Journal of Multivariate Analysis 203 (2024) 105333

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sparse subspace clustering in diverse multiplex network model

Majid Noroozi [a,*], Marianna Pensky [b]

[a] Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152, USA
[b] Department of Mathematics, University of Central Florida, Orlando, FL 32816-1364, USA

ARTICLE INFO

AMS 2020 subject classifications:
primary 62F12
secondary 62H30

Keywords:
Multilayer network
Sparse subspace clustering
Stochastic block model

ABSTRACT

The paper considers the DIverse MultiPLEx (DIMPLE) network model, where all layers of the network have the same collection of nodes and are equipped with the Stochastic Block Models. In addition, all layers can be partitioned into groups with the same community structures, although the layers in the same group may have different matrices of block connection probabilities. To the best of our knowledge, the DIMPLE model, introduced in Pensky and Wang (2021), presents the most broad SBM-equipped binary multilayer network model on the same set of nodes and, thus, generalizes a multitude of papers that study more restrictive settings. Under the DIMPLE model, the main task is to identify the groups of layers with the same community structures since the matrices of block connection probabilities act as nuisance parameters under the DIMPLE paradigm. The main contribution of the paper is achieving the strongly consistent between-layer clustering by using Sparse Subspace Clustering (SSC), the well-developed technique in computer vision. In addition, SSC allows to handle much larger networks than spectral clustering, and is perfectly suitable for application of parallel computing. Moreover, our paper is the first one to obtain precision guarantees for SSC when it is applied to binary data.

1. Introduction

Network models are an important tool for describing and analyzing complex systems in many areas such as the social, biological, physical, and engineering sciences. Originally, almost all studies of networks were focused on a single network. Many models have been introduced to describe communities in networks, with the most popular Stochastic Block Model (SBM) and its extensions (see, e.g., [1,20,28,43]).

Over the last decade, however, the focus has changed to analysis of a multilayer network [21], in which different individual networks evolve or interact with each other. In addition to a node set and an edge set, a multilayer network includes a layer set, whose each layer represents a different type of relation among those nodes. For example, a general multilayer network could be used to represent an urban transportation network, where nodes might be stations in the city and each layer might represent a mode of transportation such as buses, metro, rail, etc. While the term "multilayer network" is often used in a more general context, we focus on the multilayer networks where the same set of nodes appears on every layer, and there are no edges between two different layers. Following [31], we call this multilayer network a *multiplex network*. One such example is a collection of brain connectivity networks of several individuals, where each layer corresponds to a brain connectivity network of an individual.

In this paper, we study a multiplex network where each layer is enabled with a community structure. One of the problems in multilayer networks is community detection with many important applications. While in such networks different layers have

* Corresponding author.
E-mail address: mnoroozi@memphis.edu (M. Noroozi).

https://doi.org/10.1016/j.jmva.2024.105333
Received 2 June 2023; Received in revised form 6 May 2024; Accepted 15 May 2024
Available online 17 May 2024

different forms of connections, it is often the case that one underlying unobserved community structure is in force. For example, in the multilayer Twitter networks in [15], ground truth community memberships can be assigned to the users (nodes) based on some fundamental attributes (e.g., political views, country of origin, football clubs) that are independent of the observed twitter interactions, whereas the interactions provide multiple sources of information about the same latent community structure. Combining information from these multiple sources would then lead to enhanced performance in the consensus community detection [40].

The assumption of one common community structure may not be true in some applications. It is often the case that there are groups of layers that are similar in some sense, and layers within each group share the same community structure, but each group has different community structure. One example is the worldwide food trading networks, collected by [9], which has been widely analyzed in literature (see, e.g., [18,30], among others). The data present an international trading network, in which layers represent different food products, nodes are countries, and edges at each layer represent trading relationships of a specific food product among countries. Two types of products, e.g., unprocessed and processed foods, can be considered as two groups of layers where each group has its own pattern of trading among the countries. While some large countries import/export unprocessed food from and/or to a great number of other countries worldwide, for processed foods, countries are mainly clustered by the geographical location, i.e., countries in the same continent have closer trading ties [18].

1.1. The DIMPLE model

In what follows, we consider a multilayer network where each of the layers is equipped with the Stochastic Block Model (SBM). We also assume that the layers can be partitioned into several types, each of them is equipped with a distinct community structure, while the matrices of block probabilities can take different values in each of the layers. We call this model, first introduced in [41], the DIverse MultiPLEx (DIMPLE) network model.

Specifically, we consider an undirected multilayer network with L layers over a common set of n vertices with no self loops, where each of the layers follows the SBM. Assume that those L layers can be partitioned into $M \ll L$ groups, $S_1, \dots, S_M$, where each group is equipped with its own community structure. The latter means that there exists a clustering function $c : [L] \to [M]$ such that $c(\ell) = m$ if the $\ell \in S_m$, $m \in [M]$, where $[N] = \{1, \dots, N\}$ for any positive integer N. Nodes in the layer $\ell \in S_m$ follow SBM with the K_m communities $G_{m,1}, \dots, G_{m,K_m}$, that persist in the layers of type m. Hence, for every $m \in [M]$, there exists a clustering function $z^{(m)} : [n] \to [K_m]$ with the corresponding clustering matrix $Z^{(m)} \in \{0,1\}^{n \times K_m}$, such that $Z^{(m)}_{i,k} = 1$ if and only if $z^{(m)}(i) = k$. Nonetheless, the block connectivity matrices $B^{(\ell)} \in [0,1]^{K_m \times K_m}$ can vary from layer to layer. Therefore, the probability of connection between nodes i and j in layer ℓ is $P^{(\ell)}_{i,j} = B^{(\ell)}_{k_1,k_2}$ where $k_1 = z^{(m)}(i)$ and $k_2 = z^{(m)}(j)$. In summary, while the membership function $z^{(m)} : [n] \to [K_m]$ is completely determined by the group m of layers, the block connectivity matrices $B^{(\ell)}$ are not, and can be all different in the group m of layers. In this case, the matrix of connection probabilities in layer ℓ is of the form

$$P^{(\ell)} = Z^{(m)} B^{(\ell)} (Z^{(m)})^\top, \quad m = c(\ell), \quad \ell \in [L]. \tag{1}$$

Furthermore, we assume that symmetric adjacency matrices $A^{(\ell)} \in \{0,1\}^{n \times n}$, $\ell \in [L]$, are such that $A^{(\ell)}_{i,j} \sim \text{Bernoulli}(P^{(\ell)}_{i,j})$, $1 \le i < j \le n$, where $A^{(\ell)}_{i,j}$ are conditionally independent given $P^{(\ell)}_{i,j}$, $A^{(\ell)}_{i,j} = A^{(\ell)}_{j,i}$ and $A^{(\ell)}_{i,i} = 0$. Denote the three-way tensors with layers $A^{(\ell)}$ and $P^{(\ell)}$, $\ell \in [L]$, by $\mathbf{A}, \mathbf{P} \in \mathbb{R}^{n \times n \times L}$, respectively.

Note that in the setting (1), the main task is to identify the groups of layers with the common community structures. When this task is accomplished, one can combine the layers and find community structures with higher precision than if those structures were elicited from individual layers. On the other hand, the matrices of block connection probabilities act as nuisance parameters in (1) since they vary from one layer to another. For this reason, in this paper, we consider the problem of clustering of layers into the sets of layers with the identical community structures (the between-layer clustering) as well as identification of community structures in the groups of layers (the within-layer clustering). We do not study estimation of block probability matrices, however, the latter can be done by averaging the adjacency matrices $A^{(\ell)}$, $\ell \in [L]$, over pairs of respective communities.

In this paper, similarly to [41], we assume that the number of communities in each group of layers is the same, i.e. $K_1 = \cdots = K_M = K$. If one is unsure that each of the layers of the network has the same number of communities, one can use a different number of communities $K^{(\ell)}$ in each layer. After groups of layers are identified, the number of layers in each group should be re-adjusted, so that $K^{(\ell)} = K_m$ if $m = c(\ell)$. One can, of course, assume that the values of K_m, $m \in [M]$, are known. However, since group labels are interchangeable, in the case of non-identical subspace dimensions (numbers of communities), it is hard to choose, which of the values correspond to which of the groups. This is actually the reason why [13,19], who imposed this assumption, used it only in theory while their simulations and real data examples are all restricted to the case of equal K_m, $m \in [M]$. On the contrary, knowledge of $K^{(\ell)}$ allows one to deal with different ambient dimensions (number of communities) in the groups of layers in simulations and real data examples.

In addition, for the purpose of methodological developments, we assume that the number of communities K in each layer of the network is known. Identifying the number of clusters is a common issue in data clustering, and it is a separate problem from the process of actually solving the clustering problem with a known number of clusters. A common method for finding the number of clusters is the so called "elbow" method that looks at the fraction of the variance explained as a function of the number of clusters. The method is based on the idea that one should choose the smallest number of clusters, such that adding another cluster does not significantly improve fitting of the data by a model. There are many ways to determine the "elbow". For example, one can base its detection on evaluation of the clustering error in terms of an objective function, as in, e.g., [53]. Another possibility is to monitor the eigenvalues of the non-backtracking matrix or the Bethe Hessian matrix, as it is done in [22]. One can also employ a simple technique of checking the eigen-gaps, as it has been discussed in [29], or use a scree plot as it is done in [54].

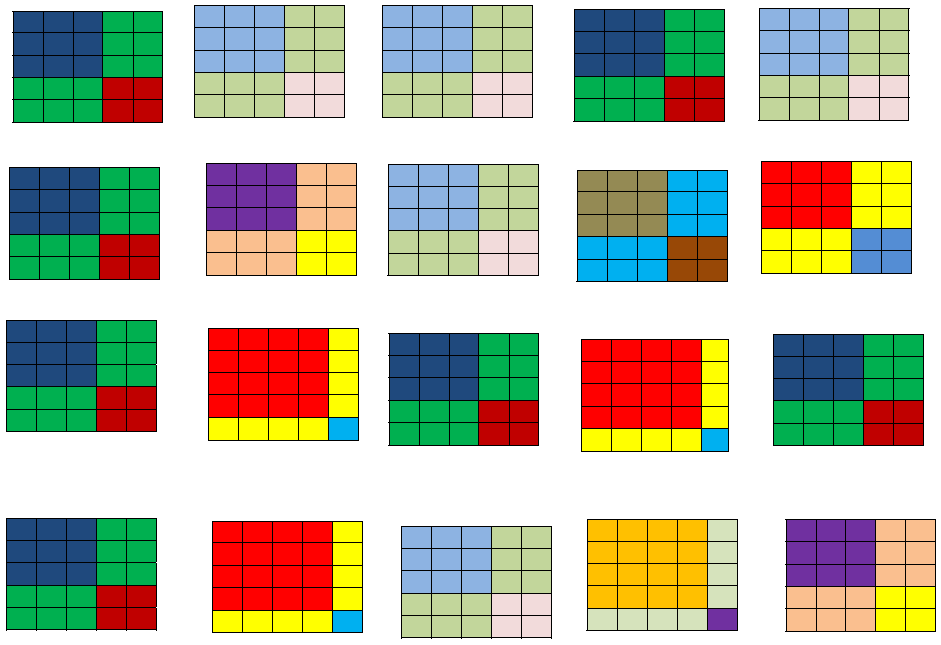

Fig. 1. Multiplex networks versions with $n = 5$, $L = 5$ and $K = 2$. and $M = 2$. First row: "checker board" model (persistent communities, two distinct block connectivity matrices). Second row: persistent communities, all block connectivity matrices are different. Third row: Mixture MultiLayer Stochastic Block Model (MMLSBM) with $M = 2$ (only two distinct layers in the network). Fourth row: DIverse MultiPLEx (DIMPLE) network model with $M = 2$ (two distinct community assignments, all block connectivity matrices are different).

1.2. Existing particular cases of the DIMPLE model

To the best of our knowledge, the DIMPLE model, introduced in Pensky and Wang [41], presents the most broad SBM-equipped binary multilayer network model on the same set of nodes and, thus, it includes as its particular cases, a variety of more restrictive settings, exhibited in Fig. 1 where different colors are used for different values of connection probabilities.

Specifically, the DIMPLE model generalizes a multitude of papers where communities persist throughout the network [4,24,25, 39,40]. In particular, it includes the simplest case of the multiplex networks where the block probabilities take only finite number of values, as it happens in "checker board" and tensor block models [8,17,52] presented in the first row of Fig. 1. The second row of Fig. 1 shows the most popular type of multiplex networks where communities persist through all layers of the network but the matrices of block probabilities vary from one layer to another (see, e.g., [4,24,25,39,40] and references therein). Another type of models that is generalized by DIMPLE is the Mixture MultiLayer Stochastic Block Model (MMLSBM), displayed in the third row of Fig. 1, where all layers can be partitioned into a few different types, with each type of layers equipped with its own community structure and a matrix of connection probabilities (see [13,19]). Finally, the last row of Fig. 1 exhibits the DIMPLE model where layers can be partitioned into groups with similar community structures like in the MMLSBM but, unlike the MMLSBM, matrices of block connection probabilities can vary from one layer to another.

It is easy to see that, for $M = 1$, the DIMPLE model reduces to the common multilayer network setting (row 2 of Fig. 1) where the community structures persist throughout the network and, hence, can be viewed as a concatenation of the latter type of networks, where the layers are scrambled. On the other hand, it becomes the MMLSBM (row 3 of Fig. 1) if the block connectivity matrices $B^{(\ell)}$ are the same for all layers in a group, i.e., $B^{(\ell)} = B^{(c(\ell))}$, $\ell \in [L]$.

1.3. Main contributions of the paper

The only other paper which studied the DIMPLE model was [41] where the between-layer clustering was based on spectral methods. Specifically, [41] used the SVDs of the layer adjacency matrices $A^{(\ell)}$, $\ell \in [L]$, to estimate the $n \times n$ membership matrices where the elements are equal to the reciprocal of the community size if nodes belong to the same community, and are equal to zero otherwise. Subsequently, the groups of layers with similar community structures were found by spectral clustering of a matrix with rows formed as the vectorized versions of the estimated membership matrices. The authors showed that the methodologies used in the networks with the persistent community structure, as well as the ones designed for the MMLSBM, cannot be applied to the DIMPLE model. The within-layer clustering procedure was based on averaging the adjacency matrices of the layers with the identical community structure or their adjusted squares.

While [41] developed the within and the between-layer clustering algorithms and studied their precision, their methodology has a number of shortcomings. To start with, the between layer clustering is not strongly consistent and, consequently, the error of the

between-layer clustering dominates the within-layer clustering error. In addition, since spectral clustering for grouping the layers is applied to the vectorized versions of the estimated membership matrices, for an n-node multilayer network, it requires clustering of vectors in $n(n-1)/2$-dimensional space. While the methodology works well for smaller n, it becomes extremely challenging when n grows. For this reason, all simulations in [41] are carried out for relatively small values of n.

The present paper uses Subspace Clustering for finding groups of layers with similar community structures. Indeed, in what follows, we shall show that the vectorized probability matrices of such layers all belong to the same low-dimensional subspace. The subspace clustering relies on self-representation of the vectors to partition them into clusters. Consequently, one has to solve a regression problem for each vector separately to find the matrix of weights, which is usually of much smaller size. Subsequently, some kind of spectral clustering is applied to the weight matrix. Subspace Clustering is a very common technique in the computer vision field. In particular, we apply Sparse Subspace Clustering (SSC) approach to identify those groups. We provide a review of the SSC technique in Section 2.

Our paper makes the following key contributions:

1. The clustering algorithms developed in the paper are not iterative and, hence, do not require provable convergence. In addition, they come with the theoretical precision guarantees.
2. Specifically, application of the SSC to clustering of layers leads to the strongly consistent between-layer clustering and, hence, to much more accurate community detection in groups of layers than in [41].
3. The SSC methodology relies on clustering the $L\times L$ matrix of weights rather than $n(n-1)/2\times L$ matrix as in [41] and, therefore, is suitable for much larger networks. The matrix of weights is obtained, column per column, by solving independent systems of linear equations which can be solved in parallel, thus, considerably speeding up the calculations.
4. Although the SSC approach has been recently used in the some network models (see, e.g., [35,37,38]), to the best of our knowledge, it has not been applied to multilayer networks. Moreover, this paper is the first one to offer assessment of clustering precision of an SSC-based algorithm applied to Bernoulli type data. This requires a different set of assumptions from a traditional application of SSC to Gaussian data, and a novel clustering algorithm.

Remark 1 (*Relation to Randomly Generated Networks*)**.** While imposing assumptions on the DIMPLE network later in Section 3.1, we postulate that the groups of layers as well as community assignments in those groups of layers are generated by random sampling, similarly to how this is done in, e.g., [5,6]. This seemingly connects our paper to a large body of literature on randomly generated networks (see, e.g., [23] and references therein). In many of such papers, the inference also exploits this randomness by using, e.g., EM algorithm (see, e.g., [2,33]) or MCMC technique (see, e.g., [10]). Our paper, however, does not impose a fully Bayesian model. In fact, the block connectivity matrices in the layers are completely arbitrary and, therefore, act as nuisance parameters. Moreover, the generative mechanism stated in Section 3.1 is not required for the validity of the inference. Instead, one can just place assumptions on the linear subspaces associated with each group of layers. These assumptions, however, may not feel intuitive to a reader, in spite of being easily satisfied when communities in the groups of layers are generated at random. For this reason, although we develop the theory for a multiplex network where communities in the groups of layers are generated at random, Algorithms 1–3 will lead to the same clustering errors under alternative assumptions (see version 1 of [36]).

Consequently, we do not employ the EM or the MCMC as an inference methodology. While those techniques have their merits, they are iterative, require convergence analysis and, unlike the methodology used in the present paper, they come without precision guarantees. In addition, the steps of our computational algorithms, such as spectral clustering and solution of the LASSO problem, are very standard techniques in computer science and, hence, allow to employ well optimized algorithms that work fast for large networks with many layers, as it is evident from our simulations in Supplementary Section S1.

Remark 2 (*Relation to Time-Varying Networks*)**.** The DIMPLE model can also be viewed as a generalization of a time-varying network where the block connectivity matrices have very erratic behavior (change from one time frame to another) but the community structures are rather stable and change with a jump. However, the inference in the DIMPLE model (and in the multiplex networks in general) is much more difficult than in the respective dynamic network since, in a dynamic network, the layers are ordered according to time instances, while in a multiplex network the enumeration of layers is completely arbitrary. The latter makes techniques designed for dynamic networks unsuitable for the inference in the DIMPLE model setting. For this reason, in this paper, we do not review approaches designed for dynamic network models.

1.4. Notations and organization of the paper

For any vector $\mathbf{v}\in\mathbb{R}^p$, denote its ℓ_2, ℓ_1, ℓ_0, and ℓ_∞ norms by $\|\mathbf{v}\|$, $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_0$ and $\|\mathbf{v}\|_\infty$, respectively. Denote by $\mathbf{1}_m$ the m-dimensional column vector with all components equal to one.

For any matrix A, denote its spectral and Frobenius norms by, respectively, $\|A\|$ and $\|A\|_F$. The column j and the row i of a matrix A are denoted by $A(:,j)$ and $A(i,:)$, respectively. Let $\mathrm{vec}(A)$ be the vector obtained from matrix A by sequentially stacking its columns. Denote by $A\otimes B$ the Kronecker product of matrices A and B. Denote the diagonal of a matrix A by $\mathrm{diag}(A)$. Also, denote the K-dimensional diagonal matrix with $a_1,\dots,a_K$ on the diagonal by $\mathrm{diag}(a_1,\dots,a_K)$.

For any matrix $A\in\mathbb{R}^{n\times m}$, denote its projection on the nearest rank K matrix or its rank K approximation by $\Pi_K(A)$, that is, if σ_k are the singular values, and $\mathbf{u}_k$ and $\mathbf{v}_k$ are the left and the right singular vectors of A, $k\in[r]$, then

$$A=\sum_{k=1}^{r}\sigma_k\mathbf{u}_k\mathbf{v}_k^\top \quad\Rightarrow\quad \Pi_K(A)=\sum_{k=1}^{\min(r,K)}\sigma_k\mathbf{u}_k\mathbf{v}_k^\top.$$

Denote

$$\mathcal{O}_{n,K} = \left\{A \in \mathbb{R}^{n\times K} : A^\top A = I_K\right\}, \quad \mathcal{O}_n = \mathcal{O}_{n,n}.$$

A matrix $X \in \{0,1\}^{n_1\times n_2}$ is a clustering matrix if it is binary and has exactly one 1 per row. Also, we denote an absolute constant independent of n, K, L and M, which can take different values at different instances, by $\mathbb{C}$.

The rest of the paper is organized as follows. Section 2 introduces the between-layer and the within-layer clustering procedures. Specifically, Section 2.1 reviews the Sparse Subspace Clustering (SSC) methodology and explains why it is a good candidate for the job. Sections 2.2 and 2.3 present a solution to the central inference task in the DIMPLE model setting: the between-layer clustering. The between-layer clustering is carried out by Algorithms 1 and 2 which, as it is shown later, ensure strongly consistent between layer clustering. Section 2.4 studies our within-layer clustering technique. Section 3.1 provides theoretical guarantees for the accuracy of the clustering algorithms in Section 2. After introducing some necessary assumptions in Section 3.1, Section 3.2 proves the strong consistency of the between-layer clustering technique employed in the paper, while Section 3.3 shows that the latter leads to very low clustering errors. Section 4 provides concluding remarks. All the proofs are given in Appendix A. Supplementary Section S1 contains a limited simulation study and numerical comparisons with the spectral clustering approach developed in [41]. Supplementary Section S2 illustrates the methodology of the paper with a real data example.

2. The between-layer and the within-layer clustering procedures

Since the between-layer clustering is the major part of the inference in the DIMPLE model, we start with this task. As it has been mentioned earlier, this task is accomplished by the Sparse Subspace Clustering (SSC) and requires a combination of two parts, Algorithms 1 and 2.

2.1. Review of sparse subspace clustering techniques

Subspace clustering has been widely used in computer vision and, for this reason, it is a very well studied and developed methodology. Subspace clustering is designed for separation of points that lie in the union of subspaces. Let $\mathbf{x}^{(\ell)} \in \mathbb{R}^D$, $\ell \in [L]$ be a given set of points drawn from an unknown union of $M \geq 1$ linear or affine subspaces $\mathcal{S}_i$, $i \in [M]$, of unknown dimensions $d_i = \dim(\mathcal{S}_i)$, $0 < d_i < D$, $i \in [M]$. In the case of linear subspaces, the subspaces can be described as $\mathcal{S}_i = \left\{\mathbf{x} \in \mathbb{R}^D : \mathbf{x} = \mathcal{U}^{(i)}\mathbf{f}\right\}$, $i \in [M]$, where $\mathcal{U}^{(i)} \in \mathbb{R}^{D\times d_i}$ is a basis for subspace $\mathcal{S}_i$ and $\mathbf{f} \in \mathbb{R}^{d_i}$ is a low-dimensional representation for point $\mathbf{x}$. The goal of subspace clustering is to find the number of subspaces M, their dimensions d_i, $i \in [M]$, the subspace bases $\mathcal{U}^{(i)}$, $i \in [M]$, and the segmentation of the points according to the subspaces.

Several methods have been developed to implement subspace clustering such as algebraic methods [49], iterative methods [47] and spectral clustering based methods [12,45,48]. In this paper, we shall use the latter group of techniques. Spectral clustering algorithms rely on construction of an affinity matrix whose entries are based on some distance measures between the points. For example, in the case of the SBM, adjacency matrix itself serves as the affinity matrix, while for the Degree Corrected Block Model (DCBM) [20], the affinity matrix is obtained by normalizing rows/columns of the adjacency matrix. In the case of the subspace clustering problem, one cannot use the typical distance-based affinity measures because two points could be very close to each other, but lie in different subspaces, while they could be far from each other, but lie in the same subspace. One of the solutions is to construct the affinity matrix using self-representation of the points, with the expectation that a point is more likely to be presented as a linear combination of points in its own subspace rather than from a different one. A number of approaches such as Low Rank Representation [27] and Sparse Subspace Clustering (SSC) [11,12] have been proposed for the solution of this problem.

In this paper we use the self-representation version of the SSC developed in [12]. The technique is based on representation of each of the vectors as a sparse linear combination of all other vectors. The weights obtained by this procedure are used to form the affinity matrix which, in turn, is partitioned using the spectral clustering methods. If vectors $\mathbf{x}^{(\ell)}$, $\ell \in [L]$, were known, the weight matrix W would be based on writing every vector as a sparse linear combination of all other vectors by minimizing the number of nonzero coefficients

$$\min_{\mathbf{w}^{(\ell)}} \left\|\mathbf{w}^{(\ell)}\right\|_0 \quad \text{s.t.} \quad \mathbf{x}^{(\ell)} = \sum_{k\neq\ell} W_{k,\ell}\mathbf{x}^{(k)}, \quad \mathbf{w}^{(\ell)} = W(:,\ell). \tag{2}$$

The affinity matrix of the SSC is the symmetrized version of the weight matrix W. Since the problem (2) is NP-hard, one usually solves its convex relaxation, with $\|\mathbf{w}^{(\ell)}\|_0$ in (2) replaced by $\|\mathbf{w}^{(\ell)}\|_1$. If the vectors $\mathbf{x}^{(\ell)}$ in (2) are unknown, one uses sample-based estimated values.

2.2. Finding the matrix of weights

In this section, we show that vectorized versions of the connection probability matrices $P^{(\ell)}$, corresponding to different groups of layers, lie in distinct subspaces and, hence, can be used for the between layer clustering.

In order to partition the layers of the network into groups with the distinct community structures, note that

$$\mathrm{vec}(P^{(\ell)}) = (Z^{(m)} \otimes Z^{(m)})\mathbf{b}^{(\ell)}, \quad \mathbf{b}^{(\ell)} = \mathrm{vec}(B^{(\ell)}), \quad m = c(\ell), \quad \ell \in [L]. \tag{3}$$

Algorithm 1 Finding the matrix of weights

Input: Tensor $\mathbf{A}$; the number of communities K in each layer; parameter λ.

Output: matrix $\widehat{\widehat{W}}$ of weights.

Steps:

1: For $\ell \in [L]$, find pre-conditioned rank $(K-1)$ approximations $\widehat{\widehat{P}}^{(\ell)}$ of $A^{(\ell)} = \mathbf{A}(:,:,\ell)$, using formula (8).

2: Construct matrix $Y \in \mathbb{R}^{n^2 \times L}$ with columns $\mathbf{y}^{(\ell)}$, $\ell \in [L]$, defined in (9).

3: Find a matrix of weights, $\widehat{W} \in \mathbb{R}^{L \times L}$ with columns $\widehat{\mathbf{w}}^{(\ell)} = \widehat{W}(:,\ell)$ and $\mathrm{diag}(\widehat{W}) = 0$, by solving the LASSO problem (10) for $\ell \in [L]$.

4: Construct matrix $\widehat{\widehat{W}} = |\widehat{W}| + |\widehat{W}^\top|$ of weights.

Hence, for $m = c(\ell)$, vectors $\mathrm{vec}(P^{(\ell)})$ belong to distinct subspaces $\overline{S}_m = \mathrm{span}(Z^{(m)} \otimes Z^{(m)})$. Denote

$$D^{(m)} = (Z^{(m)})^\top (Z^{(m)}) = \mathrm{diag}(n_1^{(m)}, \ldots, n_K^{(m)}), \quad U^{(m)} = Z^{(m)}(D^{(m)})^{-1/2}, \quad m \in [M],$$

and observe that $U^{(m)} \in \mathcal{O}_{n,K}$. Therefore, (3) can be rewritten as

$$\mathrm{vec}(P^{(\ell)}) = (U^{(m)} \otimes U^{(m)}) \left(\sqrt{D^{(m)}} \otimes \sqrt{D^{(m)}} \right) \mathbf{b}^{(\ell)}, \quad m = c(\ell), \ \ell \in [L], \tag{4}$$

so that $\overline{S}_m = \mathrm{span}(U^{(m)} \otimes U^{(m)})$. Eqs. (3) and (4) confirm that vectors $\mathrm{vec}(P^{(\ell)})$ lie in distinct subspaces $\overline{S}_m$ with $m = c(\ell)$ and, hence, possibly can be partitioned into groups using subspace clustering.

Yet, there is one potential complication in applying subspace clustering to the problem above. Indeed, the subspace clustering works well when the subspaces do not intersect or have insignificant intersection. However, each of the subspaces $\overline{S}_m$ includes $n^{-1}\mathbf{1}_{n^2}$ as its main basis vector. The latter is likely to compromise the precision of subspace clustering techniques. However, luckily, it is relatively easy to remove this vector from all subspaces. Consider a projection matrix

$$\mathscr{P} = n^{-1}\mathbf{1}_n\mathbf{1}_n^\top, \quad \mathscr{P}^2 = \mathscr{P}. \tag{5}$$

Then, for

$$\begin{aligned} \widetilde{P}^{(\ell)} &= (I - \mathscr{P})P^{(\ell)}(I - \mathscr{P}) = (I - \mathscr{P})Z^{(m)}B^{(\ell)}(Z^{(m)})^\top(I - \mathscr{P}), \\ \widetilde{U}^{(m)} &= (I - \mathscr{P})U^{(m)} = (I - \mathscr{P})Z^{(m)}(D^{(m)})^{-1/2}, \quad m \in [M], \end{aligned} \tag{6}$$

and $\mathbf{b}^{(\ell)}$ defined in (3), one has, for $m = c(\ell)$

$$\mathbf{q}^{(\ell)} = \mathrm{vec}(\widetilde{P}^{(\ell)}) = (\widetilde{U}^{(m)} \otimes \widetilde{U}^{(m)})\widetilde{\mathbf{b}}^{(\ell)}, \ \widetilde{\mathbf{b}}^{(\ell)} = \left(\sqrt{D^{(m)}} \otimes \sqrt{D^{(m)}} \right) \mathbf{b}^{(\ell)}. \tag{7}$$

Consider subspaces $S_m = \mathrm{span}(\widetilde{U}^{(m)} \otimes \widetilde{U}^{(m)})$ with dimension $(K-1)^2 = \mathrm{rank}(\widetilde{U}^{(m)} \otimes \widetilde{U}^{(m)})$. In many scenarios, the new subspaces S_m have very little or no intersection and, hence, can be well separated using the subspace clustering technique.

In the case of the DIMPLE model, vectors $\mathbf{q}^{(\ell)}$, $\ell \in [L]$, are unavailable. Instead, we use their proxies based on the adjacency matrices. Specifically, we consider matrices

$$\widehat{\widehat{P}}^{(\ell)} = \Pi_{K-1}(\widetilde{A}^{(\ell)}), \quad \widetilde{A}^{(\ell)} = (I - \mathscr{P})A^{(\ell)}(I - \mathscr{P}), \quad \mathscr{P} = n^{-1}\mathbf{1}_n\mathbf{1}_n^\top. \tag{8}$$

Here $\widehat{\widehat{P}}^{(\ell)}$ is the rank $(K-1)$ approximation of $\widetilde{A}^{(\ell)}$. Construct matrices $Y, \widehat{Q} \in \mathbb{R}^{n^2 \times L}$ with columns $\mathbf{y}^{(\ell)}$ and $\hat{\mathbf{q}}^{(\ell)}$, respectively, given by

$$\mathbf{y}^{(\ell)} = Y(:,\ell) = \hat{\mathbf{q}}^{(\ell)}/\|\hat{\mathbf{q}}^{(\ell)}\|, \quad \hat{\mathbf{q}}^{(\ell)} = \mathrm{vec}\left(\widehat{\widehat{P}}^{(\ell)}\right), \quad \ell \in [L]. \tag{9}$$

In the case of data contaminated by noise, the SSC algorithm does not attempt to write each $\mathbf{y}^{(\ell)}$ as an exact linear combination of other points. Instead, the SSC is built upon solutions of the LASSO problems

$$\widehat{\mathbf{w}}^{(\ell)} \in \underset{\mathbf{w} \in \mathbb{R}^L, \mathbf{w}_\ell = 0}{\mathrm{argmin}} \left\{ \left\| \mathbf{y}^{(\ell)} - Y\mathbf{w} \right\|^2 + 2\lambda \|\mathbf{w}\|_1 \right\}, \ \ell \in [L], \tag{10}$$

where $\lambda > 0$ is the tuning parameter of LASSO. There are many ways of identifying the LASSO parameter, see, e.g., [3,7,14] among others. We elaborate on the choice of the Lasso parameter in Supplementary Section S1 We solve (10) using a fast version of the LARS algorithm implemented in SPAMS Matlab toolbox [32].

Given $\widehat{W}$, the clustering function $\hat{c} : [L] \to [M]$ is obtained by applying spectral clustering to the affinity matrix $|\widehat{W}| + |\widehat{W}^\top|$, where, for any matrix B, matrix $|B|$ has absolute values of elements of B as its entries. Algorithm 1 summarizes the methodology described above.

Algorithm 2 The between-layer clustering

Input: Matrix $\widehat{\widehat{W}} \in \mathbb{R}^{L\times L}$ of weights; matrix Y with columns $\mathbf{y}^{(\ell)} = Y(:,\ell)$ defined in (9); the number of groups of layers M, threshold T.
Output: The clustering function $\hat{c} : [L] \to [M]$ and the corresponding clustering matrix $\widehat{C}$.
Steps:
1: Find $\hat{\hat{c}} : [L] \to [M]$ by applying spectral clustering to $\widehat{\widehat{W}}$. Find the corresponding clustering matrix $\widehat{\widehat{C}}$.
2: Find $\widehat{D} = \mathrm{diag}(\widehat{\widehat{W}}\mathbf{1})$ and the Laplacian $\mathcal{L} = \widehat{D} - \widehat{\widehat{W}}$. Find $\widetilde{M}$, the number of disconnected components of $\mathcal{L}$ and the clustering function $\phi : [L] \to [\widetilde{M}]$.
3: If $\widetilde{M} \le M$, then $\hat{c} = \hat{\hat{c}}$ and $\widehat{C} = \widehat{\widehat{C}}$.
4: If $\widetilde{M} > M$, then construct matrix $\widehat{Y} \in \mathbb{R}^{L\times L}$ with elements $\widehat{Y}_{\mathfrak{l}_1,\mathfrak{l}_2} = |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}|$, where $\mathfrak{l}_1, \mathfrak{l}_2 \in [L]$, and $\mathbf{y}^{(\ell)} = Y(:,\ell)$ are defined in (9).
5: Let $\boldsymbol{\Phi} \in \{0,1\}^{L\times\widetilde{M}}$ be the clustering matrix corresponding to the clustering function ϕ. Let $D_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}^\top\boldsymbol{\Phi}$. Construct matrix $\widehat{\widehat{Y}} = (D_{\boldsymbol{\Phi}})^{-1/2}\boldsymbol{\Phi}^\top\widehat{Y}\boldsymbol{\Phi}(D_{\boldsymbol{\Phi}})^{-1/2} \in \mathbb{R}^{\widetilde{M}\times\widetilde{M}}$ and its thresholded version $\widehat{G} \in \{0,1\}^{\widetilde{M}\times\widetilde{M}}$ with elements $\widehat{G}_{\widetilde{m}_1,\widetilde{m}_2} = I(\widehat{\widehat{Y}}_{\widetilde{m}_1,\widetilde{m}_2} > T)$, $\widetilde{m}_1, \widetilde{m}_2 \in [\widetilde{M}]$.
6: Find the SVD $\widehat{G} = U_{\widehat{G}}\Lambda_{\widehat{G}}(U_{\widehat{G}})^\top$ of $\widehat{G}$, and cluster rows of $U_{\widehat{G}}(:, 1:M)$ into M clusters. Obtain clustering function $\theta : [\widetilde{M}] \to [M]$ and the corresponding clustering matrix $\boldsymbol{\Theta}$.
7: Set $\widehat{C} = \boldsymbol{\Phi}\boldsymbol{\Theta}$ and $\hat{c}(\ell) = \theta(\phi(\ell))$, $\ell \in [L]$, superposition of θ and ϕ.

2.3. Identifying clusters of layers

As a result of Algorithms 1, one obtains a matrix $\widehat{\widehat{W}} = |\widehat{W}| + |\widehat{W}^\top|$ of weights. Then, one can apply spectral clustering to $\widehat{\widehat{W}}$, partitioning L layers into M clusters.

The success of clustering relies on the fact that the weight matrix $\widehat{W}$ is such that $\widehat{W}_{k,\ell} \neq 0$ only if points k and ℓ lie in the same subspace, which guarantees that vectors $\mathbf{y}^{(\ell)}$ are represented by vectors in their own cluster only. This notion is formalized as the Self-Expressiveness Property. Specifically, we say that the weight matrix $W \in \mathbb{R}^{L\times L}$ satisfies the Self-Expressiveness Property (SEP) if $|W(i,j)| > 0$ implies $c(i) = c(j)$, where $c : [L] \to [M]$ is the true clustering function. Hence, for the success of clustering, we would like to ensure that matrix $\widehat{W}$ with columns $\widehat{\mathbf{w}}^{(\ell)}$, $\ell \in [L]$, defined in (10), satisfies the SEP with high probability. Indeed, if SEP holds, then no two layer networks from different groups of layers can have a nonzero weight in the matrix $\widehat{\widehat{W}}$.

However, it is known that SEP alone does not guarantee perfect clustering since the similarity graph obtained on the basis of $\widehat{\widehat{W}}$ can be poorly connected (see, e.g., [34]). Indeed, if the similarity graph has $\widetilde{M} > M$ disconnected components, then one would obtain spurious clustering errors due to the incorrect grouping of those components. It is possible to have $\widetilde{M} > M$ since, within one subspace, one can have a group of vectors that can be expressed as weighted sums of each other. The connectivity issue has been addressed in, e.g., [50], where the authors proved that the SSC achieves correct clustering with high probability under the restricted eigenvalue assumption. They propose an innovative algorithm for merging subspaces by using single linkage clustering of the disconnected components. Since we cannot guarantee that the restricted eigenvalue assumption holds in our case, we suggest a different novel methodology for clustering the disconnected components into M clusters. The method is summarized in Algorithm 2. Algorithm 2 requires milder conditions and is easier to implement than the respective technique in [50].

2.4. The within-layer clustering procedure

After the groups of layers are identified by Algorithm 2, one can find the communities by some kind of averaging. Specifically, following [25,41], we average the estimated de-biased versions of the squares of the probability matrices $P^{(\ell)}$. Specifically, we introduced matrix $\widehat{\Psi}$

$$\widehat{\Psi} = \widehat{C}(\widehat{D}_{\hat{c}})^{-1/2} \in \mathcal{O}_{L,M}, \quad \text{with} \quad \widehat{D}_{\hat{c}} = \widehat{C}^\top\widehat{C}, \tag{11}$$

and subsequently construct a tensor $\widehat{\mathbf{G}} \in \mathbb{R}^{n\times n\times L}$ with layers $\widehat{G}^{(\ell)} = \widehat{\mathbf{G}}(:,:,\ell)$ of the form

$$\widehat{G}^{(\ell)} = \left(A^{(\ell)}\right)^2 - \mathrm{diag}(\widehat{\mathbf{d}}^{(\ell)}), \quad \ell \in [L],$$

where $\widehat{\mathbf{d}}^{(\ell)}$ is the vector of estimated nodes' degrees. After that, we average layers of the same types, obtaining tensor

$$\widehat{\mathbf{H}} = \widehat{\mathbf{G}} \times_3 \widehat{\Psi}^\top \in \mathbb{R}^{n\times n\times M},$$

where $\widehat{\Psi}$ is defined in (11). The communities in each group $m \in [M]$, of layers are obtained by application of the spectral clustering to layers of tensor $\widehat{\mathbf{H}}$. The procedure is summarized in Algorithm 3.

Algorithm 3 The within-layer clustering

Input: Adjacency tensor $\mathbf{A} \in \{0,1\}^{n\times n\times L}$, number of groups of layers M, number of communities K, estimated layer clustering matrix $\widehat{C} \in \mathcal{M}_{L,M}$.
Output: Estimated community assignments $\widehat{Z}^{(m)} \in \mathcal{M}_{n,K}$, $m \in [M]$.
Steps:
1: Construct tensor $\widehat{\mathbf{G}}$ with layers $\widehat{G}^{(\ell)} = \left(A^{(\ell)}\right)^2 - \mathrm{diag}(A^{(\ell)}\mathbf{1}_n)$, $\ell \in [L]$.
2: Construct tensor $\widehat{\mathbf{H}}$ using formula $\widehat{\mathbf{H}} = \widehat{\mathbf{G}} \times_3 \widehat{\Psi}^\top$.
3: Construct the SVDs of layers $\widehat{H}^{(m)} = \widetilde{U}_{\widehat{H}}{}^{(m)} \widehat{\Lambda}_{\widehat{H}}{}^{(m)} (\widetilde{U}_{\widehat{H}}{}^{(m)})^\top$, $m \in [M]$.
4: Find $\widehat{U}_{\widehat{H}}{}^{(m)} = \widetilde{U}_{\widehat{H}}{}^{(m)}(:, 1:K) = \Pi_K(\widetilde{U}_{\widehat{H}}{}^{(m)})$, $m \in [M]$.
5: Cluster rows of $\widehat{U}_{\widehat{H}}{}^{(m)}$ into K clusters using $(1+\epsilon)$-approximate K-means clustering. Obtain clustering matrices $\widehat{Z}^{(m)}$, $m \in [M]$.

3. Theoretical guarantees

3.1. Assumptions

In this paper, we assume that a DIMPLE network is generated by randomly sampling the nodes similarly to how this is done in [5,6].

There exists a large body of literature on randomly generated networks (see, e.g., [23] and references therein). In many of the papers, the inference is also carried out by some randomized procedures such as EM algorithm (see, e.g., [2,33]) or MCMC technique (see, e.g., [10]). While these techniques have their merits, the advantage of the clustering algorithms developed in this paper is that they are not iterative and, hence, do not require provable convergence. In addition, they come with the theoretical precision guarantees. Hence, the generative mechanism for the network is not required for the validity of the inference. Instead, one can just place assumptions on the linear subspaces associated with each group of layers. These assumptions, however, may not feel intuitive to a reader, in spite of being easily satisfied when communities in the groups of layers in the network are generated at random. For this reason, although we develop the theory for a multiplex network where communities in the groups of layers are generated at random, Algorithms 1–3 will lead to the same clustering errors under alternative assumptions (see version 1 of [36]).

In what follows, we consider vectors $\overline{\boldsymbol{\varpi}} = (\varpi_1, \dots, \varpi_M) \in [0,1]^M$ and $\overline{\boldsymbol{\pi}}^{(m)} = (\pi_1^{(m)}, \dots, \pi_K^{(m)}) \in [0,1]^K$, $m \in [M]$, such that

$$\sum_{m=1}^{M} \varpi_m = 1, \quad \sum_{k=1}^{K} \pi_k^{(m)} = 1, \quad m \in [M].$$

For each layer $\ell \in [L]$, we generate its group membership $c(\ell) \sim \mathrm{Multinomial}(\overline{\boldsymbol{\varpi}})$. For each node $j \in [n]$ in a layer of type $m \in [M]$, the membership function $z^{(m)}$ is generated as $z^{(m)}(j) \sim \mathrm{Multinomial}(\overline{\boldsymbol{\pi}}^{(m)})$. Hence, ϖ_m is the probability of a layer of type m, and $\pi_k^{(m)}$, $k \in [K]$, is the probability of the kth community in a layer of type m.

While, in general, the values of $\pi_k^{(m)}$ can be different for different m, in this paper, we assume that $\pi_k^{(m)} = \pi_k$, $m \in [M]$, $k \in [K]$. The latter means that for a node j in a group of layers m, its community membership can be generated as

$$\begin{aligned} &\xi_j^{(m)} \sim \mathrm{Multinomial}(\bar{\boldsymbol{\pi}}, K) \quad \text{with} \quad \bar{\boldsymbol{\pi}} = (\pi_1, \dots, \pi_K), \quad j \in [n], \\ &Z_{j,k}^{(m)} = I(\xi_j^{(m)} = k), \quad \Pr(Z_{j,k}^{(m)} = 1) = \pi_k, \quad k \in [K],\ j \in [n],\ m \in [M]. \end{aligned} \tag{12}$$

After layers' memberships and nodes' memberships in groups of layers are generated, the set of matrices $B^{(\ell)}$ is chosen independently from the groups of layers and community assignments.

In order to derive theoretical guarantees for the SEP, one needs to impose conditions that ensure that the layer networks maintain some regularity and are not too sparse. We also need to ensure that the subspaces, that represent the layer networks, are sufficiently separated, and are also well represented by the sets of vectors $\mathbf{q}^{(\ell)}$ with $c(\ell) = m$, where $\mathbf{q}^{(\ell)}$ are defined in (7). For this purpose, we introduce matrices $Q, X \in \mathbb{R}^{n^2\times L}$ with columns $\mathbf{q}^{(\ell)}$ and $\mathbf{x}^{(\ell)}$, respectively, where

$$\mathbf{x}^{(\ell)} = X(:,\ell) = \mathbf{q}^{(\ell)}/\|\mathbf{q}^{(\ell)}\|, \quad \mathbf{q}^{(\ell)} = \mathrm{vec}(\widetilde{P}^{(\ell)}), \quad \ell \in [L]. \tag{13}$$

Matrix X can be viewed as the "true" version of matrix Y in (9). We impose the following assumptions:
A1. For some positive constants $\underline{C}$ and $\bar{C}$, $0 < \underline{C} \le \bar{C} < \infty$, one has

$$B^{(\ell)} = \rho_n B_0^{(\ell)} \quad \text{with} \quad \underline{C} \le \|B_0^{(\ell)}\| \le \bar{C}. \tag{14}$$

A2. For some positive constant $C_{\sigma,0}$, one has

$$\min_{\ell\in[L]} \sigma_{\min}(B_0^{(\ell)})/\sigma_{\max}(B_0^{(\ell)}) \ge C_{\sigma,0}.$$

A3. For some positive constant C_ρ, one has

$$\rho_n \geq C_\rho n^{-1} \ln n.$$

A4. For some positive constants $\underline{c}_\varpi$, $\bar{c}_\varpi$, $\underline{c}_\pi$ and $\bar{c}_\pi$, one has

$$\underline{c}_\varpi / M \leq \varpi_m \leq \bar{c}_\varpi / M; \quad \underline{c}_\pi / K \leq \pi_k \leq \bar{c}_\pi / K; \quad k \in [K],\ m \in [M]. \tag{15}$$

A5. Matrices $B^{(\ell)}$ are such that, for any $\ell \in [L]$ with $c(\ell) = m$, there exists representation $\mathbf{x} = \widetilde{X}_* \mathbf{w}_*$ of $\mathbf{x} \equiv \mathbf{x}^{(\ell)}$ via other columns of X in $\mathcal{S}_m$, such that $\|\mathbf{w}_*\|_1 \leq \aleph_{w,K}$ where $\aleph_{w,K}$ can only depend on K.

Assumptions A1–A4 are common regularity assumptions for network papers. Since majority of networks are sparse, Assumption A1 introduces a sparsity factor ρ_n and confirms that all matrices $B^{(\ell)}$ maintain approximately the same level of sparsity. Assumption A2 requires that all matrices $B_0^{(\ell)}$, $\ell \in [L]$, are well conditioned. Assumption A3 guarantees that the eigenvectors of the subspaces constructed on the basis of the adjacency matrices are close to those that are defined by the matrices of probabilities of connections. Assumption A4 ensures that groups of layers in the network, as well as communities in each of the groups, are balanced, i.e., the number of members have the same order of magnitude when n and L grow. Denote

$$\widehat{L}_m = \sum_{\ell=1}^{M} I(c(\ell) = m), \quad \hat{n}_k^{(m)} = \sum_{j=1}^{n} I(\xi_j^{(m)} = k), \quad k \in [K],\ m \in [M]. \tag{16}$$

Then, it turns out that, under Assumption A4, there is a set Ω_t such that, for $\omega \in \Omega_t$

$$\min_m \widehat{L}_m \geq C_0\, L/M, \quad \widetilde{C}_0\, n/K \leq \hat{n}_k^{(m)} \leq \widetilde{\widetilde{C}}_0\, n/K, \quad m \in [M],\ k \in [K]. \tag{17}$$

It follows from Lemma 2 in Appendix A that, if L and n are sufficiently large, (17) holds with

$$C_0 = \underline{c}_\varpi/2, \quad \widetilde{C}_0 = \underline{c}_\pi/2, \quad \widetilde{\widetilde{C}}_0 = 3\bar{c}_\pi/2$$

on a set Ω_t with $\Pr(\Omega_t) \geq 1 - 2L^{-t} - 2K\, n^{-t}$. It turns out that Assumption A4 also ensures that groups of layers of the network are well separated.

Assumption A5 replaces much more stringent conditions, which are present in majority of papers that provide theoretical guarantees for the sparse subspace clustering, specifically, the assumption of sufficient sampling density and spherical symmetry of the residuals. While neither of these above conditions holds in our setting, Assumption A5 is much easier to satisfy. It actually requires that the low-dimensional vectors $\mathbf{b}^{(\mathfrak{l}_0)}$ are easily represented by other vectors $\mathbf{b}^{(\ell)}$, where $c(\ell) = c(\mathfrak{l}_0)$ and $\ell \neq \mathfrak{l}_0$. Assumption A5 is valid under a variety of sufficient conditions. Some examples of those conditions are presented in the following lemma.

Lemma 1. *(a) Consider vectors $\mathbf{b}_0^{(\ell)} = vec(B_0^{(\ell)})$ where matrices $B_0^{(\ell)}$ are defined in (14). Let, for any $m \in [M]$ and any $\mathfrak{l}_0$ with $c(\mathfrak{l}_0) = m$, there exist a set of indices $\mathcal{L}_0$ such that $\mathfrak{l}_0 \notin \mathcal{L}_0$, $c(\ell) = m$ for $\ell \in \mathcal{L}_0$, and matrix $\mathcal{B}_0$ with columns $\mathbf{b}_0^{(\ell)}$, $\ell \in \mathcal{L}_0$, is a full-rank matrix with the lowest singular value $\sigma_{min}(\mathcal{B}_0) \geq \sigma_{0,K}$, where $\sigma_{0,K}$ can only depend on K. Then, Assumption A5 holds with*

$$\aleph_{w,K} = \frac{(\bar{C})^2\, \widetilde{\widetilde{C}}_0}{\underline{C}\, C_{\sigma,0}\, \widetilde{C}_0} \frac{K\sqrt{K}}{\sigma_{0,K}}.$$

(b) If, for $m \in [M]$, matrices $B^{(\ell)}$ with $c(\ell) = m$ take only M_m distinct values, with at least two matrices $B^{(\ell)}$ taking identical values, then Assumption A5 holds with $\aleph_{w,K} = 1$.

Note that part (a) of Lemma 1 just prevents the situation where all but one of the vectors $\mathbf{b}^{(\ell)}$ are positioned in close proximity of one another. Part (b) of Lemma 1 includes the MMLSBM as its particular case, which means that our theoretical results also hold for the MMLSBM.

3.2. Between-layer clustering precision guarantees

The success of clustering relies on the fact that the weight matrix $\widehat{W}$ with columns $\hat{\mathbf{w}}^{(\ell)}$, $\ell \in [L]$, defined in (10), satisfies the SEP with high probability. It turns out that Assumption A3 ensures that subspaces $\mathcal{S}_m$, $m \in [M]$, corresponding to different types of layers, do not have large intersections and allow sparse representation of vectors within each subspace. The following statement guarantees that this is true for the weight matrix $\widehat{W}$ in Algorithm 1.

Theorem 1. *Let Assumptions A1–A5 hold and $t > 0$. Define*

$$\delta_{n,K,t} = C_{t,\delta}\, K\, (n\rho_n)^{-1/2}, \tag{18}$$

where $C_{t,\delta}$ is a constant that depends only on t and constants in Assumptions A1–A4. Let $\widehat{W}$ be a solution of problem (10) with $\lambda = \lambda_{n,K}$ such that

$$\lambda_{n,K} \leq (4\aleph_{w,K})^{-1}, \quad \lim_{n\to\infty} \frac{\delta_{n,K,t}\ \aleph_{w,K}}{\lambda_{n,K}} = 0, \tag{19}$$

where $\aleph_{w,K}$ is defined in Assumption A5. If n is large enough and $t > 0$ satisfies

$$t < \min\left(\underline{c}_{\varpi}^2\, L\,(2M^2 \ln L)^{-1},\quad \underline{c}_{\pi}^2\, n\,(2K^2 \ln n)^{-1}\right), \tag{20}$$

then matrix $\widehat{W}$ (and, consequently, $\widehat{\widehat{W}}$) satisfies the SEP with high probability:

$$\Pr(\widehat{W} \text{ satisfies SEP}) \geq 1 - 2\,L^{-t} - L\,n^{-t} - 2\,KM(M+2)n^{-t}.$$

We would like to point out the fact that although the statement in Theorem 1 is relatively standard, its proof follows completely different path than proofs of SEP known to us. Indeed, those proofs (see, e.g., [44,45,51]) are tailored to the case of Gaussian errors and are based on the idea that the errors are rotationally invariant. In addition, those proofs require that the sampled vectors uniformly cover each of the subspaces. It is easy to observe that rotational invariance fails in the case of the Bernoulli random vectors, so our proof is totally original. Moreover, we do not require the sampling condition as in, e.g., [45,51]. Observe that condition A5 does not require uniform sampling or sufficient sampling density. Instead, condition A5 guarantees that each vector has a sparse representation via the vectors in the same subspace.

The following theorem states that, if the threshold $T = T_{n,K}$ in Algorithm 2 satisfies certain conditions, n is large enough and the SEP holds, then Algorithm 2 leads to perfect recovery of clusters with high probability. The latter implies that our clustering procedure is strongly consistent.

Theorem 2. *Let Assumptions A1–A5 hold and the clustering function $\hat{c} : [L] \to [M]$ be obtained by Algorithm 2. Let $T \equiv T_{n,K}$ be such that*

$$\lim_{n\to\infty} T_{n,K} = 0; \quad \lim_{n\to\infty}\left(\frac{K^2 \ln n}{T_{n,K}\, n} + \frac{K}{T_{n,K}\sqrt{n\,\rho_n}}\right) = 0. \tag{21}$$

If n is large enough and $t > 0$ satisfies (20), then, the clustering procedure is strongly consistent with high probability, i.e., up to permutation of M cluster labels, one has

$$\Pr(\hat{c} = c) \geq 1 - 2\,L^{-t} - L\,n^{-t} - 2\,KM(M+2)n^{-t}.$$

Note that Algorithm 2 is very different from Algorithm 2 of [50] which relies on subspaces recovery and merging. Also, Theorem 2 above holds under milder and more intuitive assumptions than Theorem 3.2 of [50]. In conclusion, Theorem 2 establishes strong consistency of SSC for data that is not rotationally invariant.

3.3. Within-layer clustering precision guarantees

After the between-layer clustering has been accomplished, the within layer clustering can be carried out by Algorithm 3.

Since the clustering is unique only up to a permutation of clusters, denote the set of K-dimensional permutation functions of $[K]$ by $\aleph(K)$ and the set of $K \times K$ permutation matrices by $\mathfrak{F}(K)$. The local community detection error in the layer of type m is then given by

$$R_{WL}(m) = (2n)^{-1} \min_{\mathscr{P}_m \in \mathfrak{F}(K)} \|\widehat{Z}^{(m)} - Z^{(m)}\mathscr{P}_m\|_F^2, \qquad m \in [M],$$

where $Z^{(m)}$ is defined in (1). Note that, since the numbering of layers is defined also up to a permutation, the errors $R_{WL}(1), \ldots, R_{WL}(M)$ should be minimized over the set of permutations $\aleph(M)$. The average error rate of the within-layer clustering is then given by

$$R_{WL} = \frac{1}{M} \min_{\aleph(M)} \sum_{m=1}^{M} R_{WL}(m) = \frac{1}{2\,M\,n} \min_{\aleph(M)} \sum_{m=1}^{M}\left(\min_{\mathscr{P}_m \in \mathfrak{F}(K)} \|\widehat{Z}^{(m)} - Z^{(m)}\mathscr{P}_m\|_F^2\right).$$

With these definitions, one obtains the following statement.

Theorem 3. *Let Assumptions A1–A5 hold and the between-layer clustering function $\hat{c} : [L] \to [M]$ be obtained by using Algorithm 2. Let $T \equiv T_{n,K}$ satisfy condition (21) and $t > 0$ obey (20). Then, for n large enough and an absolute positive constant C_t, the average within-layer clustering error R_{WL} satisfies*

$$\Pr\left\{R_{WL} \leq C_t\left(\frac{MK^4 \ln(L+n)}{Ln\rho_n} + \frac{K^4}{n^2}\right)\right\} \geq 1 - 2L^{-t} - C_t(KM^2 + L + n^2)\,n^{1-t}. \tag{22}$$

4. Discussion

The present paper considers the DIverse MultiPLEx (DIMPLE) network model, introduced in [41]. To the best of our knowledge, the latter is the only paper which considers such a general multiplex setting. However, while [41] applied spectral clustering to the proxy of the adjacency tensor, this paper uses the SSC for identifying groups of layers with identical community structures. The latter, unlike the technique of [41] leads to the strongly consistent between-layer clustering which, in turn, results in much more precise community detection in groups of layers. Indeed, under very similar assumptions, with high probability, the spectral clustering

Algorithm S1 of [41] leads to the between layer clustering error of $O(K^2(n\rho_n)^{-1})$ while Algorithms 1 and 2 in this paper yield precise clustering. Hence, due to lack of accuracy in identification the groups of layers, the within-layer clustering error $R_{WL}^{(PW)}$ in [41] is also much higher than the within-layer clustering error R_{WL} in the present paper. Specifically, for some constants C_1 and C_2, with high probability,

$$R_{WL}^{(PW)} \le C_1\left(\frac{MK^4\ln(L+n)}{n\rho_n L}+\frac{MK^6}{n\rho_n}\right), \quad R_{WL}\le C_2\left(\frac{MK^4\ln(L+n)}{n\rho_n L}+\frac{K^4}{n^2}\right), \tag{23}$$

where the second expression in (23) is a repetition of formula (22). While the first terms in the expressions of $R_{WL}^{(PW)}$ and R_{WL} coincide, the second term in $R_{WL}^{(PW)}$, which comes from the between layer clustering error of $O(K^2(n\rho_n)^{-1})$, is significantly larger than the one in R_{WL}.

Clustering methodology in this paper has a number of advantages. Not only it is strongly consistent with high probability when the number of nodes is large, but also competitive with (and often more precise than) the spectral clustering in [41]. In addition, the algorithm of [41] requires SVD of $n(n-1)/2\times L$ matrix, which is challenging for large n, while in our case the SVD is applied to $L\times L$ matrix. Hence, the SSC-based technique allows to handle much larger networks. Moreover, the most time consuming part of the algorithm, finding the weight matrix, is perfectly suitable for application of parallel computing which can significantly reduce the computational time.

Another side benefit of the present paper is the novel theoretical development in the area of assessment of the precision of the sparse subspace clustering when it is applied to non-Gaussian data. In particular, all papers known to us, provide theoretical guarantees for the sparse subspace clustering under the assumptions of the spherical symmetry of the residuals and sufficient sampling density (see, e.g., [44,45,51]). It is easy to observe that rotational invariance fails in the case of the Bernoulli random vectors. In addition, the assumption that the sampled vectors uniformly cover each of the subspaces may not be true either (for example, it does not hold for the MMLSBM). For this reason, our paper offers a completely original proof of the clustering precision of the SSC-based technique.

In addition, although the SSC has been applied to clustering single layer networks in [35,37,38], to the best of our knowledge, our paper offers the first application of the SSC to the binary multiplex network. While the weights in Algorithm 1 are obtained in a relatively conventional manner, our between-layer clustering Algorithm 2 is entirely original and very different from the one in [50].

CRediT authorship contribution statement

Majid Noroozi: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Marianna Pensky:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization.

Acknowledgments

The second author of the paper gratefully acknowledges partial support by National Science Foundation (NSF), USA grants DMS-2014928 and DMS-2310881. The MATLAB code used to produce the results in the Supplementary Section is available online on the authors' websites.

Appendix A. Proofs

Proof of Self-Expressiveness property under the separation condition

Proof of Theorem 1 relies on the fact that the subspaces S_m, $m\in[M]$, corresponding to different types of layers do not have large intersections. Specifically, we prove the following statement from which the validity of Theorem 1 will readily follow.

Proposition 1. *Let Assumptions A1, A2, A3, and A5 hold. Let L_m and $n_k^{(m)}$ be, respectively, the number of layers of type m and the number of nodes in the kth community in the group of layers of type m, where L_m and $n_k^{(m)}$ satisfy condition (17). Assume, in addition, that there exists $\tau\equiv\tau_{n,K}\in(0,1)$ such that for any arbitrary vectors $\mathbf{x}\in S_m$ and $\mathbf{x}'\in S_{m'}$, where $m\neq m'$, one has $|\mathbf{x}^\top\mathbf{x}'|\le\tau\|\mathbf{x}\|\,\|\mathbf{x}'\|$. Let $t>0$ and $\delta=\delta_{n,K,t}$ be defined in (18) where $C_{t,\delta}$ is a constant that depends only on t and constants in Assumptions A1, A2, A3, and A5, and condition (17).*

Let $\widehat{W}$ be a solution of problem (10) with $\lambda=\lambda_{n,K}$ such that

$$\lambda_{n,K}\le(4\aleph_{w,K})^{-1}, \quad \lim_{n\to\infty}\frac{(\delta_{n,K,t}+\tau_{n,K})(1+\aleph_{w,K})}{\lambda_{n,K}}=0, \tag{A.1}$$

where $\aleph_{w,K}$ is defined in Assumption A5. If $\max_{1\le\ell\le L}\|\mathbf{y}^{(\ell)}-\mathbf{x}^{(\ell)}\|\le\delta$ and n is large enough, then, matrix $\widehat{W}$ (and, consequently, $\widehat{\widehat{W}}$) satisfies the SEP.

Proof of Proposition 1. Let matrices $Q, X \in \mathbb{R}^{n^2 \times L}$ and $\widehat{Q}, Y \in \mathbb{R}^{n^2 \times L}$ be defined in (9) and (13), respectively. Choose an arbitrary $\mathfrak{l}_0 \in [L]$ and, without loss of generality, assume that $c(\mathfrak{l}_0) = 1$, i.e., $\mathbf{x}^{(\mathfrak{l}_0)} \in \mathcal{S}_1$. Denote $\mathbf{x} = \mathbf{x}^{(\mathfrak{l}_0)}$, $\mathbf{y} = \mathbf{y}^{(\mathfrak{l}_0)}$, $\widetilde{\mathcal{S}} = \mathcal{S}_1$ and $\widetilde{\widetilde{\mathcal{S}}} = \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_M$, and present the remainder of matrix X (i.e., X with $X(:, \mathfrak{l}_0)$ removed) as $[\widetilde{X} \mid \widetilde{\widetilde{X}}]$. Here, $\widetilde{X}$ and $\widetilde{\widetilde{X}}$ are portions of X with $X(:, \mathfrak{l}_0)$ removed, that correspond to $\widetilde{\mathcal{S}}$ and $\widetilde{\widetilde{\mathcal{S}}}$, respectively. With some abuse of notations, we denote X with $X(:, \mathfrak{l}_0)$ removed by X again, i.e., $X = [\widetilde{X} \mid \widetilde{\widetilde{X}}]$.

Denote $Z = Y - X$, $\mathbf{z}^{(\ell)} = Z(:, \ell)$ and $\mathbf{z} = \mathbf{z}^{(\mathfrak{l}_0)} = \mathbf{y} - \mathbf{x}$, so that

$$\widetilde{Y} = \widetilde{X} + \widetilde{Z}, \quad \widetilde{\widetilde{Y}} = \widetilde{\widetilde{X}} + \widetilde{\widetilde{Z}}, \quad \mathbf{y} = \mathbf{x} + \mathbf{z}.$$

Let $\mathbf{w} = [\widetilde{\mathbf{w}} \mid \widetilde{\widetilde{\mathbf{w}}}]$ be the solution of problem (10) for $\ell = \mathfrak{l}_0$. Then, (10) implies that

$$\|\mathbf{y} - \widetilde{Y}\widetilde{\mathbf{w}} - \widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\|^2 + 2\lambda\|\widetilde{\mathbf{w}}\|_1 + 2\lambda\|\widetilde{\widetilde{\mathbf{w}}}\|_1 \le \|\mathbf{y} - \widetilde{Y}\widetilde{\mathbf{w}}\|^2 + 2\lambda\|\widetilde{\mathbf{w}}\|_1.$$

By simplifying the inequality, obtain

$$\Delta \stackrel{def}{=} \|\widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\|^2 - 2\langle \mathbf{y} - \widetilde{Y}\widetilde{\mathbf{w}}, \widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\rangle + \lambda\|\widetilde{\widetilde{\mathbf{w}}}\|_1 \le 0. \tag{A.2}$$

Note that the Cauchy–Schwarz inequality and Assumption A5 yield

$$\langle \mathbf{y} - \widetilde{Y}\widetilde{\mathbf{w}}, \widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\rangle \le \tau\|\mathbf{x} - \widetilde{X}\widetilde{\mathbf{w}}\|\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\| + \|\mathbf{z} - \widetilde{Z}\widetilde{\mathbf{w}}\|\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\| + \|\mathbf{x} - \widetilde{X}\widetilde{\mathbf{w}}\|\|\widetilde{\widetilde{Z}}\widetilde{\widetilde{\mathbf{w}}}\| + \|\mathbf{z} - \widetilde{Z}\widetilde{\mathbf{w}}\|\|\widetilde{\widetilde{Z}}\widetilde{\widetilde{\mathbf{w}}}\|.$$

Moreover,

$$\|\mathbf{z} - \widetilde{Z}\widetilde{\mathbf{w}}\| \le [\|\widetilde{\mathbf{w}}\|_1 + 1]\delta; \quad \|\widetilde{\widetilde{Z}}\widetilde{\widetilde{\mathbf{w}}}\| \le \|\widetilde{\widetilde{\mathbf{w}}}\|_1\delta; \quad \|\mathbf{x} - \widetilde{X}\widetilde{\mathbf{w}}\| \le \|\widetilde{\mathbf{w}}\|_1 + 1.$$

Since $\|\widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\|^2 \ge 0.5\,\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\|^2 - \|\widetilde{\widetilde{Z}}\widetilde{\widetilde{\mathbf{w}}}\|^2$, obtain

$$\Delta \ge \left\{0.5\,\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\|^2 - 2(\tau+\delta)(\|\widetilde{\mathbf{w}}\|_1 + 1)\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\|\right\} + \|\widetilde{\widetilde{\mathbf{w}}}\|_1\left\{\lambda - \delta - 2(\|\widetilde{\mathbf{w}}\|_1 + 1)\delta(1+\delta)\right\}. \tag{A.3}$$

To find an upper bound for $(\|\widetilde{\mathbf{w}}\|_1 + 1)$, consider $\widetilde{\mathbf{w}}_*$, the solution of exact problem, that is $\mathbf{x} = \widetilde{X}\widetilde{\mathbf{w}}_*$. By Assumption A6, there exists a sub-matrix $\widetilde{X}_* \in \mathbb{R}^{n^2 \times (K-1)^2}$ of $\widetilde{X}$, such that $\mathbf{x} = \widetilde{X}_*\mathbf{w}_*$ and $\|\mathbf{w}_*\|_1 \le \aleph_{w,K}$. Let $\widetilde{Y}_*$ be the portion of $\widetilde{Y}$ corresponding to $\widetilde{X}_*$ and $\widetilde{Z}_* = \widetilde{Y}_* - \widetilde{X}_*$. Since $\|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2 = \|\mathbf{z} - \widetilde{Z}\mathbf{w}_*\|^2$, derive

$$\|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2 + 2\lambda\|\mathbf{w}_*\|_1 \le \delta^2\left[\|\mathbf{w}_*\|_1 + 1\right]^2 + 2\lambda\|\mathbf{w}_*\|_1. \tag{A.4}$$

Note that, since $\mathbf{w}_*$ is not an optimal solution, one has

$$\|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2 + 2\lambda\|\mathbf{w}_*\|_1 \ge \|\mathbf{y} - \widetilde{Y}\widetilde{\mathbf{w}} - \widetilde{\widetilde{Y}}\widetilde{\widetilde{\mathbf{w}}}\|^2 + 2\lambda\|\widetilde{\mathbf{w}}\|_1 + 2\lambda\|\widetilde{\widetilde{\mathbf{w}}}\|_1 \ge 2\lambda\|\widetilde{\mathbf{w}}\|_1.$$

Thus, $\|\widetilde{\mathbf{w}}\|_1 + 1 \le (\|\mathbf{w}_*\|_1 + 1) + \|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2/(2\lambda)$, so that

$$\|\widetilde{\mathbf{w}}\|_1 + 1 \le (1 + \aleph_{w,K}) + 0.5\,\delta^2\,(1 + \aleph_{w,K})^2/\lambda. \tag{A.5}$$

Then, using (A.3) and (A.5), due to $\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\| \le \|\widetilde{\widetilde{\mathbf{w}}}\|_1$, obtain

$$\Delta \ge \frac{1}{2}\,\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\|^2 + \|\widetilde{\widetilde{\mathbf{w}}}\|_1\left\{\lambda - \delta - 2\,(1 + \aleph_{w,K})\left(1 + \frac{1}{2}\delta^2(1 + \aleph_{w,K})/\lambda\right)(2\delta + 2\tau + \delta^2)\right\}.$$

Now, observe that, due to condition (A.1), $\delta < 1$ and $\delta^2(1 + \aleph_{w,K})/\lambda$ tends to zero. Hence, for n large enough, arrive at

$$\Delta \ge \frac{1}{2}\|\widetilde{\widetilde{X}}\widetilde{\widetilde{\mathbf{w}}}\|^2 + \lambda\|\widetilde{\widetilde{\mathbf{w}}}\|_1\left\{1 - \frac{\delta}{\lambda} - \frac{12\,(\delta+\tau)(1+\aleph_{w,K})}{\lambda}\right\} > 0$$

unless $\widetilde{\widetilde{\mathbf{w}}} = 0$. Since, by (A.2), $\Delta \le 0$, one has $\widetilde{\widetilde{\mathbf{w}}} = 0$ and the SEP holds.

In order to complete the proof, we need to show that there exists λ which is not too large, so the optimization problem (10) for $\ell = \mathfrak{l}_0$ has a non-zero solution. If we show that, for some $\mathbf{w} \ne 0$, the objective function is smaller than that for $\mathbf{w} \equiv 0$, then (10) for $\ell = \mathfrak{l}_0$ yields a non-zero solution. To this end, we find a sufficient condition such that $\|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2 + 2\lambda\|\mathbf{w}_*\|_1 \le \|\mathbf{y}\|^2 = 1$ holds. It follows from (A.4) and Assumption A6 that

$$\|\mathbf{y} - \widetilde{Y}\mathbf{w}_*\|^2 + 2\lambda\|\mathbf{w}_*\|_1 \le \delta^2(1 + \aleph_{w,K})^2 + 2\lambda\aleph_{w,K}.$$

Hence,

$$\delta^2\,(1 + \aleph_{w,K})^2 + 2\lambda\aleph_{w,K} \le 1$$

is sufficient for $\mathbf{w} \ne 0$. By condition (A.1), one has $\delta\,(1+\aleph_{w,K}) \to 0$ as $n \to \infty$, so that for n large enough, $\delta\,(1+\aleph_{w,K}) \le 1/2$. Therefore, $2\lambda\aleph_{w,K} \le 1/2$ is sufficient for $\mathbf{w} \ne 0$, which is equivalent to the first inequality in (A.1). The latter completes the proof. □

Proof of Theorem 1. In order to prove that Theorem 1 holds, we show that, under assumptions of Theorem 1, (17) is true and that $\tau_{n,K} \le \mathbb{C}K^2\,n^{-1}\,\ln n$ in Proposition 1. Let $\widehat{L}_m$ and $\hat{n}_k^{(m)}$ be defined in (16). Then, the following statements are valid.

Lemma 2. *Let Assumption A4 hold. Let $t > 0$ satisfy condition (20). Then, there exists a set $\bar{\Omega}_{t1}$ with*

$$\Pr(\Omega_{t1}) \geq 1 - 2L^{-t} - 2KM\, n^{-t}$$

such that, for $\omega \in \Omega_{t1}$, one has simultaneously

$$\frac{\underline{c}_\varpi\, L}{2\,M} \leq \widehat{L}_m \leq \frac{3\,\bar{c}_\varpi\, L}{2\,M}, \quad \text{and} \quad \bigcap_{m=1}^{M}\bigcap_{k=1}^{K}\left(\omega : \frac{\underline{c}_\pi\, n}{2\,K} \leq \hat{n}_k{}^{(m)} \leq \frac{3\,\bar{c}_\pi\, n}{2\,K}\right). \tag{A.6}$$

Apply the following lemma, proved later in Appendix A, which ensures the upper bound $\max_\ell \|\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}\| \leq \delta_{n,K,t}$ in Proposition 1.

Lemma 3. *Let Assumptions of Theorem 1 hold and $t > 0$ satisfies condition (20). Let $\mathbf{q}^{(\ell)}$ and $\hat{\mathbf{q}}^{(\ell)}$ be defined in (7) and (9), respectively. Let matrices $Q, \widehat{Q} \in \mathbb{R}^{n^2 \times L}$ be defined in (13) and (9), respectively. Then,*

$$\min_\ell \|\mathbf{q}^{(\ell)}\| \geq \widetilde{C}_0 C_{\sigma,0}\, \underline{C}\, K^{-1/2}\, n\, \rho_n, \quad \max_\ell \|\mathbf{q}^{(\ell)}\| \leq \widetilde{\widetilde{C}}_0\, \bar{C}\, K^{-1/2}\, n\, \rho_n. \tag{A.7}$$

Moreover, there exists a set Ω_{t2} such that $\Pr(\Omega_{t2}) \geq 1 - Ln^{-t}$, and for $\omega \in \Omega_{t2}$, one has

$$\max_\ell \|\hat{\mathbf{q}}^{(\ell)} - \mathbf{q}^{(\ell)}\| / \|\mathbf{q}^{(\ell)}\| \leq C_{t,\rho,\sigma}\, K / \sqrt{n\rho_n}, \tag{A.8}$$

where $C_{t,\rho,\sigma}$ depends only on t and constants in Assumptions A1–A5.

In addition, the following lemma provides an upper bound on $\tau_{n,K}$ in Proposition 1.

Lemma 4. *Let Assumption A4 hold, and $Z_{j,k}^{(m)}$, $k \in [K]$, $j \in [n]$, $m \in [M]$, be generated according to (12). Let $t > 0$ satisfy condition (20). Then, there exists a set Ω_{t3} with*

$$\Pr(\Omega_{t3}) \geq 1 - 2L^{-t} - 2KM(M+1)\, n^{-t}$$

such that, for $\omega \in \Omega_{t3}$, and for any arbitrary vectors $\mathbf{x} \in S_m$ and $\mathbf{x}' \in S_{m'}$, where $m \neq m'$, one has $|\mathbf{x}^\top \mathbf{x}'| \leq \tau \|\mathbf{x}\| \|\mathbf{x}'\|$ with

$$\tau \equiv \tau_{n,K} \leq 2\,(\sqrt{2} + 3/\underline{c}_\pi)^2\, t\, K^2\, n^{-1}\, \ln n. \tag{A.9}$$

It is easy to show that, for any $\ell \in [L]$, one has $\|\mathbf{z}^{(\ell)}\| = \|\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}\| \leq 2\|\hat{\mathbf{q}}^{(\ell)} - \mathbf{q}^{(\ell)}\| / \|\mathbf{q}^{(\ell)}\|$. Hence, Lemma 3 implies that, for $\delta \equiv \delta_{n,K,t}$ defined in (18), one has

$$\Pr\left(\max_{1 \leq \ell \leq L} \|\mathbf{z}^{(\ell)}\| \leq \delta\right) \geq 1 - Ln^{-t}. \tag{A.10}$$

Now, in order to apply Proposition 1, it remains to show that condition (19) implies (A.1). For this purpose, note that, since columns of matrix X have unit norms, one has $\aleph_{w,K} \geq 1$ in A5 and, hence, (A.1) implies that $\delta_{n,K,t}/\lambda_{n,K} \to 0$ as $n \to \infty$. The latter furthermore yields that $n^{-1}\, K^2\, \ln n \to 0$, so that $\tau_{n,K} = o(\delta_{n,K,t})$ as $n \to \infty$, where $\tau_{n,K}$ and $\delta_{n,K,t}$ are defined in (A.9) and (18), respectively. This completes the proof. □

Proof of Theorem 2. Let $\widehat{W}$ be the matrix of weights and Ω_t be the set in Theorem 1, so that Ω_t is exactly the set where SEP holds. Note that Algorithm 2 allows the situation where $\widetilde{M} < M$. However, if the SEP holds, then no two network layers in different clusters can be a part of the same connected component, and hence, $\widetilde{M} \geq M$.

Consider a clustering function $\phi : [L] \to [\widetilde{M}]$ and the corresponding clustering matrix $\Phi \in \{0,1\}^{L \times \widetilde{M}}$, which partitions L layers into $\widetilde{M} \geq M$, disconnected components. Due to SEP, some of the vectors that belong to different clusters, according to ϕ, belong to the same cluster, according to c. On the other hand, if two vectors belong to different clusters according to c, they belong to different clusters according to ϕ. That is, for $\mathfrak{l}_1, \mathfrak{l}_2 \in [L]$, $\mathfrak{l}_1 \neq \mathfrak{l}_2$, one has

$$\phi(\mathfrak{l}_1) = \phi(\mathfrak{l}_2) \implies c(\mathfrak{l}_1) = c(\mathfrak{l}_2), \quad c(\mathfrak{l}_1) \neq c(\mathfrak{l}_2) \implies \phi(\mathfrak{l}_1) \neq \phi(\mathfrak{l}_2). \tag{A.11}$$

Hence, if $\widetilde{M} = M$, then $\phi = c$.

Let $\widetilde{M} > M$. Then, due to (A.11), one can partition $\widetilde{M}$ clusters into M groups. Let $\theta : [\widetilde{M}] \to [M]$ be such clustering function, and Θ be the corresponding clustering matrix. Then, for $\omega \in \Omega_t$, SEP holds and $C = \Phi\Theta$. Observe that $\theta(\widetilde{m}_1) = \theta(\widetilde{m}_2)$ if $c(\mathfrak{l}_i) = c(\mathfrak{l}_j)$ for all $\mathfrak{l}_i, \mathfrak{l}_j$ with $\phi(\mathfrak{l}_i) = \widetilde{m}_1$ and $\phi(\mathfrak{l}_j) = \widetilde{m}_2$, where $\mathfrak{l}_i, \mathfrak{l}_j \in [L]$, and $\widetilde{m}_1, \widetilde{m}_2 \in [\widetilde{M}]$. To prove the theorem, we use the following statement.

Lemma 5. *Let Assumptions A1–A5 hold and $K^2/(n\rho_n) \to 0$ as $n \to \infty$. Then, if $\omega \in \Omega_t$, for some positive constant $\check{C}$, one has*

$$|(\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}| \geq \check{C} \;\; \text{if } c(\mathfrak{l}_1) = c(\mathfrak{l}_2); \quad |(\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}| \leq \tau_{n,K} \;\; \text{if } c(\mathfrak{l}_1) \neq c(\mathfrak{l}_2). \tag{A.12}$$

Moreover, for $\omega \in \Omega_t$ and n large enough

$$\min_{\substack{\mathfrak{l}_1, \mathfrak{l}_2 \\ c(\mathfrak{l}_1) = c(\mathfrak{l}_2)}} |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}| \geq \check{C}/2, \qquad \max_{\substack{\mathfrak{l}_1, \mathfrak{l}_2 \\ c(\mathfrak{l}_1) \neq c(\mathfrak{l}_2)}} |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}| \leq \tau_{n,K} + 2\delta_{n,K,t}. \tag{A.13}$$

Consider matrices $Y, \widehat{Y} \in \mathbb{R}^{L\times L}$ with elements

$$Y_{\mathfrak{l}_1,\mathfrak{l}_2} = |(\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}|, \quad \widehat{Y}_{\mathfrak{l}_1,\mathfrak{l}_2} = |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}|, \quad \mathfrak{l}_1, \mathfrak{l}_2 \in [L].$$

Denote $D_{\boldsymbol{\Phi}} = (\boldsymbol{\Phi})^\top \boldsymbol{\Phi}$ and define matrices $\widetilde{Y}, \widehat{\widetilde{Y}} \in \mathbb{R}^{\widetilde{M}\times\widetilde{M}}$

$$\widetilde{Y} = (D_{\boldsymbol{\Phi}})^{-1/2} \boldsymbol{\Phi}^\top Y \boldsymbol{\Phi} (D_{\boldsymbol{\Phi}})^{-1/2}, \quad \widehat{\widetilde{Y}} = (D_{\boldsymbol{\Phi}})^{-1/2} \boldsymbol{\Phi}^\top \widehat{Y} \boldsymbol{\Phi} (D_{\boldsymbol{\Phi}})^{-1/2}.$$

Then, due to (21), by Lemma 5, for $\widetilde{m}_1, \widetilde{m}_2 \in [\widetilde{M}]$, $\widetilde{Y}_{\widetilde{m}_1,\widetilde{m}_2} \geq \check{C}$ if $\theta(\widetilde{m}_1) = \theta(\widetilde{m}_2)$, and $\widetilde{Y}_{\widetilde{m}_1,\widetilde{m}_2} \leq \tau_{n,K}$ if $\theta(\widetilde{m}_1) \neq \theta(\widetilde{m}_2)$. Also, for $\omega \in \Omega_t$, one has $\widehat{\widetilde{Y}}_{\widetilde{m}_1,\widetilde{m}_2} \geq \check{C}/2$ if $\theta(\widetilde{m}_1) = \theta(\widetilde{m}_2)$, and $\widehat{\widetilde{Y}}_{\widetilde{m}_1,\widetilde{m}_2} \leq \tau_{n,K} + 2\delta_{n,K,t}$ if $\theta(\widetilde{m}_1) \neq \theta(\widetilde{m}_2)$.

Now, consider matrices $G, \widehat{G} \in \{0,1\}^{\widetilde{M}\times\widetilde{M}}$ with

$$G_{\widetilde{m}_1,\widetilde{m}_2} = I\left\{\theta(\widetilde{m}_1) = \theta(\widetilde{m}_2)\right\}, \quad \widehat{G}_{\widetilde{m}_1,\widetilde{m}_2} = I\left(|\widehat{\widetilde{Y}}_{\widetilde{m}_1,\widetilde{m}_2}| \geq T\right), \quad \widetilde{m}_1, \widetilde{m}_2 \in [\widetilde{M}].$$

Then, $G = \Theta\Theta^\top$. Moreover, if n is large enough, then $\tau_{n,K} + 2\delta_{n,K,t} < T < \check{C}/2$, whenever T satisfies conditions (21). Consequently, $\widehat{G} = G$ for $\omega \in \Omega_t$ and hence, spectral clustering of $\widehat{G}$ correctly recovers M clusters given by Θ. □

Proof of Theorem 3. The proof of this theorem is very similar to the proof of Theorem 3 in [41]. In this proof, same as before, we denote by $\mathbb{C}$ an absolute constant which can be different at different instances. Consider tensors $\mathbf{G} \in \mathbb{R}^{n\times n\times L}$ and $\mathbf{H} = \mathbf{G} \times_3 (C D_c^{-1/2})^\top \in \mathbb{R}^{n\times n\times M}$ with layers, respectively, $G^{(\ell)} = \mathbf{G}(:,:,\ell)$ and $H^{(m)} = \mathbf{H}(:,:,m)$ of the forms

$$G^{(\ell)} = (P^{(\ell)})^2, \quad H^{(m)} = L_m^{-1/2} \sum_{c(\ell)=m} G^{(\ell)}, \quad \ell \in [L], \ m \in [M].$$

In order to assess R_{WL}, one needs to examine the spectral structure of matrices $H^{(m)}$ and their deviation from the sample-based versions $\widehat{H}^{(m)} = \widehat{\mathbf{H}}(:,:,m)$. We start with the first task.

It follows from (S1) and (S2) that

$$H^{(m)} = U_z{}^{(m)} \bar{Q}_D{}^{(m)} (U_z{}^{(m)})^\top \quad \text{with} \quad \bar{Q}_D{}^{(m)} = L_m^{-1/2} \sum_{c(\ell)=m} \left(B_D^{(\ell)}\right)^2.$$

Since all eigenvalues of $(B_D^{(\ell)})^2$ are positive, applying the Theorem in Complement 10.1.2 on page 327 of [42] and Assumptions A1–A5, obtain that

$$\begin{aligned}\sigma_{\min}(H^{(m)}) = \sigma_K\left(\bar{Q}_D{}^{(m)}\right) &\geq L_m^{-1/2} \sum_{c(\ell)=m} \sigma_K\left\{(B_D^{(\ell)})^2\right\} \geq L_m^{-1/2} \left\{\min_k(\hat{n}_k{}^{(m)})\right\}^2 \rho_n^2 \sum_{c(\ell)=m} \sigma_K\left\{(B_0^{(\ell)})^2\right\} \\ &\geq \mathbb{C} n^2 \rho_n^2 K^{-2} \sqrt{LM^{-1}}. \end{aligned} \tag{A.14}$$

Note that the Euclidean separation γ_m of rows of $U_H{}^{(m)}$ is the same as the Euclidean separation of rows of $U_z{}^{(m)}$, and $\gamma_m^2 \geq 2\{\min_k(\hat{n}_k{}^{(m)})\}^{-1} \geq \mathbb{C}K/n$ for $\omega \in \Omega_t$.

Therefore, by Lemma 9 of [25], derive that the total number of clustering errors Δ within all layers is bounded as

$$\Delta \leq \mathbb{C} \frac{n}{K} \sum_{m=1}^{M} \left\|\sin\Theta\left(\widehat{U}_{\widehat{H}}{}^{(m)}, U_H{}^{(m)}\right)\right\|_F^2.$$

Using Davis–Kahan theorem and formula (A.14), obtain

$$\left\|\sin\Theta\left(\widehat{U}_{\widehat{H}}{}^{(m)}, U_H{}^{(m)}\right)\right\|_F^2 \leq \frac{4K\|\widehat{H}^{(m)} - H^{(m)}\|^2}{\sigma_{\min}^2(H^{(m)})} \leq \mathbb{C} \frac{K^5 M \|\widehat{H}^{(m)} - H^{(m)}\|^2}{n^4 \rho_n^4 L},$$

where we use $\mathbb{C}$ for different constants that depend on the constants in Assumptions A1–A5. Combination of the last two inequalities yields that the total number of clustering errors within all layers is bounded by

$$\Delta \leq \mathbb{C} \frac{K^4 M}{n^3 \rho_n^4 L} \sum_{m=1}^{M} \|\widehat{H}^{(m)} - H^{(m)}\|^2. \tag{A.15}$$

Recall that $H^{(m)} = [\mathbf{G} \times_3 \Psi^\top](:,:,m)$ and $\widehat{H}^{(m)} = [\widehat{\mathbf{G}} \times_3 \widehat{\Psi}^\top](:,:,m)$. Since, by Theorem 2, for $\omega \in \Omega_t$ one has $\widehat{\Psi} = \Psi$, obtain that

$$\|\widehat{H}^{(m)} - H^{(m)}\|^2 \leq L_m^{-1} \left\|\bar{\widehat{G}}^{(m)} - \bar{G}^{(m)}\right\|^2,$$

where

$$\bar{G}^{(m)} = \sum_{c(\ell)=m} G^{(\ell)} = \sqrt{L_m}\, H^{(m)}, \quad \bar{\widehat{G}}^{(m)} = \sqrt{L_m}\, [\widehat{\mathbf{G}} \times_3 \Psi^\top](:,:,m) = \sum_{c(\ell)=m} \widehat{G}^{(\ell)}$$

use the following lemma that modifies upper bounds in [25] in the absence of the sparsity assumption $\rho_n n \leq \mathbb{C}$:

Lemma 6. *Let Assumptions A1–A5 hold, $G^{(\ell)} = (P^{(\ell)})^2$ and $\widehat{G}^{(\ell)} = (A^{(\ell)})^2 - diag(A^{(\ell)}\mathbf{1})$, where $c(\ell) = m$, $\ell \in [\tilde{L}]$. Let*

$$G = \sum_{\ell=1}^{\tilde{L}} G^{(\ell)}, \quad \widehat{G} = \sum_{\ell=1}^{\tilde{L}} \widehat{G}^{(\ell)}.$$

Then, for any $t > 0$, there exists a constant $\widetilde{C}$ that depends only on t and constants in Assumptions A1–A5, and $\widetilde{C}_{t,\epsilon}$ which depends only on t and ϵ in Algorithm 3, such that one has

$$\Pr\left[\|\widehat{G} - G\|^2 \le \widetilde{C}\left\{\rho_n^3 n^3 \tilde{L}\ln(\tilde{L}+n) + \rho_n^4 n^2 \tilde{L}^2\right\}\right] \ge 1 - \widetilde{C}_{t,\epsilon}(\tilde{L}+n)^{1-t}.$$

Applying Lemma 6 with $\tilde{L} = L_m$, obtain that there exists a set Ω_{t4}, with

$$\Pr(\Omega_{t4}) \ge 1 - \widetilde{C}_{t,\epsilon} n^{1-t},$$

and for $\omega \in \Omega_{t4}$, one has

$$\|\widehat{H}^{(m)} - H^{(m)}\|^2 \le \mathbb{C}\left\{\rho_n^3 n^3 \ln(L+n) + \rho_n^4 n^2 L/M\right\}. \tag{A.16}$$

To complete the proof, combine formulas (A.15) and (A.16), set $\tilde{\Omega}_t = \Omega_t \cap \Omega_{t4}$ and recall that $R_{WL} = \Delta/(Mn)$ and $n\rho_n \ge C_\rho \ln n$. □

Proofs of supplementary statements

Proof of Lemma 1. First, we prove part (a). Recall that, for ℓ with $c(\ell) = m$, by formula (7), one has

$$\mathbf{q}^{(\ell)} = \rho_n \left(\widetilde{U}^{(m)} \otimes \widetilde{U}^{(m)}\right)\left(\sqrt{D^{(m)}} \otimes \sqrt{D^{(m)}}\right)\mathbf{b}_0^{(\ell)}. \tag{A.17}$$

Since $\mathcal{B}_0$ is a full rank matrix, one can present $\mathbf{b}_0^{(\mathfrak{l}_0)}$ as $\mathbf{b}_0^{(\mathfrak{l}_0)} = \mathcal{B}_0\mathbf{w}$ for some vector $\mathbf{w}$. Note that, although vectors $\mathbf{b}_0^{(\ell)} \in \mathbb{R}^{K^2}$, due to symmetry of matrices $B_0^{(\ell)}$, the ambient dimension of those vectors is $K(K+1)/2 = |\mathcal{L}_0|$. Then, by Assumption A1, obtain

$$\|\mathbf{w}\|_1 \le K\|\mathbf{w}\|_2 \le K\left\{\sigma_{min}(\mathcal{B}_0)\right\}^{-1}\|\mathbf{b}_0^{(\mathfrak{l}_0)}\| \le K(\sigma_{0,K})^{-1}\|B_0^{(\mathfrak{l}_0)}\|_F \le \bar{C}(\sigma_{0,K})^{-1}K\sqrt{K}.$$

Now, (A.17) and $\mathbf{b}_0^{(\mathfrak{l}_0)} = \mathcal{B}_0\mathbf{w}$ imply that

$$\mathbf{q}^{(\mathfrak{l}_0)} = \sum_{\ell\in\mathcal{L}_0} \mathbf{q}^{(\ell)}\mathbf{w}_\ell.$$

Therefore,

$$\mathbf{x}^{(\mathfrak{l}_0)} = \sum_{\ell\in\mathcal{L}_0} \mathbf{x}^{(\ell)}(\mathbf{w}_*)_\ell,$$

where $|(\mathbf{w}_*)_\ell| = |\mathbf{w}_\ell|\|\mathbf{q}^{(\ell)}\|/\|\mathbf{q}^{(\mathfrak{l}_0)}\|$. By Lemma 3, one has $\|\mathbf{q}^{(\ell)}\|/\|\mathbf{q}^{(\mathfrak{l}_0)}\| \le (\widetilde{C}_0 C_{\sigma,0}\underline{C})^{-1}\widetilde{\widetilde{C}}_0\bar{C}$, and, hence,

$$\|\mathbf{w}_*\|_1 \le \frac{(\bar{C})^2\widetilde{\widetilde{C}}_0}{\underline{C}\widetilde{C}_0 C_{\sigma,0}}\frac{K\sqrt{K}}{\sigma_{0,K}},$$

which proves part (a).

Validity of part (b) follows from the fact that there are at least two copies of any vector $\mathbf{x}^{(\ell)}$ for any ℓ and any group of layers. □

Proof of Lemma 2. For a fixed k, note that $\hat{n}_k{}^{(m)} \sim$ Binomial(π_k, n). By Hoeffding inequality, for any $x > 0$

$$\Pr\left(\left|\hat{n}_k{}^{(m)}/n - \pi_k\right| \ge x\right) \le 2\exp\{-2nx^2\}.$$

Then, using (15), obtain

$$\Pr\left(\underline{c}_\pi\, n/K - nx \le \hat{n}_k{}^{(m)} \le \bar{c}_\pi\, n/K + nx\right) \ge 1 - 2\exp\{-2nx^2\}.$$

Now, set $x = \sqrt{t\ln n/(2n)}$ and let n be large enough, so that $K\sqrt{t\ln n/(2n)} < 1/2$, which is equivalent to $t < n/(2K^2\ln n)$. Then, combination of the union bound over k and m and

$$\Pr\left\{\frac{\underline{c}_\pi\, n}{K}\left(1 - \frac{K\sqrt{t\ln n}}{\underline{c}_\pi\sqrt{2n}}\right) \le \hat{n}_k{}^{(m)} \le \frac{\bar{c}_\pi\, n}{K}\left(1 + \frac{K\sqrt{t\ln n}}{\bar{c}_\pi\sqrt{2n}}\right)\right\} \ge 1 - 2n^{-t}$$

implies the second inequality in (A.6). The first inequality in (A.6) can be proved in a similar manner. □

Proof of Lemma 3. Denote $D = \mathrm{diag}(n_1, \dots, n_K)$, $\widehat{D}^{(m)} = \left(Z^{(m)}\right)^\top\left(Z^{(m)}\right) = \mathrm{diag}(\hat{n}_1^{(m)}, \dots, \hat{n}_K^{(m)})$, where $n_k = n\pi_k$ and $\hat{n}_k^{(m)}$ are defined in (16). Consider matrices

$$U^{(m)} = Z^{(m)}\left(\widehat{D}^{(m)}\right)^{-1/2} \in \mathcal{O}_{n,K}, \quad \widetilde{U}^{(m)} = (I - \mathscr{P})U^{(m)}, \quad m \in [M],$$

where $\mathscr{P}$ is defined in (5), and note that $\mathcal{S}_m = \text{span}\left(\widetilde{U}^{(m)} \otimes \widetilde{U}^{(m)}\right)$. For $m \in [M]$, denote

$$\mathbf{t} = n^{-1/2}\left(\sqrt{n_1}, \ldots, \sqrt{n_K}\right)^\top, \quad \hat{\mathbf{t}}^{(m)} = n^{-1/2}\left(\sqrt{\hat{n}_1^{(m)}}, \ldots, \sqrt{\hat{n}_K^{(m)}}\right)^\top, \quad \Pi_{\hat{\mathbf{t}}^{(m)}} = \hat{\mathbf{t}}^{(m)}(\hat{\mathbf{t}}^{(m)})^\top, \tag{A.18}$$

where $\Pi_{\hat{\mathbf{t}}^{(m)}}$ are the projection matrices and $\Pi^{\perp}_{\hat{\mathbf{t}}^{(m)}} = I_K - \Pi_{\hat{\mathbf{t}}^{(m)}}$. Then, for $m \in [M]$, due to $\mathbf{1}_n = Z^{(m)}\mathbf{1}_K$, one has

$$\widetilde{U}^{(m)} = U^{(m)}\left\{I_K - \left(\widehat{D}^{(m)}\right)^{1/2}\frac{\mathbf{1}_K\mathbf{1}_n^\top}{n}Z^{(m)}\left(\widehat{D}^{(m)}\right)^{-1/2}\right\}.$$

Now, since $\left(\widehat{D}^{(m)}\right)^{1/2}\mathbf{1}_K = \sqrt{n}\,\hat{\mathbf{t}}^{(m)}$ and $\mathbf{1}_n^\top Z^{(m)}\left(\widehat{D}^{(m)}\right)^{-1/2} = \sqrt{n}\,(\hat{\mathbf{t}}^{(m)})^\top$, one obtains

$$\widetilde{U}^{(m)} = (I - \mathscr{P})U^{(m)} = U^{(m)}\left\{I_K - \hat{\mathbf{t}}^{(m)}(\hat{\mathbf{t}}^{(m)})^\top\right\} = U^{(m)}\Pi^{\perp}_{\hat{\mathbf{t}}^{(m)}}. \tag{A.19}$$

Note that, $\Pi^{\perp}_{\hat{\mathbf{t}}^{(m)}} = \widehat{V}^{(m)}(\widehat{V}^{(m)})^\top$, for some matrix $\widehat{V}^{(m)} \in \mathcal{O}_{K,K-1}$. Denote

$$\widetilde{W}^{(m)} = U^{(m)}\widehat{V}^{(m)} \in \mathcal{O}_{n,K-1}, \quad m \in [M]. \tag{A.20}$$

Hence,

$$\widetilde{U}^{(m)} = \widetilde{W}^{(m)}(\widehat{V}^{(m)})^\top \quad \text{and} \quad \mathcal{S}_m = \text{span}\left\{\left(\widetilde{W}^{(m)} \otimes \widetilde{W}^{(m)}\right)\left(\widehat{V}^{(m)} \otimes \widehat{V}^{(m)}\right)^\top\right\}.$$

Consider $\mathbf{q}^{(\mathfrak{l}_i)}$ with $c(\mathfrak{l}_i) = m_i$, $i = 1, 2$. Due to (3), (6)–(7) and (A.19), obtain

$$\mathbf{q}^{(\mathfrak{l}_i)} = \left(U^{(m_i)} \otimes U^{(m_i)}\right)\left(\Pi^{\perp}_{\mathbf{t}^{(m_i)}} \otimes \Pi^{\perp}_{\mathbf{t}^{(m_i)}}\right)\widetilde{\mathbf{b}}^{(\mathfrak{l}_i)}, \quad i \in \{1, 2\}.$$

If $m_1 = m_2 = m$, then, due to $(U^{(m)})^\top U^{(m)} = I_K$ and using Theorem 1.2.22 in [16], obtain

$$\begin{aligned}(\mathbf{q}^{(\mathfrak{l}_1)})^\top\mathbf{q}^{(\mathfrak{l}_2)} &= (\widetilde{\mathbf{b}}^{(\mathfrak{l}_1)})^\top\left(\Pi^{\perp}_{\mathbf{t}^{(m)}} \otimes \Pi^{\perp}_{\mathbf{t}^{(m)}}\right)\widetilde{\mathbf{b}}^{(\mathfrak{l}_2)} = \left\{\text{vec}(\widetilde{B}^{(\mathfrak{l}_1)})\right\}^\top \text{vec}\left(\Pi^{\perp}_{\mathbf{t}^{(m)}}\widetilde{B}^{(\mathfrak{l}_2)}\Pi^{\perp}_{\mathbf{t}^{(m)}}\right) = \text{Tr}\left(\widetilde{B}^{(\mathfrak{l}_1)}\Pi^{\perp}_{\mathbf{t}^{(m)}}\widetilde{B}^{(\mathfrak{l}_2)}\Pi^{\perp}_{\mathbf{t}^{(m)}}\right)\\ &= \left\{\text{vec}(\Pi^{\perp}_{\mathbf{t}^{(m)}})\right\}^\top\left(\widetilde{B}^{(\mathfrak{l}_1)} \otimes \widetilde{B}^{(\mathfrak{l}_2)}\right)\text{vec}\left(\Pi^{\perp}_{\mathbf{t}^{(m)}}\right),\end{aligned}$$

so that

$$|(\mathbf{q}^{(\mathfrak{l}_1)})^\top\mathbf{q}^{(\mathfrak{l}_2)}| \geq \sigma_{\min}(\widetilde{B}^{(\mathfrak{l}_1)})\sigma_{\min}(\widetilde{B}^{(\mathfrak{l}_2)})\left\|\text{vec}(\Pi^{\perp}_{\mathbf{t}^{(m)}})\right\|^2.$$

Since $\widetilde{B}^{(\mathfrak{l}_i)} = \sqrt{D^{(m)}}B^{(\mathfrak{l}_i)}\sqrt{D^{(m)}}$, by Assumptions A1–A3, one has

$$\sigma_{\min}(\widetilde{B}^{(\mathfrak{l}_i)}) \geq \sigma_{\min}(D^{(m)})\,\sigma_{\min}(B_0^{(\mathfrak{l}_i)})\,\rho_n \geq \widetilde{C}_0 C_{\sigma,0}\,\sigma_{\max}(B_0^{(\mathfrak{l}_i)})n\,\rho_n/K$$

and $\left\|\text{vec}(\Pi^{\perp}_{\mathbf{t}^{(m)}})\right\|^2 = \left\|\Pi^{\perp}_{\mathbf{t}^{(m)}}\right\|_F^2 = K - 1$. Hence, for $K \geq 2$ and $c(\mathfrak{l}_1) = c(\mathfrak{l}_2)$, one has

$$|(\mathbf{q}^{(\mathfrak{l}_1)})^\top\mathbf{q}^{(\mathfrak{l}_2)}| \geq (\widetilde{C}_0\,C_{\sigma,0})^2\,\sigma_{\max}(B_0^{(\mathfrak{l}_1)})\,\sigma_{\max}(B_0^{(\mathfrak{l}_2)})n^2\rho_n^2/(2K). \tag{A.21}$$

Using (A.21) with $\mathfrak{l}_1 = \mathfrak{l}_2 = \ell$ and taking into account that $\sigma_{\max}(B_0^{(\ell)}) \geq \underline{C}$ by Assumption A1, obtain that, for $K \geq 2$,

$$\|\mathbf{q}^{(\ell)}\| \geq 0.5\,\widetilde{C}_0\,\underline{C}\,C_{\sigma,0}\,K^{-1/2}\,n\rho_n$$

which implies the first inequality in (A.7). On the other hand, if $\mathfrak{l}_1 = \mathfrak{l}_2 = \ell$, then

$$\|\mathbf{q}^{(\ell)}\| \leq \sigma_{\max}(D^{(m)})\,\rho_n\,\sigma_{\max}(B_0^{(\ell)})\,\|\text{vec}(\Pi^{\perp}_{\mathbf{t}^{(m)}})\| \leq \widetilde{\widetilde{C}}_0\,\sigma_{\max}(B_0^{(\ell)})\,n\,\rho_n\,K^{-1/2},$$

which yields the second inequality in (A.7).

In order to prove (A.8), note that, due to $\|\Pi_{(K-1)}(\widehat{A}^{(\ell)}) - \widehat{A}^{(\ell)}\|^2 \leq \|\widetilde{P}^{(\ell)} - \widehat{A}^{(\ell)}\|^2$ and $\|\widetilde{P}^{(\ell)} - \widehat{A}^{(\ell)}\| \leq \|P^{(\ell)} - A^{(\ell)}\|$, one derives

$$\|\widehat{\widetilde{P}}^{(\ell)} - \widetilde{P}^{(\ell)}\|_F^2 \leq 2K\|\Pi_{(K-1)}(\widehat{A}^{(\ell)}) - \widetilde{P}^{(\ell)}\|^2 \leq 2K\left\{2\|\Pi_{(K-1)}(\widehat{A}^{(\ell)}) - \widehat{A}^{(\ell)}\|^2 + 2\|\widehat{A}^{(\ell)} - \widetilde{P}^{(\ell)}\|^2\right\} \leq 8\,K\,\|P^{(\ell)} - A^{(\ell)}\|^2.$$

Using Theorem 5.2 of [26], for any $t > 0$, with probability at least $1 - n^{-t}$, obtain $\|P^{(\ell)} - A^{(\ell)}\| \leq C_{t,\rho}\sqrt{n\rho_n}$, where $C_{t,\rho}$ depends on $C_\rho, \bar{C}$ and t only. Hence, with probability at least $1 - n^{-t}$, one has

$$\|\hat{\mathbf{q}}^{(\ell)} - \mathbf{q}^{(\ell)}\| = \|\widehat{\widetilde{P}}^{(\ell)} - \widetilde{P}^{(\ell)}\|_F \leq 2\sqrt{2}\,C_{t,\rho}\sqrt{K\,n\,\rho_n}\,.$$

Application of the union bound and (A.7) yields that, with probability at least $1 - Ln^{-t}$,

$$\max_\ell \frac{\|\hat{\mathbf{q}}^{(\ell)} - \mathbf{q}^{(\ell)}\|}{\|\mathbf{q}^{(\ell)}\|} \leq \frac{2\sqrt{2}C_{t,\rho}\sqrt{\rho_n K n}\,\sqrt{K}}{\widetilde{C}_0\,\underline{C}\,C_{\sigma,0}\rho_n n},$$

which completes the proof. □

Proof of Lemma 4. Consider $\mathbf{x} \in \mathcal{S}_m$ and $\mathbf{x}' \in \mathcal{S}_{m'}$, where $m \neq m'$. Then $\mathbf{x} = \left(\widetilde{W}^{(m)} \otimes \widetilde{W}^{(m)}\right)\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^{(K-1)^2}$ and $\widetilde{W}^{(m)}$ is defined in (A.20), and

$$\|\mathbf{x}\|^2 = \mathbf{v}^\top \left\{ \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m)} \otimes \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m)} \right\} \mathbf{v} = \|\mathbf{v}\|^2.$$

Similarly, $\mathbf{x}' = \left(\widetilde{W}^{(m')} \otimes \widetilde{W}^{(m')}\right)\mathbf{v}'$, where $\mathbf{v}' \in \mathbb{R}^{(K-1)^2}$ and $\|\mathbf{x}'\| = \|\mathbf{v}'\|$. Then, using the Cauchy–Schwarz inequality, obtain

$$|\mathbf{x}^\top \mathbf{x}'| \le \|\mathbf{x}\| \left\| \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m')} \otimes \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m')} \right\| \|\mathbf{x}'\|.$$

Since $\widetilde{W}^{(m)} = \widetilde{U}^{(m)}\widehat{V}^{(m)}$ and $\widehat{V}^{(m)} \in \mathcal{O}_{K,K-1}$, $m \in [M]$, it is easy to see that

$$\left\| \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m')} \otimes \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m')} \right\| = \left\| \left(\widetilde{W}^{(m)}\right)^\top \widetilde{W}^{(m')} \right\|^2 \le \left\| \left(\widetilde{U}^{(m)}\right)^\top \widetilde{U}^{(m')} \right\|^2.$$

Therefore, if $\mathbf{x} \in \mathcal{S}_m$, $\mathbf{x}' \in \mathcal{S}_{m'}$, and $\|\mathbf{x}\| = \|\mathbf{x}'\| = 1$, $m \neq m'$, then

$$|\mathbf{x}^\top \mathbf{x}'| \le \left\| \left(\widetilde{U}^{(m)}\right)^\top \widetilde{U}^{(m')} \right\|^2. \tag{A.22}$$

In order to derive an upper bound for (A.22) when $m \neq m'$, note that matrix $\widetilde{U}^{(m)}$, defined in (A.19), has elements

$$\widetilde{U}^{(m)}_{j,k} = (\hat{n}^{(m)}_k)^{-1/2} \left\{ I(\xi^{(m)}_j = k) - n^{-1} \sum_{i=1}^n I(\xi^{(m)}_i = k) \right\}, \quad \sum_{j=1}^n \widetilde{U}^{(m)}_{j,k} = 0.$$

Rows of matrix $\widetilde{U}^{(m)}$ are identically distributed but not independent, which makes the analysis difficult. For this reason, we consider proxies $\widetilde{\widetilde{U}}^{(m)}$ for $\widetilde{U}^{(m)}$ with elements

$$\widetilde{\widetilde{U}}^{(m)}_{j,k} = \frac{1}{\sqrt{n_k}} I(\xi^{(m)}_j = k) - \frac{\sqrt{n_k}}{n} \equiv \frac{1}{\sqrt{n\pi_k}} \left\{ I(\xi^{(m)}_j = k) - \pi_k \right\}, \ j \in [n], \ k \in [K]$$

so that $E\widetilde{\widetilde{U}}^{(m)}_{j,k} = 0$. Rows of $\widetilde{\widetilde{U}}^{(m)}$ are i.i.d. and also $\widetilde{\widetilde{U}}^{(m)}$ and $\widetilde{\widetilde{U}}^{(m')}$ are independent when $m \neq m'$. Hence, matrices $\widetilde{\widetilde{U}}^{(m)}$ are i.i.d. with $E\widetilde{\widetilde{U}}^{(m)} = 0$. We shall use the following statements, proved later in Appendix A.

Lemma 7. *Let $\bar{\boldsymbol{\pi}} = (\pi_1, \dots, \pi_K)$ be such that $\pi_k \ge \underline{c}_\pi / K$ for $k \in [K]$. Then, there exists a set $\widetilde{\Omega}_t$ with $\Pr(\widetilde{\Omega}_t) \ge 1 - 2KM^2 n^{-t}$ such that, for any $\omega \in \widetilde{\Omega}_t$,*

$$\max_{\substack{1 \le m_1, m_2 \le M \\ m_1 \neq m_2}} \left\| \left(\widetilde{\widetilde{U}}^{(m_1)}\right)^\top \widetilde{\widetilde{U}}^{(m_2)} \right\| \le \frac{2K\sqrt{t \ln n}}{\sqrt{n}}.$$

In order to obtain an upper bound for (A.22) when $m \neq m'$, use the fact that proxies $\widetilde{\widetilde{U}}^{(m)}$ are close to $\widetilde{U}^{(m)}$. Indeed, the following statement is valid.

Lemma 8. *Let $\bar{\boldsymbol{\pi}} = (\pi_1, \dots, \pi_K)$ be such that $\pi_k \ge \underline{c}_\pi / K$, $k \in [K]$. Then, there exists a set $\widetilde{\widetilde{\Omega}}_t$ with $\Pr(\widetilde{\widetilde{\Omega}}_t) \ge 1 - 2KMn^{-t}$ such that, for any $\omega \in \widetilde{\widetilde{\Omega}}_t$, one has*

$$\Delta \equiv \max_{1 \le m \le M} \left\| \widetilde{\widetilde{U}}^{(m)} - \widetilde{U}^{(m)} \right\| \le \frac{K\sqrt{2t \ln n}}{\underline{c}_\pi \sqrt{n}}.$$

Then, due to

$$\left\| \widetilde{U}^{(m)} \right\| = \left\| (I - \mathcal{P}) U^{(m)} \right\| \le 1, \qquad \left\| \widetilde{\widetilde{U}}^{(m)} \right\| \le \left\| \widetilde{U}^{(m)} \right\| + \left\| \widetilde{\widetilde{U}}^{(m)} - \widetilde{U}^{(m)} \right\|,$$

derive for any m_1, m_2

$$\begin{aligned} \left\| \left(\widetilde{U}^{(m_1)}\right)^\top \widetilde{U}^{(m_2)} \right\| &\le \left\| \left(\widetilde{\widetilde{U}}^{(m_1)}\right)^\top \widetilde{\widetilde{U}}^{(m_2)} \right\| + \left\| \left(\widetilde{\widetilde{U}}^{(m_1)} - \widetilde{U}^{(m_1)}\right)^\top \widetilde{\widetilde{U}}^{(m_2)} \right\| + \left\| \left(\widetilde{U}^{(m_1)}\right)^\top \left(\widetilde{\widetilde{U}}^{(m_2)} - \widetilde{U}^{(m_2)}\right) \right\| \\ &\le \left\| \left(\widetilde{\widetilde{U}}^{(m_1)}\right)^\top \widetilde{\widetilde{U}}^{(m_2)} \right\| + \Delta(1 + \Delta) + \Delta. \end{aligned}$$

Now, let $\breve{\Omega}_t = \widetilde{\Omega}_t \cap \widetilde{\widetilde{\Omega}}_t$. Note that $\Delta < 1$ for n large enough. Then, $\Pr(\breve{\Omega}_t) \ge 1 - 2KM(M+1)n^{-t}$ and, for $\omega \in \breve{\Omega}_t$, one has

$$\max_{m \neq m'} \left\| \left(\widetilde{U}^{(m)}\right)^\top \widetilde{U}^{(m')} \right\| \le \frac{K\sqrt{2t \ln n}}{\sqrt{n}} \left(\sqrt{2} + \frac{3}{\underline{c}_\pi} \right)$$

which completes the proof. □

Proof of Lemma 5. In addition, the last inequality and (A.21) imply that

$$|(\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}| = \frac{|(\mathbf{q}^{(\mathfrak{l}_1)})^\top \mathbf{q}^{(\mathfrak{l}_2)}|}{\|\mathbf{q}^{(\mathfrak{l}_1)}\|\|\mathbf{q}^{(\mathfrak{l}_2)}\|} \geq \frac{(\widetilde{C}_0\, C_{\sigma,0})^2}{2(\widetilde{C}_0)^2},$$

which completes the proof of the first inequality in (A.12). The second inequality in (A.12) is true by A5.

To prove (A.13), note that, for any $\mathfrak{l}_1$ and $\mathfrak{l}_2$, by the Cauchy–Schwarz inequality and (A.10), one has

$$|(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)} - (\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}| \leq |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \left(\mathbf{y}^{(\mathfrak{l}_2)} - \mathbf{x}^{(\mathfrak{l}_2)}\right)| + |\left(\mathbf{y}^{(\mathfrak{l}_1)} - \mathbf{x}^{(\mathfrak{l}_1)}\right)^\top \mathbf{x}^{(\mathfrak{l}_2)}| \leq 2 \max_{1\leq \ell \leq L} \|\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}\| \leq 2\delta_{n,K,t}$$

for $\omega \in \Omega_t$, where Ω_t and $\delta_{n,K,t}$ are defined, respectively, in Theorem 1 and (18). Then, using (A.12), for $c(\mathfrak{l}_1) = c(\mathfrak{l}_2) = m$ and $\omega \in \Omega_t$, obtain

$$\min_{\mathfrak{l}_1,\mathfrak{l}_2} |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}| \geq |(\mathbf{x}^{(\mathfrak{l}_1)})^\top \mathbf{x}^{(\mathfrak{l}_2)}| - 2\delta_{n,K,t} \geq \check{C}/2$$

if n is large enough, due to $\delta_{n,K,t} \to 0$ as $n \to \infty$. If $c(\mathfrak{l}_1) \neq c(\mathfrak{l}_2)$, then, again by (A.12), for $\omega \in \Omega_t$, derive

$$\max_{\mathfrak{l}_1,\mathfrak{l}_2} |(\mathbf{y}^{(\mathfrak{l}_1)})^\top \mathbf{y}^{(\mathfrak{l}_2)}| \leq \tau_{n,K} + 2\delta_{n,K,t}$$

which completes the proof. □

Proof of Lemma 7. Note that $\widetilde{\widetilde{U}}^{(m)}$ are i.i.d. for $m \in [M]$, so, for simplicity, we can consider $m \in \{1,2\}$. Let $S = \left(\widetilde{\widetilde{U}}^{(1)}\right)^\top \widetilde{\widetilde{U}}^{(2)} \in \mathbb{R}^{K\times K}$. Since $\widetilde{\widetilde{U}}^{(1)}$ and $\widetilde{\widetilde{U}}^{(2)}$ are independent and $\mathbb{E}\left(\widetilde{\widetilde{U}}^{(m)}\right) = 0$, obtain $\mathbb{E}S = 0$. Now let $\mathbf{u}_j^{(m)} = \widetilde{\widetilde{U}}^{(m)}(j,:)$ be the jth row of $\widetilde{\widetilde{U}}^{(m)}$, $j \in [n]$. Then,

$$S = \sum_{j=1}^n S^{(j)}, \quad S^{(j)} = \left(\mathbf{u}_j^{(1)}\right)^\top \mathbf{u}_j^{(2)} \in \mathbb{R}^{K\times K}, \quad j \in [n].$$

Note that $S^{(j)}$ are independent, $\mathbb{E}S^{(j)} = 0$, and $\mathrm{rank}(S^{(j)}) = 1$. Hence, $\|S^{(j)}\| = \|S^{(j)}\|_F = \|\mathbf{u}_j^{(1)}\|\|\mathbf{u}_j^{(2)}\|$. Also, note that, due to $\sum_{k=1}^K I(\xi_j^{(m)} = k) = 1$ and $1/\pi_k \leq K/\underline{c}_\pi$, one has

$$\|\mathbf{u}_j^{(m)}\|^2 = \sum_{k=1}^K \left(\widetilde{\widetilde{U}}_{j,k}^{(m)}\right)^2 = \sum_{k=1}^K \frac{1}{n_k}\left\{I(\xi_j^{(m)} = k) - \pi_k\right\}^2 \leq \frac{K}{\underline{c}_\pi n}.$$

Hence, $\|S^{(j)}\| \leq K/(\underline{c}_\pi n)$.

Now, we are going to apply matrix Bernstein inequality to matrix S. Observe that

$$\mathbb{E}(S^\top S) = \mathbb{E}(SS^\top) = \sum_{j=1}^n \mathbb{E}\left\{S^{(j)}(S^{(j)})^\top\right\},$$

where $\mathbb{E}\left\{S^{(j)}(S^{(j)})^\top\right\} = \mathbb{E}\left\{(\mathbf{u}_j^{(1)})^\top \mathbf{u}_j^{(1)}\right\} \mathbb{E}\left\|\mathbf{u}_j^{(2)}\right\|^2$. Therefore,

$$\left\|\mathbb{E}\left\{S^{(j)}(S^{(j)})^\top\right\}\right\| = \mathbb{E}\left\|\mathbf{u}_j^{(2)}\right\|^2 \left\|\mathbb{E}\left\{(\mathbf{u}_j^{(1)})^\top \mathbf{u}_j^{(1)}\right\}\right\|.$$

Since the operator norm is a convex function, by Jensen inequality and due to $\mathrm{rank}\left\{(\mathbf{u}_j^{(1)})^\top \mathbf{u}_j^{(1)}\right\} = 1$, obtain

$$\left\|\mathbb{E}\left\{(\mathbf{u}_j^{(1)})^\top \mathbf{u}_j^{(1)}\right\}\right\| \leq \mathbb{E}\left\|(\mathbf{u}_j^{(1)})^\top \mathbf{u}_j^{(1)}\right\| = \mathbb{E}\left\|\mathbf{u}_j^{(1)}\right\|^2.$$

On the other hand, it is easy to show that, for any m, one has $\mathbb{E}\left\|\mathbf{u}_j^{(m)}\right\|^2 \leq K/n$. Therefore, $\left\|\mathbb{E}\left\{S^{(j)}(S^{(j)})^\top\right\}\right\| \leq n^{-2}K^2$, so that $\|\mathbb{E}(SS^\top)\| \leq n^{-1}K^2$. Now applying Theorem 1.6.2 (matrix Bernstein inequality) in [46], derive that, for any $x > 0$, one has

$$\Pr(\|S\| > x) \leq 2K \exp\left\{-\frac{x^2/2}{n^{-1}K^2 + n^{-1}Kx/(3\underline{c}_\pi)}\right\}. \tag{A.23}$$

For any $t > 0$, setting $x = 2Kn^{-1/2}\sqrt{t\ln n}$ ensures that, for n large enough, the denominator of the exponent in (A.23) is bounded above by $2K^2 n^{-1}$. Then, for any $m_1, m_2 \in [M]$, obtain

$$\Pr\left(\|S\| \geq 2Kn^{-1/2}\sqrt{t\ln n}\right) = \Pr\left\{\left\|\left(\widetilde{\widetilde{U}}^{(m_1)}\right)^\top \widetilde{\widetilde{U}}^{(m_2)}\right\| \geq 2Kn^{-1/2}\sqrt{t\ln n}\right\} \leq 2Kn^{-t}. \tag{A.24}$$

To complete the proof, apply the union bound to (A.24) and let $\widetilde{\Omega}_t$ be the set where this union bound holds. □

Proof of Lemma 8. Since $\widetilde{U}^{(m)}$ and $\widetilde{\widetilde{U}}^{(m)}$ are i.i.d. for every m, for simplicity, we drop the index m. By definition, for $k \in [K]$, one has

$$\widetilde{U}(:,k) = U(:,k) - n^{-1/2}\,\mathbf{1}_n \cdot \hat{\mathbf{t}}_k, \quad \widetilde{\widetilde{U}}(:,k) = U(:,k)\,\sqrt{\hat{n}_k}/\sqrt{n_k} - n^{-1/2}\,\mathbf{1}_n \cdot \mathbf{t}_k.$$

Hence,

$$\widetilde{U} = U - n^{-1/2}\,\mathbf{1}_n\hat{\mathbf{t}}^\top, \quad \widetilde{\widetilde{U}} = U\Lambda - n^{-1/2}\,\mathbf{1}_n\mathbf{t}^\top, \text{ with } \Lambda = \operatorname{diag}\left(\frac{\sqrt{\hat{n}_1}}{\sqrt{n_1}}, \dots, \frac{\sqrt{\hat{n}_K}}{\sqrt{n_K}}\right),$$

where $\hat{\mathbf{t}}$ and $\mathbf{t}$ are defined in (A.18). Then,

$$\left\|\widetilde{\widetilde{U}} - \widetilde{U}\right\| \le \|U(\Lambda - I)\| + n^{-1/2}\,\left\|\mathbf{1}_n(\hat{\mathbf{t}} - \mathbf{t})^\top\right\| \le \|I - \Lambda\| + \left\|\hat{\mathbf{t}} - \mathbf{t}\right\| = \max_{1\le k\le K}\left|1 - \frac{\sqrt{\hat{n}_k}}{\sqrt{n_k}}\right| + \left\{\sum_{k=1}^{K}\frac{(\sqrt{\hat{n}_k} - \sqrt{n_k})^2}{n}\right\}^{1/2}. \tag{A.25}$$

Since, for $a, b > 0$, one has $|\sqrt{a} - \sqrt{b}| \le |a-b|/\sqrt{b}$, and $n_k = n\pi_k \ge \underline{c}_\pi n/K$, one can easily show that

$$\left|1 - \frac{\sqrt{\hat{n}_k}}{\sqrt{n_k}}\right| \le \frac{K}{\underline{c}_\pi n}|\hat{n}_k - n_k|, \quad \sum_{k=1}^{K}\frac{(\sqrt{\hat{n}_k} - \sqrt{n_k})^2}{n} \le \frac{K}{\underline{c}_\pi n^2}\sum_{k=1}^{K}(\hat{n}_k - n_k)^2.$$

Now, recall that $\hat{n}_k = \sum_{j=1}^{n} I(\xi_j = k)$ and $\mathbb{E}(\hat{n}_k) = n_k$, and, using Hoeffding inequality, for any $x > 0$, obtain

$$\Pr\left(\left|\hat{n}_k - n_k\right| \ge n\,x\right) \le 2\exp\{-2nx^2\}.$$

For any $t > 0$, setting $x = \sqrt{t\ln n/(2n)}$ and taking the union bound, derive

$$\Pr\left(\max_{\substack{1\le m\le M\\ 1\le k\le K}}\left|\frac{\hat{n}_k^{(m)} - n_k^{(m)}}{n}\right| \le \sqrt{\frac{t\ln n}{2n}}\right) \ge 1 - 2KMn^{-t}. \tag{A.26}$$

Now let $\widetilde{\widetilde{\Omega}}_t$ be the set where (A.26) holds. Then for $\omega \in \widetilde{\widetilde{\Omega}}_t$, one has

$$\|I - \Lambda\| \le \frac{K}{\underline{c}_\pi}\sqrt{\frac{t\ln n}{2n}}, \quad \|\hat{\mathbf{t}} - \mathbf{t}\| \le \sqrt{\frac{K^2}{\underline{c}_\pi}}\sqrt{\frac{t\ln n}{2n}}. \tag{A.27}$$

Finally, combining (A.25) and (A.27), for $\omega \in \widetilde{\widetilde{\Omega}}_t$, we arrive at

$$\max_{1\le m\le M}\left\|\widetilde{\widetilde{U}}^{(m)} - \widetilde{U}^{(m)}\right\| \le n^{-1/2}\,K\sqrt{2t\ln n}/\underline{c}_\pi$$

which completes the proof. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2024.105333.

References

[1] E. Abbe, Community detection and stochastic block models: Recent developments, J. Mach. Learn. Res. 18 (177) (2018) 1–86.
[2] P. Barbillon, S. Donnet, E. Lazega, A. Bar-Hen, Stochastic block models for multiplex networks: An application to a multilevel network of researchers, J. R. Stat. Soc. Ser. A: Stat. Soc. 180 (1) (2016) 295–314.
[3] Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, J. Salmon, Implicit differentiation of lasso-type models for hyperparameter optimization, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, Vol. 119, PMLR, 2020, pp. 810–821.
[4] S. Bhattacharyya, S. Chatterjee, General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers, 2020, arXiv:2004.03480.
[5] P.J. Bickel, A. Chen, A nonparametric view of network models and newman–girvan and other modularities, Proc. Natl. Acad. Sci. 106 (50) (2009) 21068–21073.
[6] P. Bickel, D. Choi, X. Chang, H. Zhang, Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels, Ann. Statist. 41 (4) (2013) 1922–1943.
[7] S. Chand, On tuning parameter selection of lasso-type methods - a monte carlo study, in: Proceedings of 2012 9th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2012, pp. 120–129, http://dx.doi.org/10.1109/IBCAST.2012.6177542.
[8] E.C. Chi, B.J. Gaines, W.W. Sun, H. Zhou, J. Yang, Provable convex co-clustering of tensors, J. Mach. Learn. Res. 21 (214) (2020) 1–58.
[9] M. De Domenico, V. Nicosia, A. Arenas, V. Latora, Structural reducibility of multilayer networks, Nature Commun. 6 (1) (2015) 1–9.
[10] D. Durante, N. Mukherjee, R.C. Steorts, Bayesian learning of dynamic multilayer networks, J. Mach. Learn. Res. 18 (43) (2017) 1–29.
[11] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2790–2797, http://dx.doi.org/10.1109/CVPR.2009.5206547.
[12] E. Elhamifar, R. Vidal, Sparse subspace clustering: Algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
[13] X. Fan, M. Pensky, F. Yu, T. Zhang, ALMA: Alternating minimization algorithm for clustering mixture multilayer network, J. Mach. Learn. Res. 23 (330) (2022) 1–46.

[14] O. Fercoq, A. Gramfort, J. Salmon, Mind the duality gap: safer rules for the Lasso, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, Vol. 37, PMLR, Lille, France, 2015, pp. 333–342.
[15] D. Greene, P. Cunningham, Producing a unified graph representation from multiple social network views, in: Proceedings of the 5th Annual ACM Web Science Conference, 2013, pp. 118–121.
[16] A. Gupta, D. Nagar, Matrix Variate Distributions, Chapman and Hall/CRC, New York, 1999.
[17] R. Han, Y. Luo, M. Wang, A.R. Zhang, Exact clustering in tensor block model: Statistical optimality and computational limit, 2021, arXiv:2012.09996.
[18] B.-Y. Jing, T. Li, Z. Lyu, D. Xia, Community detection on mixture multi-layer networks via regularized tensor decomposition, 2020, arXiv preprint arXiv:2002.04457.
[19] B.-Y. Jing, T. Li, Z. Lyu, D. Xia, Community detection on mixture multilayer networks via regularized tensor decomposition, Ann. Statist. 49 (6) (2021) 3181–3205.
[20] B. Karrer, M.E.J. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E Stat., Nonlinear, Soft Matter Phys. 83 (2011) 016107.
[21] M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, M.A. Porter, Multilayer networks, J. Complex Netw. 2 (3) (2014) 203–271.
[22] C.M. Le, E. Levina, Estimating the number of communities in networks by spectral methods, 2015, arXiv:1507.00827.
[23] C. Lee, D.J. Wilkinson, A review of stochastic block models and extensions for graph clustering, Appl. Netw. Sci. 4 (122) (2019) 1–50.
[24] J. Lei, K. Chen, B. Lynch, Consistent community detection in multi-layer network data, Biometrika 107 (1) (2019) 61–73.
[25] J. Lei, K.Z. Lin, Bias-adjusted spectral clustering in multi-layer stochastic block models, 2021, arXiv:2003.08222.
[26] J. Lei, A. Rinaldo, Consistency of spectral clustering in stochastic block models, Ann. Statist. 43 (1) (2015) 215–237.
[27] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML '10, Omni Press, USA, 2010, pp. 663–670.
[28] F. Lorrain, H.C. White, Structural equivalence of individuals in social networks, J. Math. Sociol. 1 (1) (1971) 49–80.
[29] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[30] P.W. MacDonald, E. Levina, J. Zhu, Latent space models for multiplex networks with shared structure, 2020, arXiv preprint arXiv:2012.14409.
[31] P.W. MacDonald, E. Levina, J. Zhu, Latent space models for multiplex networks with shared structure, 2021, arXiv:2012.14409.
[32] J. Mairal, F. Bach, J. Ponce, G. Sapiro, R. Jenatton, G. Obozinski, SPAMS: A sparse modeling software, v2.3, 2014, URL http://spams-devel.gforge.inria.fr/downloads.html.
[33] C. Matias, V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model, J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (4) (2017) 1119–1141.
[34] B. Nasihatkon, R. Hartley, Graph connectivity in sparse subspace clustering, in: CVPR '11: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2137–2144, http://dx.doi.org/10.1109/CVPR.2011.5995679.
[35] M. Noroozi, M. Pensky, The hierarchy of block models, Sankhyā 84 (2022a) 64–107.
[36] M. Noroozi, M. Pensky, Sparse subspace clustering in diverse multiplex network model, 2022b, arXiv preprint arXiv:2206.07602.
[37] M. Noroozi, M. Pensky, R. Rimal, Sparse popularity adjusted stochastic block model, J. Mach. Learn. Res. 22 (193) (2021a) 1–36.
[38] M. Noroozi, R. Rimal, M. Pensky, Estimation and clustering in popularity adjusted block model, J. R. Stat. Soc. Ser. B Stat. Methodol. 83 (2) (2021b) 293–317.
[39] S. Paul, Y. Chen, Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel, Electron. J. Stat. 10 (2) (2016) 3807–3870.
[40] S. Paul, Y. Chen, Spectral and matrix factorization methods for consistent community detection in multi-layer networks, Ann. Statist. 48 (1) (2020) 230–250.
[41] M. Pensky, Y. Wang, Clustering of diverse multiplex networks, 2021, arXiv preprint arXiv:2110.05308.
[42] C. Rao, M. Rao, Matrix Algebra and its Applications to Statistics and Econometrics, World Scientific Publishing Co., Singapore, 1998.
[43] S. Sengupta, Y. Chen, A block model for node popularity in networks with community structure, J. R. Stat. Soc. Ser. B 80 (2) (2018) 365–386.
[44] M. Soltanolkotabi, E.J. Candes, A geometric analysis of subspace clustering with outliers, Ann. Statist. 40 (4) (2012) 2195–2238.
[45] M. Soltanolkotabi, E. Elhamifar, E.J. Candes, Robust subspace clustering, Ann. Statist. 42 (2) (2014) 669–699.
[46] J.A. Tropp, An introduction to matrix concentration inequalities, Found. Trends Mach. Learn. 8 (1–2) (2015) 1–230.
[47] P. Tseng, Nearest q-flat to m points, J. Optim. Theory Appl. 105 (1) (2000) 249–252.
[48] R. Vidal, Subspace clustering, IEEE Signal Process. Mag. 28 (2) (2011) 52–68.
[49] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1945–1959.
[50] Y. Wang, Y.-X. Wang, A. Singh, Graph connectivity in noisy sparse subspace clustering, in: A. Gretton, C.C. Robert (Eds.), Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, Vol. 51, PMLR, Cadiz, Spain, 2016, pp. 538–546.
[51] Y.-X. Wang, H. Xu, Noisy sparse subspace clustering, J. Mach. Learn. Res. 17 (1) (2016) 320–360.
[52] M. Wang, Y. Zeng, Multiway clustering via tensor block models, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, R. Garnett (Eds.), Adv. Neural Inf. Process. Syst. 32 (2019).
[53] T. Zhang, A. Szlam, Y. Wang, G. Lerman, Hybrid linear modeling via local best-fit flats, Int. J. Comput. Vis. 100 (3) (2012) 217–240.
[54] M. Zhu, A. Ghodsi, Automatic dimensionality selection from the scree plot via the use of profile likelihood, Comput. Statist. Data Anal. 51 (2) (2006) 918–930.