Blue represents larger mean for control group and red represents larger mean for schizophrenia group.*

or block-diagonal V. The dominant structure in the means is the community structure, with stronger connectivity within each functional system, which is expected.

Figure 8a and Figure 8c show the cell-level differences in connectivity between the two populations, that is, $\hat{\alpha}_1$. Negative $\hat{\alpha}_1$ (blue) corresponds to higher values for controls and positive $\hat{\alpha}_1$ (red) for schizophrenic patients. The higher control values in cell $(9, 9)$, corresponding to lower connectivity within the salience system for schizophrenic patients, match a previously reported dysfunction in schizophrenia (Palaniyappan et al. (2013)). Schizophrenia effects on the frontal and parietal brain regions (system 8) also have been reported (van den Heuvel and Pol (2010)).

Figure 8b and Figure 8d show differences between the two populations at the edge level, that is, $(\{\hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\eta}_{i,0} + \hat{\eta}_{i,1}\}) - (\{\hat{\alpha}_0 + \hat{\eta}_{i,0}\})$ for each edge i. The edge effects are quite heterogeneous, especially for the large cells, such as $(5, 5)$. The heterogeneous cell effects suggest that we do need to include edge-level effects, and in fact, interpreting cell effects without the edge effects may lead to misleading results.

3.4. *Hypothesis testing for group comparisons.* To assess the differences between groups more formally, we perform hypothesis tests. First, consider testing the edge-level effects, with

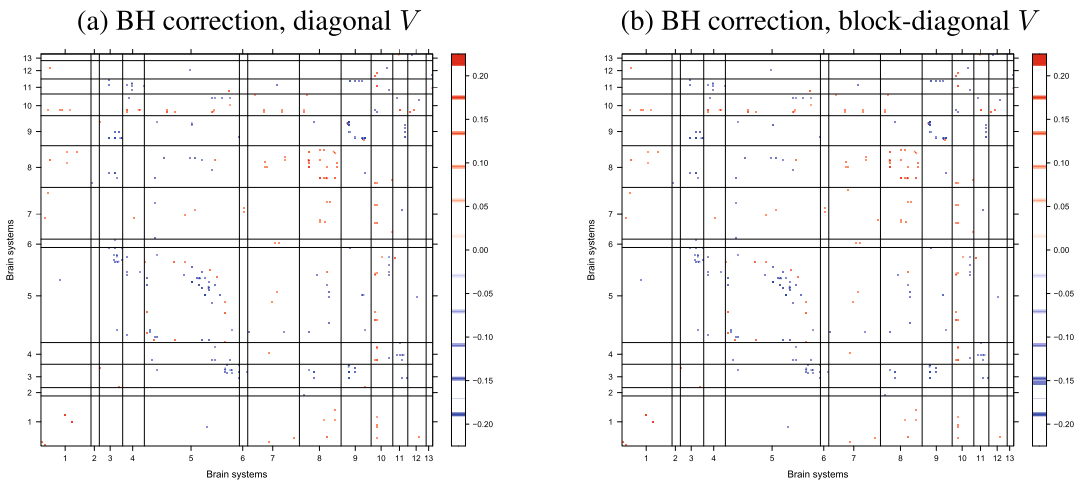


FIG. 9. *Estimates of edge-level difference between healthy vs. schizophrenic for significant edges at 5% significance level (after Benjamini–Hochberg correction).*

the null $H_{0,i} : \alpha_1 + \eta_{i,1} = 0$ vs. alternative $H_{1,i} : \alpha_1 + \eta_{i,1} \neq 0$ for every edge i. Results with the Benjamini–Hochberg multiple testing correction, controlling FDR at 5%, are presented in Figure 9. Both diagonal and block-diagonal V give very similar results with 150 and 149 significant edge-level differences, respectively; 148 edges out of these two sets are the same.

We chose the Benjamini–Hochberg correction primarily because it does well on power, but other choices are possible, including methods that control familywise error rate (FWER), such as Bonferroni's, Holm's (Holm (1979)), and Hochberg's (Hochberg (1988)), or the Benjamini-Yekutieli method for controlling FDR under dependency (Benjamini and Yekutieli (2001)). Applying the Benjamini–Yekutieli procedure to our data yielded only 22 significant edges for both diagonal V and block-diagonal V; while this procedure is valid under dependency, it is known to be conservative.

Figure 10 shows the positive and negative edge differences plotted on the brain. Our results generally agree with previous findings on schizophrenia, including a disconnection between the frontal and the temporal cortices (Bullmore and Bassett (2011), Friston and Frith (1995)) and occipito-temporal disconnections (Zalesky, Fornito and Bullmore (2010)). Figure 10 clearly shows the asymmetric connectivity difference between healthy control and schizophrenia for left and right hemispheres which is also aligned with previous studies (Angrilli et al. (2009), Mitchell and Crow (2005), Ribolsi et al. (2014)).

We next test the cell-level hypotheses $H_0^{(a,b)} : \alpha_1^{(a,b)} = 0$ vs. $H_1^{(a,b)} : \alpha_1^{(a,b)} \neq 0$ for every network cell (a, b). Results, both prior to and after correction for multiple testing, are presented in Figure 11. At the cell level, the increase in connectivity within system 8, the fronto-parietal task control system, was associated with the lowest p-value, both for diagonal and block-diagonal V, as shown in Figure 11. This is consistent with a previous study (Venkataraman et al. (2012)) which reported increased connectivity between parietal and frontal regions.

With varying significance level, we also look at the number of rejected hypotheses (out of 91) using OLS and GLS (with diagonal/ block-diagonal V) after Benjamini–Hochberg correction. From Figure 12 we can see that OLS rejects many more hypotheses (more than half, even at the smallest significance value) than either version of GLS. Our simulation results suggest this is due to the underestimation of standard errors with OLS. The GLS results are more realistic: from Figure 12 we can see that at 5% significance level, GLS with diagonal V does not find anything, and GLS with a block-diagonal V finds four significant cells, $(8, 8)$, $(2, 11)$, $(6, 12)$, and $(13, 13)$; see Figure 11.

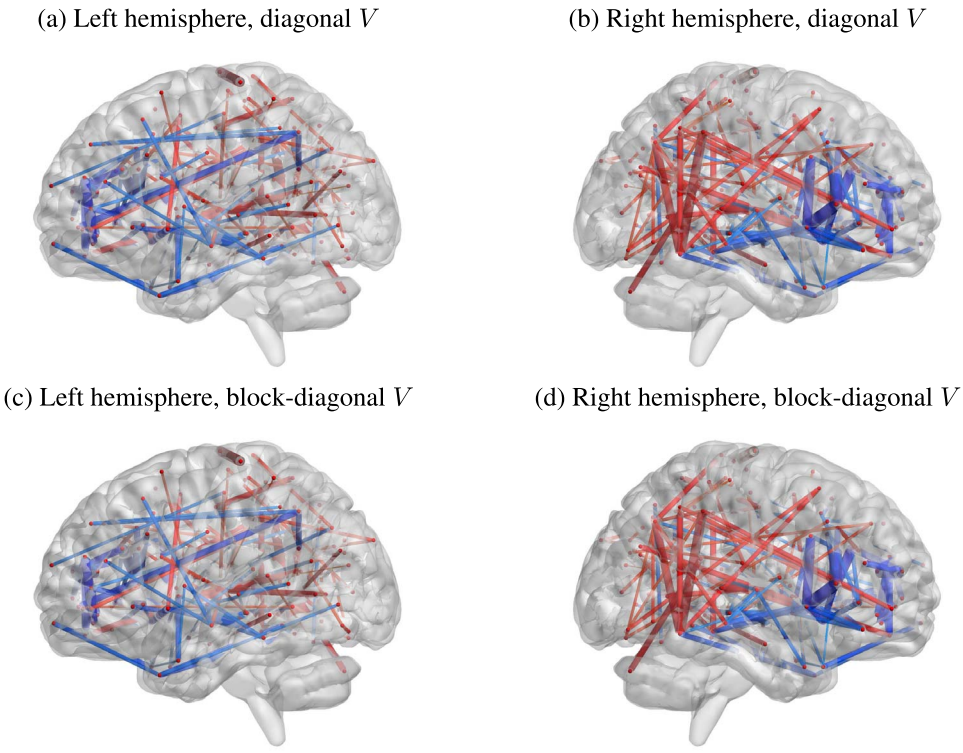


FIG. 10. *Estimates of edge-level difference on the brain. Blue edges represent larger mean for healthy control and red edges represent larger mean for schizophrenia. Line width represents the magnitude of the difference. Figures generated using BrainNet Viewer (Xia, Wang and He (2013)).*

The distributions of raw p-values from OLS and GLS are shown in Figure 13. Consistent with the simulation findings, OLS produces a large number of small p-values, and the ordering of p-values is also fairly different between OLS and GLS. On the other hand, the p-values from GLS with diagonal and block-diagonal V agree closely, and the cells with the lowest p-values match. Top five cells, ordered by p-values, are shown in Table 2.

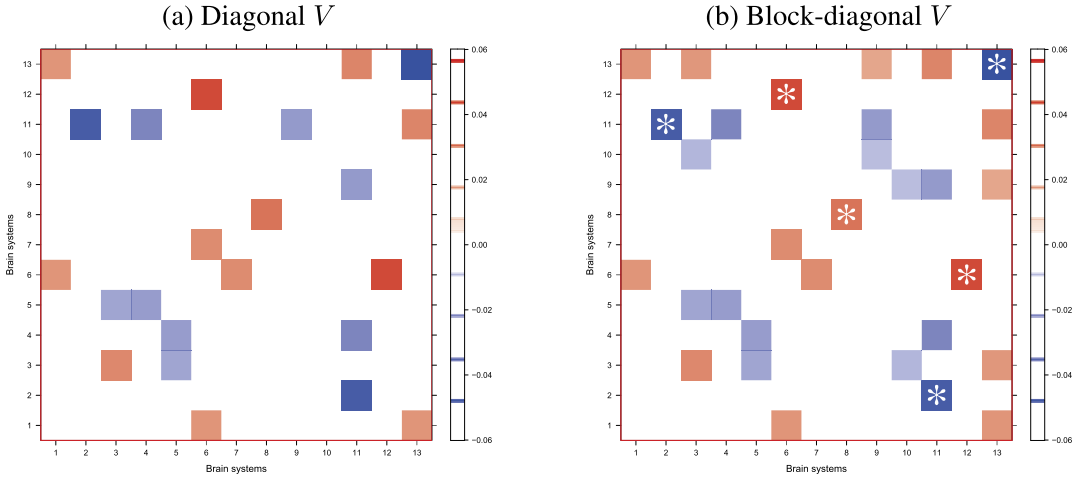


FIG. 11. *Values of $\hat{\alpha}_1^{(a,b)}$ for cells (a, b) with $\hat{\alpha}_1^{(a,b)}$ significant at 5% before multiple testing correction. Cells retaining significance after the Benjamini–Hochberg correction are marked with an asterisk (*). With diagonal V, there are no significant cells after correction.*

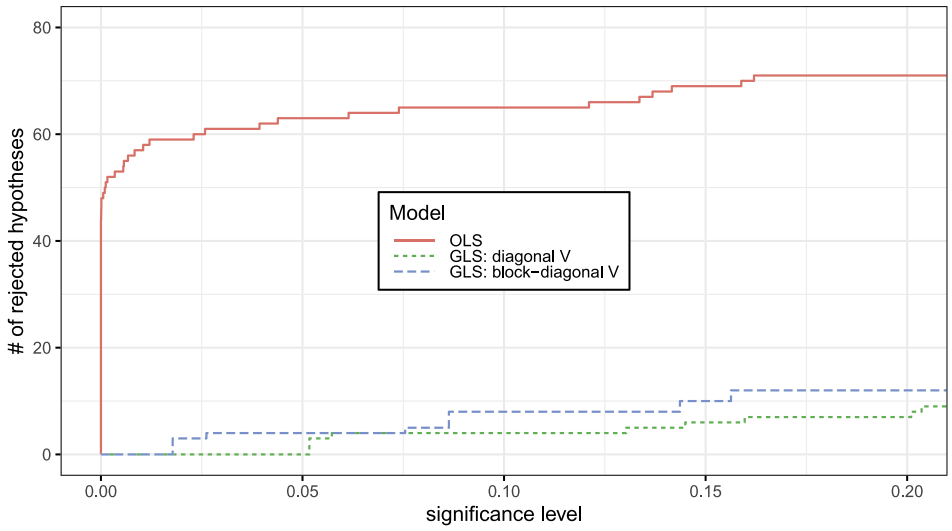


FIG. 12. *Number of rejected hypotheses as a function of significance level with a Benjamini–Hochberg multiple testing correction. Red*: *OLS*; *green*: *GLS with diagonal V*; *blue*: *GLS with block-diagonal V*.

3.5. *Comparison with Xia et al.* (2020). A related approach that can be applied in our setting was proposed by Xia et al. (2020), who fit a model of the relationship between phenotypes and brain connectivity using penalized least squares. Adjusting their notation to match ours, they solve the optimization problem

$$\underset{\Theta,\alpha}{\text{minimize}} \sum_{m=1}^{N} \left\| A^{(m)} - \Theta - d_m Z \alpha Z^T \right\|_F^2 + \lambda_1 \|\Theta\|_\star + \lambda_2 \|\alpha\|_1, \tag{3.1}$$

where $\Theta \in \mathbb{R}^{n \times n}$ is the intercept, $\alpha \in \mathbb{R}^{K \times K}$ is a matrix of cell effects, d_m is a binary indicator for schizophrenia diagnosis, $\|\Theta\|_\star$ is the nuclear norm of Θ (i.e., the sum of the singular values), $\|\alpha\|_1$ is the sum of the absolute values of the entries of the matrix α, and λ_1 and λ_2 are tuning parameters. We applied this method to the COBRE data using code available from Xia et al. (2020). We selected tuning parameters by searching over a grid, with each parameter taking possible values in the set $\{e^{-10}, e^{-9}, \ldots, e^{10}\}$, and computing the average loss from

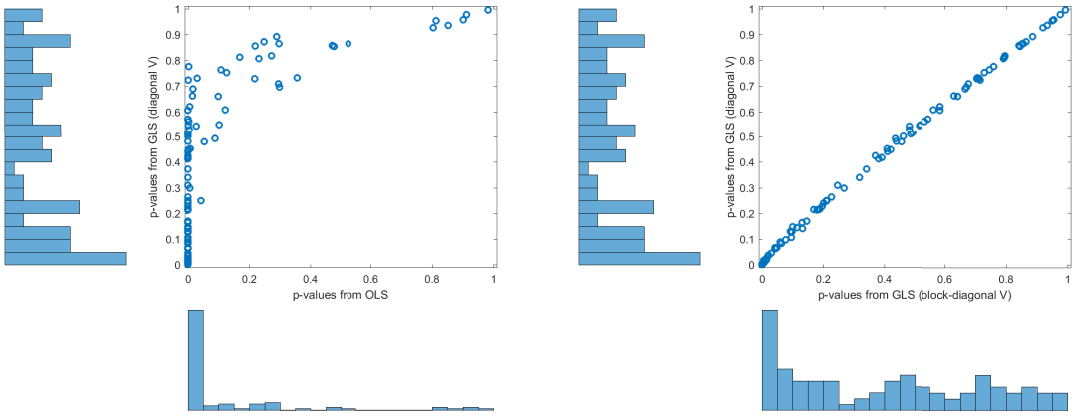


FIG. 13. *Scatter plot of raw p-values. Left*: *OLS vs. GLS with block-diagonal V*; *right*: *GLS with diagonal V vs. GLS with block-diagonal V*. *Histograms depict the marginal distribution of p-values for each setting.*

TABLE 2

Five most significant cells from GLS methods. The ordering of the cells, obtained by two GLS versions, is the same. Last three columns show the p-values from GLS methods and the Benjamini–Hochberg adjusted significance level, that is, $\frac{0.05k}{91}$ *for* $k = 1 \ldots 5$

p-value rank	Network cell	Diagonal V	Block-diagonal V	Adjusted significance level
1	(8, 8)	8.9e−04	3.0e−04	5.5e−04
2	(6, 12)	0.0012	5.1e−04	0.0011
3	(2, 11)	0.0017	5.9e−04	0.0016
4	(13, 13)	0.0025	0.0012	0.0022
5	(4, 5)	0.0072	0.0041	0.0027

five-fold cross-validation at each point. We then refitted the model with the best-performing parameters ($\lambda_1 = e^2$, $\lambda_2 = e^6$); we visualize both Θ and α in Figure 14.

The intercept Θ appears qualitatively similar to that estimated by our method (see left column of Figure 7). Since this method induces sparsity in the cell effects α, we compare the cells it selected as nonzero to the cells our method identified as statistically significant (see Figure 11). While there are some overlaps between the two sets of results (e.g., cells involving Brain System 5, the Default mode system), overall the patterns are not especially similar. These dissimilarities can largely be explained, however, by considering the objective function (3.1) which does not adjust for the relative size of the network cells induced by the varying functional system sizes. For example, the cell (5, 5) comprises 1653 edges, whereas the cell (13, 13) has only six edges. Since the penalty on α does not adjust for this, the model tends to select the largest cells. Whether or not this is a desirable property depends upon the question being asked; in the present setting we are more interested in knowing which cells show group differences, as opposed to just minimizing the fitted error of the model. Another important distinction is that the method of Xia et al. (2020) does not provide inference, so it is more difficult to assign significance to the pattern of discovered effects. Our approach, in contrast, provides p-values which can be then appropriately corrected for multiple comparisons.

4. Discussion. We introduced a new framework for modeling multiple brain networks with the goal of taking network structure into account when assessing covariate effects, in

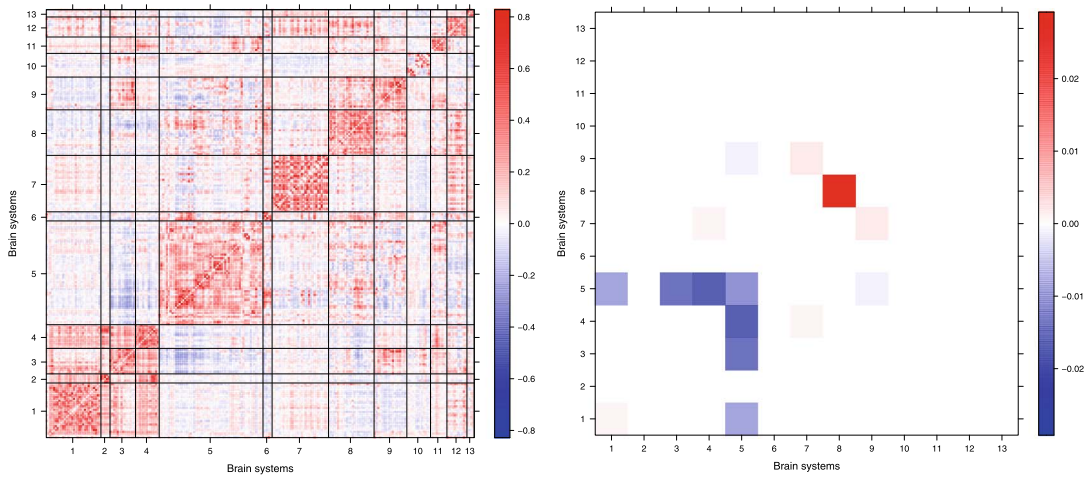


FIG. 14. *Results of applying the method of Xia et al. (2020). Left: The intercept* Θ*; right: the cell effects* α*.*

modeling both the mean and the variance. The choice of a linear mixed effects model allowed for a simple interpretable decomposition into cell-level and edge-level effects while accounting for individual variation. In an important departure from typical network models, we allowed for reasonably general dependency between a very large number of edges by allowing a general variance structure in the linear model. While our application focused on a single binary disease status covariate, the method can be applied to a general linear model in subject-level covariates.

Our empirical results suggest that, without a variance structure that allows for some edge dependence, standard errors are severely underestimated, producing unreliable and misleadingly small p-values. Modeling the variance with a network community structure, on the other hand, results in accurate standard errors, as shown both on synthetic data and on the analysis of healthy subjects from the COBRE data split into two parts at random. While the analysis comparing schizophrenia patients and healthy controls has no ground truth, our results confirm important earlier findings, such as a disconnection between frontal and temporal cortices in schizophrenia (Bullmore and Bassett (2011), Friston and Frith (1995)).

Previously proposed methods for comparing brain networks tend to perform standard two-sample inference on either global graph summaries or vectorized edge weights. While the global graph summaries do account for network structure, they tend to collapse a lot of information and, generally, do not do as well in classification tasks (Arroyo Relión et al. (2019)). Vectorized edge weights, on the other hand, preserve all the information but do not take advantage of the network structure. In contrast, our method allows inference, both at the level of cell means (as depicted, e.g., in Figure 11) and at the level of individual edges (as depicted, e.g., in Figure 9), and it does so while accounting for edge dependence. Importantly, it provides an interpretable model and, in particular, hypothesis testing for specific systems in the brain, which none of the previous methods can easily do. System-level inference also greatly reduces the number of hypotheses to be tested, compared to the massive-univariate approach (vectorized edge weights) which suggests our tests have more power after correcting for multiple testing. While some previous work (Xia et al. (2020)) does directly consider system-level effects, it does not readily provide inference for these effects, and so results must generally be interpreted qualitatively.

One limitation of any approach that works with edge weights is that these are summary statistics drawn from a sequence of time courses acquired via fMRI; shorter time courses or higher levels of noise will limit the effectiveness of the method. While we include individual edge random effects that can reflect this noise in edge weights, there is a limit to how useful the results can be in the presence of very noisy edge weights. Another potential limitation of our method is using a predetermined parcellation. While one can learn a parcellation first or even use the fitted edge effects to improve a given parcellation (e.g. split cells that show a lot of heterogeneity), validity of inference in the presence of such a model selection step is unclear. Still, we believe developing graph-aware approaches that strike a balance between massively univariate but information-preserving methods and global summaries is a fruitful direction for future work on multiple network analysis, both for the brain imaging application and other applications involving multiplex networks, such as global trade networks (in multiple commodities), transportation networks (by different means of transport), social networks (with different types of connections), and many others.

Acknowledgments. We thank Professor Stephan Taylor (Psychiatry, University of Michigan) and Professor Chandra Sripada (Psychiatry and Philosophy, University of Michigan) and members of both of their labs for many useful discussions, and the Taylor lab for providing processed schizophrenia data. We thank Jesús Arroyo Relión (Statistics, Texas A&M University) for his help with the data.

Funding. This research was supported in part by NSF DMS grants 1521551, 1646108, and 1916222, ONR grant N000141612910, a Rackham Predoctoral Fellowship from the University of Michigan awarded to D. Kessler, and a Dana Foundation grant to E. Levina as well as by computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

REFERENCES

AINE, C. J., BOCKHOLT, H. J., BUSTILLO, J. R., CAÑIVE, J. M., CAPRIHAN, A., GASPAROVIC, C., HANLON, F. M., HOUCK, J. M., JUNG, R. E. et al. (2017). Multimodal neuroimaging in schizophrenia: Description and dissemination. *Neuroinformatics* **15** 343–364.

ANGRILLI, A., SPIRONELLI, C., ELBERT, T., CROW, T. J., MARANO, G. and STEGAGNO, L. (2009). Schizophrenia as failure of left hemispheric dominance for the phonological component of language. *PLoS ONE* **4** e4507. https://doi.org/10.1371/journal.pone.0004507

ARROYO RELIÓN, J. D., KESSLER, D., LEVINA, E. and TAYLOR, S. F. (2019). Network classification with applications to brain connectomics. *Ann. Appl. Stat.* **13** 1648–1677. MR4019153 https://doi.org/10.1214/19-AOAS1252

BAHRAMI, M., LAURIENTI, P. J. and SIMPSON, S. L. (2019). A Matlab toolbox for multivariate analysis of brain networks. *Hum. Brain Mapp.* **40** 175–186. https://doi.org/10.1002/hbm.24363

BAHRAMI, M., LAURIENTI, P. J., QUANDT, S. A., TALTON, J., POPE, C. N., SUMMERS, P., BURDETTE, J. H., CHEN, H., LIU, J. et al. (2017). The impacts of pesticide and nicotine exposures on functional brain networks in Latino immigrant workers. *NeuroToxicology* **62** 138–150.

BELILOVSKY, E., VAROQUAUX, G. and BLASCHKO, M. B. (2016). Testing for differences in Gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems* 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 595–603. Curran Associates, Red Hook.

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

BULLMORE, E. T. (2012). Functional network endophenotypes of psychotic disorders. *Biol. Psychiatry* **71** 844–845.

BULLMORE, E. T. and BASSETT, D. S. (2011). Brain graphs: Graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* **7** 113–140. https://doi.org/10.1146/annurev-clinpsy-040510-143934

BULLMORE, E. T. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.

CHUNG, J., BRIDGEFORD, E., ARROYO, J., PEDIGO, B. D., SAAD-ELDIN, A., GOPALAKRISHNAN, V., XIANG, L., PRIEBE, C. E. and VOGELSTEIN, J. T. (2021). Statistical connectomics. *Annu. Rev. Stat. Appl.* **8** 463–492. MR4243556 https://doi.org/10.1146/annurev-statistics-042720-023234

CRADDOCK, R. C., HOLTZHEIMER, P. E., HU, X. P. and MAYBERG, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.* **62** 1619–1628.

CRADDOCK, R. C., JBABDI, S., YAN, C.-G., VOGELSTEIN, J. T., CASTELLANOS, F. X., MARTINO, A. D., KELLY, C., HEBERLEIN, K., COLCOMBE, S. et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* **10** 524–539. https://doi.org/10.1038/nmeth.2482

FRISTON, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2** 56–78.

FRISTON, K. J. and FRITH, C. D. (1995). Schizophrenia: A disconnection syndrome? *Clin. Neurosci.* **3** 89–97.

FIECAS, M., CRIBBEN, I., BAHKTIARI, R. and CUMMINE, J. (2017). A variance components model for statistical inference on functional connectivity networks. *NeuroImage* **149** 256–266. https://doi.org/10.1016/j.neuroimage.2017.01.051

FUJITA, A., TAKAHASHI, D. Y., BISOL BALARDIN, J., CALEBE VIDAL, M. and SATO, J. R. (2017). Correlation between graphs with an application to brain network analysis. *Comput. Statist. Data Anal.* **109** 76–92. MR3603642 https://doi.org/10.1016/j.csda.2016.11.016

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126 https://doi.org/10.1093/biomet/75.4.800

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

KIM, J., WOZNIAK, J. R., MUELLER, B. A., SHEN, X. and PAN, W. (2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage* **101** 681–694.

LANDIS, D., COURTNEY, W., DIERINGER, C., KELLY, R., KING, M., MILLER, B., WANG, R., WOOD, D., TURNER, J. A. et al. (2016). COINS data exchange: An open platform for compiling, curating, and disseminating neuroimaging data. *NeuroImage* **124** 1084–1088.

LI, J. (2015). *The Influence of Misspecification of Between-Subject and Within-Subject Covariance Structures in Hierarchical Growth Models*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Pittsburgh. MR3427391

LIANG, X., WANG, J., YAN, C., SHU, N., XU, K., GONG, G. and HE, Y. (2012). Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: A resting-state functional MRI study. *PLoS ONE* **7** e32766.

MITCHELL, R. L. C. and CROW, T. J. (2005). Right hemisphere language functions and schizophrenia: The forgotten hemisphere? *Brain* **128** 963–978. https://doi.org/10.1093/brain/awh466

NARAYAN, M., ALLEN, G. I. and TOMSON, S. (2015). Two sample inference for populations of graphical models with applications to functional connectivity. Preprint. Available at arXiv:1502.03853.

NARAYAN, M. and ALLEN, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Front. Neurosci.* **10** 108. https://doi.org/10.3389/fnins.2016.00108

PALANIYAPPAN, L., SIMMONITE, M., WHITE, T. P., LIDDLE, E. B. and LIDDLE, P. F. (2013). Neural primacy of the salience processing system in schizophrenia. *Neuron* **79** 814–828. https://doi.org/10.1016/j.neuron.2013.06.027

PAN, W., KIM, J., ZHANG, Y., SHEN, X. and WEI, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* **197** 1081–1095.

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

RIBOLSI, M., DASKALAKIS, Z. J., SIRACUSANO, A. and KOCH, G. (2014). Abnormal asymmetry of brain connectivity in schizophrenia. *Front. Human Neurosci.* **8** 1010. https://doi.org/10.3389/fnhum.2014.01010

SIMPSON, S. L., BAHRAMI, M. and LAURIENTI, P. J. (2019). A mixed-modeling framework for analyzing multitask whole-brain network data. *Netw. Neurosci.* **3** 307–324. https://doi.org/10.1162/netn_a_00065

SIMPSON, S. L. and LAURIENTI, P. J. (2015). A two-part mixed-effects modeling framework for analyzing whole-brain network data. *NeuroImage* **113** 310–319. https://doi.org/10.1016/j.neuroimage.2015.03.021

SMITH, S. M. (2012). The future of FMRI connectivity. *NeuroImage* **62** 1257–1266. https://doi.org/10.1016/j.neuroimage.2012.01.022

SMITH, S. M., MILLER, K. L., SALIMI-KHORSHIDI, G., WEBSTER, M., BECKMANN, C. F., NICHOLS, T. E., RAMSEY, J. D. and WOOLRICH, M. W. (2011). Network modelling methods for FMRI. *NeuroImage* **54** 875–891. https://doi.org/10.1016/j.neuroimage.2010.08.063

SMITH, S. M., BECKMANN, C. F., ANDERSSON, J., AUERBACH, E. J., BIJSTERBOSCH, J., DOUAUD, G., DUFF, E., FEINBERG, D. A., GRIFFANTI, L. et al. (2013). Resting-state fMRI in the human connectome project. *NeuroImage* **80** 144–168.

SOBEL, M. E. and LINDQUIST, M. A. (2014). Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Amer. Statist. Assoc.* **109** 967–976. MR3265669 https://doi.org/10.1080/01621459.2014.922886

TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. MR3640191 https://doi.org/10.1080/10618600.2016.1193505

VAN DEN HEUVEL, M. P. and POL, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20** 519–534. https://doi.org/10.1016/j.euroneuro.2010.03.008

VAN DEN HEUVEL, M. P., MANDL, R. C. W., KAHN, R. S. and POL, H. E. H. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Hum. Brain Mapp.* **30** 3127–3141. https://doi.org/10.1002/hbm.20737

VAROQUAUX, G. and CRADDOCK, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage* **80** 405–415. https://doi.org/10.1016/j.neuroimage.2013.04.007

VENKATARAMAN, A., WHITFORD, T. J., WESTIN, C.-F., GOLLAND, P. and KUBICKI, M. (2012). Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophr. Res.* **139** 7–12. https://doi.org/10.1016/j.schres.2012.04.021

WOOD, D., KING, M., LANDIS, D., COURTNEY, W., WANG, R., KELLY, R., TURNER, J. A. and CALHOUN, V. D. (2014). Harnessing modern web application technology to create intuitive and efficient data visualization and sharing tools. *Front. Neuroinform.* **8** 71.

XIA, M., WANG, J. and HE, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.

XIA, C. H., MA, Z., CUI, Z., BZDOK, D., THIRION, B., BASSETT, D. S., SATTERTHWAITE, T. D., SHINOHARA, R. T. and WITTEN, D. M. (2020). Multi-scale network regression for brain-phenotype associations. *Hum. Brain Mapp.*. https://doi.org/10.1002/hbm.24982

YEO, B. T. T., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.

ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041

ZALESKY, A., COCCHI, L., FORNITO, A., MURRAY, M. M. and BULLMORE, E. (2012). Connectivity differences in brain networks. *NeuroImage* **60** 1055–1062. https://doi.org/10.1016/j.neuroimage.2012.01.068

ZHEN, Z., TIAN, J., QIN, W. and ZHANG, H. (2007). Partial correlation mapping of brain functional connectivity with resting state fMRI. *Proc. SPIE* **6511** 651112.