Latent space models for grouped multiplex networks

Alexander Kagan *[1], Peter W. MacDonald[2], Elizaveta Levina[1], and Ji Zhu[1]

[1]Department of Statistics, University of Michigan
[2]Department of Statistics & Actuarial Science, University of Waterloo

Abstract

Complex multilayer network datasets have become ubiquitous in various applications, such as neuroscience, social sciences, economics, and genetics. Notable examples include brain connectivity networks collected across multiple patients or trade networks between countries collected across multiple goods. Existing statistical approaches to such data typically focus on modeling the structure shared by all networks; some go further by accounting for individual, layer-specific variation. However, real-world multilayer networks often exhibit additional patterns shared only within certain subsets of layers, which can represent treatment and control groups, or patients grouped by a specific trait. Identifying these group-level structures can uncover systematic differences between groups of networks and influence many downstream tasks, such as testing and low-dimensional visualization. To address this gap in existing research, we introduce the GroupMultiNeSS model, which generalizes the previously proposed MultiNeSS model of MacDonald et al. (2022). This model enables the simultaneous extraction of shared, group-specific, and individual latent structures from a sample of multiplex networks — multiple, heterogeneous networks observed on a shared node set. For this model, we establish identifiability, develop a fitting procedure using convex optimization in combination with a nuclear norm penalty, and prove a guarantee of recovery for the latent positions as long as there is sufficient separation between the shared, group-specific, and individual latent subspaces. We compare the model with MultiNeSS and other models for multiplex networks in various synthetic scenarios and observe an apparent improvement in the modeling accuracy when the group component is accounted for. Experiment with the Parkinson's disease brain connectivity dataset of Badea et al. (2017) demonstrates the superiority of GroupMultiNeSS in highlighting node-level insights on biological differences between the treatment and control patient groups.

Keywords: multiplex networks, latent space models, group structure

1 Introduction

Complex multilayer network datasets have become ubiquitous in many statistical applications, with neuroscience, genetics, social sciences, and economics, among others. Notable examples of such datasets include brain connectivity networks observed in a group of patients, protein-protein interaction networks observed across multiple tissues, or trade networks between countries for various goods. The development of statistical methodologies for modeling such data has been an active area of research in recent years. This led to natural generalizations of classical single-layer models, such as the stochastic block model (SBM) (Holland et al., 1983), the random dot product

*Corresponding author. Email: amkagan@umich.edu

graph (RDPG) (Young and Scheinerman, 2007; Athreya et al., 2017), and the unifying model of Ma et al. (2020) to their respective multilayer versions (Han et al., 2015; Jones and Rubin-Delanchy, 2021; Zhang et al., 2020). Other lines of work focused on Bayesian latent space approaches (Gollini and Murphy, 2013; Salter-Townshend and McCormick, 2017; D'Angelo et al., 2018; Sosa and Betancourt, 2021), and some attempted to develop novel frequentist approaches, such as the Common Subspace Independent Edge Model (COSIE) (Arroyo et al., 2021) and the MultiNeSS model (MacDonald et al., 2022; Tian et al., 2024). Despite their methodological differences, all of these models address the same key question of how to model the common structure shared by all layers, while incorporating individual network-specific information. For example, multilayer SBM does that by assuming the community labels of the nodes to be the same across layers but allowing the community connectivity probabilities to differ, COSIE assumes all expected layer adjacency matrices lie in a common subspace but can differ within it, and MultiNeSS allows for decomposition of each layer's latent space into a disjoint union of the shared and individual subspaces.

In this work, we consider the natural follow-up question of how to account for the additional structures specific only to subgroups of network layers within the collection and how to separate these group-level information pieces from the information shared by all the layers. For us, this question is mainly motivated by neuroscience studies, where it is of interest to identify brain regions that exhibit abnormal connectivity patterns in patients with a given disease, such as ADHD (Ghaderi et al., 2017), Alzheimer (Wu et al., 2024), or Parkinson's disease (Badea et al., 2017), compared to brain networks of healthy controls. These works, as well as other neuroscience studies, mainly base their findings on permutation tests applied to appropriate descriptive statistics of network groups, such as average connectivity or average path length. This approach has an apparent limitation, as descriptive statistics are often insufficiently expressive and cannot account for all the complex underlying connectivity patterns in the network, thus providing tests of very low power. On the other hand, most existing statistical works addressing this problem assume a latent space network-generating model and develop hypothesis tests to assess whether the latent positions of nodes in the two samples are drawn from the same distribution, usually up to an orthogonal rotation or scaling (Tang et al., 2017; Nguen et al., 2024; MacDonald et al., 2024). Although these approaches help determine whether there exists a statistically significant difference in the two network samples, they are incapable of providing quantitative or visual node-level insights on where this difference comes from – the information that could be crucial for many downstream analyses, such as the identification of malfunctioning brain regions or connections between them.

To address this question, we propose a latent space model for a collection of multiplex networks exhibiting group-level latent structures. It generalizes the previously proposed MultiNeSS model by allowing the latent positions of nodes within each layer to comprise not only a component shared across all networks and an individual component specific only to this layer, but also an additional group component shared by the networks only within its group. We develop a computationally efficient fitting procedure for the proposed model, building upon the estimation techniques for MultiNeSS, and establish theoretical guarantees for it under the Gaussian edge-entry model. Through simulation studies and applications to real-world data, we show that incorporating the group component significantly improves modeling accuracy. Our results highlight the importance of accommodating group-level structure in latent space models and provide a versatile tool for analyzing grouped network data.

The rest of this manuscript is organized as follows. In Section 2, we formally describe the proposed model and state the identifiability conditions for its parameters. In Section 3, we formulate the fitting procedure and propose a cross-validation approach to tuning the algorithm's hyperparameters. In Section 4, we establish consistency for the fitting algorithm under the

Gaussian edge entry distribution. In Section 5, we provide numerical simulations on synthetic data to justify the efficacy of our method, and in Section 6, we apply it to the real-world dataset in Badea et al. (2017) and study brain connectivity differences in patients diagnosed with Parkinson's disease. Finally, in Section 7, we discuss the future developments of our approach.

2 Model for grouped multiplex networks

2.1 GroupMultiNeSS

Here we propose a new latent space model for GROUPed MULTIplex NEtworks with Shared Structure (GroupMultiNeSS). We start by fixing the notation. Consider M undirected networks on a shared set of n nodes and assume that the networks are separated into K groups of sizes $m_k, k = 1, \ldots, K$ so that $M = \sum_{k=1}^{K} m_k$. Each of these networks is encoded via a symmetric and possibly weighted adjacency matrix

$$A_{k\ell} \in \mathbb{R}^{n\times n},\ (k,\ell) \in \mathcal{I}, \quad \text{where} \quad \mathcal{I} := \{(k,\ell) : k = 1,\ldots,K;\ \ell = 1,\ldots,m_k\},$$

and we will refer to an edge between nodes i and j in layer ℓ of group k as $A_{k\ell,ij}$. A node i within each layer $A_{k\ell}$ is associated with a fixed latent position $x_{k\ell}^{(i)} \in \mathcal{X}_{k\ell} \subset \mathbb{R}^{D_{k\ell}}$ where the latent space $\mathcal{X}_{k\ell}$ is equipped with a symmetric function $\kappa_{k\ell} : \mathcal{X}_{k\ell} \times \mathcal{X}_{k\ell} \to \mathbb{R}$ capturing similarity between the two input latent positions. Stacked latent positions of all the nodes in $A_{k\ell}$ will be denoted as a matrix $X_{k\ell} \in \mathbb{R}^{n\times D_{k\ell}}$. We assume that conditional on $X_{k\ell}$, the elements of $A_{k\ell}$ are independent and follow an edge entry distribution

$$A_{k\ell,ij} \overset{\text{ind}}{\sim} f(\cdot; \kappa_{k\ell}(x_{k\ell}^{(i)}, x_{k\ell}^{(j)}), \phi_{k\ell}), \quad 1 \le i \le j \le n, \quad (k,\ell) \in \mathcal{I}. \tag{1}$$

Here, $f(\cdot;\theta,\phi)$ depends on some possible layer-dependent nuisance parameters ϕ and a scalar parameter θ capturing the effect of latent similarity between nodes i and j on the edge connecting them in the corresponding layer. For example, if each $\kappa_{k\ell}$ is the standard Euclidean inner product on $\mathcal{X}_{k\ell}$ such that $x^\top y \in [0,1]$ for any $x, y \in \mathcal{X}_{k\ell}$, and $f(\cdot;\theta) = \text{Bernoulli}(\theta)$, we get the Multilayer RDPG model (Jones and Rubin-Delanchy, 2021). For most applications, it would usually be sufficient to restrict the edge-entry distribution to belong to a one-parameter canonical exponential family

$$f(x;\theta) \propto \exp\{\theta x - \nu(\theta)\}, \quad \theta \in \mathbb{R} \tag{2}$$

with a natural parameter θ and a log-partition function ν as it includes Bernoulli, Poisson, Gaussian, and other distributions capable of modeling nearly all possible edge entry types, including binary, count, and continuous. For now, we do not make any assumptions on the distribution family and state the model for an arbitrary edge-entry distribution for completeness.

Notice that (1) assumes the distribution of layer $A_{k\ell}$ depends on the latent positions $X_{k\ell}$ only through the Gram matrix $\Theta_{k\ell} \in \mathbb{R}^{n\times n}$ with $\Theta_{k\ell,ij} = \kappa_{k\ell}(x_{kl}^{(i)}, x_{k\ell}^{(j)})$. In particular, when $\kappa_{k\ell}$ is the standard Euclidean inner product, $\Theta_{k\ell} = X_{k\ell}X_{k\ell}^\top$ obtains the recognizable low rank matrix form that widely appears in the latent space modeling literature following the seminal work of (Hoff et al., 2002). To emphasize this observation, we rewrite (1) in the matrix form for convenience as follows

$$A_{k\ell} \overset{\text{ind}}{\sim} f(\cdot; \Theta_{k\ell}, \phi_{k\ell}), \quad (k,\ell) \in \mathcal{I}. \tag{3}$$

The assumption that the layer distribution depends on the latent positions $X_{k\ell}$ only through their Gram matrix $\Theta_{k\ell}$ naturally leads to identifiability issues which will be discussed in section 2.2 in more detail.

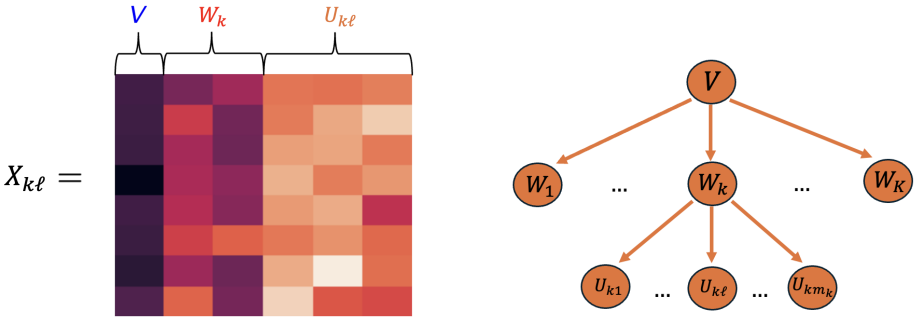

Figure 1: Latent space decomposition assumed by GroupMultiNeSS.

Until now, we have treated $X_{k\ell}$ as independent variables and have not imposed any relationship between the latent positions of the nodes between layers. The main assumption of the proposed GroupMultiNeSS model (see Figure 1) is that the latent positions $X_{k\ell} \in \mathbb{R}^{n \times D_{k\ell}}$ contain some common structure $V \in \mathbb{R}^{n \times d_0}$ shared by all the layers, a group-specific structure $W_k \in \mathbb{R}^{n \times d_k}$ shared by the layers only within a given group $k = 1, \ldots, K$, and finally, some individual structure $U_{k\ell} \in \mathbb{R}^{n \times d_{k\ell}}$ specific only to a given layer $\ell = 1, \ldots, m_k$ within group k:

$$X_{k\ell} = [V \; W_k \; U_{k\ell}], \quad (k, \ell) \in \mathcal{I}, \tag{4}$$

so that $D_{k\ell} = d_0 + d_k + d_{k\ell}$. Without the group components, that is, if $d_k = 0$ for each $k = 1, \ldots, K$, this model would transform into the previously proposed MultiNeSS model of MacDonald et al. (2022).

To allow $\kappa_{k\ell}$ capture richer dependencies between the nodes' latent positions, we assume it as the generalized inner product that takes vectors x and y in $\mathbb{R}^{p+q}$ as input and outputs

$$\kappa_{p,q}(x, y) = x_1 y_1 + \cdots + x_p y_p - x_{p+1} y_{p+1} - \cdots - x_{p+q} y_{p+q} = x^\top I_{p,q} y, \tag{5}$$

where $I_{p,q}$ is a $(p+q) \times (p+q)$ diagonal matrix with first p diagonal values equal to 1 and remaining q equal to -1. In this decomposition, the first p and last q latent dimensions are referred to as *assortative* and *disassortative*, respectively. Intuitively, similarity along an assortative dimension increases the edge weight between two nodes, whereas similarity along a disassortative dimension decreases it. For full generality, we also allow each latent component to contain a different number of assortative and disassortative dimensions. That is, we assume $\Theta_{k\ell}$ admits the decomposition

$$\Theta_{k\ell} = V I_{p_0,q_0} V^\top + W_k I_{p_k,q_k} W_k^\top + U_{k\ell} I_{p_{k\ell},q_{k\ell}} U_{k\ell}^\top, \tag{6}$$

where $p_k + q_k = d_k$ for $k = 0, 1, \ldots, K$ and $p_{k\ell} + q_{k\ell} = d_{k\ell}$ for $(k, \ell) \in \mathcal{I}$.

2.2 Identifiability

Here, we formulate a sufficient condition for the identifiability of the GroupMultiNeSS parameters. For this section only, we assume a slightly more general form of the similarity function κ than we did in (5). That is, we assume $\kappa(x, y) = \psi(x^\top I_{p,q} y)$ is an invertible scalar function of the generalized inner product. In the context of identifiability, this is a natural extension since two edge-entry distributions $f(\cdot; \theta_1)$ and $f(\cdot; \theta_2)$ that do not coincide for any $\theta_1 \neq \theta_2$, will also not coincide at $\psi(\theta_1) \neq \psi(\theta_2)$ by invertibility.

Assume first we observed just one network layer $A_{k\ell}$ (and thus just one group k). Then it would be natural to assume column-wise linear independence of the latent position matrix $X_{k\ell} = [V\ W_k\ U_{k\ell}]$ to guarantee identifiability of the latent dimension $D_{k\ell}$, a standard condition in the context of inner-product latent models. Apparently, it is not enough as we need to observe at least two groups to separate V from W_k and at least two networks in each group to be able to separate W_k from $U_{k\ell}$, implying $K \geq 2$ and $m_k \geq 2, k = 1, \ldots, K$. Indeed, if we had only one group, mixing the columns of V and W_k would not change the Gram matrix in (6) and thus the edge-entry distribution. Similarly, if we observed only one layer ℓ in the group k, mixing the columns of $U_{k\ell}$ and W_k would also preserve the distribution. Unfortunately, these conditions are still insufficient as we can simultaneously put a subset of W_k's columns to each $U_{k\ell},\ \ell = 1, \ldots, m_k$ and a subset of V's columns to each $W_k, k = 1, \ldots, K$ or to each $\{U_{k\ell}\}_{\mathcal{I}}$ without changing the distribution. To avoid this, intuitively, we need V to contain the maximum information shared by all the layers, and W_k to have the maximum information shared by the layers in group k but not present in V. Finally, even if all this holds, each latent component would be only identifiable up to an *indefinite orthogonal transformation* in $\mathcal{O}_{p,q} = \{O \in \mathbb{R}^{(p+q)\times(p+q)} : O^\top I_{p,q} O = I_{p,q}\}$, where p and q are the numbers of assortative and disassortative dimensions in the corresponding component, respectively. For example, it is easy to notice that the Gram matrix in (6), would not change from substituting V by VO where $O \in \mathcal{O}_{p_0,q_0}$. The following proposition formulates a sufficient condition that gathers all these considerations in one statement. The proof can be found in section B.1 of Appendix.

Proposition 2.1. *Assume $\{f(\cdot;\theta,\phi),\ \theta \in \mathbb{R}\}$ is an identifiable parametric family and $\kappa(x,y) = \psi(x^\top I_{p,q} y)$ is an invertible function of the generalized inner product. Under the following conditions, parameters of the GroupMultiNeSS model with edge-entry distribution in* (1) *are identifiable up to indefinite orthogonal transformations:*

1. *For each $(k,\ell) \in \mathcal{I}$, columns of the matrix $X_{k\ell} = [V\ W_k\ U_{k\ell}]$ are linearly independent.*
2. *For each $k = 1, \ldots, K$, there exist layers (k,s) and (k,t) such that the columns of the matrix*
$$[V\ W_k\ U_{ks}\ U_{kt}]$$
are linearly independent.
3. *There exist $(k_1,\ell_1), (k_2,\ell_2) \in \mathcal{I}$ with $k_1 \neq k_2$ such that the columns of the matrix*
$$[V\ W_{k_1}\ W_{k_2}\ U_{k_1\ell_1}\ U_{k_2\ell_2}]$$
are linearly independent.

That is, under these conditions, if the probability distributions induced by two different parameterizations $\left(V, \{W_k\}_{k=1}^K, \{U_{k\ell}\}_{\mathcal{I}}\right)$ and $\left(V', \{W'_k\}_{k=1}^K, \{U'_{k\ell}\}_{\mathcal{I}}\right)$ coincide, then

$$\begin{aligned} V &= V' O_0 \\ W_k &= W'_k O_k, && k = 1, \ldots, K, \\ U_{k\ell} &= U'_{k\ell} O_{k\ell}, && (k,\ell) \in \mathcal{I}. \end{aligned}$$

for some indefinite orthogonal rotations $O_k \in \mathcal{O}_{p_k,q_k}, k = 0, \ldots, K$ and $O_{k\ell} \in \mathcal{O}_{p_{k\ell},q_{k\ell}}, (k,\ell) \in \mathcal{I}$.

Remark 1. Intuitively, Condition 2 ensures that, for each group k, there exist two individual components U_{ks} and U_{kt} whose columns are linearly independent of those in the "total shared" component $[V,\ W_k]$. Condition 3 ensures that the shared component V has linearly independent columns with those in the "total individual" components $[W_{k_1},\ U_{k_1\ell_1}]$ and $[W_{k_2},\ U_{k_2\ell_2}]$ for some groups k_1 and k_2.

3 Estimation

3.1 Likelihood maximization with nuclear norm penalization

Assuming the ground truth latent positions $V, \{W_k\}_{k=1}^K, \{U_{k\ell}\}_{\mathcal{I}}$ satisfy the identifiability condition from Proposition 2.1, each of them is still identifiable only up to an indefinite orthogonal transformation. To avoid this issue, we propose to estimate their Gram matrices instead, and we denote them as follows:

$$\begin{aligned} &S = V I_{p_0,q_0} V^\top, \\ &Q_k = W_k I_{p_k,q_k} W_k^\top, && k = 1, \dots, K, \\ &R_{k\ell} = U_{k\ell} I_{p_{k\ell},q_{k\ell}} U_{k\ell}^\top, && (k,\ell) \in \mathcal{I}. \end{aligned} \tag{7}$$

In practice, given the low-rank estimates $\hat{S}, \hat{Q}_k$, and $\hat{R}_{k\ell}$, we can further extract $\hat{V}, \hat{W}_k$, and $\hat{U}_{k\ell}$ with Adjacency Spectral Embedding (ASE) (Rubin-Delanchy et al., 2017). For example, to estimate $\hat{V}$, this method would suggest computing the truncated eigen-decomposition $\hat{S} = \hat{\hat{V}}\hat{\Gamma}\hat{\hat{V}}^\top$, where $\hat{\Gamma} \in \mathbb{R}^{\hat{d}_0 \times \hat{d}_0}$ is a diagonal matrix containing the top $\hat{d}_0$ eigenvalues (in absolute magnitude), which are sorted on the diagonal in decreasing order with respect to their signed value, and $\hat{\hat{V}} \in \mathbb{R}^{n \times \hat{d}_0}$ contains the corresponding eigenvectors. In this case, $\hat{p}_0$ and $\hat{q}_0$ with $\hat{d}_0 = \hat{p}_0 + \hat{q}_0$ correspond to the number of positive and negative entries in $\hat{\Gamma}$, respectively. Then, we can set $\hat{V} = \hat{\hat{V}}|\hat{\Gamma}|^{1/2}$ where $|\cdot|$ takes absolute values of all entries in a matrix. Estimates $\hat{W}_k$ and $\hat{U}_{k\ell}$ can be computed similarly.

With the notations in (7) at hand, a natural way to fit the GroupMultiNeSS model would be to minimize the negative log-likelihood by exploiting the independence of adjacency matrix entries stated in (1):

$$\mathcal{L}\left(S, \{Q_k\}_{k=1}^K, \{R_{k\ell}\}_{\mathcal{I}};\ \{A_{k\ell}\}_{\mathcal{I}}\right) = \sum_{k=1}^K \mathcal{L}_k(S + Q_k, \{R_{k\ell}\}_{\ell=1}^{m_k};\ \{A_{k\ell}\}_{\ell=1}^{m_k}) \tag{8}$$

where we denoted the negative likelihood of the k-th group by

$$\mathcal{L}_k(S + Q_k, \{R_{k\ell}\}_{\ell=1}^{m_k};\ \{A_{k\ell}\}_{\ell=1}^{m_k}) = -\sum_{\ell=1}^{m_k} \sum_{i \le j} \log f\left(A_{k\ell,ij}; S_{ij} + Q_{k,ij} + R_{k\ell,ij}, \phi\right) \tag{9}$$

with the inner summation restricted to node pairs with $i < j$ in case the network is loop-free. Notice that the group log-likelihood $\mathcal{L}_k$ depends on variables $S + Q_k$ and $\{R_{k\ell}\}_{\ell=1}^{m_k}$ that appear only in the k-th group, giving a hint about a potential block structure in the optimization procedure. Indeed, we could first estimate $\widehat{S + Q_k}$ and $\{\hat{R}_{k\ell}\}_{\ell=1}^{m_k}$ within each group and then separate $\hat{S}$ from $\hat{Q}_k, k = 1, \dots, K$ by optimizing the full likelihood $\mathcal{L}$ with all $\{\hat{R}_{k\ell}\}_{\mathcal{I}}$ fixed. An issue of direct optimization of (8) is that the parameters introduced in (7) are not restricted to be of low rank and can violate the required rank constraints:

$$\begin{aligned} &\operatorname{rank}^+(S) = p_0, && \operatorname{rank}^-(S) = q_0 && \\ &\operatorname{rank}^+(Q_k) = p_k, && \operatorname{rank}^-(Q_k) = q_k && k = 1, \dots, K, \\ &\operatorname{rank}^+(R_{k\ell}) = p_{k\ell}, && \operatorname{rank}^-(R_{k\ell}) = q_{k\ell} && (k,\ell) \in \mathcal{I}, \end{aligned}$$

where $\operatorname{rank}^+(Z)$ and $\operatorname{rank}^-(Z)$ denote respectively the number of strictly positive and strictly negative eigenvalues of a symmetric matrix Z. Unfortunately, even in an oracle scenario where the

dimensions of latent components are known, including the rank constraints in the optimization makes the problem non-convex and usually intractable. Instead, we can perform convex relaxation using a nuclear penalty regularization, resulting in the following two-stage optimization problem, where in the first stage, for each $k = 1, \ldots, K$, we estimate $\widehat{S + Q_k}, \{\hat{R}_{k\ell}\}_{\ell=1}^{m_k}$ by solving the following optimization problem:

$$\min_{S+Q_k, R_{k\ell}} \left\{ \mathcal{L}_k \left(S + Q_k, \{R_{k\ell}\}_{\ell=1}^{m_k} ; \ \{A_{k\ell}\}_{\ell=1}^{m_k} \right) + \lambda_{1k} \|S + Q_k\|_* + \sum_{\ell=1}^{m_k} \lambda_{1k} \alpha_{1k\ell} \|R_{k\ell}\|_* \right\} \tag{10}$$

and in the second stage, we get $\hat{S}, \{\hat{Q}_k\}_{k=1}^K$ by solving:

$$\min_{S, Q_k} \left\{ \mathcal{L} \left(S, \{Q_k\}_{k=1}^K, \{\hat{R}_{k\ell}\}_{\mathcal{I}} ; \ \{A_{k\ell}\}_{\mathcal{I}} \right) + \lambda_2 \|S\|_* + \sum_{k=1}^K \lambda_2 \alpha_{2k} \|Q_k\|_* \right\}. \tag{11}$$

Here, $\|\cdot\|_*$ denotes the nuclear norm of a matrix equal to the sum of its singular values, and $\{\lambda_{1k}\}_{k=1}^K, \lambda_2, \{\alpha_{2k}\}_{k=1}^K, \{\alpha_{1k\ell}\}_{\mathcal{I}}$ are tunable hyperparameters (see details on tuning in Section 3.2). Since the nuclear norm is convex, these optimization problems are convex if the edge distribution density $f(\cdot; \theta, \phi)$ is log-concave in θ.

Remark 2. The formulated two-stage fitting procedure can be thought of as a bottom-to-top pruning of the GroupMultiNeSS tree in Figure 1. Indeed, estimation of the individual components in the first stage can be interpreted as "cutting off" the leaves from the tree, since they are no longer updated in the second stage. Similarly, the second-stage separation of the group and shared components can be interpreted as cutting off the tree's second-level nodes, which act as new leaves after the first-stage completion. This analogy suggests a natural extension of the proposed fitting procedure to the model, which generalizes GroupMultiNeSS by allowing the tree structure of the network layers to involve a more complex hierarchy, for example, by including additional subgroup latent components.

To minimize the loss functions in (10) and (11), we perform a coordinate descent by iteratively updating one of the optimized matrices while keeping the others fixed. In turn, each matrix update can be performed using a proximal gradient method, specifically to optimize a loss function L combined with the non-differentiable nuclear norm penalty (Fithian and Mazumder, 2018). Given the gradient step-size $\eta > 0$, the update at iteration $t \geq 1$ is performed as follows

$$Z^{(t+1)} = \underset{Z \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \left\{ \frac{1}{2\eta} \left\| Z - \left[Z^{(t)} - \eta \nabla L(Z^{(t)}) \right] \right\|_F^2 + \rho \|Z\|_* \right\} \tag{12}$$

where L, Z, and ρ are place-holders respectively, for a loss function, an optimized parameter matrix, and a hyperparameter associated with the nuclear norm penalty. In the first stage, the loss is $\mathcal{L}_k$ and the matrices are $S + Q_k$ and $R_{k\ell}$ with respective thresholds λ_{1k} and $\lambda_{1k}\alpha_{1k\ell}$. In the second stage, the loss is $\mathcal{L}$ and the matrices are S and Q_k with thresholds λ_2 and $\lambda_2\alpha_{2k}$, respectively. It is well-known that the solution to the optimization problem in (12) has an explicit form in terms of the so-called *soft-thresholding operator*

$$Z^{(t+1)} = \mathcal{T}_{\eta\rho} \left(Z^{(t)} - \eta \nabla L(Z^{(t)}) \right). \tag{13}$$

For a diagonal matrix $\Sigma \in \mathbb{R}^{n \times n}$, the soft-thresholding operator $\mathcal{T}_s$ with parameter s is defined as

$$\mathcal{T}_s(\Sigma) = \operatorname{diag}[(\Sigma_{11} - s)_+, \ldots, (\Sigma_{nn} - s)_+]$$

with $(x)_+ := \max\{0, x\}$ and for a non-diagonal matrix with singular value decomposition $Z = U\Sigma V^\top$ as

$$\mathcal{T}_s(Z) = U\mathcal{T}_s(\Sigma)V^\top.$$

Intuitively, the soft-thresholding operator in the update rule (13) mimics a projector onto the space of low-rank matrices by truncating the singular values of a parameter matrix after each gradient step. In case the ground-truth rank d of Z was known, soft-thresholding could be replaced by hard-thresholding

$$Z^{(t+1)} = \left[Z^{(t)} - \eta\nabla L(Z^{(t)})\right]_d \tag{14}$$

where

$$[Z]_d := \underset{\text{rank}(Z')\leq d}{\text{argmin}} \|Z - Z'\|_F \tag{15}$$

is well-defined by the Eckart-Young Theorem as the truncation of the SVD of Z to the largest d singular values.

With the derived update rule for each parameter matrix at hand, we can define the end-to-end optimization procedure as Algorithm 1. In the first stage of the algorithm, we solve Problem (10) for every $k = 1, \ldots, K$ by alternating the updates of $S + Q_k$ and $\{R_{k\ell}\}_{l=1}^{m_k}$ until convergence. In the second stage, we solve Problem (11) by alternating the updates of S and $\{Q_k\}_{k=1}^K$. Each matrix update is performed according to (13) but with stage-specific learning rate $\eta_1, \eta_2 > 0$ and parameter-specific normalization for convergence stability. In the first stage, we normalize η_1 by m_k for updates of $S + Q_k$ and in the second stage, we normalize η_2 by M and m_k for updates of S and Q_k, respectively, that is, we divide by the number of times these parameters appear in the joint log-likelihood (8). The parameter matrices obtained by running the alternating updates within each optimization subproblem are further passed to the optional *refitting step* aimed to debias the learned non-zero eigenvalues of the fitted matrices and so compensate for their excessive truncation caused by the nuclear norm penalization (see details in Section 3.4). The first and second stage refitting procedures will exhibit some principal differences, arising from the fact that the individual components $\hat{R}_{k\ell}$ fitted in the first stage should be kept fixed during the second stage refitting. Therefore, we denote the corresponding algorithms as *FirstStageRefit* and *SecondStageRefit*, respectively.

For now, we do not explicitly state how to initialize parameters in both stages and denote the corresponding procedures for the first and second stages as *FirstStageInit* and *SecondStageInit*, respectively. The detailed investigation of initialization options is postponed to Section 3.3. Notice that for initializing $S + Q_k$ in the first stage, we only use the k-th group layers $\{A_{k\ell}\}_{\ell=1}^{m_k}$ as $S + Q_k$ is essentially an independent variable that does not appear in any other group except for the group k itself and so other layers cannot give extra information on its possible value.

Remark 3. The computational bottleneck within every iteration of the first and second stages of Algorithm 1 is the computation of the SVD for each updated parameter matrix, which is needed for soft thresholding. If we computed a full SVD, each iteration within group k of the first stage would require $O(m_k n^3)$ operations, and each iteration within the second stage would take $O(Kn^3)$. In addition to the discussed parallelization of the first-stage subproblems, we speed up the computations by using a truncated SVD for matrix update, which only finds its leading $r \ll n$ singular values, reducing the complexity to $O(m_k n^2 r)$ and $O(Kn^2 r)$, respectively. In our implementation, we use $r = \lceil \sqrt{n} \rceil$ as the default value.

Remark 4. Algorithm 1 allows for many similar optimization alternatives, such as updating all three types of components (S, Q_k, $R_{k\ell}$) in a single loop, substituting the second stage loss from

Algorithm 1 Stepwise proximal gradient descent for optimizing (8)

Input: Adjacency matrices $\{A_{k\ell}\}_{\mathcal{I}}$, penalty coefficients $\{\lambda_{1k}\}_{k=1}^K, \lambda_2, \{\alpha_{2k}\}_{k=1}^K, \{\alpha_{1k\ell}\}_{\mathcal{I}}$, learning rates $\eta_1, \eta_2 > 0$.

First Stage

for $k = 1, \dots, K$ **do**

$(S+Q_k)^{(0)}, \{R_{k\ell}^{(0)}\}_{\ell=1}^{m_k} \leftarrow \text{FirstStageInit}(\{A_{k\ell}\}_{\ell=1}^{m_k})$

for iteration $t = 1, 2, \dots$ until convergence **do**

$$R_{k\ell}^{(t)} \leftarrow \mathcal{T}_{\eta_1\lambda_{1k}\alpha_{1k\ell}}\left[R_{k\ell}^{(t-1)} - \eta_1 \tfrac{\partial}{\partial R_{k\ell}}\mathcal{L}_k((S+Q_k)^{(t-1)}, \{R_{k\ell}^{(t-1)}\})\right], \quad \text{for } \ell = 1, \dots, m_k,$$
$$(S+Q_k)^{(t)} \leftarrow \mathcal{T}_{\eta_1\lambda_{1k}/m_k}\left[(S+Q_k)^{(t-1)} - \tfrac{\eta_1}{m_k}\tfrac{\partial}{\partial Q_k}\mathcal{L}_k((S+Q_k)^{(t-1)}, \{R_{k\ell}^{(t)}\})\right].$$

end for

$\widehat{S+Q}_k, \{\hat{R}_{k\ell}\}_{\ell=1}^{m_k} \leftarrow \text{FirstStageRefit}\big((S+Q_k)^{(t)}, \{R_{k\ell}^{(t)}\}_{\ell=1}^{m_k}; \{A_{k\ell}\}_{\ell=1}^{m_k}\big)$

end for

Second Stage

$S^{(0)}, \{Q_k^{(0)}\}_{k=1}^K \leftarrow \text{SecondStageInit}(\{A_{k\ell}\}_{\mathcal{I}}, \{\hat{R}_{k\ell}\}_{\mathcal{I}}, \{\widehat{S+Q}_k\}_{k=1}^K)$

for iteration $t = 1, 2, \dots$ until convergence **do**

$$Q_k^{(t)} \leftarrow \mathcal{T}_{\eta_2\lambda_2\alpha_{2k}/m_k}\left[Q_k^{(t-1)} - \tfrac{\eta_2}{m_k}\tfrac{\partial}{\partial Q_k}\mathcal{L}(S^{(t-1)}, \{Q_k^{(t-1)}\}, \{\hat{R}_{k\ell}\})\right], \quad \text{for } k = 1, \dots, K,$$
$$S^{(t)} \leftarrow \mathcal{T}_{\eta_2\lambda_2/M}\left[S^{(t-1)} - \tfrac{\eta_2}{M}\tfrac{\partial}{\partial S}\mathcal{L}(S^{(t-1)}, \{Q_k^{(t)}\}, \{\hat{R}_{k\ell}\})\right].$$

end for

$\hat{S}, \{\hat{Q}_k\}_{k=1}^K \leftarrow \text{SecondStageRefit}\big(S^{(t)}, \{Q_k^{(t)}\}_{k=1}^K, \{\hat{R}_{k\ell}\}_{\mathcal{I}}; \{A_{k\ell}\}_{\mathcal{I}}\big)$

the negative log-likelihood to the quadratic loss between $S+Q_k$ and $\widehat{S+Q}_k$, or pre-estimating component ranks $\{d_k\}_{k=0}^K$ and $\{d_{k\ell}\}_{\mathcal{I}}$ first and then substituting the soft-thresholding updates by the hard-thresholding ones as described in (14). The discussion and comparison of these and other alternatives is postponed to Section A of the Appendix.

Notice that the optimization procedure suggested for the first-stage problem (10) essentially applies the fitting procedure of the MultiNeSS model to the layers $\{A_{k\ell}\}_{\ell=1}^{m_k}$ (as defined in Equation 5 of MacDonald et al. (2022)) aiming to separate the individual components $\{R_{k\ell}\}_{\ell=1}^{m_k}$ from the "total shared" component $S+Q_k$. In contrast, the second-stage optimization is no longer so interpretable since we fixed the estimates of the individual components. To build some intuition about what happens in the second stage of Algorithm 1, we explicitly state its relationship to the MultiNeSS fitting procedure in the special case of the Gaussian edge-entry distribution. Intuitively, we demonstrate that in the Gaussian case, the second stage of the GroupMultiNeSS essentially fits the MultiNeSS model to the group-wise averaged residuals between the layers and their estimated individual components. The proof of this proposition can be found in Section B.2 of the Appendix.

Proposition 3.1. *With the edge entry distribution $f(\cdot;\theta,\sigma) = \mathcal{N}(\theta,\sigma^2)$, Problem (11) coincides with the objective of the MultiNeSS model fitted to the layers $\tilde{A}_k = \frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell} - \hat{R}_{k\ell})$, $k = 1, \dots, K$ under the assumption that their edge entries are independent and Gaussian with layer-*

dependent variances:

$$\tilde{A}_{k,ij} \overset{\text{ind}}{\sim} \mathcal{N}(S_{ij} + Q_{k,ij}, \sigma^2/m_k), \quad 1 \le i \le j \le n, \quad k = 1, \dots, K.$$

Remark 5. In Section 4, we will derive theoretical guarantees for the estimators produced by Algorithm 1 under the assumption of Gaussian edge entry distribution and unit learning rate $\eta_1 = \eta_2 = 1$. In this case, the proximal gradient descent updates are especially convenient for theoretical analysis. For the k-th problem in the first stage, the update rule at iteration $t \ge 1$ obtains the form

$$\begin{aligned} R_{k\ell}^{(t)} &= \mathcal{T}_{\lambda_{1k}\alpha_{1k\ell}}\Big[A_{k\ell} - (S + Q_k)^{(t-1)}\Big], \qquad \ell = 1, \dots, m_k, \\ (S + Q_k)^{(t)} &= \mathcal{T}_{\lambda_{1k}/m_k}\Big[\frac{1}{m_k}\sum_{\ell=1}^{m_k}\Big(A_{k\ell} - R_{k\ell}^{(t)}\Big)\Big]. \end{aligned} \tag{16}$$

Similarly, the second stage updates can be written for $t \ge 1$ as

$$\begin{aligned} Q_k^{(t)} &= \mathcal{T}_{\lambda_2\alpha_{2k}/m_k}\Big[\frac{1}{m_k}\sum_{\ell=1}^{m_k}\Big(A_{k\ell} - \hat{R}_{k\ell} - S^{(t-1)}\Big)\Big], \qquad k = 1, \dots, K, \\ S^{(t)} &= \mathcal{T}_{\lambda_2/M}\Big[\frac{1}{M}\sum_{(k,\ell)\in\mathcal{I}}\Big(A_{k\ell} - \hat{R}_{k\ell} - Q_k^{(t)}\Big)\Big]. \end{aligned} \tag{17}$$

Notice that with unit learning rates, the update rule for each parameter does not depend on its value in the previous iteration. While these properties help us simplify our theoretical analysis, tuning of the learning rate is highly recommended in practice to achieve a reasonable convergence speed.

3.2 Hyperparameter tuning

In this section, we propose a procedure for tuning the hyperparameters $\{\lambda_{1k}\}_{k=1}^K, \lambda_2, \{\alpha_{2k}\}_{k=1}^K$, and $\{\alpha_{1k\ell}\}_{\mathcal{I}}$ used in Algorithm 1. Our approach is inspired by the commonly used edge cross-validation method of Li et al. (2020).

At first glance, the total number of hyperparameters that need to be tuned in Algorithm 1 may seem overwhelming. However, there are several important considerations that significantly simplify our cross validation procedure. Note that λ_{1k} with $\{\alpha_{1k\ell}\}_{\ell=1}^{m_k}$ only appear in (10) and λ_2 with $\{\alpha_{2k\ell}\}_{\ell=1}^{m_k}$ only in (11), which means we can tune these groups of hyperparameter separately. Since each group is still too large for tuning, we can fix $\alpha_{1k} := \alpha_{1k\ell}$ for $(k, \ell) \in \mathcal{I}$ and $\alpha_2 := \alpha_{2k}$ for $k = 1, \dots, K$. Then, it is enough to tune only two parameters in each of the subproblems. As a simpler alternative, it is also possible to optimize only the more influential $\lambda_2, \{\lambda_{1k}\}_{k=1}^K$ parameters appearing in all the nuclear norm penalties of the corresponding problems. In this case, $\{\alpha_{1k\ell}\}_{\mathcal{I}}$ and $\{\alpha_{2k}\}_{k=1}^K$ are assigned following the theoretically optimal values derived for the Gaussian edge entry distribution in Corollary 4.1. In our simulations, this approach produced errors comparable to the more expensive tuning of two hyperparameters per subproblem, so we use it as the default option in our implementation.

To tune the hyperparameters of the k-th subproblem in the first stage, we propose sample a random subset of "training" edge entries in layers of the k-th group:

$$\mathcal{A}_{train}^{(k)} \subset \mathcal{A}^{(k)} := \{A_{k\ell,ij} : \ell = 1, \dots, m_k,\ 1 \le i \le j \le n\},$$

solve Problem (10) with log-likelihood terms in $\mathcal{L}_k$ restricted to edge entries in $\mathcal{A}_{train}^{(k)}$ and then evaluate the non-penalized likelihood (9) on the remaining "test" entries $\mathcal{A}_{test}^{(k)} = \mathcal{A}^{(k)} \setminus \mathcal{A}_{train}^{(k)}$.

This procedure is repeated for each combination of hypeparameters in the predefined grid, with the test log-likelihood possibly averaged across multiple cross-validation folds for stability. Similarly, for tuning λ_2 and $\alpha_2 := \alpha_{2k},\ k = 1, \ldots, K$ in Problem (11), we can randomly sample a subset of edge entries from all the layers

$$\mathcal{A}_{train} \subset \mathcal{A} := \{A_{k\ell,ij} : k = 1, \ldots, K,\ \ell = 1, \ldots, m_k,\ 1 \le i \le j \le n\},$$

solve the problem with log-likelihood terms restricted to $\mathcal{A}_{train}$, and evaluate the non-penalized log-likelihood (8) on $\mathcal{A}_{test} = \mathcal{A} \setminus \mathcal{A}_{train}$.

3.3 Parameter initialization

In this section, we discuss different ways to define the initialization procedures *FirstStageInit* and *SecondStageInit* in Algorithm 1. Due to the convexity of the problems (10) and (11), we expect initialization to impact only the number of iterations needed for convergence but not the algorithm's attaining the global optimum. From the perspective of our theoretical framework, initialization is more crucial as our consistency result relies on the initializer being sufficiently close to the ground truth.

For the Gaussian edge-entry model, we propose to initialize the shared component of a collection of layers as their average and the individual components as the difference between the layer and the shared component initializer. When the ranks of the latent components are known or preliminarily estimated, we can also truncate the initializers, as defined in (15). In our case, due to the first identifiability assumption in Proposition 2.1, the rank of the total shared component for layers in group k is $d_0 + d_k$, implying the initializer:

$$(S + Q_k)^{(0)} = \Big[\frac{1}{m_k}\sum_{\ell=1}^{m_k} A_{k\ell}\Big]_{d_0+d_k}, \quad R_{k\ell}^{(0)} = \Big[A_{k\ell} - (S + Q_k)^{(0)}\Big]_{d_{k\ell}}, \quad \ell = 1, \ldots, m_k. \tag{18}$$

For exponential family distributions with a non-linear link function g, we propose first getting a proxy of $\Theta_{k\ell}$ by apply g^{-1} to each layer $A_{k\ell}$ element-wise, then truncate the result to $[-5, 5]$, and then average the corresponding transformations instead of the layers themselves.

Alternatively, one can use the recently developed "shared space hunting" (SSH) approach (see Algorithm 1 in Tian et al. (2024)). In contrast to the averaging initialization, this method is computationally more expensive as it requires first estimating the latent positions $X_{k\ell}$ of the nodes in each layer and thus their latent ranks $D_{k\ell}$ in particular. Moreover, our simulations (not presented in this work, but available on GitHub) suggest that the SSH can produce very poor initializers if the latent dimensions are inaccurately estimated, which turned out to be the case for small n and large ground-truth ranks. On the other hand, even when n is large and ranks are small, the averaging approach appears to produce only marginally worse results than SSH. So due to its robustness and low computational cost, we use the averaging initializer as our default option.

For the *SecondStageInit*, we have more options to choose from since it can use the first-stage estimates $\{\widehat{S + Q_k}\}_{k=1}^K$ and $\{\hat{R}_{k\ell}\}_{\mathcal{I}}$ besides the layers $\{A_{k\ell}\}_{\mathcal{I}}$ themselves. The natural candidates for the initializer of the shared component S would be to average the estimates $\widehat{S + Q_k}$

$$S_{\text{sq}}^{(0)} = \Big[\frac{1}{K}\sum_{k=1}^{K} \widehat{S + Q_k}\Big]_{d_0} \tag{19}$$

or the group-wise mean residuals between the layer and the individual component estimate

$$S_{\text{resid}}^{(0)} = \Big[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell} - \hat{R}_{k\ell})\Big]_{d_0}. \tag{20}$$

Similarly to the first-stage initializer in (18), each layer $A_{k\ell}$ in (20) has to be preliminary transformed by the inverse link function g^{-1} and then truncated for non-Gaussian distributions. Alternatively, the SSH approach can be used for the initial estimation of the shared component S from either $\{\widehat{S+Q}_k\}_{k=1}^K$ or $\left\{\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell}-\hat{R}_{k\ell})\right\}_{k=1}^K$.

In our implementation of Algorithm 1, we used (20) as the second-stage initializer, as it draws a more direct parallel to the MultiNeSS fitting procedure per Proposition 3.1. However, empirically, we observed that both (20) and (19) imply almost indistinguishable convergence speeds for Problem (11). To give a possible theoretical explanation to this observation, we derive an explicit relationship between the corresponding terms in the two initializing averages under the Gaussian edge entry distribution and a mild extra assumption on the self-loop distribution. In particular, the following proposition demonstrates that in the Gaussian case, $\widehat{S+Q}_k$ is the soft-thresholded version of the averaged residuals $\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell}-\hat{R}_{k\ell})$. The proof of this proposition can be found in Section B.2 of the Appendix.

Proposition 3.2. *Assume that the edge entry distribution is $f(\cdot;\theta,\sigma)=\mathcal{N}(\theta,\sigma^2)$ for the layers' off-diagonal entries and $\mathcal{N}(\theta,2\sigma^2)$ for the diagonal entries (loops). Then, the estimates produced by Problem* (10) *are related as follows:*

$$\widehat{S+Q}_k = \mathcal{T}_{2\sigma^2\lambda_{1k}/m_k}\Big[\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell}-\hat{R}_{k\ell})\Big]. \tag{21}$$

3.4 Refitting step

The drawback of using nuclear norm penalization in optimization is the excessive shrinkage it applies to the non-zero eigenvalues of the fitted matrices as a "payment" for recovering their eigenvalue supports. To reconstruct the correct magnitudes of the non-zero eigenvalues after solving Problems (10) and (11), we use the *refitting* procedure developed by Mazumder et al. (2010) for these purposes.

We begin by describing the *FirstStageRefit*, that is, the refitting step for Problem (10). Consider the eigen-decompositions of the solutions to Problem (10):

$$\widehat{S+Q}_k = \hat{\tilde{V}}_{0k}\hat{\Gamma}_{0k}\hat{\tilde{V}}_{0k}^\top, \quad \hat{R}_{k\ell} = \hat{\tilde{U}}_k\hat{\Gamma}_{k\ell}\hat{\tilde{U}}_{k\ell}^\top, \quad \text{where} \quad \ell=1,\dots,m_k.$$

Element-wise, we have

$$[\widehat{S+Q}_k]_{ij} = \sum_{r=1}^{\widehat{d_0+d_k}} \gamma_r(\widehat{S+Q}_k)\hat{\tilde{V}}_{0k,ir}\hat{\tilde{V}}_{0k,jr} \quad i=1,\dots,n;\ j=1,\dots n,$$

where $\widehat{d_0+d_k}$ is the estimated rank of $\widehat{S+Q}_k$, and $\gamma_r(\cdot)$ denotes the r-th eigenvalue of the input matrix, ordered by magnitude. We can write the element-wise decomposition of $\hat{R}_{k\ell}$ similarly. Since $\widehat{S+Q}_k$ and $\{\hat{R}_{k\ell}\}_{\ell=1}^{m_k}$ are expected to be low-rank up to numerical precision (due to the nuclear norm penalization), in our implementation, we compute the ranks by counting the number of eigenvalues with magnitude above a small positive threshold, set to 10^{-6}. Fixing the eigenvectors corresponding to such eigenvalues, *FirstStageRefit* solves the following convex

problem with variables $\hat{\Gamma}_{0k}$ and $\{\hat{\Gamma}_{k\ell}\}_{\ell=1}^{m_k}$:

$$\min \Big\{ -\sum_{\ell=1}^{m_k} \sum_{i \le j} \log f\Big(A_{k\ell,ij};\ \sum_{r=1}^{\widehat{d_0+d_k}} \gamma_r(\widehat{S+Q_k})\hat{\tilde{V}}_{0k,ir}\hat{\tilde{V}}_{0k,jr}\ + \sum_{r=1}^{\hat{d}_{k\ell}} \gamma_r(\hat{R}_k)\hat{\tilde{U}}_{k\ell,ir}\hat{\tilde{U}}_{k\ell,jr},\ \phi\Big)\Big\} \tag{22}$$

Similarly, consider the eigen-decompositions of the solutions to Problem (11) that are truncated up to the first $\hat{d}_0$ and $\hat{d}_k, k = 1, \dots, K$ eigenvectors, respectively:

$$\hat{S} = \hat{\tilde{V}}\hat{\Gamma}_0\hat{\tilde{V}}^\top, \quad \hat{Q}_k = \hat{\tilde{W}}_k\hat{\Gamma}_k\hat{\tilde{W}}_k^\top, \quad \text{where} \quad k = 1, \dots, K.$$

Then, *SecondStageRefit* solves the following convex problem with variables $\hat{\Gamma}_0, \{\hat{\Gamma}_k\}_{k=1}^K$ and the refitted estimates $\{\hat{R}_{k\ell}\}_{\mathcal{I}}$ of the individual components kept fixed:

$$\min \Big\{ -\sum_{k=1}^{K} \sum_{\ell=1}^{m_k} \sum_{i \le j} \log f\Big(A_{k\ell,ij}; \sum_{r=1}^{\hat{d}_0} \gamma_r(\hat{S})\hat{\tilde{V}}_{ir}\hat{\tilde{V}}_{jr}\ + \sum_{r=1}^{\hat{d}_k} \gamma_r(\hat{Q}_k)\hat{\tilde{W}}_{k,ir}\hat{\tilde{W}}_{k,jr}\ + \hat{R}_{k\ell,ij},\ \phi\Big)\Big\}. \tag{23}$$

If f is from a one-parameter exponential family as in (2), *FirstStageRefit* is equivalent to fitting a GLM with $\widehat{d_0+d_k} + \sum_{\ell=1}^{m_k} \hat{d}_{k\ell}$ predictors (no intercept) and $n(n+1)m_k/2$ responses. In turn, *SecondStageRefit* is equivalent to fitting a GLM with $\hat{d}_0 + \sum_{k=1}^K \hat{d}_k$ predictors and $n(n+1)M/2$ responses. Notice that the first-stage GLMs are fitted without any intercept, while the second-stage GLM is fitted with a fixed vector of observation-dependent offsets $\hat{R}_{k\ell,ij}$.

4 Theory for Gaussian edge-entry distribution

This section establishes theoretical guarantees for the convergence of Algorithm 1 with unit learning rates $\eta_1 = \eta_2 = 1$ under the Gaussian edge entry distribution $f(\cdot; \theta, \sigma) = \mathcal{N}(\theta, \sigma^2)$. Throughout this section, we assume parameters M, K, d_0 as well as m_k, d_k, and $d_{k\ell}$ for all $(k, \ell) \in \mathcal{I}$, are possibly increasing functions of n. The variance σ^2 is assumed to be known and fixed. Proofs of all statements in this section can be found in Section B.3 of the Appendix.

We begin by introducing additional notation. For $a, b \in \mathbb{R}$, denote $a \vee b = \max(a, b)$. For real-valued sequences g_n, h_n , we write $g_n \asymp h_n$ if for sufficiently large n, there exist constants $c_1, c_2 > 0$ such that $c_2 h_n \leqslant g_n \leqslant c_1 h_n$, and we write $g_n \lesssim h_n$ if there is a constant $c > 0$ such that eventually $g_n \le c h_n$. For soft thresholds in the update rules of Remark 5, we will use the following notations for brevity

$$\rho_{1k} := \lambda_{1k}/m_k, \quad \rho_{1k\ell} := \lambda_{1k}\alpha_{1k\ell}, \quad \rho_2 := \lambda_2/M, \quad \rho_{2k} := \lambda_2\alpha_{2k}/m_k, \qquad \text{for} \quad (k, \ell) \in \mathcal{I}.$$

Notice that the original threshold parameters $\{\lambda_{1k}\}_{k=1}^K, \lambda_2, \{\alpha_{2k}\}_{k=1}^K, \{\alpha_{1k\ell}\}_{\mathcal{I}}$ are uniquely reconstructed from the newly defined ones. For symmetric matrices $Z_1, Z_2 \in \mathbb{R}^{n\times n}$ with eigen-decompositions $Z_1 = V_1\Gamma_1V_1^\top$ and $Z_2 = V_2\Gamma_2V_2^\top$, we will denote the cosine similarity between their eigenspaces as

$$\cos(Z_1, Z_2) = \|V_1^\top V_2\|_2.$$

This definition depends on the choice of the orthogonal basis in column spaces $\mathrm{col}(Z_1)$ and $\mathrm{col}(Z_2)$, but can be alternatively reformulated in a basis-free fashion as

$$\cos(Z_1, Z_2) = \max_{x \in \mathrm{col}(Z_1),\ y \in \mathrm{col}(Z_2)} \frac{|x^\top y|}{\|x\|_2 \|y\|_2}.$$

For convenience, we also introduce the following maximal angles between every pair of latent component types.

$$\begin{aligned}
s_{v,w} &= \max_{1 \le k \le K} \cos(S, Q_k) \\
s_{v,u}^{(k)} &= \max_{1 \le \ell \le m_k} \cos(S, R_{k\ell}), && k = 1, \dots, K, \\
s_{w,u}^{(k)} &= \max_{1 \le \ell \le m_k} \cos(Q_k, R_{k\ell}), && k = 1, \dots, K. \\
s_{u,u}^{(k)} &= \max_{1 \le \ell_1 < \ell_2 \le m_k} \cos(R_{k\ell_1}, R_{k\ell_2}), && k = 1, \dots, K, \\
s_{u,u} &= \max_{(k_1,\ell_1) \ne (k_2,\ell_2)} \cos(R_{k_1\ell_1}, R_{k_2\ell_2}), \\
s_{w,w} &= \max_{1 \le k_1 < k_2 \le K} \cos(Q_{k_1}, Q_{k_2})
\end{aligned} \tag{24}$$

These quantities will further appear in the derived error bounds on the latent components, providing the intuition that accurate separation of latent spaces is possible only when their cosine similarity is not too large.

Before we proceed to formulate the main result, we would like to highlight the advantageous properties of the Gaussian edge-entry distribution, which will be crucial for our theoretical framework. The first property is the additive noise structure of the Gaussian distribution, allowing us to conveniently separate each layer as the sum of a deterministic mean $\Theta_{k\ell}$ and a symmetric centered noise matrix $E_{k\ell} \in \mathbb{R}^{n \times n}$ with $E_{k\ell,ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$ for $1 \le i \le j \le n$ and $(k, \ell) \in \mathcal{I}$:

$$A_{k\ell} = \Theta_{k\ell} + E_{k\ell} = (S + Q_k + R_{k\ell}) + E_{k\ell}. \tag{25}$$

Secondly, Gaussian matrices enjoy a convenient concentration bound on their operator norm as well as the averages of their independent copies. In particular, our theoretical analysis will be restricted to a highly probable event where all individual errors $E_{k\ell}, (k, \ell) \in \mathcal{I}$, their groupwise averages $\bar{E}_k = \frac{1}{m_k} \sum_{\ell=1}^{m_k} E_{k\ell}, k = 1, \dots, K$, and total average $\bar{E} = \frac{1}{M} \sum_{(k,\ell) \in \mathcal{I}} E_{k\ell}$ are bounded as follows:

$$\mathcal{E}_{noise} := \left\{ \|E_{k\ell}\|_2 \le 3\sigma\sqrt{n},\ (k,\ell) \in \mathcal{I};\ \|\bar{E}_k\|_2 \le 3\sigma\sqrt{n/m_k},\ k = 1, \dots, K;\ \|\bar{E}\|_2 \le 3\sigma\sqrt{n/M} \right\} \tag{26}$$

In Lemma B.1, we will prove $\mathbb{P}(\mathcal{E}_{noise}) \ge 1 - (M + K + 1)ne^{-C_0 n}$ for a universal constant $C_0 > 0$. Importantly, both the additive noise and operator norm concentration properties are satisfied for any sub-Gaussian distribution, suggesting a natural relaxation of the Gaussian assumption. This will be discussed in more detail in Remark 7.

Next, we formulate several assumptions needed to establish consistency. Our first assumption states that the group sizes are asymptotically of the same order.

Assumption 1. *For each group $k = 1, \dots, K$, it holds $m_k \asymp M/K$, that is, the proportion of layers in each group is asymptotically of order $1/K$.*

For consistency, we naturally want the signal strength to dominate the noise asymptotically. Since in the set $\mathcal{E}_{noise}$, we have $\|E_{k\ell}\|_2 \le 3\sigma\sqrt{n}$, the natural constraint on signal strength rate

would be to require the singular values of $\Theta_{k\ell}$ to have asymptotic order n^τ for some $\tau > 1/2$. This leads us to the following asymptotic regime:

Assumption 2. *There are constants $0 < b_S < B_S$, $0 < b_Q < B_Q$, $0 < b_R < B_R$ and $\tau \in (1/2, 1]$, such that*

$$\begin{aligned} &b_S n^\tau \le |\gamma_{d_0}(S)| \le |\gamma_1(S)| = \|S\|_2 \le B_S n^\tau, \\ &b_Q n^\tau \le |\gamma_{d_k}(Q_k)| \le |\gamma_1(Q_k)| = \|Q_k\|_2 \le B_Q n^\tau \qquad k = 1, \dots, K, \\ &b_R n^\tau \le |\gamma_{d_{k\ell}}(R_{k\ell})| \le |\gamma_1(R_{k\ell})| = \|R_{k\ell}\|_2 \le B_R n^\tau \quad k = 1, \dots, K,\ \ell = 1, \dots, m_k. \end{aligned} \tag{27}$$

To make our theoretical analysis applicable to both averaging and SSH initialization, we do not specify which method we use for *FirstStageInit* and *SecondStageInit* in Algorithm 1, but assume that these are deterministic functions of the input with generic error rates for the produced initializers. These error rates are assumed to be dominated by the corresponding ground truth parameter rates in Assumption 2 to emphasize that the initializers are sufficiently better than random initialization.

Assumption 3. *There exist deterministic $o(n^\tau)$ functions $\{r_k^{(I)}\}_{k=1}^K, r^{(II)}$ that dominate the errors of the first- and second-stage initializers on $\mathcal{E}_{noise}$ almost surely, that is, the event*

$$\mathcal{E}_{init} := \left\{ \|S + Q_k - (S+Q_k)^{(0)}\|_2 \le r_k^{(I)} \text{ for } k = 1, \dots, K; \quad \|S - S^{(0)}\|_2 \le r^{(II)} \right\} \tag{28}$$

satisfies $\mathbb{P}\big(\mathcal{E}_{init} \mid \mathcal{E}_{noise}\big) = 1$ with n sufficiently large.

In Theorem 4.1, we state explicit conditions for the averaging initializers in (18) and (20) to have $r_k^{(I)}, k = 1, \dots, K$ and $r^{(II)}$ of order $\sqrt{n}$. As an alternative, Theorem 1 of Tian et al. (2024) provides different conditions under which the Frobenius norm (and thus the operator norm) of the initializer's error is of order $\sqrt{n}\log n$. We do not present these conditions in this work and refer the reader to the result of Tian et al. (2024). With Assumptions 1, 2, and 3 at hand, we are ready to state the following main result.

Theorem 4.1. *Suppose the edge-entry distribution is $f(\cdot; \theta, \sigma) = \mathcal{N}(\theta, \sigma^2)$ with known σ^2. Then under Assumptions 1, 2, 3, with probability greater than $1 - (M + K + 1)ne^{-C_0 n}$, where $C_0 > 0$ is a universal constant, and for n sufficiently large, the estimates produced by Algorithm 1 with learning rates $\eta_1 = \eta_2 = 1$ and omitted refitting step satisfy*

$$\|\hat{R}_{k\ell} - R_{k\ell}\|_F \le 4 d_{k\ell}^{1/2} \rho_{1k\ell}, \qquad \|\hat{Q}_k - Q_k\|_F \le 4 d_k^{1/2} \rho_{2k}, \qquad \|\hat{S} - S\|_F \le 4 d_0^{1/2} \rho_2,$$

if the parameters $\{\rho_{1k\ell}\}_{\mathcal{I}}, \{\rho_{1k}\}_{k=1}^K, \{\rho_{2k}\}_{k=1}^K$, and ρ_2 have the following rates with multiplicative constants dependent on $(B_R, b_R, B_Q, b_Q, B_S, b_S, \sigma)$ only:

$$\begin{aligned} &\rho_{1k\ell} \asymp r_k^{(I)} \vee n^{1/2}, && (k,\ell) \in \mathcal{I} \\ &\rho_{1k} \asymp \max_{1 \le \ell \le m_k} \rho_{1k\ell} \Big[\frac{1}{m_k} \vee s_{u,u}^{(k)} \vee n^{-\tau} \max_{1 \le \ell \le m_k} \rho_{1k\ell} \Big]^{1/2}, && k = 1, \dots, K \\ &\rho_{2k} \asymp r^{(II)} \vee \rho_{1k}, && k = 1, \dots, K \\ &\rho_2 \asymp \max_{1 \le k \le K} \rho_{2k} \Big[\frac{1}{K} \vee s_{w,w} \vee n^{-\tau} \max_{1 \le k \le K} \rho_{2k} \Big]^{1/2} \vee \max_{(k,\ell) \in \mathcal{I}} \rho_{1k\ell} \Big[\frac{1}{M} \vee s_{u,u} \vee n^{-\tau} \max_{(k,\ell) \in \mathcal{I}} \rho_{1k\ell} \Big]^{1/2}. \end{aligned} \tag{29}$$

If additionally S, Q_k, and $R_{k\ell}$ are positive semi-definite (PSD), then $\hat{S}, \hat{Q}_k, \hat{R}_{k\ell}$ are also PSD.

Remark 6. An important computational advantage of Algorithm 1 is the convexity of the underlying optimization problems. From a theoretical perspective, we would hope that the convexity assumption would allow us to demonstrate latent component consistency independently of the initialization procedure. However, the structure of the derived error bounds in Theorem 4.1 implies that we still have to impose some restrictions on the initialization error to establish consistency. The main restriction disallowing the free choice of the initializer is the alternating structure of the latent component updates, which is hard to analyze after more than one gradient descent step and thus requires that the initializer is already a sufficiently good proxy of the ground truth. As for future work, it would be interesting to find a way to prove the result while avoiding the additional initialization restrictions.

Remark 7. Our proof of Theorem 4.1 essentially relies on the three important properties of the Gaussian edge-entry distribution: (i) the convenient form of the proximal gradient updates arising from the quadratic structure of the Gaussian log-likelihood, (ii) its additive noise property stated in (25), and (iii) a concentration result of Bandeira and van Handel (2016) used to establish in (26). Importantly, the same work (see Corollary 3.3) generalizes this concentration bound to the sub-Gaussian case, showing that if the first property is satisfied, we would be able to generalize Theorem 4.1 to the case of an arbitrary sub-Gaussian edge-entry distribution. This, however, can be resolved by running Algorithm 1 with the mean square loss instead of the log-likelihood in (8), so that the update rule perfectly mimics the one stated in Remark 5 for the Gaussian log-likelihood. In particular, this consideration implies consistency for an RDPG-like binary edge model with $\Theta_{k\ell} \in [0,1]^{n\times n}, (k,\ell) \in \mathcal{I}$ due to sub-Gaussianity of the Bernoulli distribution.

Importantly, Theorem 4.1 does not assume particular rates for any of the key GroupMultiNeSS parameters such as $r_k^{(I)}$, $r^{(II)}$, m_k, or K, thus making our theoretical analysis applicable to a wide range of scenarios. For instance, we can derive the conditions under which Algorithm 1 produces estimates matching the error rates in the oracle scenario where each parameter $S, \{Q_k\}_{k=1}^K, \{R_{k\ell}\}_{\mathcal{I}}$ is estimated with the knowledge of its rank and the values of all remaining parameters:

$$\begin{aligned}
\hat{S}^{(oracle)} &= \Big[\frac{1}{M}\sum_{\mathcal{I}}(A_{k\ell} - Q_k - R_{k\ell})\Big]_{d_0},\\
\hat{Q}_k^{(oracle)} &= \Big[\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell} - S - R_{k\ell})\Big]_{d_k},\\
\hat{R}_{k\ell}^{(oracle)} &= [A_{k\ell} - S - Q_k]_{d_{k\ell}}.
\end{aligned} \tag{30}$$

With (25), the oracle estimators can be respectively rewritten as $\big[S+\bar{E}\big]_{d_0}$, $\big[Q_k+\bar{E}_k\big]_{d_k}$, and $\big[R_{k\ell}+E_{k\ell}\big]_{d_{k\ell}}$, showing that their errors are dominated by the rates for the corresponding noise components in (26). We state sufficient conditions ensuring that Algorithm 1 achieves the oracle rates in the following straightforward corollary of Theorem 4.1.

Corollary 4.1 (Oracle rate conditions)**.** *Let assumptions of Theorem 4.1 hold together with* $s_{u,u} \lesssim n^{1/2-\tau}$, $r^{(II)} \lesssim (n/M)^{1/2}$, *and* $r_k^{(I)} \lesssim n^{1/2}$, $m_k \lesssim n^{\tau-1/2}$ *for each* $k = 1,\dots,K$. *Then, we have*

$$\|\hat{R}_{k\ell} - R_{k\ell}\|_F \le C_R d_{k\ell}^{1/2} n^{1/2}, \quad \|\hat{Q}_k - Q_k\|_F \le C_Q d_k^{1/2} n^{1/2} m_k^{-1/2}, \quad \|\hat{S} - S\|_F \le C_S d_0^{1/2} n^{1/2} M^{-1/2},$$

if with a sufficiently large constants $c_1, c_2 > 0$, *the hyperparameters are set as*

$$\lambda_{1k} = c_1\sqrt{n m_k}, \quad \alpha_{1k\ell} = 1/\sqrt{m_k}, \quad \lambda_2 = c_2\sqrt{nM}, \quad \alpha_{2k} = \sqrt{m_k/M}, \qquad (k,\ell) \in \mathcal{I}. \tag{31}$$

We conclude this section by establishing the error rates for the averaging initializers of $\{S+Q_k\}_{k=1}^K$ in the first stage and of S in the second stage.

Proposition 4.1. *Assume $m_k \to \infty$ for every $k = 1, \ldots, K$ and that there is a sufficient separation between the minimal eigenvalue of S and the largest eigenvalue of Q_k's, that is, there exists $\delta > 0$ such that*

$$\frac{|\gamma_{d_S}(S)|}{\max_k |\gamma_1(Q_k)|} \geq \frac{b_S}{B_Q} > (1+\delta)|1+2s_{v,w}|\Big[\frac{1}{K} + s_{w,w}\Big]^{1/2}. \tag{32}$$

Then, under no extra assumptions, with n sufficiently large, the first- and second-stage averaging initializers in (18) *and* (20)*, respectively, satisfy the following error bounds conditional on $\mathcal{E}_{init}$:*

$$\|(S+Q_k) - (S+Q_k)^{(0)}\|_2 \lesssim n^{\tau}(s_{w,u}^{(k)} \vee s_{v,u}^{(k)})\big[m_k^{-1} \vee s_{u,u}^{(k)}\big]^{1/2}, \tag{33}$$

$$\|S - S^{(0)}\|_2 \lesssim n^{\tau} s_{v,w}\big[K^{-1} \vee s_{w,w}\big]^{1/2} \tag{34}$$

In particular, $r_k^{(I)} \lesssim \sqrt{n}$ and $r^{(II)} \lesssim (n/M)^{1/2}$ as required in Corollary 4.1 if $s_{u,u}, s_{w,u}^{(k)}, s_{v,u}^{(k)} \lesssim n^{1/2-\tau}, m_k \lesssim n^{\tau-1/2}, s_{v,w} \lesssim n^{1/2-\tau}(K/M)^{1/2}$, and $s_{w,w} \lesssim K^{-1}$.

Remark 8. Notice that the condition in (32) essentially requires a small overlap between the spectrum of S and Q_k's. This condition, is always satisfied asymptotically when $K \to \infty$, but, should be introduced when the number of groups K is assumed finite and fixed. On the other hand, in this scenario, the condition on $s_{w,w}$ is no longer required.

5 Synthetic networks experiments

In this section, we empirically study the properties of Algorithm 1 in various synthetic scenarios. In particular, in Section 5.2, we demonstrate how the quality of fitting the GroupMultiNeSS model depends on key parameters such as the size of the networks, the number of layers, and the similarities of the latent components defined in (24). In Section 5.3, we compare GroupMultiNeSS with other models for multiplex networks and demonstrate its advantage over existing methods if a latent group structure is present in the layers.

5.1 Experiment setup

For layer generation in all subsequent synthetic experiments, we assume that the ranks $\{d_k\}_{k=0}^K$, $\{d_{k\ell}\}_{\mathcal{I}}$ of all latent components are the same and denote them by d. We also only consider the standard Euclidean inner product as the embedding similarity measure κ, so that the GroupMultiNeSS latent space decomposition in (6) has all $\{q_k\}_{k=0}^K$ and $\{q_{k\ell}\}_{\mathcal{I}}$ equal to zero. To generate latent components with varying pairwise maximum cosine similarities, we use the procedure described in Algorithm 2, inspired by a similar sampling idea in Tian et al. (2024). For simplicity of notation, we do not allow $s_{u,u}^{(k)}, s_{v,u}^{(k)}$, and $s_{w,u}^{(k)}$ to vary over $k = 1, \ldots, K$, and denote them respectively as

$$s_{u,u} = \max_{k=1,\ldots,K} s_{u,u}^{(k)}, \qquad s_{w,u} = \max_{k=1,\ldots,K} s_{w,u}^{(k)}, \qquad \text{and} \qquad s_{v,u} = \max_{k=1,\ldots,K} s_{v,u}^{(k)}.$$

The proposed sampling approach ensures that all d columns of each generated latent component and the columns with distinct indices of any two distinct latent components are orthogonal. On the contrary, columns with identical indices for any two components would have a fixed similarity

Algorithm 2 Latent component sampling procedure

1: **Input:** number of nodes n, latent dimension d, group sizes $\{m_k\}_{k=1}^K$, maximum cosine angles $s_{v,w}, s_{v,u}, s_{w,w}, s_{w,u}, s_{u,u}$ (all zero by default)
2: Collect all angles into a single matrix:

$$\Omega = \begin{pmatrix} 1 & s_{v,w}\mathbf{1}_K^\top & s_{v,u}\mathbf{1}_M^\top \\ s_{v,w}\mathbf{1}_K & \Sigma_K(s_{w,w}) & s_{w,u}\mathbf{1}_K\mathbf{1}_M^\top \\ s_{v,u}\mathbf{1}_M & s_{w,u}\mathbf{1}_M\mathbf{1}_K^\top & \Sigma_M(s_{u,u}) \end{pmatrix} \tag{35}$$

with $\Sigma_m(s)$ denoting an $m \times m$ matrix with ones on the diagonal and s everywhere else.
3: Generate initial latent positions

$$\tilde{L} = [\tilde{V}, \tilde{W}_1, \ldots, \tilde{W}_K, \tilde{U}_{11}, \ldots, \tilde{U}_{1m_1}, \ldots, \tilde{U}_{K1}, \ldots, \tilde{U}_{Km_K}] \in \mathbb{R}^{n \times d(1+K+M)}$$

with entries sampled independently from the standard normal distribution
4: Compute the Gram matrices of the initialized and target latent positions, that is, respectively, $\tilde{\mathcal{G}} = \frac{1}{n}\tilde{L}^\top\tilde{L}$ and $\mathcal{G} = \Omega \otimes I_d$
5: Compute final latent positions

$$\tilde{L}\tilde{\mathcal{G}}^{-1/2}\mathcal{G}^{1/2} = [V, W_1, \ldots, W_K, U_{11}, \ldots, U_{1m_1}, \ldots, U_{K1}, \ldots, U_{Km_K}]$$

6: **Output:** $S = VV^\top, Q_k = W_kW_k^\top, R_{k\ell} = U_{k\ell}U_{k\ell}^\top$ for $(k,\ell) \in \mathcal{I}$.

depending only on the types of the two input components. In particular, this implies that this procedure is valid only if $d(1+K+M) \le n$, as otherwise the number of angle constraints is larger than the number of available degrees of freedom n.

When fitting the GroupMultiNeSS model with Algorithm 1, we use a five-fold edge cross-validation to tune $\{\lambda_{1k}\}_{k=1}^K$ and λ_2 as described in Section 3.2, so that the train and test sets within each fold satisfy $|\mathcal{A}_{train}| = 0.8|\mathcal{A}|$ and $|\mathcal{A}_{test}| = 0.2|\mathcal{A}|$. For hyperparameter tuning, we define the grid $c_\lambda \in [0.03, 0.1, 0.3, 1, 3, 10]$ and use the oracle rates in (31) as multiplicative constants for these ranges, that is, we tune λ_{1k} over the grid $c_\lambda\sqrt{nm_k}$ and λ_2 over $c_\lambda\sqrt{nM}$. In all further experiments, we use the oracle values for $\{\alpha_{1k\ell}\}_{\mathcal{I}}$ and $\{\alpha_{2k}\}_{k=1}^K$. To track the convergence of the algorithm, after each iteration, we note the relative difference between the current loss and its best value achieved so far, that is, compute $1 - \mathcal{L}_k^{(t)}/\min_{\tau<t}\mathcal{L}_k^{(\tau)}$ and $1 - \mathcal{L}^{(t)}/\min_{\tau<t}\mathcal{L}^{(\tau)}$ in the first and second stages, respectively. We stop if this difference has not exceeded the tolerance 10^{-5} during the last ten steps. The learning rates used in Algorithm 1 are set to $\eta_1 = \eta_2 = 1$ for the Gaussian model and to $\eta_1 = \eta_2 = 3$ for the logistic.

Throughout this section, we measure the quality of a matrix estimate using the *Relative Frobenius Error (RFE)* defined for a ground-truth matrix $Z \in \mathbb{R}^{n\times n}$ and its compatible estimate $\hat{Z}$ as

$$\mathrm{RFE}(Z, \hat{Z}) := \frac{\|Z - \hat{Z}\|_F}{\|Z\|_F}. \tag{36}$$

If we have multiple matrices $\{Z_\ell\}_{\ell=1}^m$ and we want to evaluate the quality of corresponding estimates $\{\hat{Z}_\ell\}_{\ell=1}^m$ with a single metric value, we use the *Average Relative Frobenius Error (ARFE)* defined as:

$$\mathrm{ARFE}(\{Z_\ell\}_{\ell=1}^m, \{\hat{Z}\}_{\ell=1}^m) := \frac{1}{m}\sum_{\ell=1}^m \mathrm{RFE}(Z_\ell, \hat{Z}_\ell). \tag{37}$$

In our simulations, we will use this metric to measure the average estimation accuracy for a given type of latent components, such as $\{Q_k\}_{k=1}^K$ or $\{R_{k\ell}\}_{\mathcal{I}}$.

All the following simulation studies can be found in our GitHub repository

`https://github.com/AlexanderKagan/GroupmultinessExperiments`.

Python implementations of GroupMultiNeSS (Algorithm 1), MultiNeSS (MacDonald et al., 2022), and Shared Space Hunting with refinement (Tian et al., 2024) fitting procedures, as well as the synthetic data sampler (Algorithm 2), are available as a `GroupMultiNeSS` package `https://github.com/AlexanderKagan/GroupMultiNeSS`.

5.2 Dependency of the estimation error on the key parameters

In this section, we study how estimation accuracy of the GroupMultiNeSS latent components $S, \{Q_k\}_{k=1}^K, \{R_{k\ell}\}_{\mathcal{I}}$ and the latent space matrices $\{\Theta_{k\ell}\}_{\mathcal{I}}$ depends on the number of nodes n and the number of layers M. Additional experiments showing the error dependency on cosine similarities of the latent components can be found in Section C.1 of the Appendix.

Throughout all simulations within this experiment, we consistently generate latent components using Algorithm 2 with $K = 4$ balanced groups of networks, $s_{v,u} = s_{w,u} = 0.1$, and latent dimensions $d = 3$. Denoting $\Theta_{k\ell} = S + Q_k + R_{k\ell}$ for $(k, \ell) \in \mathcal{I}$, the observed layers are further sampled via

$$A_{k\ell,ij} \overset{\text{ind}}{\sim} \mathcal{N}(\Theta_{k\ell,ij}, 1) \quad \text{or} \quad A_{k\ell,ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\big(\sigma(\Theta_{k\ell,ij})\big), \qquad (k,\ell) \in \mathcal{I}, \quad i \leq j, \tag{38}$$

in case of the Gaussian and Logistic edge-entry distributions, respectively. Figure 2 presents the dependency of the ARFE metric on the number of layers $M \in [8, 16, 24, 32, 40, 48, 56]$ with $n = 200$ and the number of nodes $n \in [100, 200, 300, 400, 500]$ with $M = 16$ for the Gaussian and Logistic models. All four experiments were repeated ten times with different random seeds to assess the empirical standard errors. Ten replications was sufficient for Monte Carlo errors to be negligible, so error bars are omitted in the plots.

As expected from our theoretical analysis, the estimation error of all components improves with the number of nodes n. In contrast, increasing the number of layers M improves S, Q, and Θ but does not significantly improve the individual component error R. This is expected both from the error bound of the individual component in Theorem 4.1 and intuition that the new layers do not add information on the individual component of a given layer $A_{k\ell}$. A slight visual improvement in R can be explained by the fact that new layers improve the estimates of the shared and group components S and Q_k and thus indirectly help separate $R_{k\ell}$ from $S + Q_k$.

As an apparent difference between the Logistic and Gaussian models, we can see that the relative error of Θ dominates all the other components in the Logistic case and lies between R and (Q, S) in the Gaussian case. We explain it by the fact that separating additive latent components is a much easier task under a linear link function than a logistic one. This also explains the much higher relative errors of all components in the logistic case.

5.3 Comparison to other methods

Here, we compare Algorithm 1 with several baseline methods that can be used to fit the GroupMultiNeSS model. The first method we chose for comparison is the MultiNeSS model (MacDonald et al., 2022) fitted with refitting and cross-validation of nuclear norm penalty hyperparameters. In addition, we also compare our optimization approach to its non-convex

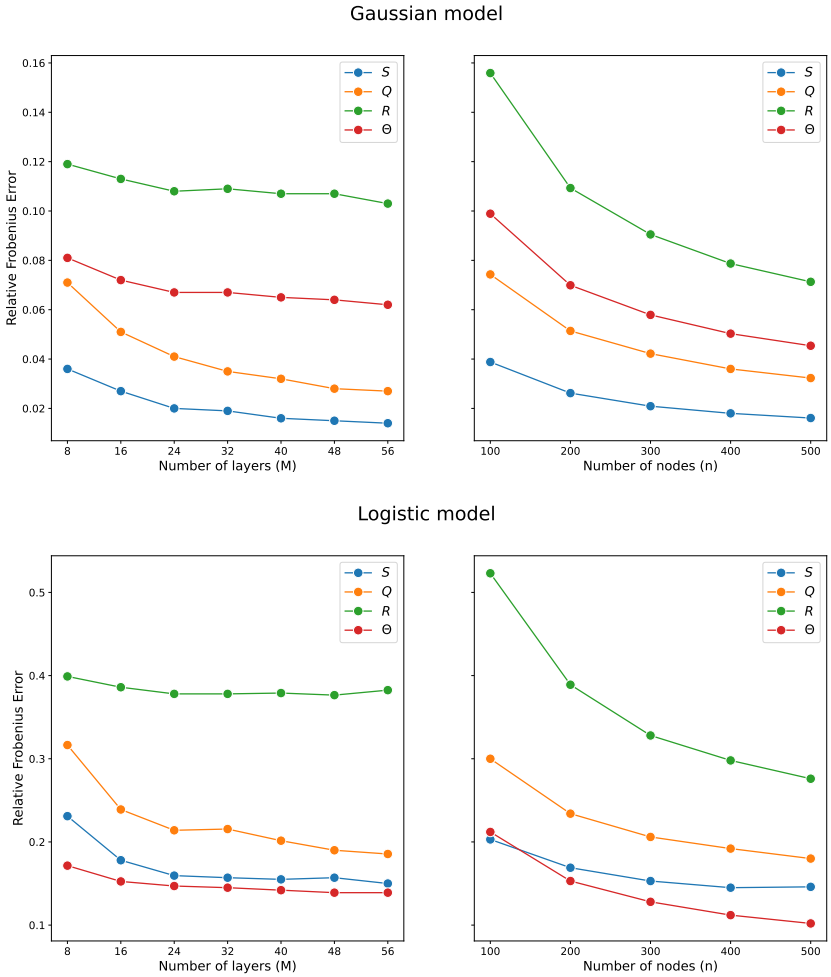

Figure 2: Dependency of ARFE on the key parameters of the Gaussian and Logistic GroupMultiNeSS models. Unless one of the parameters is varied, the networks are generated according to Algorithm 2 with $M = 16$ layers on $n = 200$ nodes, $K = 4$ balanced groups, and $s_{v,w} = s_{w,u} = 0.1$.

oracle version, which substitutes all hyperparameter-dependent soft-thresholding updates by hard-thresholding with the ground-truth latent component ranks as in (14). In particular, for updating $\{S + Q_k\}_{k=1}^K$ in the first stage, it uses the ranks $\{d_0 + d_k\}_{k=1}^K$. The algorithm also does not perform the refitting step after convergence of the alternating steps within each subproblem, since hard-thresholding does not suffer from overshrinking the non-zero eigenvalues.

Finally, we compare our method to the Multiple Adjacency Spectral Embedding (MASE) algorithm, proposed by Arroyo et al. (2021) for fitting the COSIE multiplex network model introduced by the authors in the same work. COSIE provides estimates of the expected adjacency matrices for each layer but does not separate these estimates into common and individual components, meaning we can only assess the accuracy of the overall expectation matrices $\{\Theta_{k\ell}\}_{\mathcal{I}}$. Although COSIE is designed for the RDPG model, it can also be directly applied to the Gaussian model. We consider an oracle version of COSIE that assumes knowledge of the true component

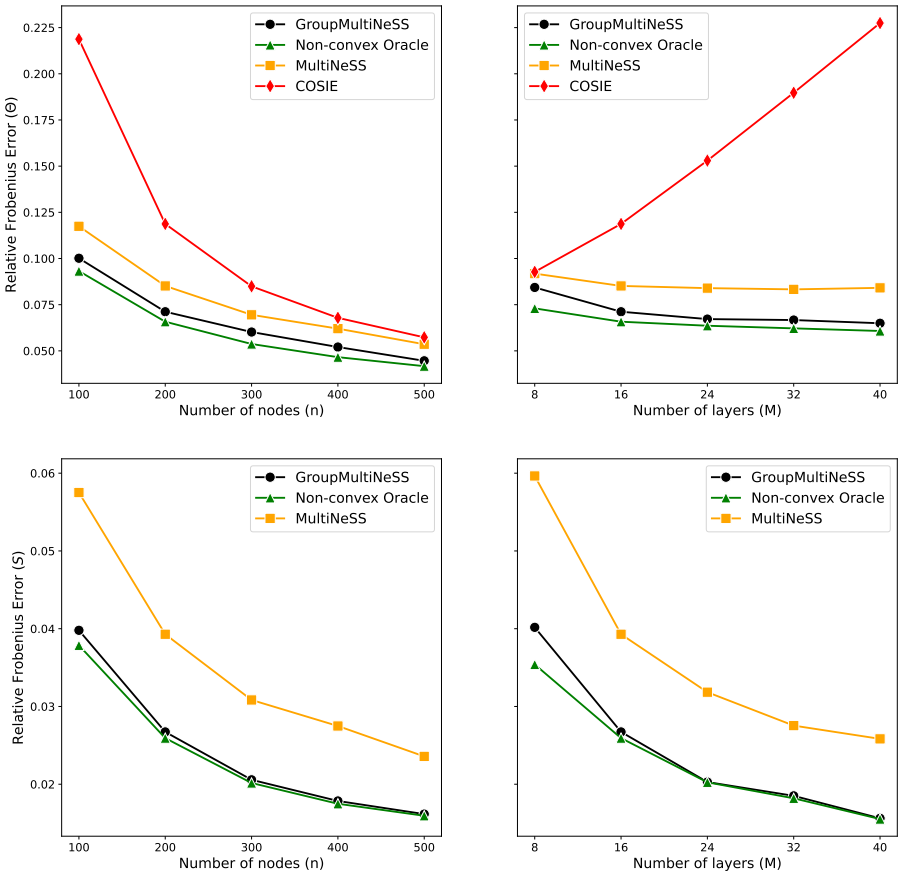

Figure 3: Change in ARFE of Θ (upper row) and S (lower row) with the increase of (left column) the number of nodes n with $M = 16$ and (right column) the number of layers M with $n = 200$. In all simulations, the number of groups is $K = 4$ and the latent dimension is $d = 3$. Layers are sampled from the Gaussian edge-entry model.

ranks $\{d_k\}_{k=0}^K$ and $\{d_{k\ell}\}_{\mathcal{I}}$. To ensure a fair comparison with our method, we first select the leading $d_0 + d_k + d_{k\ell}$, $(k, \ell) \in \mathcal{I}$ eigenvectors for each layer, and then use COSIE to fit a common invariant subspace of dimension $d_0 + \sum_{k=1}^K d_k + \sum_{(k,\ell)\in\mathcal{I}} d_{k\ell}$, corresponding to the total number of latent dimensions in the GroupMultiNeSS model.

For comparison, we fit the four candidate models (GroupMultiNeSS, Non-convex Oracle, MultiNeSS, and COSIE) to a fixed number of network samples ($M = 16$) with an increasing number of nodes $n \in [100, 200, 300, 400, 500]$ and to a fixed number of node $n = 200$ with an increasing number of network samples $M \in [8, 16, 24, 32, 40]$. In both experiments, we set $K = 4$

groups and sample the ground-truth latent positions according to Algorithm 2 with $d = 3$ and balanced group sizes. The observed network layers are further sampled from the Gaussian model as in (38). The resulting ARFE of the joint latent space matrices Θ and the RFE of the shared component S are presented, respectively, in the first and second rows of Figure 5.3. Similar experiments for the Logistic model are postponed to Section C.2 of Appendix.

We can see that the GroupMultiNeSS fitting procedure consistently matches the rate of the oracle estimator, especially for larger values of n and M. In terms of accurate estimation of Θ, MultiNeSS is also doing a good job with ARFE being roughly 20% higher than GroupMultiNeSS. Importantly, the difference in the accuracy of S estimation is much more apparent with MultiNeSS being far behind GroupMultiNeSS and the oracle. This observation aligns with the plots in the lower row of Figure 2, showing that the "total" latent space Θ can be accurately estimated even when the algorithm poorly separates its components from each other.

In contrast, the COSIE model demonstrates a more significant error improvement as the number of nodes increases, but reaches the performance level of its competitors only at $n = 500$ with $M = 16$. Contrary to other models, the accuracy of the COSIE model deteriorates as the number of layers M increases. This behavior is a well-known property of this model, related to the fact that the MASE algorithm starts by estimating the layers' joint latent dimension $d_0 + \sum_k d_k + \sum_{\mathcal{I}} d_{k\ell}$, a task with a complexity growing as the number of layers or groups increases.

6 Application to the Parkinson's disease brain networks

To demonstrate the practical utility of the proposed GroupMultiNeSS model, we analyze a publicly available functional connectivity dataset originally curated by Badea et al. (2017). It consists of resting-state fMRI data from 40 subjects, 17 women and 23 men aged between 57 and 75 years, among whom 20 patients are diagnosed with Parkinson's disease and 20 are healthy controls. The functional brain network of each subject is represented by a Pearson correlation matrix computed across the 116 regions defined by the AAL116 atlas (Tzourio-Mazoyer et al., 2002). According to this atlas, these 116 regions are distributed across several major brain lobes: 28 in the frontal lobe, 26 in the cerebellum, 14 in the occipital lobe, 14 in the parietal lobe, 12 in the limbic system, 12 in the temporal lobe, 8 in subcortical gray matter (SCGM), and 2 in the insula.

Our further analysis will attempt to use the GroupMultiNeSS model to identify the latent structure differences between the diseased and control groups. With the original edge entries representing correlations, it is convenient for us to put them onto the real line range to mimic the Gaussian edge-entry distribution. We do that by applying the Fisher z-transformation to off-diagonal elements in each layer. As the unit diagonal elements are uninformative, we further ignore them during the optimization by assuming the input networks to be free of self-loops. So, following the notation of the present work, we encode this dataset as $M = 40$ symmetric adjacency matrices $A_{k\ell} \in \mathbb{R}^{n\times n},\ k \in \{1, 2\},\ \ell = 1, \dots, 20$ on $n = 116$ nodes, separated into $K = 2$ groups of sizes $m_1 = m_2 = 20$.

Before proceeding to model fitting, we need to account for additional network covariates, such as patients' age and sex, as they can confound the group effect of interest. A possible solution to that would be to consider more network groups, for example, by additionally stratifying over sex and/or age group. However, this would lead to too small group sample sizes, which would not allow for accurate extraction of the group latent positions. Therefore, we propose using a standard neuroscience technique, which suggests regressing out the effects of additional covariates separately for each edge entry. That is, for every $1 \le i < j \le n$, we consider the linear model

$$A_{k\ell,ij} = \beta_{0,ij} + \beta_{\text{age},ij} \cdot \text{Age}_{k\ell} + \beta_{\text{sex},ij} \cdot \mathbf{1}[\text{Sex}_{k\ell} = \text{Male}] + \varepsilon_{k\ell,ij},$$

where $\text{Age}_{k\ell}$ and $\text{Male}_{k\ell}$ represent the (k,ℓ)-patient's age and indicator of being male, respectively. We then fit the coefficients of the (i,j)-th model with ordinary least squares (OLS) and calculate the residuals via:

$$\hat{\varepsilon}_{k\ell,ij} = A_{k\ell,ij} - \hat{\beta}_{0,ij} - \hat{\beta}_{\text{age},ij} \cdot \text{Age}_{k\ell} - \hat{\beta}_{\text{sex},ij} \cdot \mathbf{1}[\text{Sex}_{k\ell} = \text{Male}]$$

and further apply GroupMultiNeSS to the residual matrices $\hat{A}^{(r)}_{k\ell} = [\hat{\varepsilon}_{ij}]_{i,j=1}^{n}, (k,\ell) \in \mathcal{I}$ under the assumption of the Gaussian edge-entry distribution.

For model fitting, we used Algorithm 1 coupled with edge cross-validation to tune the hyperparameters $\{\lambda_{1k}\}_{k=1}^{K}$ and λ_2. After fitting the latent components $\hat{S}, \{\hat{Q}_k\}_{k=1}^{K}$, and $\{\hat{R}_{k\ell}\}_{\mathcal{I}}$, we extracted the latent positions $\hat{V}, \{\hat{W}_k\}_{k=1}^{K}$, and $\{\hat{U}_{k\ell}\}_{\mathcal{I}}$ using ASE as described at the beginning of Section 3. In Figure 4, we plot the brain regions in the three leading latent dimensions of $\{\hat{W}_k\}_{k=1}^{K}$ with regions (nodes) colored according to the brain lobe they lie in. For better visual comparison, we align the embeddings of the two groups by applying the appropriate three-dimensional right rotation (obtained using Procrustes alignment) to the PD latent positions. The three leading dimensions in both components appeared to be disassortative, which roughly means that two regions with coordinate vectors forming a smaller angle tend to exhibit less connection strength. Together, the three dimensions explain roughly 43% and 47% of the variance (sum of all singular values), respectively. To simplify visual interpretability, we only plot 94 regions corresponding to the five most represented lobes in the AAL116 atlas (frontal, occipital, parietal, temporal, and cerebellum). While plotting the shared component may also be of interest, we do not include it here as it appeared less visually interpretable. However, as we discuss later, estimating it is crucial to eliminate the common latent patterns from the group components, thereby simplifying their visual comparison. The plots of the first four leading dimensions of the shared as well as the first two dimensions of the group components can be found in Section C.3 of the Appendix.

Comparing the two node embedding plots in Figure 4, we can observe a more apparent clustering of the cerebellum (purple) and occipital (orange) regions in the right and left plots, respectively, which potentially indicates their less coherent connectivity (due to the latent dimensions being disassortative). Weaker connections in the cerebellum lobe (responsible for balance and muscle control) are expected, as these mechanisms are most commonly impaired in the presence of Parkinson's. The fact that the PD group exhibits less clustering of the occipital regions (responsible for visual reception/interpretation) may suggest the presence of functional changes in the visual cortex of diseased patients. This aligns with previous evidence of occipital dysfunction in PD patients (Weil et al., 2016), linking structural changes within the visual cortex to impaired visual processing, spatial disorientation, and visual hallucinations. As another possible explanation, Göttlich et al. (2013) suggests that abnormal connectivity in the visual network may be related to adaptation and compensation processes as a consequence of altered motor function.

Except for clustering of the occipital regions in the left plot, it is also worth noting their stronger collocation with the yellow temporal (responsible for memory and hearing) and green frontal lobe regions (responsible for motor functions) in the right plot, suggesting a weaker connectivity between these lobes in patients with Parkinson's. This phenomenon is well-known in the literature (Lucas-Jiménez et al., 2016; Baggio et al., 2014) and is usually associated with visuospatial impairment and cognitive decline in patients with PD.

Finally, in contrast to the left plot, we can observe complete separation between the cerebellum and frontal/temporal lobe regions in the PD group, suggesting increased connectivity between these areas. This has been observed in the literature (Tessitore et al., 2019) and is widely treated as a compensatory mechanism aimed at maintaining motor performance in the presence of basal ganglia dysfunction.

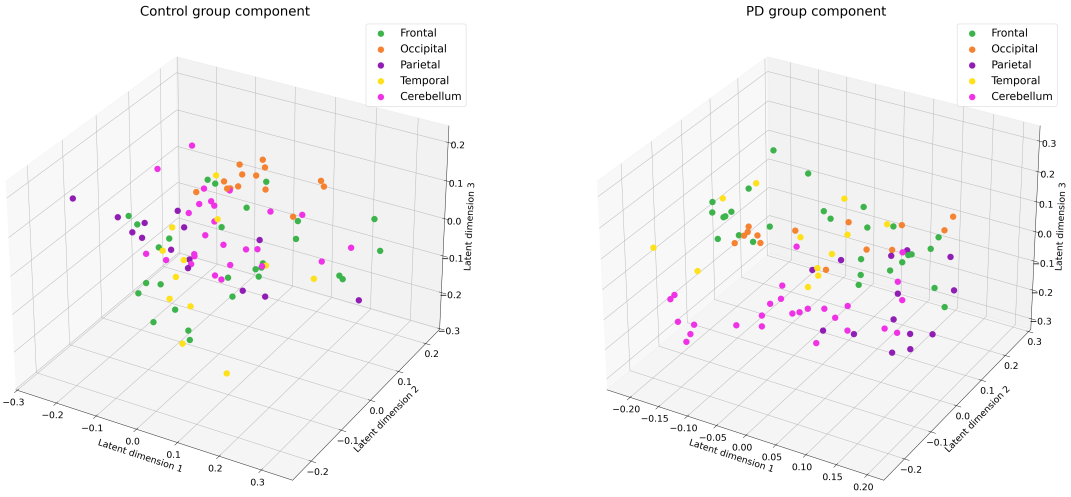

Figure 4: Group latent positions fitted with GroupMultiNeSS and plotted in the leading three latent dimensions (all disassortative).

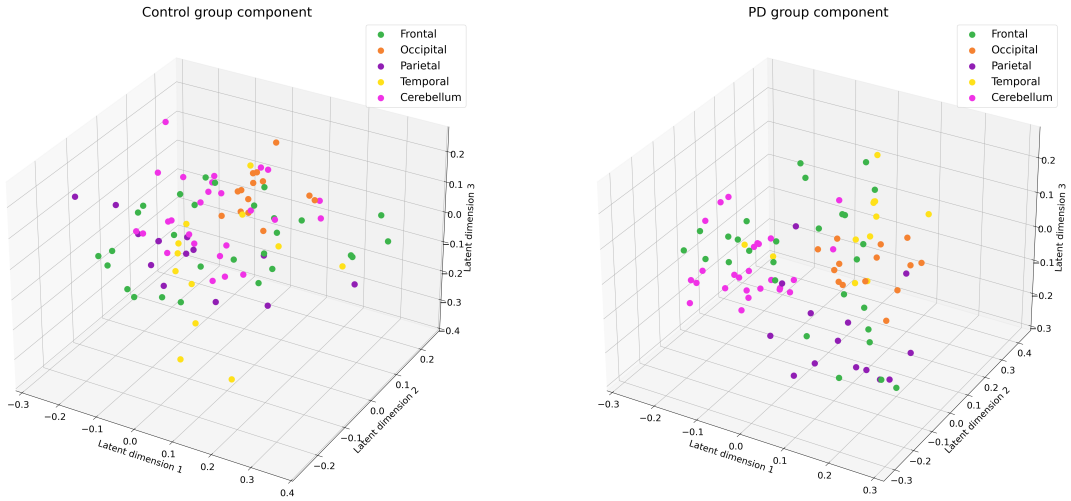

Figure 5: Group latent positions obtained by fitting separate MultiNeSS models on the layers of the two groups and plotted in the leading three latent dimensions (all disassortative).

To evaluate the quality of the GroupMultiNeSS group embeddings, we also compare them to the embeddings obtained by applying ASE to $\widehat{S+Q_k},\ k=1,K$ fitted during the first stage of Algorithm 1. The obtained latent positions then coincide with those one would get by running two separate MultiNeSS models (with refitting and cross-validation to tune λ_{1k}) on the layers of the control and PD groups, respectively. Due to a more similar latent structure caused by the common S term, we expect the corresponding latent position matrices to be less helpful in uncovering group-wise biological differences than the GroupMultiNeSS embeddings that essentially eliminated the common latent component during the second stage of Algorithm 1. This is exactly

what we see in Figure 5. Compared to Figure 4, we can see that both groups exhibit clustering of the occipital regions and their mixing with temporal (yellow) and frontal (green) regions. Additionally, the cerebellum (purple) regions in the PD group demonstrate weaker clustering and less separation from the frontal regions.

Finally, to quantify the connectivity differences between the two groups using the group components extracted by the GroupMultiNeSS model, we employ the following approach. For each pair of lobes (a, b), we consider all possible pairs $(r_1, r_2), r_1 \in a, r_2 \in b$ of their distinct regions and compute their unnormalized pairwise cosine similarity in the latent space of each group:

$$h_k^{(a,b)}(r_1, r_2) = \hat{W}_{k,r_1}^\top \hat{W}_{k,r_2}, \qquad k = 1, 2.$$

The similarity average $\bar{h}_k^{(a,b)}$ across all region pairs gives us a proxy for the lobes' connectivity in group $k = 1, 2$. Then the group-wise difference in the average connectivity of lobes (a, b) can be measured as $\bar{h}_2^{(a,b)} - \bar{h}_1^{(a,b)}$, where we assume $k = 1$ is the control and $k = 2$ is the PD group. The heatmaps of such cosine similarity differences across all pairs of lobes are presented in Figure 6. Note that a positive difference corresponds to an increased average cosine similarity in the latent space of the PD group compared to the control, meaning lower lobe connectivity due to the disassortativeness of the latent dimensions. To develop intuition about the significance of these values, we perform permutation testing by shuffling the group labels 100 times. For each run, we repeat our estimation procedure to extract the group latent positions and note the average similarity difference for each of the 15 lobe pairs. For each pair, we then construct the empirical distribution from the obtained 100 differences, calculate the p-value of the true difference based on this distribution, and apply the Benjamini-Hochberg correction to adjust for multiple testing. If an adjusted p-value appears below 0.01, 0.05, or 0.1, we respectively put ***, **, or * next to the corresponding difference value in the heatmap.

In Figure 6, we can see some additional quantitative support for the observations previously made based on the visual analysis of Figure 4. Although these differences lack high statistical significance, we still can observe some moderate changes in occipital-occipital, cerebellum-cerebellum, cerebellum-occipital, and frontal-cerebellum lobe connectivities. Importantly, several additional lobe interactions, such as temporal-temporal, temporal-parietal, and parietal-parietal, can be hypothesized from this analysis at a very high significance level, even after the multiple-testing correction.

7 Discussion

In this work, we proposed a latent space model for grouped multiplex networks with shared structure. We developed a fully adaptive procedure for estimating the latent components within this model and provided a theoretical analysis of the algorithm's performance under the Gaussian edge-entry distribution. Importantly, the proposed estimation and theoretical frameworks can be directly extended to multiplex network models assuming a more complex hierarchical structure of the network layers, for example, by adding a subgroup level to the GroupMultiNeSS tree in Figure 1. This consideration gives us hope that the proposed latent space framework can lay the basis for a universal modeling approach to complex multiplex networks. Despite its universality in decomposing the latent spaces of network layers into hierarchically structured components, our framework still has several important limitations that should be addressed in future developments of our work to achieve a truly universal multiplex network model.

In Remark 7, we demonstrate how the Gaussianity assumption used in our theoretical analysis can be relaxed to the sub-Gaussian case under the assumption of the linear link function g. Unfortunately, the chosen theoretical framework appears to be hardly generalizable to non-linear

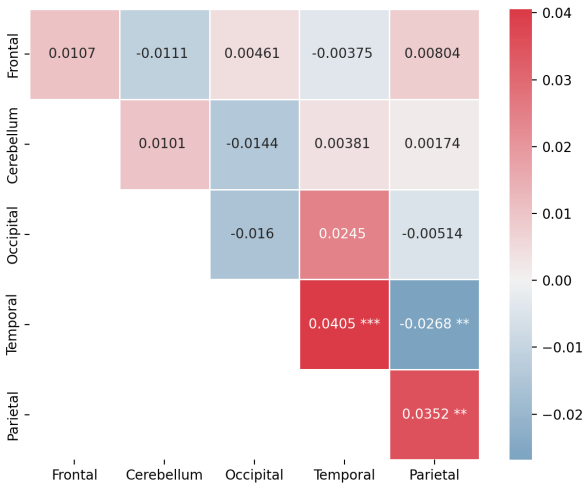

Figure 6: Each cell of the heatmaps represents the difference in the average cosine similarity of the two corresponding lobes' regions. The differences are computed between the PD and Control group latent components extracted by the GroupMultiNeSS model. Asterisks represent significance with respect to the permutation testing.

link functions. A natural extension of our work would be to apply the more general framework of Tian et al. (2024) to establish consistency for a broader class of exponential edge distribution models. Another important assumption of the GroupMultiNeSS model is that the group labels of layers are observed. A useful generalization would be to extend the current framework to the case of unknown (and possibly mixed) group memberships.

On a different note, the current framework is only applicable to undirected networks due to the symmetric inner product structure of the latent embeddings. It would be of interest to further generalize the MultiNeSS and GroupMultiNeSS models to the case of directed networks and propose a methodology for their estimation and theoretical analysis.

Finally, with the proposed model being able to extract the group-wise latent components from the network layers, a natural subsequent step would be to propose a more powerful non-permutation-based testing procedure that would be able to evaluate if the difference between these components is statistically significant.

References

Arroyo, J., A. Athreya, J. Cape, G. Chen, C. Priebe, and J. Vogelstein (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *Journal of machine learning research 22*, 1–49.

Athreya, A., D. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. Sussman, M. Tang, J. Vogelstein, and C. Priebe (2017). Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research 18*.

Badea, L., M. Onu, T. Wu, A. Roceanu, and O. Bajenaru (2017). Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLOS ONE 12*(11), 1–21.

Baggio, H.-C., R. Sala-Llonch, B. Segura, M.-J. Marti, F. Valldeoriola, Y. Compta, E. Tolosa, and C. Junqué (2014). Functional brain networks and cognitive deficits in Parkinson's disease. *Hum. Brain Mapp. 35*(9), 4620–4634.

Bandeira, A. S. and R. van Handel (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability 44*(4).

Cai, T. and A. Zhang (2016). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics 46*.

D'Angelo, S., T. Murphy, and M. Alfo (2018). Latent space modeling of multidimensional networks with application to the exchange of votes in eurovision song contest. *The Annals of Applied Statistics 13*.

Fithian, W. and R. Mazumder (2018). Flexible low-rank statistical modeling with missing data and side information. *Statistical Science 33*, 238–260.

Ghaderi, A. H., M. A. Nazari, H. Shahrokhi, and A. H. Darooneh (2017). Functional brain connectivity differences between different ADHD presentations: Impaired functional segregation in ADHD-combined presentation but not in ADHD-inattentive presentation. *Basic Clin. Neurosci. 8*(4), 267–278.

Gollini, I. and T. Murphy (2013). Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics 25*.

Göttlich, M., T. F. Münte, M. Heldmann, M. Kasten, J. Hagenah, and U. M. Krämer (2013). Altered resting state brain networks in Parkinson's disease. *PLoS One 8*(10), e77336.

Han, Q., K. S. Xu, and E. M. Airoldi (2015). Consistent estimation of dynamic and multi-layer block models. arXiv:1410.8597.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*(460), 1090–1098.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social Networks 5*(2), 109–137.

Jones, A. and P. Rubin-Delanchy (2021). The multilayer random dot product graph. arXiv:2007.10455.

Koltchinskii, V., A. Tsybakov, and K. Lounici (2010). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics 39*.

Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

Lucas-Jiménez, O., N. Ojeda, J. Peña, M. Díez-Cirarda, A. Cabrera-Zubizarreta, J. C. Gómez-Esteban, M. Á. Gómez-Beldarrain, and N. Ibarretxe-Bilbao (2016). Altered functional connectivity in the default mode network is associated with cognitive impairment and brain anatomical changes in Parkinson's disease. *Parkinsonism Relat. Disord. 33*, 58–64.

Ma, Z., Z. Ma, and H. Yuan (2020). Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research 21*(4), 1–67.

MacDonald, P. W., E. Levina, and J. Zhu (2022). Latent space models for multiplex networks with shared structure. *Biometrika 109*(3), 683–706.

MacDonald, P. W., E. Levina, and J. Zhu (2024). Mesoscale two-sample testing for network data. arXiv:2410.17046.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research 11*, 2287–2322.

Nguen, C. K., O. H. M. Padilla, and A. A. Amini (2024). Network two-sample test for block models. arXiv:2406.06014.

Rubin-Delanchy, P., C. Priebe, and M. Tang (2017). The generalised random dot product graph. arXiv:1709.05506.

Salter-Townshend, M. and T. H. McCormick (2017). Latent space models for multiview network data. *The Annals of Applied Statistics 11*(3), 1217–1244.

Sosa, J. and B. Betancourt (2021). A latent space model for multilayer network data. arXiv:2102.09560.

Tang, M., A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics 26*(2), 344–354.

Tessitore, A., M. Cirillo, and R. De Micco (2019). Functional connectivity signatures of Parkinson's disease. *J. Parkinsons. Dis. 9*(4), 637–652.

Tian, Y., J. Sun, and Y. He (2024). Efficient analysis of latent spaces in heterogeneous networks. arXiv:2412.02151.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage 15*(1), 273–289.

Weil, R. S., A. E. Schrag, J. D. Warren, S. J. Crutch, A. J. Lees, and H. R. Morris (2016). Visual dysfunction in Parkinson's disease. *Brain 139*(11), 2827–2843.

Wu, S., P. Zhan, G. Wang, X. Yu, H. Liu, and W. Wang (2024). Changes of brain functional network in alzheimer's disease and frontotemporal dementia: a graph-theoretic analysis. *BMC Neurosci. 25*(1), 30.

Young, S. J. and E. R. Scheinerman (2007). Random dot product graph models for social networks. *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, 138–149.

Zhang, X., S. Xue, and J. Zhu (2020). A flexible latent space model for multilayer networks. *Proceedings of the 37th International Conference on Machine Learning*.

A Alternative optimization approaches

In this section, we discuss several optimization alternatives to Algorithm 1 and the reasons for choosing it as our main fitting approach. All simulations referred to in this section are not included in the manuscript and can be found on our GitHub.

The first seemingly more natural alternative to optimizing the joint log-likelihood in (8) would be to combine it with a nuclear norm penalty for each variable type $S, Q_k, R_{k\ell}$ and update all the variables in a single loop. Compared to this alternative, an important advantage of Algorithm 1 is that the loop over groups $k = 1, \dots, K$ in the first stage can be run in parallel. The price we pay for distributing the group-specific problems is the appearance of the additional second-stage optimization task. However, in practice, we expect the second stage to be much less computationally expensive than each first-stage problem. This is because the number of parameter matrices to optimize in the second stage is $K + 1$, a number typically much smaller than $m_k + 1, k = 1, \dots, K$ in the first stage, as there are usually more layers in each group than the number of groups. Empirically, we observed that this approach is not only significantly slower in terms of CPU time and the number of SVD computations required but also exhibits much less stable convergence. We explain this with the intuition that separating only two types of components at a time (first, $S + Q_k$ and $R_{k\ell}$, then S and Q_k) results in much less undesirable component mixing than separating all three types at once.

Another reasonable alternative would be to substitute the joint log-likelihood optimized in the second stage of Algorithm 1 with the quadratic discrepancies between $S + Q_k$ and $\widehat{S + Q_k}$ estimated within the first stage:

$$\mathcal{L}_{sq}(S, \{Q_k\}_{k=1}^K) = \sum_{(k,\ell)\in\mathcal{I}} \|S + Q_k - \widehat{S + Q_k}\|_F^2.$$

The theoretical results of Propositions 3.1 and 3.2 suggest that this optimization problem would imply very similar results under the Gaussian edge-entry distribution. Our simulations supported this claim but showed that likelihood-based optimization provides much better results for non-Gaussian distributions, such as the logistic model, suggesting that accounting for the form of the edge-entry distribution is useful.

Finally, for the first stage problems (essentially fitting a MultiNeSS model ot each group), there is a fundamentally different nonconvex approach inspired by the recent work of Tian et al. (2024), which pre-estimates the latent component ranks via the Shared Space Hunting (SSH) approach and then fixes the ranks of the updated parameter matrices during the gradient descent. For GroupMultiNeSS, we can do that by substituting the soft thresholding updates with the hard thresholding as defined in (14). The key advantage of this method is that it does not require using cross-validation for tuning the thresholding hyperparameters, thus being more appealing computationally, and that it estimates the ranks in a likelihood-free fashion, making the procedure general for arbitrary link functions. On the negative side, the hard singular value thresholds, theoretically derived by the authors to determine the latent ranks, are only guaranteed to produce accurate rank estimates asymptotically and often lead to unstable behavior in low n or high ground-truth rank regimes. With SSH providing inaccurate rank estimates, Algorithm 1 significantly outperforms this alternative version, while showing comparable or only marginally worse results when SSH perfectly estimates the ranks. In other words, compared to SSH tuning of the hard thresholds (latent ranks), cross-validation used to tune the soft thresholds in Algorithm 1 appears to have a more controllable and robust effect on the subsequent optimization procedure.

B Proofs

Here, we present proofs of all the statements made throughout this manuscript.

B.1 Section 2.2 proofs

Proof of Proposition 2.1. By the first condition and Lemma 1 in (MacDonald et al., 2022), for any $k = 1, \dots, K$ and $\ell = 1, \dots, m_k$, we have

$$[V \ W_k \ U_{k\ell}] \ O_{k\ell} = [V' \ W'_k \ U'_{k\ell}] \tag{39}$$

where $O_{k\ell}$ satisfies $S_{k\ell} O_{k\ell} S_{k\ell}^\top \in \mathcal{O}_{p_0+p_k+p_{k\ell},\ q_0+q_k+q_{k\ell}}$ with a permutation matrix

$$S_{k\ell} = \begin{bmatrix} I_{p_0} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{q_0} & 0 & 0 \\ 0 & I_{p_k} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{q_k} & 0 \\ 0 & 0 & I_{p_{k\ell}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{q_{k\ell}} \end{bmatrix}.$$

It is sufficient to demonstrate that $O_{k\ell}$ is block-diagonal for any layer, i.e., we need to verify that

$$O_{k\ell} = \begin{bmatrix} O_{11,k\ell} & O_{12,k\ell} & O_{13,k\ell} \\ O_{21,k\ell} & O_{22,k\ell} & O_{23,k\ell} \\ O_{31,k\ell} & O_{32,k\ell} & O_{33,k\ell} \end{bmatrix} = \begin{bmatrix} O_{11,k\ell} & 0 & 0 \\ 0 & O_{22,k\ell} & 0 \\ 0 & 0 & O_{33,k\ell} \end{bmatrix}$$

with $O_{11,k\ell} \in \mathcal{O}_{p_0,q_0}, O_{22,k\ell} \in \mathcal{O}_{p_k,q_k}, O_{33,k\ell} \in \mathcal{O}_{p_{k\ell},q_{k\ell}}$. We will refer to the components of the r-th column of $O_{k\ell}$ as follows

$$o_{k\ell,r} = \begin{bmatrix} o^{(1)}_{k\ell,r} \\ o^{(2)}_{k\ell,r} \\ o^{(3)}_{k\ell,r} \end{bmatrix} \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_k} \times \mathbb{R}^{d_{k\ell}}$$

Consider the layers (k_1, ℓ_1) and (k_2, ℓ_2) satisfying the third identifiability condition. Writing (39) for them and equating two expressions for V', we obtain for every $r = 1, \dots, d_0$

$$0 = V(o^{(1)}_{k_1\ell_1,r} - o^{(1)}_{k_2\ell_2,r}) + W_{k_1} o^{(2)}_{k_1\ell_1,r} - W_{k_2} o^{(2)}_{k_2\ell_2,r} + U_{k_1\ell_1} o^{(3)}_{k_1\ell_1,r} - U_{k_2\ell_2} o^{(3)}_{k_2\ell_2,r}.$$

By linear independence, it implies that $o^{(2)}_{k_1\ell_1,r} = o^{(3)}_{k_1\ell_1,r} = 0$. Since it holds for any $r = 1, \dots, d_0$, we have $O_{21,k_1\ell_1} = O_{31,k_1\ell_1} = 0$. By (39), it means $V' = V O_{k_1,\ell_1}$ and so for any layer (k, ℓ), we have $O_{21,k\ell} = O_{31,k\ell} = 0$. Using the fact that $S_{k\ell} O_{k\ell} S_{k\ell}^\top$ is an indefinite orthogonal transformation, we can derive that the symmetric upper-diagonal blocks $O_{12,k\ell}, O_{13,k\ell}$ are also zeros.

Now, consider layers s and t in group k satisfying the second identifiability condition. Writing (39) for them and equating two expressions for W'_k, we have for every $r = d_0 + 1, \dots, d_0 + d_k$

$$0 = V(o^{(1)}_{ks,r} - o^{(1)}_{kt,r}) + W_k(o^{(2)}_{ks,r} - o^{(2)}_{kt,r}) + U_{ks} o^{(3)}_{ks,r} - U_{ks} o^{(3)}_{kt,r}.$$

By linear independence, it implies that $o^{(3)}_{ks,r} = o^{(3)}_{kt,r} = 0$. Since it holds for any $r = d_0 + 1, \dots, d_0 + d_k$, we deduce $O_{32,ks} = 0$. Combining with the previously established $O_{12,ks} = 0$ this implies that $W'_k = W_k O_{22,ks}$ and so for any layer (k, ℓ), we have $O_{32,k\ell} = 0$. Since the lower right 2×2 block sub-matrix of $O_{k\ell}$ is also an indefinite orthogonal rotation (up-to conjugation with a permutation matrix), we deduce $O_{23,k\ell} = 0$, which completes the proof.

□

B.2 Section 3 Proofs

Proof of Proposition 3.1. Under the Gaussian distribution, for each $1 \le i \le j \le n$, we can rewrite the joint negative log-likelihood of the (i,j)-th entries (up to $1/2\sigma^2$ multiplier) as

$$\sum_{(k,\ell)\in\mathcal{I}} (A_{k\ell,ij} - S_{ij} - Q_{k,ij} - \hat{R}_{k\ell,ij})^2 = \sum_{k=1}^{K}\Big[m_k\big(S_{ij}+Q_{k,ij}\big)^2 - 2m_k\tilde{A}_{k,ij}(S_{ij}+Q_{k,ij})\Big] + const$$
$$= \sum_{k=1}^{K} m_k\big(S_{ij}+Q_{k,ij}-\tilde{A}_{k,ij}\big)^2 + const,$$

where *const* denotes terms independent of S and Q_k. Therefore, the solution to Problem (11) coincides with the solution of the following program:

$$\min_{S,Q_k}\Big\{\sum_{k=1}^{K}\frac{m_k}{2\sigma^2}\sum_{i\le j}\big[S_{ij}+Q_{k,ij}-\tilde{A}_{k,ij}\big]^2 + \lambda_2\|S\|_* + \sum_{k=1}^{K}\lambda_2\alpha_{2k}\|Q_k\|_*\Big\}.$$

□

Proof of Proposition 3.2. With the assumed edge-entry distribution, Problem (10) solves

$$\widehat{S+Q}_k, \{\hat{R}_{k\ell}\}_{\ell=1}^{m_k} = \underset{S+Q_k, R_{k\ell}}{\operatorname{argmin}} \frac{1}{4\sigma^2}\sum_{\ell=1}^{m_k}\Big\|A_{k\ell}-(S+Q_k)-R_{k\ell}\Big\|_F^2 + \lambda_{1k}\|S+Q_k\|_* + \sum_{\ell=1}^{m_k}\lambda_{1k}\alpha_{1k\ell}\|R_{k\ell}\|_*, \tag{40}$$

where the off-diagonal entries of each matrix inside the Frobenius norm are counted twice compared to (9), leading to an additional $1/2$ multiplier in front of the sum. By optimality, $\widehat{S+Q}_k$ should minimize the target function in (40) over $S+Q_k$ with each $R_{k\ell}$ fixed to $\hat{R}_{k\ell}$, that is, solve the optimization problem

$$\widehat{S+Q}_k = \underset{Z\in\mathbb{R}^{n\times n}}{\operatorname{argmin}}\Big\{\frac{1}{4\sigma^2}\sum_{\ell=1}^{m_k}\big\|(A_{k\ell}-\hat{R}_{k\ell})-Z\big\|_F^2 + \lambda_{1k}\big\|Z\big\|_*\Big\}$$
$$= \underset{Z\in\mathbb{R}^{n\times n}}{\operatorname{argmin}}\Big\{\frac{1}{2}\big\|\frac{1}{m_k}\sum_{\ell=1}^{m_k}(A_{k\ell}-\hat{R}_{k\ell})-Z\big\|_F^2 + \frac{2\sigma^2\lambda_{1k}}{m_k}\big\|Z\big\|_* + const\Big\}$$

where *const* denotes terms independent of Z. By the variational definition of soft-thresholding, the optimal Z in the last problem is achieved at (21). □

B.3 Section 4 Proofs

Throughout this section, we will use the following additional notations. For $a, b \in \mathbb{R}$, denote $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$. For a matrix $V \in \mathbb{R}^{n\times d}$, denote the projector on its column space by $\mathcal{P}_V = V(V^\top V)^\dagger V^\top$ where $(\cdot)^\dagger$ is the Moore-Penrose pseudoinverse. For a symmetric matrix $Z \in \mathbb{R}^{n\times n}$ with the eigen-decomposition $Z = V\Gamma V^\top$, we will assume by default that the eigenvalues in $\Gamma = \operatorname{diag}[\gamma_1(Z),\dots,\gamma_n(Z)]$ are sorted in descending order wrt their absolute value:

$$\|\Gamma\|_2 = |\gamma_1(Z)| \ge \ldots \ge |\gamma_n(Z)|$$

where the smallest (in absolute value) nonzero eigenvalue of Z is denoted by $\gamma_{\min}(Z)$.

In this section, we will use multiple measures of distance between subspaces spanned by the columns of two $n \times d$ orthogonal matrices V and $\hat{V}$. Suppose the singular values of $V^\top \hat{V}$ are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$. Then, following standard notations, we define the $\sin\Theta$-matrix as

$$\sin\Theta(V, \hat{V}) = \operatorname{diag}\left[\sin(\cos^{-1}(\sigma_1)), \cdots, \sin(\cos^{-1}(\sigma_d))\right].$$

and use $\|\sin\Theta(V, \hat{V})\|_2$ to measure the distance. For symmetric matrices $Z, \hat{Z} \in \mathbb{R}^{n\times n}$ with top $d < n$ eigenvectors denoted respectively as $V, \hat{V} \in \mathbb{R}^{n\times d}$, it would be convenient to define:

$$\sin_d(Z, \hat{Z}) := \|\sin\Theta(V, \hat{V})\|_2.$$

Notice that V and $\hat{V}$ (and thus the defined quantity above) are uniquely defined only if there is a gap between the d-th and $(d+1)$-st eigenvalues of Z and $\hat{Z}$ or these two eigenvalues are zero. In all further applications, this will be satisfied for sufficiently large n, as d will denote the rank of Z and $\hat{Z} = \mathcal{T}_\rho(Z + E)$ will be a soft-thresholded perturbation of Z with an additive noise matrix E satisfying $\|E\|_2 = o(\gamma_{\min}(Z))$.

Lemma 1 in Cai and Zhang (2016) establishes the equivalence between the sin distance and several other natural metrics. Here, we only state several useful corollaries from this Lemma that we will further need throughout this section:

$$\|\hat{V}\hat{V}^\top - VV^\top\|_2 \leq 2\|\sin\Theta(\hat{V}, V)\|_2, \tag{41}$$

$$\inf_{O\in\mathcal{O}_d} \|\hat{V} - VO\|_2 \leq \sqrt{2}\|\sin\Theta(\hat{V}, V)\|_2. \tag{42}$$

Another important result we will need further is Theorem 1 of Cai and Zhang (2016), which provides a useful bound on the sin-distance between the eigenspaces of a matrix and its perturbation. Since originally this result was formulated for arbitrary matrices and in this work, we only deal with symmetric matrices, we restate it in Lemma B.2 with less notational load and a slightly loosened but much more algebraically convenient bound.

We start by giving an outline of the proof for Theorem 4.1. The initial step is to rewrite the first iteration update for each parameter in Remark 5 as a soft-thresholding operator applied to the parameter's ground-truth value, say Z, plus the error term, which comprises Gaussian noise and the weighted sum of an update-dependent set of latent components errors:

$$\hat{Z} = \mathcal{T}_\rho(Z + \text{other components' errors} + \text{Gaussian noise}). \tag{43}$$

For example, the first stage update rule for the parameters $S + Q_k$ and $\{R_{k\ell}\}_{\ell=1}^{m_k}$ can be rewritten

$$\begin{aligned} R_{k\ell}^{(1)} &= \mathcal{T}_{\rho_{1k\ell}}\Big[R_{k\ell} - \Delta^{(0)}_{S+Q_k} + E_{k\ell}\Big], \qquad \text{for } \ell = 1, \ldots, m_k, \\ (S+Q_k)^{(1)} &= \mathcal{T}_{\rho_{1k}}\Big[(S+Q_k) - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta^{(1)}_{R_{k\ell}} + \bar{E}_k\Big], \end{aligned} \tag{44}$$

where by $\Delta_Z^{(t)} := Z^{(t)} - Z$ we denote the error of parameter Z after t iterations. The second stage update rule at the first iteration can be written similarly:

$$\begin{aligned} Q_k^{(1)} &= \mathcal{T}_{\rho_{1k}}\Big[Q_k - \Delta_S^{(0)} - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta_{R_{k\ell}} + \bar{E}_k\Big], \qquad k = 1, \ldots, K, \\ S^{(1)} &= \mathcal{T}_{\rho_0}\Big[S - \sum_{k=1}^{K}\frac{m_k}{M}\Delta^{(1)}_{Q_k} - \frac{1}{M}\sum_{(k,\ell)\in\mathcal{I}}\Delta_{R_{k\ell}} + \bar{E}\Big], \end{aligned} \tag{45}$$

where $\Delta_{R_{k\ell}}, (k,\ell) \in \mathcal{I}$ denotes the individual component error after the last iteration of the k-th group updates in the first stage. We do not consider further ($t > 1$) iterations in the update procedures of both stages, since in the theoretical framework we use, it is impossible to demonstrate the rate improvement after more than one iteration update. However, we can demonstrate that the error would not increase after additional iterations. Thus, by induction, the final error rate obtained by running the updates until numerical convergence will match the rate achieved after the first iteration.

The key to our proof will be Lemma B.3 that shows that the soft threshold operator applied to a matrix perturbed by additive error as in (43) should have the threshold ρ with the rate of the total error's spectral norm to guarantee that the estimate $\hat{Z}$ is close to the ground truth Z in the Frobenius norm. In particular, it should dominate both the spectral norm of the other components' errors, which we control using Lemma B.4, and the average of the appropriate Gaussian noise matrices, which we control by restricting the analysis to the set $\mathcal{E}_{noise}$ in (26). Importantly, the latter event occurs with an overwhelming probability according to Lemma B.1.

We postpone the proof of the mentioned technical lemmas to Section B.4 and give a formal proof of Theorem 4.1.

Proof of Theorem 4.1. Assume that the events $\mathcal{E}_{noise}$ in (26) and $\mathcal{E}_{init}$ in (28) occur simultaneously. By Lemma B.4 and Assumption 3, probability of that is at least $1-(M+K+1)ne^{-C_0 n}$ as needed.

First stage (group k): We first establish the properties of $R^{(1)}_{k\ell}$, $\ell = 1, \dots, m_k$ by applying Lemma B.3 with the following correspondence of notations due to (44)

$$E := E_{k\ell} - \Delta^{(0)}_{S+Q_k}, \quad \rho := \rho_{1k\ell}, \quad d := d_{k\ell}.$$

Define $\rho_{1k\ell} = C_1(r^{(I)}_k \vee n^{1/2})$ with $C_1 := 2(1+3\sigma)$, By Assumption 3 and the triangular inequality, we have

$$\|E_{k\ell} - \Delta^{(0)}_{S+Q_k}\|_2 \le r^{(I)}_k + 3\sigma n^{1/2} \le \rho_{1k\ell}/2 \tag{46}$$

Therefore, Properties 1 and 2 imply

$$\|\Delta^{(1)}_{R_{k\ell}}\|_2 \le 2\rho_{1k\ell} \quad \text{and} \quad \|\Delta^{(1)}_{R_{k\ell}}\|_F \le 4\rho_{1k\ell} d^{1/2}_{k\ell}, \tag{47}$$

and by Property 4, if $R_{k\ell}$ is PSD, then $R^{(1)}_{k\ell}$ is also PSD. Combining Corollary B.1 and (46), we have for sufficiently large n that

$$\theta_{k\ell} := \sin_{d_{k\ell}}(R_{k\ell}, R^{(1)}_{k\ell}) \le \frac{\rho_{1k\ell}/2}{b_R n^\tau/2} = \frac{\rho_{1k\ell}}{b_R n^\tau},$$

which implies together with (46) that for sufficiently large n it holds

$$\|E_{k\ell} - \Delta^{(0)}_{S+Q_k}\|_2 + 2\|R_{k\ell}\|_2\theta^2_{k\ell} \le \rho_{1k\ell}/2 + o(\rho_{1k\ell}) < \rho_{1k\ell}. \tag{48}$$

So, by Lemma B.4 with $s := s^{(k)}_{u,u}$, $\theta := \max_{1\le\ell\le m_k}\theta_{k\ell}$, and $\rho := \max_{1\le\ell\le m_k}\rho_{1k\ell}$, we have

$$\Big\|\frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta^{(1)}_{R_{k\ell}}\Big\|_2 \le 11\frac{B_R}{b_R}\max_{1\le\ell\le m_k}\rho_{1k\ell}\Big[\frac{1}{m_k} \vee s^{(k)}_{u,u} \vee \frac{1}{b_R n^\tau}\max_{1\le\ell\le m_k}\rho_{1k\ell}\Big]^{1/2} \tag{49}$$

Now, let us establish an error bound for $\Delta^{(1)}_{S+Q_k}$. With $C_2 := \frac{11(1+3\sigma)B_R}{b_R(1\wedge b_R)}$, define

$$\rho_{1k} = C_2\max_{1\le\ell\le m_k}\rho_{1k\ell}\Big[\frac{1}{m_k} \vee s^{(k)}_{u,u} \vee n^{-\tau}\max_{1\le\ell\le m_k}\rho_{1k\ell}\Big]^{1/2} \tag{50}$$

so that it holds due to $\rho_{1k\ell} \geq 3\sigma(n/m_k)^{1/2}$

$$\Big\|\bar{E}_k - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta^{(1)}_{R_{k\ell}}\Big\|_2 \leq 3\sigma(n/m_k)^{1/2} + \Big\|\frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta^{(1)}_{R_{k\ell}}\Big\| < \rho_{1k}. \tag{51}$$

Then, Properties 1 and 2 in Lemma B.3 imply

$$\|\Delta^{(1)}_{S+Q_k}\|_2 \leq 2\rho_{1k} \quad \text{and} \quad \|\Delta^{(1)}_{S+Q_k}\|_F \leq 4\rho_{1k}\sqrt{d_0+d_k}. \tag{52}$$

Combining (50) and (52), we can see that $\|\Delta^{(1)}_{S+Q_k}\|_2 = o(r_k^{(I)} \vee n^{1/2})$. So, treating $(S+Q_k)^{(1)}$ as a new initializer for $S+Q_k$, we can see that all bounds dependent on it, that is, (46) and (48), are still satisfied. Thus, by induction, the next iteration errors $\Delta^{(t)}_{R_{k\ell}}$ and $\Delta^{(t)}_{S+Q_k}$ for any $t \geq 1$ satisfy the error bounds in (47) and (52), respectively.

Second stage: We first establish the properties of $Q_k^{(1)},\ k = 1, \ldots, K$ by applying Lemma B.3 with the following correspondence of notations due to (45)

$$E := \bar{E}_k - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta_{R_{k\ell}} - \Delta_S^{(0)}, \quad \rho := \rho_{2k}, \quad d := d_k.$$

Define $\rho_{2k} = 4(r^{(II)} \vee \rho_{1k})$, so that by (51) and the triangular inequality, we have

$$\|\bar{E}_k - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta_{R_{k\ell}} - \Delta_S^{(0)}\|_2 \leq \|\bar{E}_k - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta_{R_{k\ell}}\|_2 + \|\Delta_S^{(0)}\|_2 \leq \rho_{1k} + r^{(II)} \leq \rho_{2k}/2. \tag{53}$$

Therefore, by Properties 1 and 2 in Lemma B.3, we have

$$\|\Delta^{(1)}_{Q_k}\|_2 \leq 2\rho_{2k} \quad \text{and} \quad \|\Delta^{(1)}_{Q_k}\|_F \leq 4\rho_{2k}\sqrt{d_k} \tag{54}$$

and by Property 4, if Q_k is PSD, then $Q_k^{(1)}$ is also PSD. Combining Corollary B.1 and (53), we get for sufficiently large n:

$$\theta_k := \sin_{d_k}(Q_k, Q_k^{(1)}) \leq \frac{\rho_{2k}}{b_Q n^\tau},$$

which implies together with (53)that for sufficiently large n it holds

$$\Big\|\bar{E}_k - \frac{1}{m_k}\sum_{\ell=1}^{m_k}\Delta_{R_{k\ell}} - \Delta_S^{(0)}\Big\|_2 + 2\|Q_k\|_2\theta_k^2 \leq \rho_{2k}/2 + o(\rho_{2k}) < \rho_{2k}.$$

So, Lemma B.4 with $s := s_{w,w}$, $\theta := \max_{1\leq k\leq K}\theta_k$, and $\rho := \max_{1\leq k\leq K}\rho_{2k}$, we have for sufficiently large n such that $\pi_{\max} \geq Km_k/M$ (guaranteed by Assumption 1):

$$\Big\|\frac{1}{M}\sum_{k=1}^{K} m_k\Delta^{(1)}_{Q_k}\Big\|_2 \leq \pi_{\max}\Big\|\frac{1}{K}\sum_{k=1}^{K}\Delta^{(1)}_{Q_k}\Big\|_2 \leq \frac{11B_Q\pi_{\max}}{b_Q(1\wedge b_Q)}\max_{1\leq k\leq K}\rho_{2k}\Big[\frac{1}{K}\vee s_{w,w}\vee n^{-\tau}\max_{1\leq k\leq K}\rho_{2k}\Big]^{1/2}.$$

Now, let us establish an error bound for $\Delta_S^{(1)}$. Notice that (49) holds for the average of any iteration errors $\Delta^{(t)}_{R_{k\ell}}$, and thus also for the average of errors $\Delta_{R_{k\ell}}$ output at the end of the first stage. Thus, by Lemma B.4 with $s := s_{u,u}, \theta := \max_{(k,\ell)\in\mathcal{I}}\theta_{k\ell}$, and $\rho := \max_{(k,\ell)\in\mathcal{I}}\rho_{1k\ell}$:

$$\Big\|\frac{1}{M}\sum_{(k,\ell)\in\mathcal{I}}\Delta_{R_{k\ell}}\Big\|_2 \leq 11\frac{B_R}{b_R}\max_{(k,\ell)\in\mathcal{I}}\rho_{1k\ell}\Big[\frac{1}{M}\vee s_{u,u}\vee\frac{1}{b_R n^\tau}\max_{(k,\ell)\in\mathcal{I}}\rho_{1k\ell}\Big]^{1/2}$$

With $C_3 := 12\big(\frac{B_Q \pi_{\max}}{b_Q(1\wedge b_Q)} + \frac{B_R}{b_R(1\wedge b_R)}\big)$, define

$$\rho_2 = C_3\Big(\max_{1\le k\le K}\rho_{2k}\big[K^{-1}\vee s_{w,w}\vee n^{-\tau}\max_{1\le k\le K}\rho_{2k}\big]^{1/2}\vee\max_{(k,\ell)\in\mathcal{I}}\rho_{1k\ell}\Big[\frac{1}{M}\vee s_{u,u}\vee n^{-\tau}\max_{(k,\ell)\in\mathcal{I}}\rho_{1k\ell}\Big]^{1/2}\Big),$$

so that due to $\rho_{1k\ell}/M^{1/2} \ge 3\sigma(n/M)^{1/2} \ge \|\bar{E}\|_2$ it holds:

$$\Big\|\frac{1}{M}\sum_{(k,\ell)\in\mathcal{I}}\Delta^{(1)}_{R_{k\ell}}\Big\|_2 + \Big\|\frac{1}{M}\sum_{k=1}^{K} m_k\Delta^{(1)}_{Q_k}\Big\|_2 + \|\bar{E}\|_2 < \rho_2$$

Then, for sufficiently large n, Properties 1 and 2 in Lemma B.3 imply

$$\|\Delta^{(1)}_S\|_2 \le 2\rho_2 \quad \text{and} \quad \|\Delta^{(1)}_S\|_F \le 4\rho_2\sqrt{d_0}. \tag{55}$$

and by Property 4, if S is PSD, then $S^{(1)}$ is also PSD. □

B.4 Technical lemmas

In this section, we present the proofs of the technical lemmas used to establish the main result of Theorem 4.1 and Proposition 4.1.

Lemma B.1. *Let $E_{k\ell} \in \mathbb{R}^{n\times n}, (k,\ell)\in\mathcal{I}$ be symmetric Gaussian matrices with independent entries following*

$$E_{k\ell,ij} \overset{\text{ind}}{\sim} \mathcal{N}(0,\sigma^2), \qquad 1\le i\le j\le n, \qquad (k,\ell)\in\mathcal{I}.$$

Then, the following concentration bounds hold simultaneously for the operator norms of the individual matrices $E_{k\ell}$, $(k,\ell)\in\mathcal{I}$, their group-wise averages $\bar{E}_k = \frac{1}{m_k}\sum_{\ell=1}^{m_k} E_{k\ell}$, $k=1,\dots,K$, and the grand-total average $\bar{E} = \frac{1}{M}\sum_{(k,\ell)\in\mathcal{I}} E_{k\ell}$:

$$\mathbb{P}\Big(\|E_{k\ell}\|_2 \le 3\sigma\sqrt{n},\ (k,\ell)\in\mathcal{I};\ \|\bar{E}_k\|_2 \le 3\sigma\sqrt{n/m_k};\ \|\bar{E}\|_2 \le 3\sigma\sqrt{n/M}\Big) \ge 1-(M+K+1)ne^{-C_0 n},$$

where $C_0 > 0$ is some universal constant.

Proof. Analogously to Lemma 3 in (MacDonald et al., 2022). □

Lemma B.2 (Version of Theorem 1, (Cai and Zhang, 2016))**.** *Consider symmetric matrices $Z, E \in \mathbb{R}^{n\times n}$. Define the eigen-decompositions of Z and its perturbation $\tilde{Z} = Z+E$ as, respectively,*

$$Z = \begin{bmatrix} V & V_\perp \end{bmatrix}\cdot\begin{bmatrix}\Gamma_1 & 0\\ 0 & \Gamma_2\end{bmatrix}\cdot\begin{bmatrix} V^\top \\ V_\perp^\top\end{bmatrix}$$

and

$$\tilde{Z} = \begin{bmatrix} \tilde{V} & \tilde{V}_\perp \end{bmatrix}\cdot\begin{bmatrix}\tilde{\Gamma}_1 & 0\\ 0 & \tilde{\Gamma}_2\end{bmatrix}\cdot\begin{bmatrix} \tilde{V}^\top \\ \tilde{V}_\perp^\top\end{bmatrix},$$

where $V^\top V = I_d$, $V_\perp^\top V_\perp = I_{n-d}$, $\Gamma_1 = \mathrm{diag}[\gamma_1(Z),\dots,\gamma_d(Z)]$, $\Gamma_2 = \mathrm{diag}[\gamma_{d+1}(Z),\dots,\gamma_n(Z)]$ for some $d<n$, and $\tilde{\Gamma}_1, \tilde{\Gamma}_2, \tilde{V}, \tilde{V}_\perp$ defined similarly. Let also

$$\alpha = |\gamma_{\min}(V^\top\tilde{Z}V)|,\ \beta = \|V_\perp^\top\tilde{Z}V_\perp\|_2,\ e_{21} = \|\mathcal{P}_{V_\perp}E\mathcal{P}_V\|_2,\ e_{12} = \|\mathcal{P}_V E\mathcal{P}_{V_\perp}\|_2.$$

Then, if

$$\alpha > \beta + e_{12}\wedge e_{21}, \tag{56}$$

it holds

$$\|\sin\Theta(V,\hat{V})\|_2 \le \frac{e_{12}\vee e_{21}}{\alpha-\beta-e_{21}\wedge e_{12}} \tag{57}$$

Proof. Condition (56) is stronger than the one in Theorem 1, (Cai and Zhang, 2016) since it implies

$$\alpha^2 > (\beta + e_{12} \wedge e_{21})^2 \geq \beta^2 + e_{12}^2 \wedge e_{21}^2,$$

and we can further loosen the original upper-bound on $\|\sin\Theta(V, \hat{V})\|_2$ as follows

$$\begin{aligned}\|\sin\Theta(V, \hat{V})\|_2 &\leq \frac{\alpha e_{12} + \beta e_{21}}{\alpha^2 - \beta^2 - e_{12}^2 \wedge e_{21}^2} \\ &\leq \frac{(\alpha + \beta + e_{12} \wedge e_{21})(e_{12} \vee e_{21})}{\alpha^2 - (\beta + e_{12} \wedge e_{21})^2} \\ &= \frac{e_{12} \vee e_{21}}{\alpha - \beta - e_{12} \wedge e_{21}}.\end{aligned}$$

□

Corollary B.1. *With notations of Lemma B.2, if d is the rank of Z and $|\gamma_d(Z)| > 3\|E\|_2$, then*

$$\|\sin\Theta(V, \tilde{V})\|_2 \leq \frac{\|E\|_2}{|\gamma_d(Z)| - 3\|E\|_2}.$$

Proof. Notice that $e_{12}, e_{21} \leq \|E\|_2$ and by Weyl's inequality combined with submultiplicativity,

$$\alpha = |\gamma_{\min}(V^\top \tilde{Z} V)| \geq |\gamma_d(V^\top Z V + V^\top E V)| \geq |\gamma_d(Z)| - \|E\|_2$$

$$\beta = \|V_\perp^\top \tilde{Z} V_\perp\|_2 \leq \|V_\perp^\top Z V_\perp\|_2 + \|V^\top E V\|_2 \leq \|E\|_2.$$

Then $\alpha - \beta - e_{12} \wedge e_{21} \geq |\gamma_d(Z)| - 3\|E\|_2 > 0$ by our assumption, and the needed bound follows by Lemma B.2. □

The next lemma summarizes various relationships between a matrix and the soft-thresholding of its perturbation.

Lemma B.3 (Soft thresholding with noise)**.** *Consider symmetric matrices $Z, E \in \mathbb{R}^{n\times n}$. Define the perturbation $\tilde{Z} = Z + E$ and its soft-thresholding $\hat{Z} = \mathcal{T}_\rho(\tilde{Z})$ for some positive number $\rho > 0$. Let d be the rank of Z and define $V, \tilde{V} \in \mathbb{R}^{n\times d}$ as the top d eigenvectors of Z and $\tilde{Z}$, respectively. Then, we can relate Z and $\hat{Z}$ as follows:*

1. *The spectral norm of their difference can be bounded as*

$$\|\hat{Z} - Z\|_2 \leq \rho + \|E\|_2.$$

2. *If $\rho \geq \|E\|_2$, the Frobenius norm of their difference satisfies*

$$\|\hat{Z} - Z\|_F \leq 4\rho\sqrt{d}$$

3. *If $\rho \geq \|E\|_2$, it holds $\|\hat{Z}\|_2 \leq \|Z\|_2$,*
4. *If $\rho \geq \|E\|_2$ and Z is positive semi-definite (PSD), then $\hat{Z}$ is also PSD.*
5. *If $\rho \geq \|E\|_2 + 2\|Z\|_2\|\sin\Theta(V, \tilde{V})\|_2^2$, the rank of $\hat{Z}$ is at most d.*

Proof. 1. Consider the eigendecomposition $Z + E = \tilde{V}\Sigma\tilde{V}^\top$. Then, the eigendecomposition of $\hat{Z}$ can be written as $\hat{Z} = \tilde{V}\Sigma_\rho\tilde{V}^\top$ with

$$\Sigma_\rho = \operatorname{diag}\left[\operatorname{sign}(\Sigma_{11})(|\Sigma_{11}| - \rho)_+, \ldots, \operatorname{sign}(\Sigma_{nn})(|\Sigma_{nn}| - \rho)_+\right].$$

So, we can decompose the error as

$$\hat{Z} - Z = \tilde{U}\Sigma_\rho\tilde{V}^\top - \left(\tilde{U}\Sigma\tilde{V}^\top - E\right) = \tilde{U}\left(\Sigma_\rho - \Sigma\right)\tilde{V}^\top + E$$

where $\Sigma_\rho - \Sigma$ is diagonal with

$$[\Sigma_\rho - \Sigma]_{ii} = \operatorname{sign}(\Sigma_{ii})(|\Sigma_{ii}| - \rho)_+ - \Sigma_{ii} = -\operatorname{sign}(\Sigma_{ii})\min(|\Sigma_{ii}|, \rho),$$

From that, the needed bound on the spectral norm of the error follows immediately

$$\|\hat{Z} - Z\|_2 \le \|\Sigma_\rho - \Sigma\|_2 + \|E\|_2 = \max_{1\le i\le n} |\min\{\rho, |\Sigma_{ii}|\}| + \|E\|_2 \le \rho + \|E\|_2.$$

2. Consider the variational definition of soft thresholding:

$$\hat{Z} = \operatorname*{argmin}_{y\in\mathbb{R}^{n\times n}}\Big\{\frac{1}{2}\|Y - Z\|_F^2 + \rho\|Y\|_*\Big\}$$

By optimality, there exists a subgradient $G_Z \in \partial\|\hat{Z}\|_*$ such that

$$\langle\hat{Z} - Z - E + \rho G_Z, \hat{Z} - \check{Z}\rangle \le 0 \tag{58}$$

for any matrix $\check{Z}$ with the same column space and row space as Z. Let $\check{G}_Z \in \partial\|\check{Z}\|_*$ be arbitrary. Adding and subtracting $\rho\langle\check{G}_Z, \hat{Z} - \check{Z}\rangle$ gives

$$\langle\hat{Z} - Z, \hat{Z} - \check{Z}\rangle + \rho\langle G_Z - \check{G}_Z, \hat{Z} - \check{Z}\rangle \le \langle -\rho\check{G}_Z + E, \hat{Z} - \check{Z}\rangle \tag{59}$$

On the other hand, by convexity of the nuclear norm, we have

$$\|\hat{Z}\|_* - \|\check{Z}\|_* \ge \left\langle\check{G}_Z, \hat{Z} - \check{Z}\right\rangle$$
$$\|\check{Z}\|_* - \|\hat{Z}\|_* \ge \left\langle G_Z, \check{Z} - \hat{Z}\right\rangle$$

which together imply

$$\langle G_Z - \check{G}_Z, \hat{Z} - \check{Z}\rangle \ge 0. \tag{60}$$

Combining (59) and (60), we obtain

$$\langle\hat{Z} - Z, \hat{Z} - \check{Z}\rangle \le -\rho\langle\check{G}_Z, \hat{Z} - \check{Z}\rangle + \langle E, \hat{Z} - \check{Z}\rangle. \tag{61}$$

which is the inequality of the same form as (58) but now, instead of a fixed G_Z, we have a free choice of $\check{G}_Z$. By Koltchinskii et al. (2010), subgradient $\check{G}_Z$ can be expressed as

$$\check{G}_Z = \sum_{i=1}^{d} u_i v_i^\top + \check{\mathcal{P}}_Z^\perp W \check{\mathcal{P}}_Z^\perp$$

where $\check{\mathcal{P}}_Z$ is the projector on the column space of $\check{Z}$, vectors u_i, v_i are left and right singular vectors of $\check{Z}$ respectively, and W is an arbitrary matrix satisfying $\|W\|_2 \le 1$. So, we can

specify W to attain the upper bound of the following expression obtained using the duality of nuclear and operator norms

$$\langle \check{\mathcal{P}}_Z^{\perp} W \check{\mathcal{P}}_Z^{\perp}, \hat{Z} - \check{Z} \rangle = \langle W, \check{\mathcal{P}}_Z^{\perp} \hat{Z} \check{\mathcal{P}}_Z^{\perp} \rangle \leq \| \check{\mathcal{P}}_Z^{\perp} \hat{Z} \check{\mathcal{P}}_Z^{\perp} \|_*. \tag{62}$$

By trace duality, we can also bound

$$|\langle \sum_{i=1}^{d} u_i v_i^{\top}, \hat{Z} - \check{Z} \rangle| \leq \sqrt{d} \| \hat{Z} - \check{Z} \|_F \tag{63}$$

Finally, we bound the second term in the RHS of (61) again by trace duality

$$\begin{aligned} \langle E, \hat{Z} - \check{Z} \rangle &= \langle E - \check{\mathcal{P}}_Z^{\perp} E \check{\mathcal{P}}_Z^{\perp}, \hat{Z} - \check{Z} \rangle + \langle \check{\mathcal{P}}_Z^{\perp} E \check{\mathcal{P}}_Z^{\perp}, \hat{Z} - \check{Z} \rangle \\ &= \langle \check{\mathcal{P}}_Z^{\perp} E \check{\mathcal{P}}_Z + \check{\mathcal{P}}_Z E \check{\mathcal{P}}_Z^{\perp} + \check{\mathcal{P}}_Z E \check{\mathcal{P}}_Z, \hat{Z} - \check{Z} \rangle + \langle E, \check{\mathcal{P}}_Z^{\perp} \hat{Z} \check{\mathcal{P}}_Z^{\perp} \rangle \\ &\leq 3\sqrt{d} \, \|E\|_2 \, \|\hat{Z} - \check{Z}\|_F + \|E\|_2 \, \|\mathcal{P}_Z^{\perp} \hat{Z} \mathcal{P}_Z^{\perp}\|_*. \end{aligned} \tag{64}$$

Plugging the bounds (62), (63), and (64) into (61) and specifying $\check{Z} = Z$, we get

$$\|\hat{Z} - Z\|_F^2 + (\rho - \|E\|_2) \, \|\check{\mathcal{P}}_Z^{\perp} \hat{Z} \check{\mathcal{P}}_Z^{\perp}\|_* \leq \sqrt{d} \, (\rho + 3 \, \|E\|_2) \, \|\hat{Z} - Z\|_F$$

Using the assumption $\rho \geq \|E\|_2$ and dividing both sides by $\|\hat{Z} - Z\|_F$, we deduce the needed bound

$$\|\hat{Z} - Z\|_F \leq \sqrt{d} \, (\rho + 3 \, \|E\|_2) \leq 4\sqrt{d}\rho.$$

3. By definition of soft-thresholding and triangular inequality, we have

$$\|\hat{Z}\|_2 \leq \|\tilde{Z}\|_2 - \rho \leq \|Z\|_2 + \|E\|_2 - \rho \leq \|Z\|_2$$

4. If Z is PSD, it holds $v^{\top} Z v \geq 0$ for any $v \in \mathbb{R}^n$ and we can bound

$$v^{\top}(Z + E)v \geq v^{\top} E v \geq -\|E\|_2 \geq -\rho,$$

so only positive eigenvalues can survive after applying soft-thresholding $\mathcal{T}_{\rho}$ to $Z + E$.

5. Let v be a unit vector from the orthogonal complement of $\mathrm{col}(\tilde{V})$. Then, by orthogonality and (42)

$$\left\|V^{\top} v\right\|_2 = \inf_{O \in \mathcal{O}_d} \left\|V^{\top} v - O\tilde{V}^{\top} v\right\|_2 \leq \inf_{O \in \mathcal{O}_d} \left\|V - O\tilde{V}\right\|_2 \leq \sqrt{2} \|\sin\Theta(V, \tilde{V})\|_2$$

and

$$\left|v^{\top} Z v\right| = \left|v^{\top} V V^{\top} Z V V^{\top} v\right| \leq \|Z\|_2 \left\|V^{\top} v\right\|_2^2 \tag{65}$$

Therefore,

$$\left|v^{\top} \left(Z + E\right) v\right| \leq \left|v^{\top} Z v\right| + \|E\|_2 \leq \rho.$$

So, at most first d eigenvalues of $\hat{Z}$ can survive the soft-thresholding step

$$\left|\gamma_{d+1}\left(Z + E\right)\right| = \max_{v \perp \mathrm{col}\left(\tilde{V}\right)} \left|v^{\top} \left(Z + E\right) v\right| \leq \rho.$$

□

Finally, in Lemma B.4, we establish a generic bound on the spectral norm of the average difference between matrices $Z_\ell, \ell = 1, \dots, m$ and their soft thresholding estimators $\hat{Z}_\ell$ defined similarly to Lemma B.3.

Lemma B.4 (Average of soft-thresholding estimators)**.** *Consider symmetric matrices $Z_\ell, E_\ell \in \mathbb{R}^{n\times n}, \ell = 1, \dots, m$ and define the perturbations $\tilde{Z}_\ell = Z_\ell + E_\ell$. Let d_ℓ be the rank of Z_ℓ, and define $V_\ell, \tilde{V}_\ell \in \mathbb{R}^{n\times d_\ell}$ as the top d_ℓ eigenvectors of Z_ℓ and $\tilde{Z}_\ell$, respectively. Consider estimators $\hat{Z}_\ell = \mathcal{T}_{\rho_\ell}(\tilde{Z}_\ell)$ with thresholds ρ_ℓ satisfying*

$$\rho_\ell > \|E_\ell\|_2 + 2\|Z_\ell\|_2 \|\sin\Theta(V_\ell, \tilde{V}_\ell)\|_2^2. \tag{66}$$

Define also

$$\rho = \max_{1\le\ell\le m} \rho_\ell, \quad \theta = \max_{1\le\ell\le m} \|\sin\Theta(V_\ell, \tilde{V}_\ell)\|_2, \quad s = \max_{1\le\ell_1<\ell_2\le m} \|V_{\ell_1}^\top V_{\ell_2}\|_2, \quad \gamma = \max_{1\le\ell\le m} \|Z_\ell\|_2.$$

Then

$$\Big\|\frac{1}{m}\sum_{\ell=1}^m (\hat{Z}_\ell - Z_\ell)\Big\|_2 \le 11(\gamma\theta \vee \rho)\Big[\frac{1}{m} \vee s \vee \theta\Big]^{1/2} \tag{67}$$

Proof. With the choice of the thresholds in (66), we deduce by property 5 in Lemma B.3 that $\operatorname{rank}(\hat{Z}_\ell) \le d_\ell$. Therefore, $\hat{Z}_\ell$ satisfies

$$\hat{Z}_\ell = \tilde{V}_\ell \tilde{V}_\ell^\top \hat{Z}_\ell \tilde{V}_\ell \tilde{V}_\ell^\top.$$

Thus, we can decompose $\Delta_\ell := \hat{Z}_\ell - Z_\ell$ as

$$\begin{aligned}
\Delta_\ell &= \tilde{V}_\ell \tilde{V}_\ell^\top \hat{Z}_\ell \tilde{V}_\ell \tilde{V}_\ell^\top - V_\ell V_\ell^\top Z_\ell V_\ell V_\ell^\top \\
&= \left(\tilde{V}_\ell \tilde{V}_\ell^\top - V_\ell V_\ell^\top\right) \hat{Z}_\ell \tilde{V}_\ell \tilde{V}_\ell^\top + V_\ell V_\ell^\top \Delta_\ell V_\ell V_\ell^\top + V_\ell V_\ell^\top \hat{Z}_\ell \left(\tilde{V}_\ell \tilde{V}_\ell^\top - V_\ell V_\ell^\top\right) \\
&=: (\mathcal{I})_\ell + (\mathcal{II})_\ell + (\mathcal{III})_\ell.
\end{aligned}$$

We bound the operator norm of the sum over ℓ for each of these three terms separately.

Term $(\mathcal{I})$. We start by bounding the cosine similarity between perturbed eigenspaces in terms of the similarity of the true eigenspaces. By (42) and the triangular inequality, we have

$$\begin{aligned}
\left\|\tilde{V}_{\ell_1}^\top \tilde{V}_{\ell_2}\right\|_2 &= \inf_{O_1\in\mathcal{O}_{d_{\ell_1}},\, O_2\in\mathcal{O}_{d_{\ell_2}}} \left\|\tilde{V}_{\ell_1}^\top \tilde{V}_{\ell_2} - \tilde{V}_{\ell_1}^\top V_{\ell_2} O_1 + \tilde{V}_{\ell_1}^\top V_{\ell_2} O_1 - O_2 V_{\ell_1}^\top V_{\ell_2} O_1 + O_2 V_{\ell_1}^\top V_{\ell_2} O_1\right\|_2 \\
&\le \inf_{O_1\in\mathcal{O}_{d_{\ell_1}}} \left\|\tilde{V}_{\ell_2} - V_{\ell_2} O_1\right\|_2 + \inf_{O_2\in\mathcal{O}_{d_{\ell_2}}} \left\|\tilde{V}_{\ell_1} - V_{\ell_1} O_2\right\|_2 + \left\|V_{\ell_1}^\top V_{\ell_2}\right\|_2 \\
&\le 2\sqrt{2}\theta + s
\end{aligned} \tag{68}$$

Notice that by property 3 in Lemma B.3 and the threshold choice in (66), we have $\|\hat{Z}_\ell\|_2 \le \gamma$. Thus, by (41) and submultiplicativity

$$\left\|\left(\tilde{V}_\ell \tilde{V}_\ell^\top - V_\ell V_\ell^\top\right) \hat{Z}_\ell \tilde{V}_\ell \tilde{V}_\ell^\top\right\|_2 \le \|\tilde{V}_\ell \tilde{V}_\ell^\top - V_\ell V_\ell^\top\|_2 \|\hat{Z}_\ell\|_2 \le 2\theta\gamma$$

Similarly, by (68)

$$\begin{aligned}
&\left\|\left(\tilde{V}_{\ell_1} \tilde{V}_{\ell_1}^\top - V_{\ell_1} V_{\ell_1}^\top\right) \hat{Z}_{\ell_1} \tilde{V}_\ell \tilde{V}_{\ell_1}^\top \cdot \tilde{V}_{\ell_2} \tilde{V}_{\ell_2}^\top \hat{Z}_{\ell_2} \left(\tilde{V}_{\ell_2} \tilde{V}_{\ell_2}^\top - V_{\ell_2} V_{\ell_2}^\top\right)\right\|_2 \le \\
&\|\tilde{V}_{\ell_1}^\top \tilde{V}_{\ell_2}\|_2 \cdot \max_\ell \|\hat{Z}_\ell\|_2^2 \cdot \max_{\ell_1<\ell_2} \|\tilde{V}_{\ell_1} \tilde{V}_{\ell_1}^\top - V_{\ell_1} V_{\ell_1}^\top\|_2^2 \;\le\; (2\sqrt{2}\theta + s)\cdot 4\theta^2\gamma^2.
\end{aligned} \tag{69}$$

Therefore, by Lemma 4 in MacDonald et al. (2022)

$$\Big\|\frac{1}{m}\sum_{\ell=1}^{m}(\mathcal{I})_\ell\Big\|_2 \le 2m^{-1/2}\theta\gamma\Big[1+m(2\sqrt{2}\theta+s)\Big]^{1/2} \le 5\theta\gamma\Big[\frac{1}{m}\vee\theta\vee s\Big]^{1/2},$$

where in the last line we used $(2+2\sqrt{2})^{1/2} < 2.5$.

Term $(\mathcal{II})$. With the choice of the thresholds in (B.3), we have by property 1 in Lemma B.3 that $\|\Delta_\ell\|_2 \le 2\rho$

$$\|V_{\ell_1}V_{\ell_1}^\top\Delta_{\ell_1}V_{\ell_1}V_{\ell_1}^\top\cdot V_{\ell_2}V_{\ell_2}^\top\Delta_{\ell_2}V_{\ell_2}V_{\ell_2}^\top\|_2 \le \|\Delta_{\ell_1}\|_2\|V_{\ell_1}^\top V_{\ell_2}\|_2\|\Delta_{\ell_2}\|_2 \le 4s\rho^2$$

Thus, by Lemma 4 in MacDonald et al. (2022),

$$\Big\|\frac{1}{m}\sum_{k=1}^{m}(\mathcal{II})_k\Big\|_2 \le 2m^{-1/2}\rho\Big[1+ms\Big]^{1/2} \le 3\rho\Big[\frac{1}{m}\vee s\Big]^{1/2}$$

Term $(\mathcal{III})$. Similarly to Term $(\mathcal{I})$, we can establish by substituting $\|\tilde{V}_{\ell_1}\tilde{V}_{\ell_2}\|_2$ with $\|V_{\ell_1}V_{\ell_2}\|_2$ in (69):

$$\Big\|\frac{1}{m}\sum_{\ell=1}^{m}(\mathcal{III})_\ell\Big\|_2 \le 2m^{-1/2}\theta\gamma\Big[1+ms\Big]^{1/2} \le 3\theta\gamma\Big[\frac{1}{m}\vee s\Big]^{1/2}.$$

□

C Additional experiments

C.1 Dependency of estimation errors on the cosine similarities of the components

In this section, we present experiments exploring how the latent component estimation error depends on their cosine similarities. Similarly to Section 5, we generate latent components following Algorithm 2 with $n = 200$, $M = 16$, $K = 4$, and $s_{v,u} = s_{w,u} = 0.1$ unless one of these two parameters is varied. Here, we only present the results for the Gaussian edge-entry distribution, as the error trends are very similar between the two model types. In Figure 7, we vary $s_{v,w}, s_{w,u}, s_{w,w}, s_{u,u} \in [0.1, 0.2, \ldots, 0.9]$, where cosine similarities, which were not yet introduced, are defined similarly to (24).

As the most important general trend, we can see that an increase in any of the similarities either does not affect or improves the error of Θ. Indeed, a high similarity of latent components makes it harder to separate them, but improves the effective sample size needed to estimate their sum because of the similarity of the summed terms. As an illustrative corner case, when all components are perfectly collinear with each other, it becomes sufficient to fit just one low-rank matrix to get an accurate estimator of Θ, even though separation of the latent components becomes impossible in this case due to violation of the identifiability condition (see Proposition 2.1).

Probably, the most unexpected behavior can be observed in the plot of $s_{w,u}$, showing that individual and group latent component errors improve slightly before reaching $s_{w,u} = 0.4$ and then rapidly grow for larger $s_{w,u}$. We believe that the initial error improvement occurs because, by definition, $s_{w,u}$ induces correlation between each group component and individual components not only within but also outside of the corresponding group. Based on the group-wise separation

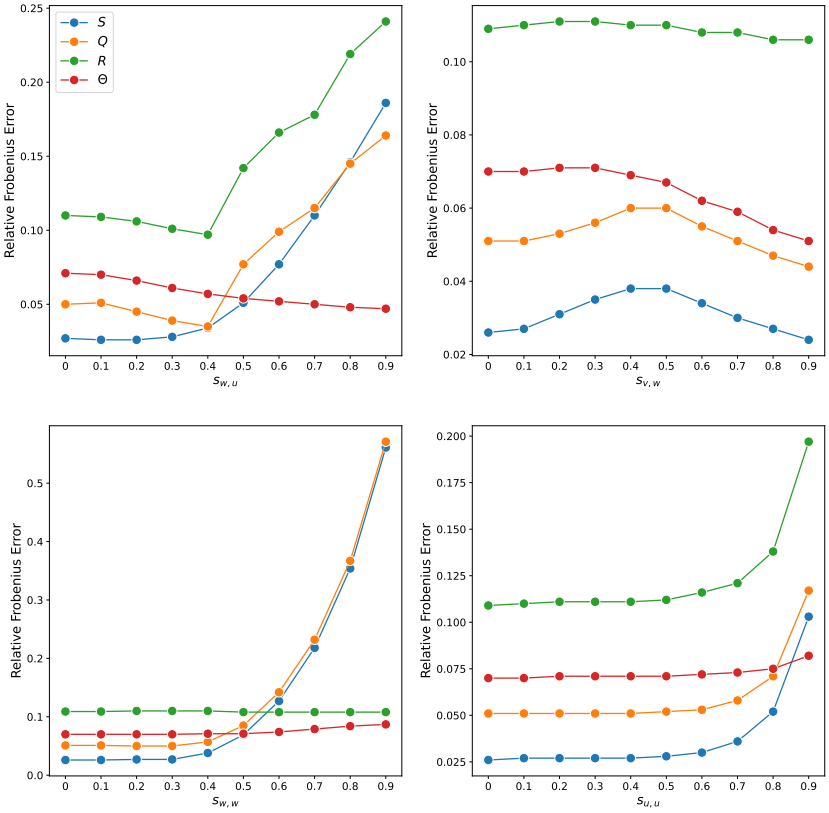

Figure 7: Dependency of ARFE on the cosine similarities between different types of latent components. Unless one of the parameters is varied, the networks are generated according to Algorithm 2 with $M = 16$ layers on $n = 200$ nodes, $K = 4$ balanced groups, and $s_{v,w} = s_{w,u} = 0.1$.

structure of the first stage of Algorithm 1, it is clear that the correlation of a group component with its own individual components makes their separation harder, while the correlation with other components can induce additional information, which implies that the trend change around $s_{w,u} = 0.4$ is due to reaching the critical point in the trade-off between the within-group and outside-of-the-group correlations.

Another interesting artifact is the inverted U-shape dependency of the shared and group errors from $s_{v,w}$. This can be explained by the fact that for $s_{v,w}$ close to one, group components add information about the latent space basis of the shared component due to their significant overlap. However, because of the difference in the bases of the group components themselves, the overlap with the shared component does not lead to their inferior estimates. As another corner case, for $s_{v,w}$ close to zero, the group components contain no additional information on the shared one. However, each of them can be easily separated from the shared component by orthogonality. On

the other hand, the closer $s_{v,w}$ approaches the middle ($s_{v,w} \approx 0.5$) between these two cases, the more difficult the separation task becomes since both simplifying assumptions are no longer valid.

In the $s_{w,w}$ and $s_{v,w}$ plots, we can observe that the similarity of the group components does not affect the individual error R. This is expected since the separation of individual components occurs in the first stage and only involves the sum of the shared and group components, which is not affected by the correlation of the additive terms defining it.

It is also worth mentioning that the errors are robust to the increase in both $s_{w,w}$ and $s_{u,u}$, with the estimation accuracy of S and Q starting to decrease only for $s_{w,w} \geq 0.4$ and $s_{u,u} \geq 0.6$. This observation aligns with the form of the theoretical bounds in Theorem 4.1, suggesting that $s_{w,w}$ and $s_{u,u}$ only play a role in the component errors if they, respectively, dominate $\frac{1}{K} \vee n^{1/2-\tau}$ and $\frac{1}{m_k} \vee n^{1/2-\tau}$. While the individual component error bound in Theorem 4.1 does not suggest direct dependency of R on $s_{u,u}$, we can see that $s_{u,u}$ appears in the bound on $r^{(I)}$ derived in Proposition 4.1.

C.2 Method comparison for the Logistic edge-entry model.

Here, we repeat the method comparison experiment in Section 5.3, but this time, under the logistic network-generating model. Figure 8 presents the corresponding results. We omit the points for $n = 100$ and $M = 8$ due to the instability of the Logistic model fitting for smaller data samples.

Based on the upper row of Figure 8, we can see that with both n and M varying, GroupMultiNeSS is very close to matching the oracle rates while permanently dominating the MultiNeSS model. Similar to the Gaussian results in Figure 5.3, we can see that the COSIE model is significantly inferior to both the MultiNeSS and GroupMultiNeSS models.

Looking at the lower row, we can see that contrary to the Gaussian model, GroupMultiNeSS exhibits higher error in estimating the shared component S in the logistic case, with both models significantly less accurate than the oracle estimator. This behavior is related to the fact that latent space models with a non-linear link function, such as the sigmoid, produce adjacency matrices of much higher rank than the original latent space and thus usually suffer from overestimated ranks. In our case, the more latent components are estimated simultaneously, the more freedom the model has to overestimate their latent dimensions and thus overfit during the refitting step.

C.3 Additional real data analysis plots.

In this section, we present additional plots mentioned in the real data experiment described in Section 6. In particular, the upper row of Figure 9 shows the four leading latent dimensions of the shared component, and the lower row contains the first two dimensions of the group components aligned with the same Procrustes rotation as the one used in Figure 4.

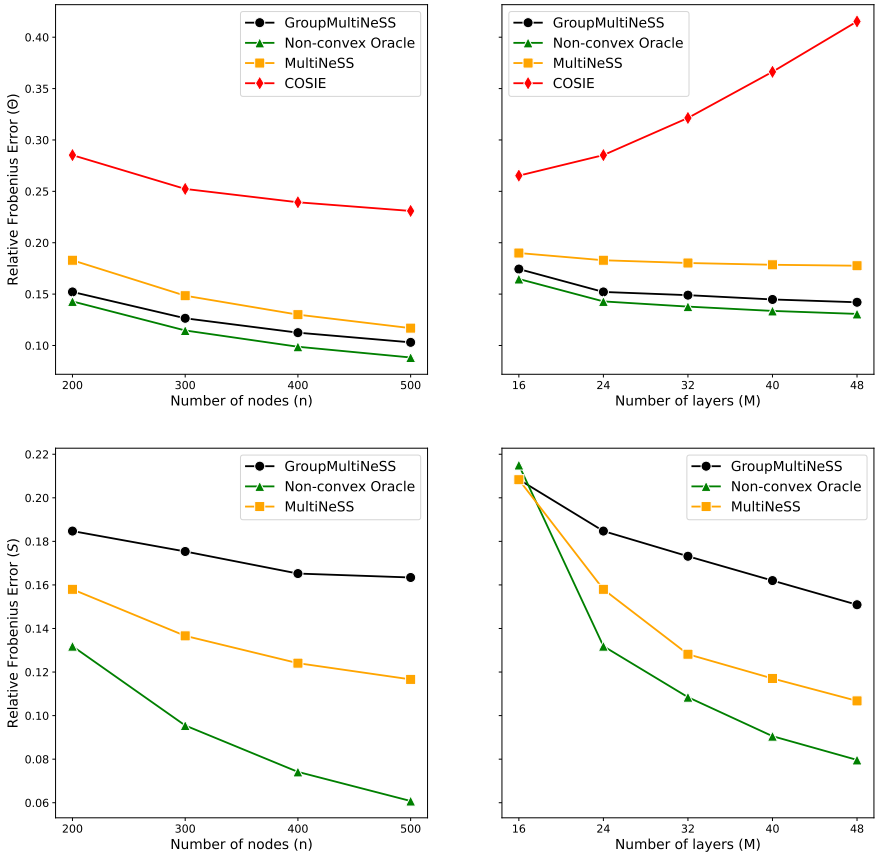

Figure 8: Change in ARFE of Θ (upper row) and S (lower row) with the increase of (left column) the number of nodes n with $M = 16$ and (right column) the number of layers M with $n = 200$. In all simulations, the number of groups is $K = 4$ and the latent dimension is $d = 3$. Layers are sampled from the Logistic edge-entry model.

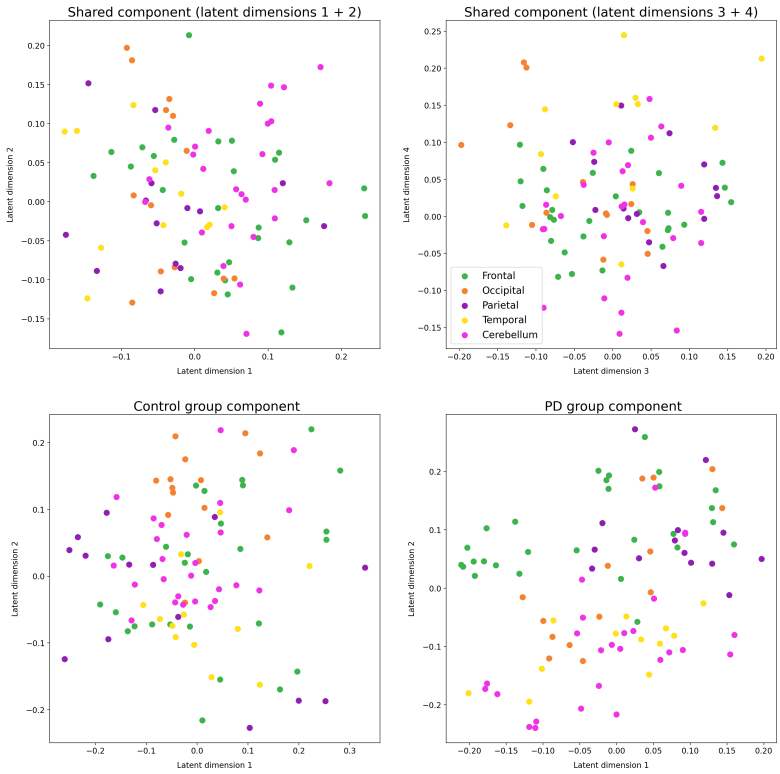

Figure 9: First four dimensions of the shared component and first two dimensions of the group components, all fitted with the GroupMultiNeSS (all dimensions in the plots are disassortative).