Generalized Connectivity Matrix Response Regression with Applications in Brain Connectivity Studies

Jingfei Zhang [1], Will Wei Sun [2] and Lexin Li [3]

[1] *Department of Management Science, Miami Herbert Business School,*

University of Miami, Miami, FL, 33146.

[2] *Krannert School of Management,*

Purdue University, West Lafayette, IN, 47906.

[3] *Department of Biostatistics and Epidemiology, School of Public Health,*

University of California at Berkeley, Berkeley, CA, 94720.

Abstract

Multiple-subject network data are fast emerging in recent years, where a separate connectivity matrix is measured over a common set of nodes for each individual subject, along with subject covariates information. In this article, we propose a new generalized matrix response regression model, where the observed network is treated as a matrix-valued response and the subject covariates as predictors. The new model characterizes the population-level connectivity pattern through a low-rank intercept matrix, and the effect of subject covariates through a sparse slope tensor. We develop an efficient alternating gradient descent algorithm for parameter estimation, and establish the non-asymptotic error bound for the actual estimator from the algorithm, which quantifies the interplay between the computational and statistical errors. We further show the strong consistency for graph community recovery, as well as the edge selection consistency. We demonstrate the efficacy of our method through simulations and two brain connectivity studies.

Keywords: Computational and statistical errors; Generalized linear model; High-dimensional regression; Neuroimaging; Tensors.

1 Introduction

Network data are now ubiquitous in a wide range of scientific applications. More recently, multiple-subject network data are fast emerging, in which a separate connectivity network is measured over a common set of nodes for each individual subject. Examples include social cognitive science (Brands, 2013), genetics (Dai et al., 2019), and our motivating brain connectivity analysis. Brain connectivity concerns functional and structural architectures of the brain (Varoquaux and Craddock, 2013). A typical connectivity study collects imaging scans, e.g., functional magnetic resonance imaging (fMRI), or diffusion tensor imaging (DTI), from multiple subjects. Based on the scan, a connectivity network is constructed for each subject, with the nodes corresponding to a common set of brain regions, and the edges encoding functional or structural associations between the regions. In addition, the study collects subject features such as age, sex and other traits. A fundamental scientific question of interest is to characterize the brain connectivity at both the population-level and subject-level, and to ascertain how subject features modulate the subject-level connectivity changes. Characterizing such individualized brain connectivity networks is central in developing personalized treatment for neurological disorders (Sylvester et al., 2020).

There have been some recent proposals on modeling a collection of networks (Chen et al., 2015; Kang et al., 2016; Wang et al., 2016; Zhang and Cao, 2017; Kundu et al., 2018; Wang and Guo, 2019). However, these methods may not be able to capture complex associations between the network connectivity and external covariates. Wang et al. (2017); Durante et al. (2017) proposed Bayesian network models with covariates, which are flexible, but can be computationally intensive, especially for large networks or a large number of covariates. There is another line of related work in matrix and tensor data analysis. Notably, Sun and Li (2017) developed a tensor response regression model, Kong et al. (2019) proposed a matrix response linear regression model, and Hu et al. (2020) considered a matrix response regression based on nonlinear kernels. These models were designed to handle a continuous-valued response, and imposed different structures on the coefficients compared to our model. Relatedly, Zhang and Li (2017); Li and Zhang (2017); Tang et al. (2019) studied tensor

predictor models, where the tensor was treated as a predictor and the response variable was a scalar.

In this article, we propose a new connectivity matrix response generalized linear regression model for a collection of network samples with network-level covariates. We represent the observed network as a matrix-valued response variable, and the subject covariates as predictors. We then adopt the form of generalized linear model (GLM), and formulate the population-level connectivity, after a proper transformation, as the sum of two high-dimensional components. The first component is the intercept matrix and is assumed to possess a low-rank structure. The second component involves the slope coefficient tensor, which models the effects of covariates on the connectivity and is assumed to be sparse. These structural assumptions substantially reduce the number of free parameters, as well as the subsequent modeling and computation complexity. Moreover, they are scientifically plausible, and are frequently employed in scientific applications (Bi et al., 2018).

Our proposal makes some useful contributions to both methodology and theory. As to the methodology, we develop a systematic approach to model the associations between connectivity matrices and covariates. The proposed model framework preserves the intrinsic characteristics of networks, facilitates a scalable computation, and allows an explicit quantification of the computational and statistical errors. Besides, although our motivating application is the brain connectivity study, our method is applicable to other problems, e.g., the genetic study that investigates the gene regulatory relationships among gene-gene networks based on single-cell samples (Dai et al., 2019). As to the theory, we establish several useful statistical properties. We obtain an explicit non-asymptotic error bound for the iterates of our algorithm. This error bound reveals an interesting interplay between the computational efficiency and statistical rate of convergence. It shows that the computational error decays geometrically with the number of iterations, while the statistical error matches with the existing rates for sparse regressions and low-rank regressions. Built on this error bound, we further establish the consistency of a community detection procedure and the selection consistency, in that we can consistently identify the edges that are affected by the covariates, and exclude those that are not. These theoretical analyses are highly

nontrivial, involving alternating gradient descent, factorization of the low-rank component, hard-thresholding operator for sparsity, and non-quadratic form of the loss function.

The rest of the article is organized as follows. Section 2 introduces the generalized matrix response model. Section 3 develops the estimation algorithm, and Section 4 investigates the statistical properties. Section 5 presents the simulations, and Section 6 illustrates with two studies of brain functional and structural connectivity. Section 7 concludes the paper with a short discussion. All technical proofs are relegated to the supplement.

2 Generalized Connectivity Matrix Response Model

We start with some notation. Let $\boldsymbol{I}_{n\times n}$ denote the $n \times n$ identity matrix. For a vector $\boldsymbol{b} \in \mathbb{R}^{d_1}$, let $\|\boldsymbol{b}\|_2$ denote its ℓ_2 norm. For a matrix $\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$, let $\boldsymbol{B}_{i\cdot}$ and $\boldsymbol{B}_{\cdot j}$ denote its ith row and jth column, and let $\|\boldsymbol{B}\|_2$, $\|\boldsymbol{B}\|_*$, $\|\boldsymbol{B}\|_F$, and $\|\boldsymbol{B}\|_\infty$ denote its spectral norm, nuclear norm, Frobenius norm, and entry-wise infinity norm, respectively. Let $\mathrm{SVD}_r(\boldsymbol{B})$ denote the rank-r singular value decomposition of $\boldsymbol{B}$ such that $\mathrm{SVD}_r(\boldsymbol{B}) = [\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}]$, where $\boldsymbol{\Sigma}_{r\times r}$ is a diagonal matrix with the largest r singular values and $\boldsymbol{U}_{d_1\times r}, \boldsymbol{V}_{d_2\times r}$ collect the left and right singular vectors, respectively. For a tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1\times d_2\times d_3}$, let $\boldsymbol{\mathcal{B}}_{ijk}$, $\boldsymbol{\mathcal{B}}_{ij\cdot}$ and $\boldsymbol{\mathcal{B}}_{\cdot\cdot k}$ denote its (i, j, k)th entry, (i, j)th tube fiber, and kth frontal slice, respectively. Let $\|\boldsymbol{\mathcal{B}}\|_F = \sqrt{\sum_{ijk} \boldsymbol{\mathcal{B}}_{ijk}^2}$ and $\|\boldsymbol{\mathcal{B}}\|_0$ denote the number of nonzero entries. Lastly, define the tensor matrix product $\langle \boldsymbol{B}, \boldsymbol{\mathcal{B}} \rangle = \sum_{ijk} \boldsymbol{\mathcal{B}}_{ijk}\boldsymbol{B}_{ij}$ for $\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1\times d_2\times d_3}$.

2.1 Model formulation

Consider a network with n nodes, and the $n\times n$ adjacency matrix $\boldsymbol{A}$, where $\boldsymbol{A}_{jj'}$ denotes the edge from node j to j', $1 \le j, j' \le n$. If the edge is undirected, then $\boldsymbol{A}_{jj'} = \boldsymbol{A}_{j'j}$. The edge value can be binary, i.e., $\boldsymbol{A}_{jj'} \in \{0, 1\}$, or count, i.e., $\boldsymbol{A}_{jj'}$ is a nonnegative integer. We consider independent network samples observed from N individuals, with corresponding $n \times n$ adjacency matrices $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}$. Here we assume all N networks share a common set of n nodes. Additionally, for each subject, we observe a vector of p covariates, denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$.

Denote $\boldsymbol{\mu}^{(i)} = \mathbb{E}\{\boldsymbol{A}^{(i)}|\boldsymbol{x}_i\}$, where the expectation $\mathbb{E}(\cdot)$ is applied element-wise to the entries in $\boldsymbol{A}^{(i)}$. We assume that $\boldsymbol{A}^{(i)}$ conditional on $\boldsymbol{x}_i$ follows an exponential family distribution with a canonical link function, i.e.,

$$f(\boldsymbol{A}^{(i)}|\boldsymbol{\mu}^{(i)}) = \prod_{j \neq j'}^{n} h(\boldsymbol{A}^{(i)}_{jj'}) \exp\left[\boldsymbol{A}^{(i)}_{jj'}\boldsymbol{\eta}^{(i)}_{jj'} - \psi\left\{\boldsymbol{\eta}^{(i)}_{jj'}\right\}\right], \tag{1}$$

where $\boldsymbol{\eta}^{(i)} = g(\boldsymbol{\mu}^{(i)})$, $g(\cdot)$ is a known invertible link function in usual GLM and is applied element-wise to the entries of $\boldsymbol{\mu}^{(i)}$, and $\psi(\cdot)$ is the cumulant function with its first derivative $\psi'(\cdot) = g(\cdot)^{-1}$. Furthermore, we postulate that,

$$g\left\{\boldsymbol{\mu}^{(i)}\right\} = \boldsymbol{\Theta} + \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i, \quad i = 1, \ldots, N, \tag{2}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{n \times n}$ is the intercept matrix that characterizes the population level connectivity, $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{n \times n \times p}$ is the slope tensor that encodes the effects of subject covariates on the connectivity matrix, and $\boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i = \sum_{l=1}^{p} x_{il} \boldsymbol{\mathcal{B}}_{\cdot\cdot l}$.

We postulate that the population level connectivity $\boldsymbol{\Theta}$ to be low-rank, which reduces the number of free parameters and is plausible in neuroscience applications (Bi et al., 2018; Kong et al., 2019). In addition, we assume $\boldsymbol{\mathcal{B}}$ is sparse, i.e., the effects of covariates concentrate only on a subset of connections. This sparsity assumption again reduces the number of free parameters, greatly facilitates the model interpretation, and is also well supported by empirical neurological studies (Vounou et al., 2010). As every subject has a unique sparse deviation $\boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i$ from the low-rank $\boldsymbol{\Theta}$, model (2) is identifiable. We note that it is possible to impose more complex structures on $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$; e.g., $\boldsymbol{\mathcal{B}}$ is low-rank and sparse, or $\boldsymbol{\mathcal{B}}$ is slice sparse. Accommodating these structures requires some straightforward modification to the estimation procedure. We choose to focus on the current setup as it offers a good balance between model complexity and model flexibility.

To ensure the low-rank structure of $\boldsymbol{\Theta}$, we adopt the Burer-Monteiro factorization (Burer and Monteiro, 2003), in which the low-rank matrix is reparameterized as the product of two factor matrices, $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n \times r}$, and r is the rank of $\boldsymbol{\Theta}$. This reparameterization avoids repeatedly performing the computationally expensive SVD, which is often required in optimization with the low-rank constraint. If the adjacency matrix

is symmetric, we reparameterize $\boldsymbol{\Theta}$ as $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\Lambda}$ is an $r \times r$ diagonal matrix with diagonal entries $\{-1, 1\}$. If $\boldsymbol{\Theta}$ is positive semi-definite (PSD), then $\boldsymbol{\Lambda}$ becomes the identity matrix. We note that the intercept matrix $\boldsymbol{\Theta}$ may not be PSD. To enforce the sparsity of $\boldsymbol{\mathcal{B}}$, we adopt the hard-thresholding sparsity constraint, by setting $\|\boldsymbol{\mathcal{B}}\|_0 \leq s$ for some positive integer s. Compared to the lasso type soft-thresholding constraint, the hard-thresholding constraint reduces bias and has been shown to enjoy superior performance in many high-dimensional problems (Zhang et al., 2018).

We briefly discuss the benefits and necessity of imposing separate structures on $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$. At first glance, it seems that one could stack $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$ into one coefficient tensor of size $n \times n \times (p + 1)$, and require it to be both low-rank and sparse. However, assuming $\boldsymbol{\Theta}$ to be sparse may not be plausible in the GLM setting. For instance, when the edges are binary and $g(\cdot)$ is the logit link, $g(0)$ yields a connecting probability of 0.5; when the edges are counts and $g(\cdot)$ is the log link, $g(0)$ is not well defined. As such, a sparse $\boldsymbol{\Theta}$ does not necessarily imply the sparsity in connectivity at the population-level, and may not even be well defined. This is a unique challenge in using GLM to model discrete-valued connectivity matrices. In addition, based on the Burer-Monteiro reparameterization of the low-rank $\boldsymbol{\Theta}$, we can detect clusters, or communities, of nodes, so that the nodes are more densely connected within the clusters and less so between the clusters. This also contributes to the network community detection literature, as the existing spectral clustering methods cannot handle network heterogeneity induced by network-level covariates.

2.2 Connections with existing models for a single network

Our proposed model (2), when applied to a single network sample, is connected to several prevalent network models, including the stochastic blockmodel (Holland et al., 1983), the latent space model (Hoff et al., 2002), and the latent factor model (Minhas et al., 2016). Similar to those models, our model (2) also assumes the low-rank structure, but is more general in that it imposes no additional structural constraint, e.g., the block structure.

Consider a single observed adjacency matrix $\boldsymbol{A}$ and $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{A})$. The stochastic block-

model is one of the most popular network models. It imposes that the nodes form K communities, and the edges are determined by the community memberships of the two end nodes and are independent given the community assignment. Accordingly, the model can be written as

$$g(\boldsymbol{\mu}) = \boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^\top,$$

where $\boldsymbol{C}$ is a $n \times K$ community assignment matrix, with $\boldsymbol{C}_{jk} = 1$ if node j belongs to the kth community, and 0 otherwise, and $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ characterizes the connecting probabilities within and between the K communities. It is seen that the rank of the matrix $\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^\top$ at most K, and may be viewed as a special case of model (2).

The latent space model (Hoff et al., 2002) is another well-studied network model. It assumes the nodes are positioned in a K-dimensional latent space, and two nodes are likely to form a tie if their latent positions are close. The model can be written as

$$g(\boldsymbol{\mu}) = \alpha \mathbf{1}_n \mathbf{1}_n^\top + \boldsymbol{C}(\boldsymbol{M}\boldsymbol{C})^\top,$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones, $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ has its jth row $\boldsymbol{c}_j^\top \in \mathbb{R}^{K\times 1}$ encoding the latent position of node j, and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with its jth diagonal entry equal to $1/\|\boldsymbol{c}_j\|_2$, $1 \le j \le n$. We see the rank of the matrix $\alpha\mathbf{1}\mathbf{1}^\top + \boldsymbol{C}(\boldsymbol{M}\boldsymbol{C})^\top$ is $(K+1)$, and thus this model is again a special case of (2). Relatedly, the latent factor model (Minhas et al., 2016), similar to the latent space model, imposes

$$g(\boldsymbol{\mu}) = \boldsymbol{\alpha} \otimes \mathbf{1}_n^\top + \boldsymbol{\alpha}^\top \otimes \mathbf{1}_n + \boldsymbol{C}\boldsymbol{C}^\top,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$ encodes the additive effect, and $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ encodes the multiplicative effect. In this case, the rank of the matrix $\boldsymbol{\alpha} \otimes \mathbf{1}_n^\top + \boldsymbol{\alpha}^\top \otimes \mathbf{1}_n + \boldsymbol{C}\boldsymbol{C}^\top$ is $(K+1)$.

3 Estimation

Denote the negative log-likelihood function of the connectivity matrix response model (2) by $\ell(\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}})$, which, up to a constant, is of the form (McCullagh and Nelder, 1989),

$$\ell(\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq j'}^{n}\left[\boldsymbol{A}_{jj'}^{(i)}\boldsymbol{\eta}_{jj'}^{(i)} - \psi\left\{\boldsymbol{\eta}_{jj'}^{(i)}\right\}\right], \tag{3}$$

Algorithm 1 Optimization algorithm for (4)

Step 1: compute $\bar{\boldsymbol{A}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{A}^{(i)}$ and let $\mathrm{SVD}_r\{g(\bar{\boldsymbol{A}})\} = [\bar{\boldsymbol{U}}_0, \bar{\boldsymbol{\Sigma}}_0, \bar{\boldsymbol{V}}_0]$. Set $\boldsymbol{U}^{(0)} = \bar{\boldsymbol{U}}_0\bar{\boldsymbol{\Sigma}}_0^{1/2}$, $\boldsymbol{V}^{(0)} = \bar{\boldsymbol{V}}_0\bar{\boldsymbol{\Sigma}}_0^{1/2}$, and $\boldsymbol{\mathcal{B}}^{(0)} = \boldsymbol{0}$.

repeat

Step 2: update $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \delta\nabla_{\boldsymbol{U}}\tilde{\ell}\left\{\boldsymbol{U}\boldsymbol{V}^{(t)\top}, \boldsymbol{\mathcal{B}}^{(t)}\right\}\Big|_{\boldsymbol{U}=\boldsymbol{U}^{(t)}}$;

Step 3: update $\boldsymbol{V}^{(t+1)} = \boldsymbol{V}^{(t)} - \delta\nabla_{\boldsymbol{V}}\tilde{\ell}\left\{\boldsymbol{U}^{(t+1)}\boldsymbol{V}^{\top}, \boldsymbol{\mathcal{B}}^{(t)}\right\}\Big|_{\boldsymbol{V}=\boldsymbol{V}^{(t)}}$;

Step 4: update $\boldsymbol{\mathcal{B}}^{(t+1)} = \mathrm{Truncate}\left[\boldsymbol{\mathcal{B}}^{(t)} - \tau\nabla_{\boldsymbol{\mathcal{B}}}\tilde{\ell}\left\{\boldsymbol{U}^{(t+1)}\boldsymbol{V}^{(t+1)\top}, \boldsymbol{\mathcal{B}}\right\}\Big|_{\boldsymbol{\mathcal{B}}=\boldsymbol{\mathcal{B}}^{(t)}}, s\right]$.

until the objective function converges.

where $\boldsymbol{\eta}^{(i)} = \boldsymbol{\Theta} + \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i$. We propose to estimate the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$ through a non-convex regularized optimization. We first develop the optimization algorithm for the general case without the symmetry constraint, which is an easier scenario. Building upon this procedure, we further develop the algorithm for the symmetric case.

For the general case that $\boldsymbol{\Theta}$ is low-rank but not necessarily symmetric, we consider the factorization $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^{\top}$ and the corresponding optimization problem,

$$\min_{\substack{\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{n\times r}\\ \boldsymbol{\mathcal{B}}\in\mathbb{R}^{n\times n\times p}}} \tilde{\ell}\left(\boldsymbol{U}\boldsymbol{V}^{\top}, \boldsymbol{\mathcal{B}}\right), \quad \text{subject to} \quad \|\boldsymbol{\mathcal{B}}\|_0 \le s, \tag{4}$$

where we augment the loss function $\ell(\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}})$ with an additional regularizer, $\tilde{\ell}\left(\boldsymbol{U}\boldsymbol{V}^{\top}, \boldsymbol{\mathcal{B}}\right) = \ell\left(\boldsymbol{U}\boldsymbol{V}^{\top}, \boldsymbol{\mathcal{B}}\right) + \frac{1}{8}\|\boldsymbol{U}^{\top}\boldsymbol{U} - \boldsymbol{V}^{\top}\boldsymbol{V}\|_F^2$. In our theoretical analysis, we write the true intercept matrix $\boldsymbol{\Theta}^*$ as $\boldsymbol{\Theta}^* = \boldsymbol{U}^*\boldsymbol{V}^{*\top}$, and assume $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ have the same set of singular values; see Section S2 of the supplement. The regularizer $\|\boldsymbol{U}^{\top}\boldsymbol{U} - \boldsymbol{V}^{\top}\boldsymbol{V}\|_F^2/8$ is added to guarantee the uniqueness of solutions to $\boldsymbol{U}$ and $\boldsymbol{V}$ in the optimization. In low-rank matrix factorization, a regularizer of this type, i.e., $\lambda_0\|\boldsymbol{U}^{\top}\boldsymbol{U} - \boldsymbol{V}^{\top}\boldsymbol{V}\|_F^2$ with $\lambda_0 > 0$, has been commonly used (Tu et al., 2016; Zheng and Lafferty, 2016; Park et al., 2018), whereas the scalar λ_0 is sometimes treated as a tuning parameter (Park et al., 2018). Assuming that $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ have the same set of singular values, our theoretical analysis establishes the linear convergence of the proposed algorithm when $\lambda_0 = 1/8$, and thus we treat λ_0 as a fixed constant in our method. To enforce sparsity along the solution path, we employ a truncation operator $\mathrm{Truncate}(\boldsymbol{\mathcal{B}}, s)$,

$$[\text{Truncate}(\boldsymbol{\mathcal{B}}, s)]_{jj'l} = \begin{cases} \boldsymbol{\mathcal{B}}_{jj'l} & \text{if } (j, j', l) \in \text{supp}(\boldsymbol{\mathcal{B}}, s), \\ 0 & \text{otherwise,} \end{cases}$$

for $\boldsymbol{\mathcal{B}} \in R^{d_1 \times d_2 \times d_3}$ and $s \leq d_1 d_2 d_3$. Here $\text{supp}(\boldsymbol{\mathcal{B}}, s)$ is the set of indices of $\boldsymbol{\mathcal{B}}$ corresponding to its largest s absolute values. We then develop an alternating gradient descent algorithm for (4) to iteratively update $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\mathcal{B}}$. We summarize the optimization procedure in Algorithm 1. In this algorithm, $\nabla_{\boldsymbol{U}} \tilde{\ell}(\boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{\mathcal{B}})$ denotes the gradient of the objective function $\tilde{\ell}(\boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{\mathcal{B}})$ with respect to $\boldsymbol{U}$, and $\nabla_{\boldsymbol{V}} \tilde{\ell}(\boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{\mathcal{B}})$, $\nabla_{\boldsymbol{\mathcal{B}}} \tilde{\ell}(\boldsymbol{U}\boldsymbol{V}^\top, \boldsymbol{\mathcal{B}})$ are defined similarly. Explicit forms of these gradients are given in the supplement. In Section 4, some theoretical conditions are placed on δ and τ to ensure the linear convergence rate of the algorithm, based on which we discuss their empirical choices.

Next, for the case that $\boldsymbol{\Theta}$ is low-rank and symmetric, we consider the factorization $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and the corresponding optimization problem,

$$\min_{\substack{\boldsymbol{U} \in \mathbb{R}^{n \times r}, \boldsymbol{\Lambda} \in \mathcal{D}_r \\ \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n \times n \times p}}} \ell(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \boldsymbol{\mathcal{B}}), \quad \text{subject to} \quad \|\boldsymbol{\mathcal{B}}\|_0 \leq s, \tag{5}$$

where $\mathcal{D}_r$ denotes the set of all $r \times r$ diagonal matrices with diagonal entry values $\{-1, 1\}$. The alternating gradient descent algorithm for (5) is summarized in Algorithm 2. In this algorithm, we have chosen not to update the estimate of $\boldsymbol{\Lambda}$. This is because we initialize by first solving the optimization problem (4), treating $\boldsymbol{\Theta}$ as a general matrix without the symmetry constraint. From solving (4), the obtained $[\widetilde{\boldsymbol{U}}; \widetilde{\boldsymbol{V}}]$ consistently estimates $[\boldsymbol{U}^*; \boldsymbol{\Lambda}\boldsymbol{U}^{*\top}]$, as we show in Proposition 1 in the supplement, where $\boldsymbol{\Theta}^* = \boldsymbol{U}^*\boldsymbol{\Lambda}\boldsymbol{U}^{*\top}$ is the true coefficient. As such, the diagonal entries of $\boldsymbol{\Lambda}$ can be accurately estimated using $\boldsymbol{\Lambda}_{ii} = \text{sign}(\widetilde{\boldsymbol{U}}_{.i}^\top \widetilde{\boldsymbol{V}}_{.i})$.

The rank r and the sparsity s in (4) and (5) are two tuning parameters. We select these parameters via the eBIC criterion (Chen and Chen, 2012) that was first developed for variable selection in the diverging dimension regime. As a heuristic criterion to balance model fitting and complexity, the eBIC function has been used in low-rank estimation problems and has been found to give a good performance (Srivastava et al., 2017; Cai et al., 2021). Specifically, among a set of working ranks and sparsity levels, we choose the

Algorithm 2 Optimization algorithm for (5)

Step 1: first solve (4) using Algorithm 1 and denote the output as $\widetilde{\boldsymbol{U}}, \widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{\mathcal{B}}}$. Set $\boldsymbol{\Lambda}_{ii} = \text{sign}(\widetilde{\boldsymbol{U}}_{.i}^\top \widetilde{\boldsymbol{V}}_{.i})$, $i = 1, \ldots, r$, $\boldsymbol{U}^{(0)} = (\widetilde{\boldsymbol{U}} + \boldsymbol{\Lambda}\widetilde{\boldsymbol{V}}^\top)/2$ and $\boldsymbol{\mathcal{B}}^{(0)} = \widetilde{\boldsymbol{\mathcal{B}}}$.

repeat

Step 2: update $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \delta \nabla_{\boldsymbol{U}} \ell \left\{ \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \boldsymbol{\mathcal{B}}^{(t)} \right\} \Big|_{\boldsymbol{U}=\boldsymbol{U}^{(t)}}$;

Step 3: update $\boldsymbol{\mathcal{B}}^{(t+1)} = \text{Truncate}\left[\boldsymbol{\mathcal{B}}^{(t)} - \tau \nabla_{\boldsymbol{\mathcal{B}}} \ell \left\{ \boldsymbol{U}^{(t+1)} \boldsymbol{\Lambda} \boldsymbol{U}^{(t+1)\top}, \boldsymbol{\mathcal{B}} \right\} \Big|_{\boldsymbol{\mathcal{B}}=\boldsymbol{\mathcal{B}}^{(t)}}, s \right]$.

until the objective function converges.

combination of (r, s) that minimizes

$$\text{eBIC} = 2N \times \ell\big(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\mathcal{B}}}\big) + \left[\log(n^2 N) + \log\left\{n^2(p+1)\right\}\right] \times (2nr + s)\,, \tag{6}$$

where ℓ is the loss function in (3), and $\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\mathcal{B}}}$ are the estimates of $\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}}$ under the working rank and sparsity level. The eBIC criterion for the symmetric case is computed similarly.

4 Theory

We first derive the non-asymptotic error bound of the actual estimator from our algorithm, then establish the community detection consistency and edge selection consistency. We focus on the symmetric case and leave the results for the asymmetric case to the supplement.

4.1 Non-asymptotic error bound

Suppose the parameter space for $\{\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}}\}$ is compact. Let $\boldsymbol{\Theta}^*$ denote the true coefficient matrix with rank r^* and $\boldsymbol{\mathcal{B}}^*$ the true coefficient tensor with s^* nonzero entries. Denote the nonzero singular values of $\boldsymbol{\Theta}^*$ as $\sigma_1^* \geq \ldots \geq \sigma_{r^*}^* > 0$. Write $\boldsymbol{\Theta}^* = \boldsymbol{U}^* \boldsymbol{\Lambda} \boldsymbol{U}^{*\top}$, where $\boldsymbol{U}^* \in \mathbb{R}^{n \times r^*}$ and $\boldsymbol{\Lambda}$ is a $r^* \times r^*$ diagonal matrix with diagonal entries in $\{-1, 1\}$, collecting signs of the eigenvalues of $\boldsymbol{\Theta}^*$. Let $\mathbb{B}_{\boldsymbol{\Theta}^*}(\kappa_1) \subset \mathbb{R}^{n \times n}$ and $\mathbb{B}_{\boldsymbol{\mathcal{B}}^*}(\kappa_2) \subset \mathbb{R}^{n \times n \times p}$ denote the Frobenius-norm ball around $\boldsymbol{\Theta}^*$ with radius $\kappa_1 > 0$ and around $\boldsymbol{\mathcal{B}}^*$ with radius $\kappa_2 > 0$, respectively. We next introduce several regularity conditions on the model.

(B1) The samples $\boldsymbol{x}_i$'s are i.i.d. from a zero-mean distribution with covariance $\boldsymbol{\Sigma}_x$ satisfying

$b_l \leq \lambda_{\min}(\boldsymbol{\Sigma}_x) \leq \lambda_{\max}(\boldsymbol{\Sigma}_x) \leq b_u$ for some positive constants b_l, b_u, where $\lambda_{\min}(\boldsymbol{\Sigma}_x)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_x)$ denote the smallest and largest eigenvalues of $\boldsymbol{\Sigma}_x$, respectively.

(B2) The covariates are bounded by some constant $M_x > 0$, i.e., $|\boldsymbol{x}_{is}| \leq M_x$.

(B3) Each element of $\boldsymbol{A}^{(i)}$ conditional on $\boldsymbol{x}_i$ follows an exponential family distribution with continuous $\psi''(\cdot)$. For $\boldsymbol{\Theta} \in \mathbb{B}_{\boldsymbol{\Theta}^*}(\sqrt{\sigma^*_{r^*}}/3)$ and $\boldsymbol{\mathcal{B}} \in \mathbb{B}_{\boldsymbol{\mathcal{B}}^*}(\sqrt{\sigma^*_{r^*}}/3)$, it holds that $\nu_0^{-1} \leq \psi''(\boldsymbol{\Theta}_{jj'} + \boldsymbol{x}_i^\top \boldsymbol{\mathcal{B}}_{jj'}) \leq \nu_0$, for any j and some large constant $\nu_0 > 0$.

(B4) For $\boldsymbol{\Theta} \in \mathbb{B}_{\boldsymbol{\Theta}^*}(\sqrt{\sigma^*_{r^*}}/3)$, $\boldsymbol{\mathcal{B}} \in \mathbb{B}_{\boldsymbol{\mathcal{B}}^*}(\sqrt{\sigma^*_{r^*}}/3)$, we have $|\mathbb{E}\langle \psi''(\boldsymbol{\eta}^{(i)}) \circ \boldsymbol{\Theta}, \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i \rangle| \leq \kappa_0 \|\boldsymbol{\Theta}\|_F \cdot \|\boldsymbol{\mathcal{B}}\|_F$, where $\circ$ denotes Hadamard product and $\kappa_0 = \sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_x)}/(18\nu_0)$.

Condition (B1) places a regularity condition on the design matrix and Condition (B2) is to bound the Hessian of the cumulant function in the neighborhood of $\boldsymbol{\mathcal{B}}^*$. These two conditions are commonly assumed in high-dimensional generalized linear models (Negahban et al., 2012). Condition (B3) is satisfied by most generalized linear models. In particular, the boundedness of $\psi''(\boldsymbol{\Theta}_{jj'} + \boldsymbol{x}_i^\top \boldsymbol{\mathcal{B}}_{jj'})$ is directly implied by Condition (B2) as well as the compactness of the parameter space for $\{\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}}\}$. Condition (B4) is to bound the Lipschitz gradient parameter. In the case of a linear model, (B4) is easily satisfied with $|\mathbb{E}\langle \psi''(\boldsymbol{\eta}^{(i)}) \circ \boldsymbol{\Theta}, \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i \rangle| = 0$, since $\psi''(\cdot)$ is a constant and $\boldsymbol{x}_i$ has mean zero. For a GLM, such as a logistic or multinomial model, $\psi''(\cdot)$ is not a constant, and (B4) requires the inner product of a sparse matrix and a low-rank matrix to be bounded. This is satisfied if the sparse entries are spread out so that $\boldsymbol{\mathcal{B}}$ is not exactly low-rank, and the low-rank matrix is not spiky so that $\boldsymbol{\Theta}$ is not sparse. Such a conditions has been commonly assumed in the matrix factorization literature; see, e.g., Zhang et al. (2018).

For any $\boldsymbol{U}$ and $\boldsymbol{\mathcal{B}}$, we define the distance

$$D\{\boldsymbol{U}, \boldsymbol{\mathcal{B}}\} = d^2(\boldsymbol{U}, \boldsymbol{U}^*) + \|\boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{B}}^*\|_F^2/\sigma_1^*, \quad \text{where } d(\boldsymbol{U}, \boldsymbol{U}^*) = \min_{\boldsymbol{\Gamma} \in \mathbb{Q}_{r^*}} \|\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{\Gamma}\|_F,$$

and $\mathbb{Q}_{r^*}$ denotes the set of $r^* \times r^*$ orthonormal matrices. The factor $1/\sigma_1^*$ in the distance metric comes from the difference between $\boldsymbol{\Theta}$ and $\boldsymbol{U}$, as it holds that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F^2 \leq c\sigma_1^* d^2(\boldsymbol{U}, \boldsymbol{U}^*)$ for a constant c (Zhang et al., 2018). The next theorem gives the non-asymptotic error bound of $\boldsymbol{U}^{(t)}$ and $\boldsymbol{\mathcal{B}}^{(t)}$ from Algorithm 2 at the tth iteration.

Theorem 1 *Assume (B1)-(B4) and define* $\mu_1 = \nu_0^{-1}$, $\mu_2 = \lambda_{\min}(\boldsymbol{\Sigma}_x)/(4\nu_0)$, $\alpha_1 = \nu_0$ *and* $\alpha_2 = 7\lambda_{\max}(\boldsymbol{\Sigma}_x)\nu_0/4$. *Let* c_1 *and* c_2 *be positive constants such that* $c_1 \leq \mu_1/(96\alpha_1^2)$, *and* $3c_1\alpha_2 \leq c_2 \leq \min\left\{1/3, \sqrt{\mu_1/(5\alpha_1)}\right\}$. *Let the step sizes* $\delta = c_1/\sigma_1^*$, $\tau = c_2/\alpha_2$, *and* $s = \gamma s^*$, *where* $\gamma \geq 1 + \{(3\alpha_2 + \mu_2 c_2)/(\mu_2 c_2)\}^2$. *When* $N \geq c_3(r^* n \log n + s^* \log n)$ *for some positive constants* c_3 *and* c_4, *for any initial estimator* $\left\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\right\}$ *satisfying* $D\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\} \leq c_2^2 \sigma_{r^*}^*$, *we have, with probability of at least* $1 - c_4/n$,

$$D\left\{\boldsymbol{U}^{(t)}, \boldsymbol{\mathcal{B}}^{(t)}\right\} \leq \rho^t D\left\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\right\} + \phi_1 \frac{r^* n \log n}{N} + \phi_2 \frac{s^* \log n}{N}, \tag{7}$$

where $\rho = \max\{1 - \delta\mu_1\sigma_{r^*}^*/16, 1 - \tau\mu_2/18\} \in (0,1)$ *is a contraction parameter, and* ϕ_1 *and* ϕ_2 *are positive constants that depend on* c_1, c_2, ν_0, $\lambda_{\min}(\boldsymbol{\Sigma}_x)$ *and* $\lambda_{\max}(\boldsymbol{\Sigma}_x)$.

Theorem 1 portrays the estimation error at each iteration. The error bound consists of two terms that correspond to the computational error and the statistical error, respectively. It reveals an interesting interplay between the computational efficiency and the statistical rate of convergence. Note that the computational error decays geometrically with the iteration number t, whereas the statistical error remains the same. Therefore, as the iteration number increases, the computational error is to be dominated by the statistical error and the resulting estimator falls within the statistical precision of the true parameter. Theorem 1 also offers some useful guidance on the choice of the step sizes δ and τ. Their bounds hinge on ν_0, $\lambda_{\min}(\boldsymbol{\Sigma}_x)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_x)$. Specifically, we suggest to estimate $\boldsymbol{\Sigma}_x$ using the sample covariance of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, and estimate ν_0 by identifying a ν_0 such that $\nu_0^{-1} \leq \psi''(\boldsymbol{\Theta}_{jj'}^{(0)} + \boldsymbol{x}_i^\top \boldsymbol{\mathcal{B}}_{jj'}^{(0)}) \leq \nu_0$, where $\boldsymbol{\Theta}^{(0)}$ and $\boldsymbol{\mathcal{B}}^{(0)}$ are initial estimates. As the true intercept matrix is unknown, we suggest to upper bound σ_1^* with $\|\boldsymbol{\Theta}^{(0)}\|_F$. Meanwhile, in practice, we often find the estimates of the above step sizes from the data very small, leading to a slow convergence of the algorithm. We thus further suggest to multiply a large constant, e.g., 1000, with those calculated step sizes. We can dynamically vary this constant depending on the convergence behavior of the algorithm. We also investigate in Section S4.3 of the supplement the backtracking line search for step size, which yields a similar performance.

We make some additional remarks on the computational error, statistical error and the initial condition. First, the computational error $\rho^t D\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\}$ directly relies on the

contraction parameter ρ, in that a smaller value of ρ leads to a faster convergence. Given the conditions on the steps sizes δ and τ and the definitions of μ_1 and μ_2, some direct calculation shows that ρ is ensured to be positive. When the step sizes δ and τ increase, ρ decreases. Second, the term $r^* n \log n/N$ is the statistical error from the low-rank matrix estimation, which, up to a logarithmic factor, matches with the minimax rate for multi-response regression with a low-rank constraint (Raskutti et al., 2011), and the term $s^* \log n/N$ is the statistical error from the sparse tensor estimation, which matches with the minimax rate in sparse regressions (Negahban et al., 2012). While the above minimax rates are established under different settings compared to ours, we expect the error rate in Theorem 1 to be sharp up to a logarithmic factor. Third, in Theorem 1, we require the initialization error to be bounded. Such an assumption is often needed in non-convex optimizations (Zhang and Xia, 2018). In Algorithm 1, we initialize with the truncated singular value decomposition for the low-rank component and a zero tensor for the sparse component, which we have found to enjoy a good empirical performance. Meanwhile, it is useful to devise an initialization procedure that can ensure the conditions in Theorem 1. We leave its investigation as future research.

4.2 Consistency of community detection and edge selection

One implication of our model is that we may recover the community structure of the nodes given the low-rank parameterization of $\boldsymbol{\Theta}$. We show that our solution can correctly recover the true community labels for all nodes with probability $1 - O(K/n)$, while allowing the number of communities K to grow sub-linearly with the number of nodes n.

We first formally define the true underlying community structure. Based on $\boldsymbol{U}^*$ from the decomposition $\boldsymbol{\Theta}^* = \boldsymbol{U}^* \boldsymbol{\Lambda} \boldsymbol{U}^{*\top}$, the true community structure is determined by the rows of $\boldsymbol{U}^*$ in that there are K distinct groups of rows, such that

$$\boldsymbol{U}^* = (\boldsymbol{U}^*_{1\cdot}, \ldots, \boldsymbol{U}^*_{n\cdot})^\top = \Big(\underbrace{\boldsymbol{u}^*_1, \ldots, \boldsymbol{u}^*_1}_{l \text{ nodes}}, \underbrace{\boldsymbol{u}^*_2, \ldots, \boldsymbol{u}^*_2}_{l \text{ nodes}}, \ \ldots, \ \underbrace{\boldsymbol{u}^*_K, \ldots, \boldsymbol{u}^*_K}_{l \text{ nodes}} \Big)^\top \in \mathbb{R}^{n \times r^*},$$

where $\boldsymbol{u}^*_k \in \mathbb{R}^{1 \times r^*}$, $k = 1, \ldots, K$. Here for notational simplicity, we assume there is an

equal number of nodes, $l = n/K$, in each community. Accordingly, we define the true community assignments as $\mathcal{A}_1^* := \{1, \ldots, l\}, \ldots, \mathcal{A}_K^* := \{n - l, \ldots, n\}$.

We propose to recover community labels by applying a distance-based clustering procedure, such as K-means to rows of the final estimate $\boldsymbol{U}^{(t)}$ obtained from Algorithm 2. We show that the resulting clustering output achieves strong consistency, under the following regularity conditions.

(C1) Assume that $\sigma_{r^*}^* > c_5$ for some constant $c_5 > 0$, where $\sigma_{r^*}^*$ is the smallest non-zero singular value of $\boldsymbol{\Theta}^*$.

(C2) Assume that $\min_{k \neq k'} \|\boldsymbol{u}_k^* - \boldsymbol{u}_{k'}^*\|_2^2 > c_6 e_0$ for some constant $c_6 > 0$, where $e_0 = \phi_1 r^* n \log n/N + \phi_2 s^* \log n/N$, and ϕ_1, ϕ_2 are defined as in Theorem 1.

Condition (C1) requires that the minimum non-zero singular value of $\boldsymbol{\Theta}^*$ is bounded below by a positive constant. Condition (C2) ensures that the minimal gap between different cluster centers does not tend to zero too fast.

Theorem 2 *Suppose the conditions in Theorem 1 and (C1)-(C2) hold. Then after t iterations, with $t \geq \log_\rho \left(e_0 / D\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\} \right)$, we have, with probability of at least $1 - c_4 K/n$, $\widehat{\mathcal{A}}_k^{(t)} = \mathcal{A}_k^*$, for all $k = 1, \ldots, K$, where c_4 is the constant defined as in Theorem 1.*

Theorem 2 shows that our community detection procedure achieves the strong consistency as long as $K = o(n)$. Note that $\hat{\boldsymbol{\Theta}}$ is estimated after the covariate effects have been removed from the connectivity matrix. Existing spectral clustering methods, either for a single network or for multiple networks, cannot handle heterogeneity due to the network-level covariates. Our result allows K to grow at a sub-linear rate with n, which is achievable as we have N network samples, which provides more information than a single sample.

Another property of our estimator is that we can select the edges that are affected by the covariates consistently.

Corollary 1 *Assume all conditions in Theorem 1 hold and $\min_{ijk} |\boldsymbol{\mathcal{B}}_{ijk}^*| > 2\sqrt{\sigma_1^* e_0}$. Then after t iterations with $t \geq \log_\rho \left(e_0 / D\{\boldsymbol{U}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\} \right)$, we have, with probability of at least*

$1 - c_4/n$, *for any* $\boldsymbol{\mathcal{B}}^*_{ijk} \neq 0$, *the estimate* $\boldsymbol{\mathcal{B}}^{(t)}_{ijk} \neq 0$, *and for any* $\boldsymbol{\mathcal{B}}^*_{ijk} = 0$, *the estimate* $\boldsymbol{\mathcal{B}}^{(t)}_{ijk} = 0$.

Corollary 1 is a direct consequence of Theorem 1, and thus we omit its proof. This result has an important implication in practice, as it ensures that our model can correctly select the edges that are affected by the subject covariates. The condition on $\min_{ijk} |\boldsymbol{\mathcal{B}}^*_{ijk}|$ is a minimal signal condition, which is commonly employed to establish selection consistency (Kong et al., 2019). It allows the minimal signal to tend to zero as the sample size N increases. Nevertheless, its optimality remains unclear, and we leave the search for the optimal minimal signal condition as future research.

5 Simulations

We carry out simulations to investigate the finite-sample performance of our proposed method, and to compare with some competing solutions. We focus on symmetric matrices throughout the simulations. We first consider our proposed model (2), then the CISE model of Wang et al. (2019) where our model structure is not satisfied. We further consider a stochastic blockmodel (Holland et al., 1983), and a latent factor model (Minhas et al., 2016), and we report the results in the supplement. We have found our method performs competitively in all settings, even under potential model misspecification. In all simulations, we tune the rank r and sparsity s using the eBIC criterion.

5.1 Generalized matrix response model

We first simulate the connectivity matrix with binary edges from (2), $g\{\boldsymbol{\mu}^{(i)}\} = \boldsymbol{\Theta} + \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i$, where $g(\cdot)$ is the logit link function. We generate the covariates from $\mathcal{N}(0, 1)$ and standardize the columns of the design matrix to have zero mean and unit standard deviation. For $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we set $\boldsymbol{\Lambda}$ as an $r \times r$ identity matrix, and generate the entries of $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ from $\mathcal{N}(0, 1)$. For $\boldsymbol{\mathcal{B}}$, we randomly set a proportion of its entries to be 2, and the rest to zero; let $s_0 = s/(n^2 p)$ denote this proportion of the nonzero entries. We set the

number of nodes $n = 50$, the number of covariates $p = 10$, and vary the number of subjects $N = 200, 400$, the rank $r = 2, 5$, and the sparsity proportion $s_0 = 0.1, 0.3$, respectively.

We compare with three alternative methods. The first is the element-wise penalized GLM method of Firth (1993), which fits a penalized GLM to each entry of $\boldsymbol{A}_{jj'}$, for all j, j'. This approach has been shown to be effective in reducing the small sample bias (Firth, 1993). The second method is similar to the first one, except that it uses an elastic-net penalty (Zou and Hastie, 2005). The third is the common and individual structure explained method proposed by Wang et al. (2019) coupled with a GLM, and the tuning is done using the elbow method as described in Wang et al. (2019).

To evaluate the estimation accuracy, we report the estimation errors, $N^{-1}\sum_{i=1}^{N}\|\boldsymbol{\mu}^{(i)} - \hat{\boldsymbol{\mu}}^{(i)}\|_F$, $\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_F$, and $\|\boldsymbol{\mathcal{B}} - \hat{\boldsymbol{\mathcal{B}}}\|_F$, where $\hat{\boldsymbol{\mu}}^{(i)} = g^{-1}(\hat{\boldsymbol{\Theta}} + \hat{\boldsymbol{\mathcal{B}}} \times_3 \boldsymbol{x}_i)$. To evaluate the edge selection accuracy, we report the F1 score, which is calculated as 2TP/(2TP+FP+FN), where TP is the true positive count, FP is the false positive count, and FN is the false negative count. Since the method of Firth (1993) does not consider entry-wise sparsity, its F1 score is not reported. Since the method of Wang et al. (2019) could only estimate $\boldsymbol{\mu}^{(i)}$, the estimation errors for $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$ are not reported. Table 1 reports the average criteria, with the standard errors in the parentheses, over 50 data replications. The four methods under comparison are: the element-wise penalized GLM with the Jeffreys invariant prior penalty (denoted as GLM_{JP}), the element-wise penalized GLM with the elastic-net penalty (GLM_{EN}), the common and individual structure explained method (CISE), and the proposed generalized connectivity matrix response model (GLSNet). Our proposed method is seen to achieve the best performance among all competing methods, in terms of both estimation accuracy and selection accuracy, and this holds true for different sample sizes N, ranks r and sparsity levels s_0. Moreover, we see the estimation error of our method decreases as N increases, or as r and s_0 decrease. Such observations agree with our theoretical results in Theorem 1. We further report the heat map of the eBIC over varying r and s_0 values in Section S4.2 and results with a larger network size n in Section S4.4 of the supplement.

Finally, we report the computation time of the proposed method with varying sample

Table 1: Simulation results under the low-rank and sparse model, with the varying sample size N, rank r and sparsity proportion s_0. Marked in boldface are those achieving the best evaluation criteria in each setting.

N	r	s_0	Method	Error of $\boldsymbol{\mu}^{(i)}$	Error of $\boldsymbol{\Theta}$	Error of $\boldsymbol{\mathcal{B}}$	F1 score
200	2	0.1	GLM_{JP}	1.106 (0.009)	47.09 (1.531)	35.09 (0.389)	-
			GLM_{EN}	1.063 (0.011)	47.50 (1.865)	28.38 (0.201)	0.709 (0.002)
			CISE	0.638 (0.001)	-	-	-
			GLSNet	**0.152 (0.002)**	**3.490 (0.129)**	**25.79 (0.426)**	**0.964 (0.002)**
		0.3	GLM_{JP}	1.101 (0.008)	45.06 (1.454)	52.53 (0.696)	-
			GLM_{EN}	1.062 (0.008)	45.61 (1.561)	46.73 (0.345)	0.905 (0.001)
			CISE	0.818 (0.001)	-	-	-
			GLSNet	**0.207 (0.002)**	**4.15 (0.175)**	**35.35 (0.415)**	**0.994 (0.001)**
	5	0.1	GLM_{JP}	1.353 (0.005)	93.38 (1.604)	35.55 (0..486)	-
			GLM_{EN}	1.328 (0.006)	94.64 (1.599)	29.40 (0.215)	0.736 (0.002)
			CISE	0.631 (0.001)	-	-	-
			GLSNet	**0.154 (0.001)**	**6.51 (0.232)**	**26.86 (0.346)**	**0.960 (0.002)**
		0.3	GLM_{JP}	1.311 (0.005)	88.39 (1.648)	52.30 (0.613)	-
			GLM_{EN}	1.287 (0.005)	87.92 (1.625)	47.63 (0.295)	0.916 (0.001)
			CISE	0.838 (0.001)	-	-	-
			GLSNet	**0.211 (0.001)**	**9.79 (0.488)**	**37.27 (0.337)**	**0.981 (0.001)**
400	2	0.1	GLM_{JP}	0.774 (0.007)	39.04 (1.892)	25.05 (0.132)	-
			GLM_{EN}	0.756 (0.007)	40.58 (1.815)	18.35 (0.101)	0.700 (0.002)
			CISE	0.457 (0.001)	-	-	-
			GLSNet	**0.055 (0.000)**	**2.44 (0.110)**	**13.61 (0.185)**	**0.997 (0.000)**
		0.3	GLM_{JP}	0.769 (0.006)	36.47 (1.388)	33.27 (0.130)	-
			GLM_{EN}	0.752 (0.006)	38.48 (1.135)	30.47 (0.117)	0.901 (0.001)
			CISE	0.577 (0.001)	-	-	-
			GLSNet	**0.088 (0.000)**	**3.04 (0.131)**	**22.51 (0.151)**	**0.998 (0.000)**
	5	0.1	GLM_{JP}	0.974 (0.004)	81.80 (1.923)	26.82 (0.110)	-
			GLM_{EN}	0.964 (0.004)	82.52 (1.893)	18.86 (0.087)	0.716 (0.002)
			CISE	0.452 (0.001)	-	-	-
			GLSNet	**0.061 (0.000)**	**4.57 (0.161)**	**14.97 (0.234)**	**0.993 (0.001)**
		0.3	GLM_{JP}	0.939 (0.004)	77.18 (1.466)	34.65 (0.156)	-
			GLM_{EN}	0.927 (0.004)	77.94 (1.435)	31.13 (0.136)	0.908 (0.001)
			CISE	0.513 (0.001)	-	-	-
			GLSNet	**0.094 (0.001)**	**7.87 (0.274)**	**25.47 (0.273)**	**0.995 (0.000)**

size N, network size n and covariate dimension p, while we adopt the simulation setting with rank $r = 2$ and sparsity level $s_0 = 0.1$. All simulations were run on an iMac with a 3.6 GHz

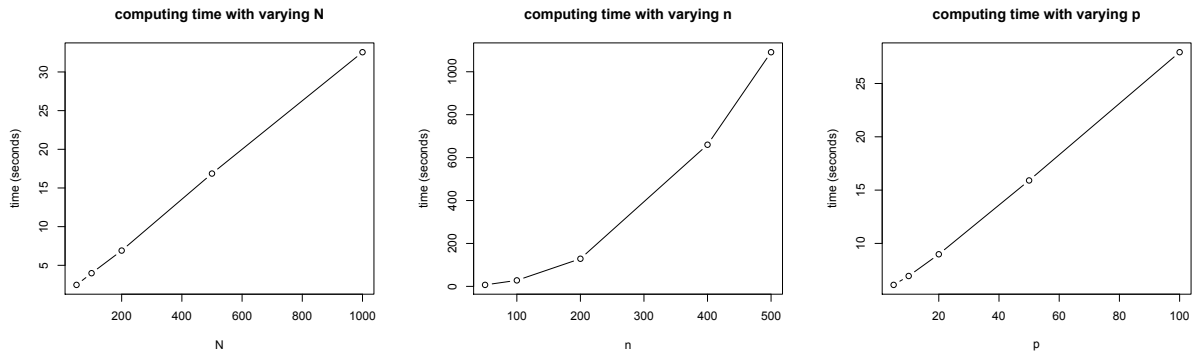


Figure 1: Average computing time over 50 data replications with varying sample size N, network size n and covariate dimension p.

Intel Core i9 processor. Figure 1 shows the average computing time over 50 replications when $n = 50$, $p = 10$ with varying N (left panel), $N = 200$, $p = 10$ with varying n (middle), and $N = 200$, $n = 50$ with varying p (right), when the working rank and sparsity level are set at the truth. We observe that the computing time is approximately linear with respect to N and p, and quadratic to n. When n is large, the number of parameters involved in the computation is large too. For instance, there are 1,375,000 parameters involved when $n = 500$ and $p = 10$, and correspondingly, the computing time is longer.

5.2 Common and individual structure explained model

Next we consider the performance of our method under a potentially misspecified model, and compare with the individual structure explained method of Wang et al. (2019). The CISE model assumes the entries in $\boldsymbol{A}^{(i)}$ are independent Bernoulli random variables with

$$\text{logit}\{\boldsymbol{\mu}^{(i)}\} = \boldsymbol{\Theta} + \boldsymbol{D}_i, \quad i = 1, \ldots, N, \tag{8}$$

where $\boldsymbol{\mu}^{(i)}$ is as defined in (2), $\boldsymbol{\Theta}$ characterizes the common connectivity pattern, and $\boldsymbol{D}_i$ represents the subject-specific deviation; the subject-specific deviation $\boldsymbol{D}_i$ is assumed to be low-rank while no structure assumption is placed on $\boldsymbol{\Theta}$. For $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we set $\boldsymbol{\Lambda} = \boldsymbol{I}_{r\times r}$, and generate the entries of $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ from $\mathcal{N}(0, 1)$. We set $\boldsymbol{D}_i = \boldsymbol{d}_i \otimes \boldsymbol{d}_i$, where $\otimes$ is outer product, and generate the entries of $\boldsymbol{d}_i \in \mathbb{R}^n$ from $\mathcal{N}(0, 1)$.

We simulate binary networks from the CISE model in (8) with details given in Table

Table 2: Simulation results under the common and individual structure explained model, with the varying sample size N and rank r. Marked in boldface are those achieving the best evaluation criteria in each setting.

r	Method	$N = 200$		$N = 400$	
		Error of $\boldsymbol{\mu}^{(i)}$	Error of $\boldsymbol{\Theta}$	Error of $\boldsymbol{\mu}^{(i)}$	Error of $\boldsymbol{\Theta}$
5	CISE	**0.435 (0.000)**	46.08 (0.553)	**0.301 (0.000)**	44.74 (0.651)
	GLSNet	0.506 (0.002)	**16.50 (0.256)**	0.359 (0.002)	**16.27 (0.289)**
20	CISE	0.306 (0.000)	157.2 (1.130)	0.440 (0.000)	155.8 (1.195)
	GLSNet	**0.294 (0.001)**	**104.8 (1.367)**	**0.423 (0.002)**	**100.6 (1.463)**

2. The two methods under comparison are: the common and individual structure explained method (CISE) and the proposed generalized connectivity matrix response model (GLSNet). The CISE model cannot incorporate subject covariates, and hence $\boldsymbol{\mathcal{B}}$ is not included. Moreover, $\boldsymbol{\Theta} + \boldsymbol{D}_i$ is not necessarily low-rank. As such, our model assumption may not be satisfied. We set $n = 50$, $N = 200, 400$, and $r = 5, 20$. Table 2 reports the estimation errors based on 50 data replications for the CISE method and our proposed method. It is seen that, under this potentially misspecified model, our method still achieves a comparable performance as Wang et al. (2019).

6 Applications to Brain Connectivity Analysis

We apply the proposed method to two brain connectivity studies. The first is a study of brain functional connectivity based on resting-state fMRI, where the edge is *binary* resulting from a thresholded partial correlation matrix. The second is a study of brain structural connectivity based on DTI, where the edge is the *count* of white matter fibers between pairs of brain regions.

6.1 Functional connectivity analysis

We first analyze an fMRI dataset from ADHD-200 (`http://neurobureau.projects.nitrc.org/ADHD200/Data.html`). We focus on $N = 319$ healthy control subjects, aging between

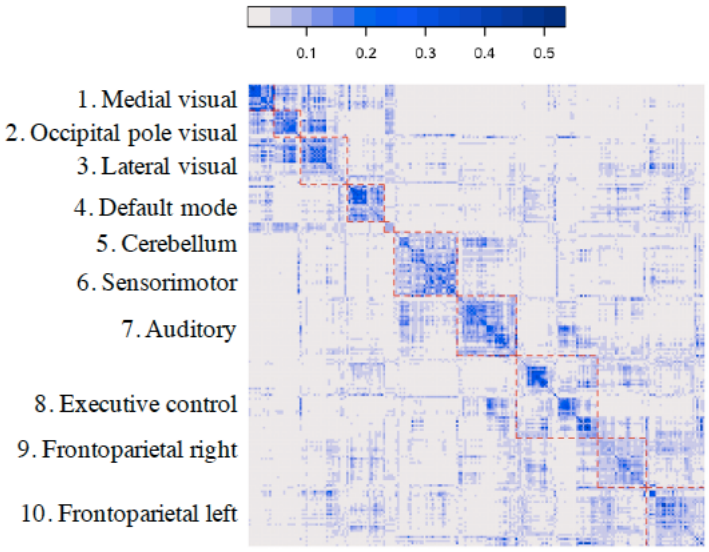


Figure 2: The functional connectivity study. Heatmap of the 264×264 matrix $g^{-1}(\widehat{\boldsymbol{\Theta}})$ with rows and columns ordered according to the pre-specified functional module membership. The red dashed lines mark the boundaries of the ten functional modules.

7.09 to 21.8 years old, with 46.4% females and 53.6% males. Each subject received a resting-state fMRI scan, and the image was preprocessed following the usual Athena pipeline, including slice timing correction, motion correction, spatial smoothing, denoising by regressing out motion parameters, white matter, and cerebrospinal fluid time, and band-pass filtering. Each fMRI image was then summarized in the form of a binary network, with the nodes corresponding to 264 seed regions of interest in the brain Power et al. (2011), and the edges recording the binary indicator of the thresholded partial correlations. We apply our proposed model to this data with a logit link function. We standardize the covariates, age and sex, to have mean zero and variance one. The rank is selected as $r = 9$ and the sparsity proportion as $s_0 = 0.02$ based on eBIC.

We first examine the estimate $\hat{\boldsymbol{\Theta}}$. In the neuroscience literature, those 264 nodes have been partitioned into 10 functional modules (Smith et al., 2009). Each module possesses a relatively autonomous functionality, and complex brain tasks are carried out through coordinated collaborations among those modules. Figure 2 shows the heatmap of $g^{-1}(\hat{\boldsymbol{\Theta}})$, with the nodes ordered according to the functional modules. Here the function $g^{-1}(\cdot)$ maps

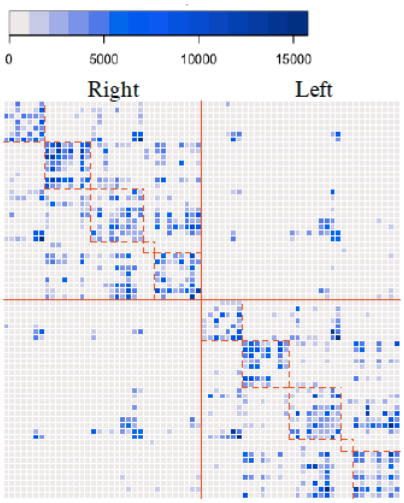
a value from the real line to $[0, 1]$ so to facilitate data visualization. From this figure, we see that our estimate agrees reasonably well with the pre-specified functional modules by Smith et al. (2009). We observe larger values of $\hat{\boldsymbol{\Theta}}$ located within the diagonal blocks, which indicates higher functional connectivities within those functional modules. Furthermore, there are high connectivities among modules 1-3, namely, the medial visual, occipital pole visual and lateral visual modules. These visual modules appear to have high connectivities with the cerebellum, but generally low connectivities with the rest of functional modules. We observe a high connectivity between modules 9-10, namely, the frontoparietal right and frontoparietal left modules. These two modules are important in attention control and can generate a diverse range of control signals depending on task demands (Scolari et al., 2015).

We next examine the estimate $\hat{\boldsymbol{\mathcal{B}}}$. In $\hat{\boldsymbol{\mathcal{B}}}_{\cdot\cdot 1}$, i.e., the coefficient matrix for the sex covariate. The non-sparse entries are located within the lateral visual module, and those values are negative, ranging from -0.777 to -0.506. This indicates that male subjects have lower connectivities in those regions within the lateral visual module. This result agrees with the existing finding that developing females outperform developing males on tasks related to emotion identification and reasoning (Satterthwaite et al., 2014). The non-sparse entries of $\hat{\boldsymbol{\mathcal{B}}}$ mostly concentrate in $\hat{\boldsymbol{\mathcal{B}}}_{\cdot\cdot 2}$, i.e., the coefficient matrix for the age covariate. In $\hat{\boldsymbol{\mathcal{B}}}_{\cdot\cdot 2}$, the positive entries are located within the occipital pole visual, default mode, executive control and frontoparietal left modules, with values ranging from 0.454 to 0.902, indicating the connectivities within those modules increase with age. We also observe positive entries located in the default model to executive control and the default mode to frontoparietal right, which agrees with the literature that the default mode module has increasingly synchronized connections to other modules with increasing age (Grayson and Fair, 2017). We also find negative entries located in the medial visual to lateral visual, the executive control to frontoparietal right, and the default mode to auditory modules, with values ranging from -1.350 to -1.095, suggesting the connectivities between those modules also decrease with age. These findings suggest some interesting patterns that warrant further investigation and validation.

6.2 Structural connectivity analysis

We next analyze a structural DTI dataset from KKI-42 (`http://mrneurodata.s3-website-us-east-1.amazonaws.com/KKI2009/ndmg_0-0-48/graphs/desikan/`). We focus on 21 subjects with no history of neurological conditions, aging from 22 to 61 years old, with 47.6% females and 52.4% males. Each subject received a resting-state DTI scan, which is a magnetic resonance imaging technique that enables measurement of the diffusion of water. Estimates of white matter connectivity patterns can be obtained using the diffusion anisotropy and the principal diffusion directions. In the KKI-42 study, a scan-rescan imaging session was conducted on each subject, leading to two images for each subject, and a total of $N = 42$ for the study. For simplicity, we treat those images as if they formed independent samples. Each DTI image was preprocessed, and summarized in the form of a count network, with $n = 68$ nodes defined following the Desikan Atlas, and the edges recording the total number of white matter fibers between the pair of nodes. See Landman et al. (2011) for more information about data collection and brain networks construction using DTI scans. We apply our proposed method to this data, with a log link function. We standardize the covariates, age and sex, to have mean zero and variance one. The rank is selected as $r = 5$ and the sparsity proportion as $s_0 = 0.31$ based on eBIC.

We first examine the estimate $\hat{\boldsymbol{\Theta}}$. To the best of our knowledge, communities in structural connectivity networks have not been studied before. We applied the K-means clustering algorithm to the estimate $\boldsymbol{U}^{(t)}$ from $\hat{\boldsymbol{\Theta}}$, and identified five clusters among the 68 anatomic regions of interest (ROIs). We selected the number of clusters based on the elbow plot. Figure 3, right panel, reports the members of each cluster in the table. From an anatomical perspective, the first group of nodes are entirely contained in the frontal lobe, the second group are mostly contained in the temporal lobe, the fourth group are entirely contained in the temporal lobe, and the third and fifth groups contain nodes from the frontal, parietal, occipital and temporal lobes. Many of the 68 anatomic ROIs in the Desikan Atlas overlap with the resting-state functional modules. By exploring this overlap, we gained further insights of potential functions of those five groups. We found that group 1



1	frontalpole, parsopercularis, lateralorbitalfrontal, parstriangularis, medialorbitalfrontal, rostralanteriorcingulate, rostralmiddlefrontal
2	bankssts, fusiform, inferiortemporal, lingual, middletemporal, parsorbitalis, superiortemporal, transversetemporal
3	caudalanteriorcingulate, caudalmiddlefrontal, parahippocampal, posteriorcingulate, precentral, isthmuscingulate, precuneus, corpuscallosum, superiorfrontal
4	entorhinal, temporalpole
5	inferiorparietal, lateraloccipital, paracentral, cuneus, pericalcarine, postcentral, superiorparietal, supramarginal

Figure 3: The structural connectivity study. Left panel: heatmap of the 68×68 matrix $g^{-1}(\widehat{\boldsymbol{\Theta}})$ with rows and columns ordered according to the K-means clustering result. Right and left hemispheres are marked in the plot. The red dashed lines mark the boundaries of the identified groups. Right panel: the anatomic regions of interest in the identified groups.

is related to the dorsal attention and default mode modules, group 2 is related to the visual and auditory, group 3 is related to the default mode, and group 5 is related to the visual module. The resting-state functions of the nodes in Groups 4 are unidentified. Figure 3, left panel, shows the heatmap of the estimated $\hat{\boldsymbol{\Theta}}$, with the nodes reordered according to the cluster membership.

We next examine the estimate $\hat{\boldsymbol{\mathcal{B}}}$. Figure 4 shows the estimated subject covariates effect coefficients. From the left panel of Figure 4, we see that, as age increases, the structural connectivity generally decreases both within and between the two hemispheres. This result agrees with existing neurological finding (Betzel et al., 2014). From the right panel of Figure 4, we see that male and female subjects have different structural connectivity patterns. Such differences are observed in the between-group connections within and between hemispheres, and in the within-group connections within each hemisphere. For instance, we see males have lower between-hemisphere connectivities for the ROIs in Group 1. This observation agrees with the literature that males have lower connectivities between the left and right frontal regions (Ingalhalikar et al., 2014).

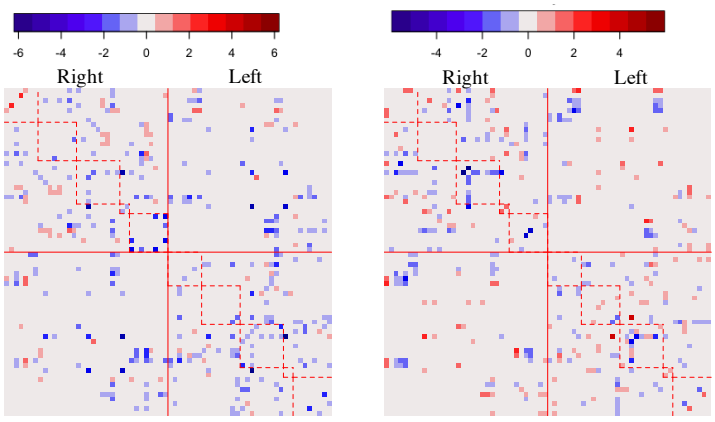


Figure 4: The structural connectivity study. Left panel: the age coefficient matrix. Right panel: the sex coefficient matrix.

7 Discussion

In this article, we propose a generalized connectivity matrix response regression model that relates subject-specific connectivity matrix to external covariates. We briefly comment on potential future research.

One direction is to incorporate random effect into our model formulation, by considering

$$g\left\{\boldsymbol{\mu}^{(i)}\right\} = \boldsymbol{\Theta} + \boldsymbol{\mathcal{B}} \times_3 \boldsymbol{x}_i + \boldsymbol{R}^{(i)}, \quad i = 1, \ldots, N,$$

where $g(\cdot)$, $\boldsymbol{\Theta}$ and $\boldsymbol{\mathcal{B}}$ are as defined in (2), and $\boldsymbol{R}^{(i)} \in \mathbb{R}^{n \times n}$ is the subject-specific random effect matrix. In this model, the edges for the same subject, i.e., the edges in $\boldsymbol{A}^{(i)}$, are correlated. To improve estimability and interpretability, we may assume that $\boldsymbol{R}^{(i)}$ follows a low-dimensional structure. For instance, we may impose that $\boldsymbol{R}^{(i)} = \boldsymbol{v}^{(i)} \circ \boldsymbol{v}^{(i)}$, where $\boldsymbol{v}^{(i)}$ follows a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma}_0$ that needs to be estimated. Model estimation in this case is expected to be more challenging, as it requires integration over the random vector $\boldsymbol{v}^{(i)} \in \mathbb{R}^n$.

Another direction is to consider alternative community structure specifications. In our current community detection setup, we assume that the communities are fully determined by the population level connectivity matrix $\boldsymbol{\Theta}$, which is closely related to the stochastic

blockmodel. Meanwhile, other specifications of the community structure are possible. For instance, the slope tensor $\boldsymbol{\mathcal{B}}$ may have a community structure, in that the covariate effect on the connectivity between nodes j and j' is determined by the community labels of those two nodes. In this case, the coefficient matrix for the lth covariate, i.e., $\boldsymbol{\mathcal{B}}_{\cdot\cdot l}$, becomes a block matrix. This new structure requires a new set of estimation algorithm and theory.

Acknowledgments

We are very grateful to two anonymous referees, an associate editor, and the Editor for their valuable comments that have greatly improved the manuscript. Zhang's research was partially supported by NSF grant DMS-2015190. Sun's research was partially supported by ONR grant N00014-18-1-2759. Li's research was partially supported by NSF grant CIF-2102227, and NIH grant R01AG061303.

References

Betzel, R. F., Byrge, L., He, Y., Goñi, J., Zuo, X.-N., and Sporns, O. (2014), "Changes in structural and functional connectivity among resting-state networks across the human lifespan," *Neuroimage*, 102, 345–357.

Bi, X., Qu, A., and Shen, X. (2018), "Multilayer tensor factorization with applications to recommender systems," *The Annals of Statistics*, 46, 3308–3333.

Brands, R. A. (2013), "Cognitive social structures in social network research: A review," *Journal of Organizational Behavior*, 34, S82–S103.

Burer, S. and Monteiro, R. D. (2003), "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, 95, 329–357.

Cai, B., Zhang, J., and Sun, W. W. (2021), "Jointly Modeling and Clustering Tensors in High Dimensions," *arXiv preprint arXiv:2104.07773*.

Chen, J. and Chen, Z. (2012), "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*, 555–574.

Chen, S., Kang, J., Xing, Y., and Wang, G. (2015), "A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks," *Human Brain Mapping*, 36, 5196–5206.

Dai, H., Li, L., Zeng, T., and Chen, L. (2019), "Cell-specific network constructed by single-cell RNA sequencing data," *Nucleic acids research*.

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2017), "Nonparametric Bayes modeling of populations of networks," *Journal of the American Statistical Association*, 112, 1516–1530.

Firth, D. (1993), "Bias reduction of maximum likelihood estimates," *Biometrika*, 80, 27–38.

Grayson, D. S. and Fair, D. A. (2017), "Development of large-scale functional networks from birth to adulthood: A guide to the neuroimaging literature," *NeuroImage*, 160, 15–31.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic blockmodels: First steps," *Social networks*, 5, 109–137.

Hu, W., Pan, T., Kong, D., and Shen, W. (2020), "Nonparametric matrix response regression with application to brain imaging data analysis," *Biometrics*.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson, H., Gur, R. E., Gur, R. C., and Verma, R. (2014), "Sex differences in the structural connectome of the human brain," *Proceedings of the National Academy of Sciences*, 111, 823–828.

Kang, J., Bowman, F. D., Mayberg, H., and Liu, H. (2016), "A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs," *NeuroImage*, 141, 431–441.

Kong, D., An, B., Zhang, J., and Zhu, H. (2019), "L2RM: Low-rank Linear Regression Models for High-dimensional Matrix Responses," *Journal of the American Statistical Association*, 1–47.

Kundu, S., Ming, J., Pierce, J., McDowell, J., and Guo, Y. (2018), "Estimating dynamic brain functional networks using multi-subject fMRI data," *NeuroImage*, 183, 635–649.

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., et al. (2011), "Multi-parametric neuroimaging reproducibility: a 3-T resource study," *Neuroimage*, 54, 2854–2866.

Li, L. and Zhang, X. (2017), "Parsimonious tensor response regression," *Journal of the American Statistical Association*, 112, 1131–1146.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 37, CRC press.

Minhas, S., Hoff, P. D., and Ward, M. D. (2016), "Inferential Approaches for Network Analyses: AMEN for Latent Factor Models," *arXiv preprint arXiv:1611.00460*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27, 538–557.

Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. (2018), "Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably," *SIAM Journal on Imaging Sciences*, 11, 2165–2204.

Power, J. D., Cohen, A., Nelson, S., Wig, G. S., Barnes, K., Church, J., Vogel, A., Laumann, T., Miezin, F., Schlaggar, B., and Petersen, S. (2011), "Functional Network Organization of the Human Brain," *Neuron*, 72, 665–78.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE transactions on information theory*, 57, 6976–6994.

Satterthwaite, T. D., Wolf, D. H., Roalf, D. R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E. D., Elliott, M. A., Smith, A., Hakonarson, H., et al. (2014), "Linked sex differences in cognition and functional connectivity in youth," *Cerebral cortex*, 25, 2383–2394.

Scolari, M., Seidl-Rathkopf, K. N., and Kastner, S. (2015), "Functions of the human frontoparietal attention network: Evidence from neuroimaging," *Current opinion in behavioral sciences*, 1, 32–39.

Smith, S. D., Fox, P. T., Miller, K., Glahn, D., Fox, P., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A., and Beckmann, C. F. (2009), "Correspondence of the brain; functional architecture during activation and rest." *Proceedings of the National Academy of Sciences*, 106, 13040–5.

Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017), "Expandable factor analysis," *Biometrika*, 104, 649–663.

Sun, W. and Li, L. (2017), "STORE: Sparse Tensor Response Regression and Neuroimaging Analysis," *Journal of Machine Learning Research*, 18, 4908–4944.

Sylvester, C. M., Yu, Q., Srivastava, A. B., Marek, S., Zheng, A., Alexopoulos, D., Smyser, C. D., Shimony, J. S., Ortega, M., Dierker, D. L., et al. (2020), "Individual-specific functional connectivity of the amygdala: A substrate for precision psychiatry," *Proceedings of the National Academy of Sciences*, 117, 3808–3818.

Tang, X., Bi, X., and Qu, A. (2019), "Individualized Multilayer Tensor Learning With an Application in Imaging Analysis," *Journal of the American Statistical Association*, 1–26.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016), "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference on Machine Learning*, PMLR, pp. 964–973.

Varoquaux, G. and Craddock, R. C. (2013), "Learning and comparing functional connectomes across subjects," *NeuroImage*, 80, 405–415.

Vounou, M., Nichols, T. E., Montana, G., and Initiative, A. D. N. (2010), "Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach," *Neuroimage*, 53, 1147–1159.

Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017), "Bayesian network–response regression," *Bioinformatics*, 33, 1859–1866.

Wang, L., Zhang, Z., and Dunson, D. (2019), "Common and individual structure of brain networks," *The Annals of Applied Statistics*, 13, 85–112.

Wang, Y. and Guo, Y. (2019), "A hierarchical independent component analysis model for longitudinal neuroimaging studies," *NeuroImage*, 189, 380–400.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. (2016), "An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation," *Frontiers in Neuroscience*, 10, 1–17.

Zhang, A. and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338.

Zhang, J. and Cao, J. (2017), "Finding Common Modules in a Time-Varying Network with Application to the Drosophila Melanogaster Gene Regulation Network," *Journal of the American Statistical Association*, 112, 994–1008.

Zhang, X. and Li, L. (2017), "Tensor envelope partial least-squares regression," *Technometrics*, 59, 426–436.

Zhang, X., Wang, L., and Gu, Q. (2018), "A Unified Framework for Nonconvex Low-Rank plus Sparse Matrix Recovery," in *International Conference on Artificial Intelligence and Statistics*, pp. 1097–1107.

Zheng, Q. and Lafferty, J. (2016), "Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.