Biometrika (2020), **107**, 1, *pp.* 61–73
Printed in Great Britain

doi: 10.1093/biomet/asz068
Advance Access publication 22 December 2019

Consistent community detection in multi-layer network data

By JING LEI

Department of Statistics and Data Science, Carnegie Mellon University, Baker Hall 132,
Pittsburgh, Pennsylvania 15213, U.S.A.
jinglei@andrew.cmu.edu

KEHUI CHEN and BRIAN LYNCH

Department of Statistics, University of Pittsburgh, 1800 Wesley W. Posvar Hall,
Pittsburgh, Pennsylvania 15260, U.S.A.
khchen@pitt.edu bcl28@pitt.edu

Summary

We consider multi-layer network data where the relationships between pairs of elements are reflected in multiple modalities, and may be described by multivariate or even high-dimensional vectors. Under the multi-layer stochastic block model framework we derive consistency results for a least squares estimation of memberships. Our theorems show that, as compared to single-layer community detection, a multi-layer network provides much richer information that allows for consistent community detection from a much sparser network, with required edge density reduced by a factor of the square root of the number of layers. Moreover, the multi-layer framework can detect cohesive community structure across layers, which might be hard to detect by any single-layer or simple aggregation. Simulations and a data example are provided to support the theoretical results.

Some key words: Community detection; Consistency; Sparse network; Tensor concentration bound.

1. Introduction

A single-layer network consists of a set of n elements and a measure of pairwise interaction between them. The observed data is represented by an adjacency matrix or a more general relationship matrix $A \in \mathbb{R}^{n\times n}$, where A_{jk} $(j,k = 1,\ldots,n)$ is a measure of interaction between element j and element k. Recently, many examples have shown that the relationships between different elements are reflected in multiple modalities, and that the observations may contain multivariate or even high-dimensional vectors that describe the relationship between each pair of elements. For example, in multimodal or multi-task brain connectivity studies, among a set of brain regions one may have one source of linkage inferred from electroencephalography measures during a working memory task, and a second source of linkage inferred from resting state functional magnetic resonance imaging measures. Other examples include social networks, where the interaction between two people could be inferred from Facebook, LinkedIn and more intimate connections such as cell phone contacts. Moreover, time-evolving networks can also be considered as multi-layer network data when the set of elements remains the same over time.

A multi-layer network can be represented by a tensor object $Y \in \mathbb{R}^{m\times n\times n}$, where each layer $Y_{i\cdot\cdot}$ $(i = 1, \ldots, m)$ represents a different aspect of the relationship between elements. In this paper we will focus on multi-layer relational data with community structures, and utilize a multi-layer stochastic block model point of view. The stochastic block model and its variants are powerful tools for modelling large networks with community structures. A single-layer stochastic block model (Holland et al., 1983) can be parameterized by (g, B). The observed adjacency matrix A satisfies $A_{jk} \sim \mathrm{Ber}(B_{g_j g_k})$ with $g \in \{1, \ldots, K\}^n$ and $B \in [0, 1]^{K\times K}$. A natural extension to multi-layer stochastic block models allows the community-wise connectivity parameter B to depend on the layer. In this paper we aim to find the overall clustering pattern of the nodes that are characterized by multiple modalities in a network structure, not the individual clustering pattern in each modality. Therefore, in our setting the membership g is the same across layers. Allowing memberships to change in different layers could also be of interest in some applications. For example, there are works on dynamic stochastic block models where the memberships are allowed to change smoothly or follow some parametric patterns over time (Ghasemian et al., 2016; Pensky & Zhang, 2019).

In recent years a considerable amount of work has emerged on community detection for multi-layer networks, including the weighted modularity approach, spectral methods based on various versions of aggregation or tensor singular value decomposition, and probability model based approaches (Tang et al., 2009; Dong et al., 2012; Kivelä et al., 2014; Xu & Hero, 2014; Han et al., 2015; Ghasemian et al., 2016; Paul & Chen, 2016, 2018; Chen & Hero, 2017; Matias & Miele, 2017; Zhang & Cao, 2017; Bhattacharyya & Chatterjee, 2018; Liu et al., 2018). Despite an explosion of disparate terminology and algorithms, there is limited work on the theoretical analysis of detection limits and consistency results for the multi-layer network structure.

Han et al. (2015) studied the consistency of clustering for the same multi-layer stochastic block model studied here, but under a different asymptotic regime where the number of nodes n is fixed and the number of layers m grows to infinity. In the physics literature, Taylor et al. (2016) considered an identical edge probability matrix B across layers, in which case a signal boost can be achieved by working on the average adjacency matrix of all layers. A recent manuscript by Bhattacharyya & Chatterjee (2018) provided consistency results for spectral methods under sparse multi-layer stochastic block models, but required each layer of the connectivity matrix B to be positive definite, with the smallest singular values bounded away from zero uniformly. Paul & Chen (2018) considered other estimation methods and achieved a similar signal boost under the same positivity conditions. Pensky & Zhang (2019) considered estimating the membership for each individual layer in a dynamic stochastic block model. In the special case that memberships do not change, the method works on the average adjacency matrix and allows a similar signal boost under the positivity assumption.

The main contribution of this paper is that we propose to work with the m-layer network data directly without first averaging the adjacency matrices across layers, and the theoretical results are for general structures of B in m layers. We derive a least squares estimation of memberships, and show that an m-layer network provides much richer information, allowing consistent estimation to be achieved for a sparser network, roughly by a factor of $m^{1/2}$, in each layer. The multi-layer framework only requires a well-defined block structure on the overall m layers, i.e., it is possible that none of the individual layers contain a full block structure. The theoretical analysis involves the development of a new spectral bound for tensor network data, which uses a tensor adaptation of the combinatorial approach developed in Lei & Rinaldo (2015). This new tensor concentration bound is crucial to developing the consistency result for multi-layer networks under weaker conditions on B.

2. Tensor stochastic block models

We use the symbol $\circ$ to denote the outer product of vectors. For example, if x, y and z are vectors, then $T = x \circ y \circ z$ is the three-way tensor with $T_{ijk} = x_i y_j z_k$. For two tensors A and B, both of dimension (m, n_1, n_2), the symbol $A * B$ denotes the $m \times 1$ vector obtained by taking elementwise products of A and B, and then summing over the second and third dimensions. Finally, $\| \cdot \|^2$ denotes the sum of squares for all entries of a vector, matrix or tensor.

A traditional single-layer stochastic block model with n nodes and K communities is parameterized by (g, B), where $g \in \{1, \ldots, K\}^n$ is a membership vector and the $K \times K$ matrix B determines the community-wise connectivity. We also define the $n \times K$ membership matrix G such that the jth row of G is 1 in the g_jth column and 0 otherwise. The observed data A_{jk} $(j, k = 1, \ldots, n)$ has expectation $P_{jk} = B_{g_j g_k}$. The key feature of a stochastic block model is that the expectation P can be reorganized as a blockwise constant matrix by grouping nodes in the same community. In the most commonly seen Bernoulli model, A_{jk} only takes two values, 0 and 1. The edges form independently with $\text{pr}(A_{jk} = 1) = P_{jk}$.

This generative model can be naturally extended to a multi-layer network. Let Y be an $m \times n \times n$ tensor with each layer $Y_{i\cdot\cdot}$ being a random graph generated by a stochastic block model parameterized by $(g, B_{i\cdot\cdot})$. The expectation, P, is an $m \times n \times n$ tensor with $P_{i\cdot\cdot} = GB_{i\cdot\cdot}G^{\mathrm{T}}$. In our setting the membership vector is assumed to be common to all layers, while the connectivity parameter $B_{i\cdot\cdot}$ could be different across layers, reflecting different aspects of node interactions. Here, $B_{ig_jg_k}$ denotes the g_jth row and g_kth column of the connectivity matrix for the ith layer, which equals P_{ijk}.

For example, consider a three-layer network where each layer is generated from a three-block stochastic block model, with connectivity matrices

$$B_{1\cdot\cdot} = \begin{pmatrix} 0.6 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.2 \end{pmatrix}, \quad B_{2\cdot\cdot} = \begin{pmatrix} 0.2 & 0.4 & 0.2 \\ 0.4 & 0.6 & 0.4 \\ 0.2 & 0.4 & 0.2 \end{pmatrix}, \quad B_{3\cdot\cdot} = \begin{pmatrix} 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.6 \end{pmatrix}. \qquad (1)$$

In this three-layer network the ith community is more active in layer i, as reflected in the community-wise connectivity matrix $B_{i\cdot\cdot}$. From the ith layer we can only separate the ith community from the rest. If we average the three layers to form a single-layer network then the community structure cannot be detected at all, since the average $(B_{1\cdot\cdot} + B_{2\cdot\cdot} + B_{3\cdot\cdot})/3$ is a matrix with the same value in all entries. By contrast, using the multi-layer network method developed in this paper we are able to obtain consistent community recovery based on the tensor observation Y generated from this type of stochastic block model.

3. A least squares approach and its statistical properties

A popular and accurate estimation method for stochastic block models is the maximum likelihood estimator, for which the consistency of membership estimation in single-layer networks has been studied by many authors under the condition that a global maximum can be achieved (Bickel & Chen, 2009; Choi et al., 2012; Zhao et al., 2012; Amini et al., 2013; Abbe et al., 2016). Here we focus on its variant, the least squares estimator (Borgs et al., 2015; Gao et al., 2015; Chen & Lei, 2018). In practice, we have found that the least squares estimator performs at least as well as the maximum likelihood estimator. Our proposed theoretical analysis based on the least squares estimator will reveal unique features of multi-layer data.

For multi-layer network data, given an observation $Y \in \mathbb{R}^{m\times n\times n}$ and the number of communities K, the least squares estimator is

$$(\hat{g}, \hat{B}) = \mathop{\arg\min}_{h,\tilde{B}} \sum_{i=1}^{m} \omega_i \sum_{1\leqslant j \neq l \leqslant n} (Y_{ijl} - \tilde{B}_{ih_jh_l})^2, \tag{2}$$

where $\omega = (\omega_1, \ldots, \omega_m)$ are user-defined weights for each layer. The minimization is over all possible $h \in \{1, \ldots, K\}^n$ and all possible $\tilde{B} \in \mathbb{R}^{m\times K\times K}$. At first sight there are a large number of parameters to estimate, but some derivation reveals that the optimal $\tilde{B}$ is uniquely determined given a membership vector. This is analogous to the profile likelihood perspective used in Stephens (2000) and Bickel & Chen (2009).

In the following, we present consistency results for the optimal solution to problem (2). For a fixed community vector h, the optimization problem (2) over $\tilde{B}$ is a simple least squares fit, and the optimal $\tilde{B}$ is obtained by averaging the corresponding entries of Y according to the membership given by h. After profiling out $\tilde{B}$ in (2), the optimization problem essentially involves finding a partition h such that the within-group residual sum of squares is minimized when the entries of each layer of Y are partitioned into $K(K+1)/2$ blocks according to h. Thus, the well-known total variance decomposition implies that the profiled optimization problem over h is equivalent to maximizing the between-group sum of squares,

$$\begin{aligned} f(h; Y) = &\sum_{k=1}^{K} \binom{n_k(h)}{2} \left\| \frac{Y * (\omega \circ H_k \circ H_k)}{n_k(h)\{n_k(h)-1\}} \right\|^2 \\ &+ \sum_{1\leqslant j<k\leqslant K} n_j(h)n_k(h) \left\| \frac{Y * (\omega \circ H_j \circ H_k)}{n_j(h)n_k(h)} \right\|^2, \end{aligned} \tag{3}$$

where H is the membership matrix corresponding to the membership vector h, i.e., the ith row of H is 1 at the location h_i and zero otherwise. We let H_k be the kth column of H and $n_k(h) = \|H_k\|_1$.

In the rest of the paper we use $\omega_i = 1$ $(i = 1, \ldots, m)$. The theory can be easily extended to the more general case of unequal ω_i. To accurately describe the community recovery error in terms of the network sparsity and community separation, we adopt the following settings, which we state as assumptions:

Assumption 1. Network sparsity: $B = \rho_n B^0$, where ρ_n controls the overall network sparsity and the entries of B^0 are of constant order with maximum entry equal to 1.

Assumption 2. Community separation: $\delta^2 = \min_{1\leqslant j\neq j'\leqslant K} \|B^0_{\cdot j \cdot} - B^0_{\cdot j' \cdot}\|^2 > 0$.

Assumption 3. Assume $m \leqslant cn$ for some constant c.

Assumption 1 is common in the literature of network community detection. Assumption 2 is a minimum requirement for the K-community structure. Assumption 3 is made merely for presentational simplicity in the main theorem. The analysis covers the case of larger m as well; see Remark 1 at the end of this section for further details.

Our theoretical analysis of the least squares estimator in (2) consists of three main steps, which we outline next as Lemmas 1–3. The key technical component is a new spectral concentration bound for tensor data, which may be of independent interest. We state the tensor concentration theorem at the end of this section. All proofs are given in the Supplementary Material.

LEMMA 1. *Under Assumption 2, $f(h;P)$ is uniquely maximized by $h = g$, up to a label permutation.*

This lemma ensures parameter identification as the population optimum.

LEMMA 2. *Under Assumptions 1 and 3, with probability tending to one as $n \to \infty$, we have, for some universal constant c_1,*

$$\sup_{h \in \{1,\ldots,K\}^n} |f(h;Y) - f(h;P)| \leqslant c_1 \kappa_n,$$

where $\kappa_n = K(\log n)\{(n\rho_n) \vee \log n\}^{1/2} \left[n\rho_n m^{1/2} + K(\log n)\{(n\rho_n) \vee \log n\}^{1/2}\right]$.

This lemma gives a uniform upper bound for the sampling errors. This is a main technical contribution of this paper. The proof of Lemma 2 relies on Theorem 2 stated at the end of this section, which extends recent spectral bounds for single-layer network data (Chin et al., 2015; Lei & Rinaldo, 2015; Le et al., 2017). The proof technique is a tensor adaptation of the combinatorial approach developed in Feige & Ofek (2005) and Lei & Rinaldo (2015).

Next, we need to analyse the population optimality gap for incorrect membership vectors. Let g be the true membership and h be any membership vector. For $k = 1, \ldots, K$ define $C_k = \{1 \leqslant j \leqslant n : g_j = k\}$, and for any other membership vector h define $\hat{C}_k = \{1 \leqslant j \leqslant n : h_j = k\}$. Define $e_h(k,l)$ as the number of nodes labelled as k in h and l in g:

$$e_h(k,l) = |C_l \cap \hat{C}_k|.$$

For $k = 1, \ldots, K$ and the membership vector h, let $e_h(k)$ be the second largest value in $\{e_h(k,l) : 1 \leqslant l \leqslant K\}$. Define

$$\eta_h = \max_k e_h(k)/n_{\min},$$

where $n_{\min}$ is the smallest community size in g.

Intuitively, the largest value in $\{e_h(k,l) : 1 \leqslant l \leqslant K\}$ corresponds to the correctly clustered nodes, so $(K-1)e_h(k)$ can be viewed as an upper bound on the number of incorrectly clustered nodes in $\hat{C}_k$. When η_h is small, each of the true communities C_l must assign a majority proportion of its nodes to some $\hat{C}_{k(l)}$, and the mapping from l to $k(l)$ must be one to one since otherwise η_h will be large. Therefore, our basic strategy of proving consistency of membership estimation is to show that $\eta_h = o_P(1)$ if h is an optimal solution to (3).

LEMMA 3. *Under Assumption 2, there exists a universal constant c_2 such that*

$$f(g;P) - f(h;P) \geqslant c_2 \eta_h n_{\min}^2 \rho_n^2 \delta^2 K^{-1}.$$

Combining the above three lemmas, we have the following main result:

THEOREM 1 (Main theorem). *Let h be a solution to the least squares problem. Then, under Assumptions 1, 2 and 3, there exists a universal constant $C > 0$ such that the following statements hold with probability tending to one.*

In the sparse case, where $n\rho_n < \log n$,

$$\eta_h \leqslant CK \left(\frac{n}{n_{\min}}\right)^2 \left(\frac{m}{\delta^2}\right) \left\{\frac{K(\log n)^{3/2}}{n\rho_n m^{1/2}}\right\} \left\{1 + \frac{K(\log n)^{3/2}}{n\rho_n m^{1/2}}\right\}.$$

In the moderately dense case, where $n\rho_n \geqslant \log n$,

$$\eta_h \leqslant CK\left(\frac{n}{n_{\min}}\right)^2\left(\frac{m}{\delta^2}\right)\left\{\frac{K\log n}{(n\rho_n m)^{1/2}}\right\}\left\{1+\frac{K\log n}{(n\rho_n m)^{1/2}}\right\}.$$

A prototypical case in the study of stochastic block models is the balanced community case, where $K = O(1)$ and $n_{\min} \asymp n$. Moreover, it is natural to assume that $\delta^2 \asymp m$. This is the case if a constant fraction of layers in the multi-layer stochastic block model exhibits the same scale of between-community connectivity difference, or, more precisely, if there is a constant fraction of is in $\{1,\ldots,m\}$ such that $\max_{j,j'}\|B^0_{ij\cdot} - B^0_{ij'\cdot}\| \geqslant c$ for some positive constant c. In particular, if the layers of B^0 are generated independently from a nondegenerate distribution, we have $\delta^2 \asymp m$ with high probability when m is large.

The state-of-the-art one-layer stochastic block model result requires $\rho_n n \to \infty$ for consistent community recovery. Under the standard assumptions made above, in the sparse case where $n\rho_n < \log n$ we only require $n\rho_n m^{1/2}/(\log n)^{3/2} \to \infty$ to guarantee a vanishing proportion of misclustered nodes. Roughly speaking, the m-layer least squares estimator combines the signal from all layers and enhances the signal strength by a factor of $m^{1/2}$.

Finally, we introduce the key technical component in our proof: the tensor concentration result. Let A be an $n \times n \times m$ tensor, with each layer $A_{\cdot\cdot l}$ being an inhomogeneous Erdős–Rényi random graph with expectation P. The maximum entry of P is of order ρ_n. For presentational simplicity we assume that $m = n$. The more general case can be treated by padding the tensor with zeros when $m < n$. The case of $m > n$ is discussed in Remark 1 below.

Theorem 2. *For any* $(x,y,z) \in \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$, *let* $W = A - P$. *We have* $|\langle W, x\circ y\circ z\rangle| \leqslant c(\log n)\{(n\rho_n) \vee \log n\}^{1/2}$ *for some universal constant* c *with all but vanishing probability as* $n \to \infty$.

Remark 1. When $m > n$ the tensor spectral concentration bound in Theorem 2 becomes $(\log m)\{(m\rho_n) \vee \log m\}^{1/2}$. When $m > n$ but $m = O(n^2)$, using the same analysis as in the proof of Theorem 1, the least squares estimator can achieve consistent community estimation when $n\rho_n$ grows faster than $(\log n)^{3/2}/m^{1/2}$, which is still an $m^{1/2}$ improvement over the density requirement for single-layer networks. A larger m beyond the order of n^2 can further reduce the required sparsity level ρ_n, but the rate of signal boost is no longer $m^{1/2}$. This regime is of less practical interest, because $\rho_n \gg n^{-2}$ is a minimum requirement for each layer to have at least one edge.

4. Numerical experiments

4.1. *Algorithm* 1

The theoretical results developed above pertain to a global optimum of problem (2). In practice, how to achieve or approximate the global optimum remains a challenging and interesting algorithmic question. A full search over K^n possible labellings takes exponential computing time. Greedy search methods such as the label-switching algorithm (Bickel & Chen, 2009) and the tabu search (Zhao et al., 2012) have been used in network clustering. The algorithm that we used in our numerical experiments, Algorithm 1, can be viewed as a label-switching method with batch updates, and can also be viewed as an adaptation of Lloyd's algorithm for k-means clustering to network data. The algorithm proceeds as follows:

Step 1. Initialize by k-means on n slices of the data, where the jth data point is a column slice $Y_{\cdot\cdot j}$, viewed as a vector of length mn.

Step 2. Assume that the current iteration starts with a membership vector g^{old}. Find a new community vector g^{new} where

$$g_j^{\text{new}} = \arg\min_{k\in\{1,\ldots,K\}} \sum_{i=1}^{m} \omega_i \sum_{l\neq j} \{Y_{ijl} - B^{\text{old}}_{ikg_l^{\text{old}}}\}^2.$$

Step 3. Compute B^{new}:

$$B^{\text{new}}_{ikk'} = \frac{\sum_{j\neq l} Y_{ijl} \mathbb{1}_{\{g_j^{\text{new}}=k\}} \mathbb{1}_{\{g_l^{\text{new}}=k'\}}}{\sum_{j\neq l} \mathbb{1}_{\{g_j^{\text{new}}=k\}} \mathbb{1}_{\{g_l^{\text{new}}=k'\}}}.$$

Step 4. Compute the least squares loss function with respect to g^{new} and B^{new}, and update with $g^{\text{old}} \leftarrow g^{\text{new}}$ and $B^{\text{old}} \leftarrow B^{\text{new}}$ if the loss function reduces.

Step 5. Repeat Steps 2–4 until the objective function cannot be further reduced.

Lloyd's algorithm is arguably the most popular approach for k-means clustering, due to its simplicity and fast convergence. It is not guaranteed to converge to a local minimum, though, if the local minimum is defined as a partition of the data where moving any single point to a different cluster increases the objective function. Arthur & Vassilvitskii (2007) showed that Lloyd's algorithm combined with a good starting point instead of a purely random start can provide an accurate approximate solution with small optimality gap. Here, we initialize our algorithm by k-means on slices of data, known as marginal clustering, which has been proved to be a very good initial point in the co-clustering literature (Anagnostopoulos et al., 2008). In our numerical studies we repeat the algorithm three times and retain the choice with the smallest objective value, and use weights $\omega_i \equiv 1$. The algorithm performs very satisfactorily in our simulations shown in § 4.2.

4.2. *Simulations*

In this section we illustrate the performance of our proposed method using simulations, where we generate multi-layer networks $Y \in \mathbb{R}^{m\times n\times n}$ given membership vector g and community connectivity tensor B. We compare our multi-layer clustering method to a single-layer method based on the same least squares criterion as in (2), either applied to only the first layer of Y or to the average over all layers of Y. In addition, we compare our method with spectral clustering applied to the average of the layers. In each simulation trial, given a membership matrix G and a B with each layer symmetric, the upper triangular entries of $Y_{i\cdot\cdot}$ are independently generated as Bernoulli random variables with probabilities given by $P_{i\cdot\cdot} = GB_{i\cdot\cdot}G^{\mathrm{T}}$. The nodes are divided into clusters such that the number of nodes in each of the first $K-1$ communities is $\lfloor n/K \rfloor$, and the Kth community contains the remaining nodes. When we instead used unequal community sizes, our method performed similarly.

In Simulation I the entries of B are randomly generated in each trial. To do this we first generate $B_0 \in \mathbb{R}^{m\times K\times K}$, where the upper triangular and diagonal entries of each layer $B_{0,i\cdot\cdot}$ are generated independently from Un$(0, 0.5)$, and the lower triangular entries are set equal to their corresponding upper triangular entries. We then set $B = rB_0$, where $r \leqslant 1$ is a preselected positive parameter that controls sparsity. If a layer of B_0 has Kth singular value less than a preset cut-off value, that layer is regenerated to ensure that we have well-formed K-block structures.

We consider $n \in \{50, 100, 200\}$ as the number of nodes, $m = 2^j$ $(j = 1, \ldots, 6)$ as the number of layers, $K \in \{2, 3, 4\}$ as the number of communities and $r \in \{0.05, 0.1, 0.25, 0.5, 1\}$ as the

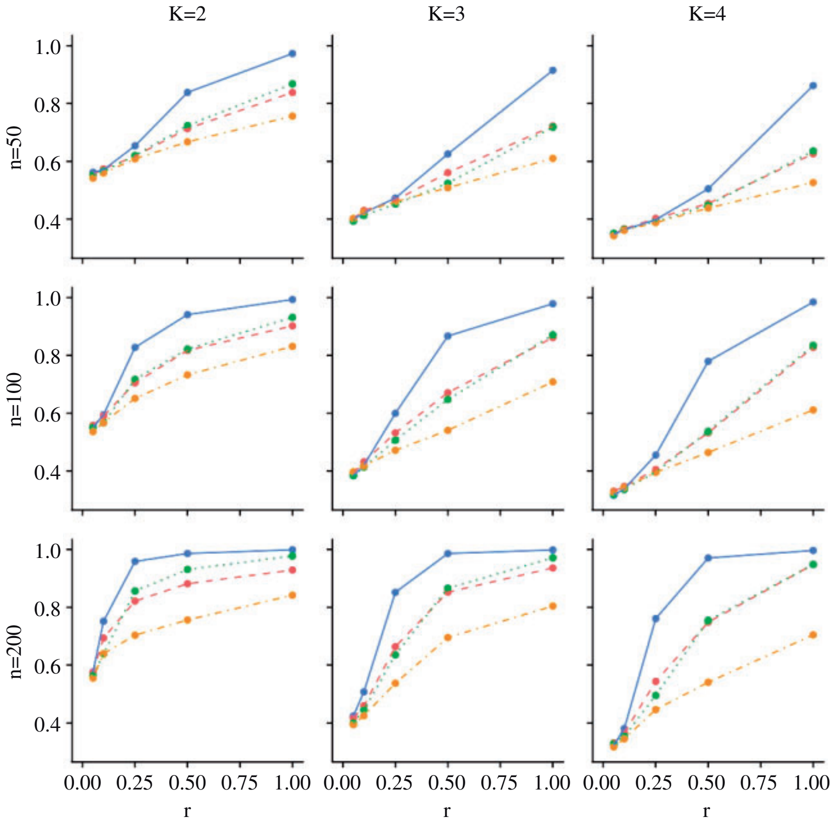

Fig. 1. Simulation I: Proportion of nodes correctly assigned for $m = 2$ layers and $r \in \{0.05, 0.1, 0.25, 0.5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer least squares method applied to the average of the layers, the green dotted line is for the single-layer least squares method applied to the first layer, and the orange dotted and dashed line is for the spectral method on the average of the layers.

level of sparsity. For each combination of values (n, m, r, K) we run 100 simulation trials. The proportion of nodes correctly clustered by a given method is averaged over all trials. In Figs. 1 and 2 we plot the success rates against different values of r, and we only show the results for $m = 2$ and $m = 8$, respectively, as the success rates for relatively dense networks are close to 1. The performance for $m = 4$ is in between these and we omit it to save space. In Fig. 3 we show the success rate of our method against different values of $m \in \{2, 4, 8, 16, 32, 64\}$ for relatively sparse networks with $r \in \{0.05, 0.1, 0.25\}$. Here, we see that sufficiently large m and n can allow for nearly 100% correct assignment, even for small r. This supports the results of Theorem 1 and the discussion following it. The performance of our multi-layer method is clearly superior to that of the single-layer methods. All of the methods show improved performance as r increases, n increases or K decreases. The performance of our multi-layer method also improves as m increases, while the performance of the single-layer methods remains relatively static with m. Under the layer-wise positivity assumption, the spectral method on the average of the layers has been proven to have superior performance in Taylor et al. (2016), Paul & Chen (2018) and Bhattacharyya & Chatterjee (2018). However, in our simulations the $B_{i\cdot\cdot}$ are generated randomly without any imposed positivity, so there can be signal cancellation due to averaging. Moreover, in our simulations the spectral method on the average of the layers does not work as well as the least squares method on the average of the layers. The reason for this is that our B matrices

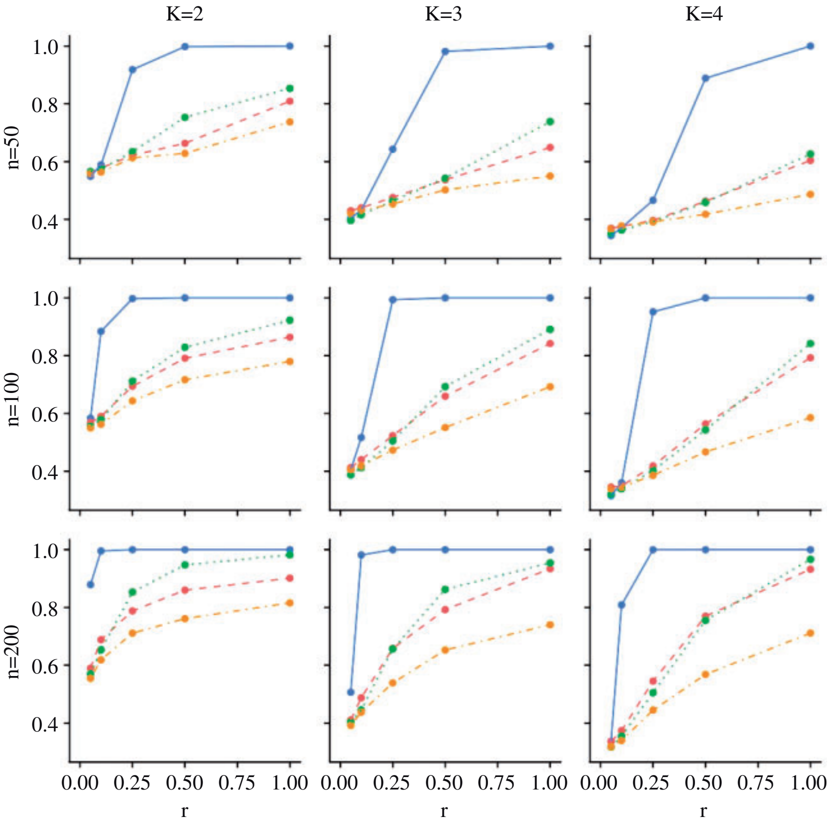

Fig. 2. Simulation I: Proportion of nodes correctly assigned for $m = 8$ layers and $r \in \{0.05, 0.1, 0.25, 0.5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer least squares method applied to the average of the layers, the green dotted line is for the single-layer least squares method applied to the first layer, and the orange dotted and dashed line is for the spectral method on the average of the layers.

are randomly generated and the kth singular value of the averaged layers could be very small. Spectral methods do not work well in these scenarios.

In Simulation II, B_0 is assigned a constant value over all trials by using $m = K = 3$ and defining each layer of B_0 as in (1). In this case, the average over all layers of $B = rB_0$ is a constant matrix. Each individual layer can only identify two unique clusters, although there are three clusters when all layers are considered. Figure 4 shows simulation results in this scenario for the three least squares methods considered above, calculating the proportion of nodes correctly identified over 100 simulation trials. As before, our multi-layer method performs the best, increasing toward 100% correct clustering as r and n increase. As expected, the averaging method performs poorly irrespective of r and n. Using only the first layer results in a performance that improves slightly with r and n, but as expected never approaches 100% accuracy.

4.3. *Application to a gene network study*

In this section we apply our method to a gene network dataset. The data have been described and used by Liu et al. (2018), and include gene co-expression data in the medial prefrontal cortex from studies of rhesus monkeys at different stages of development. For the prenatal period, they consider a six-layer network corresponding to six age categories, labelled E40 to E120 to

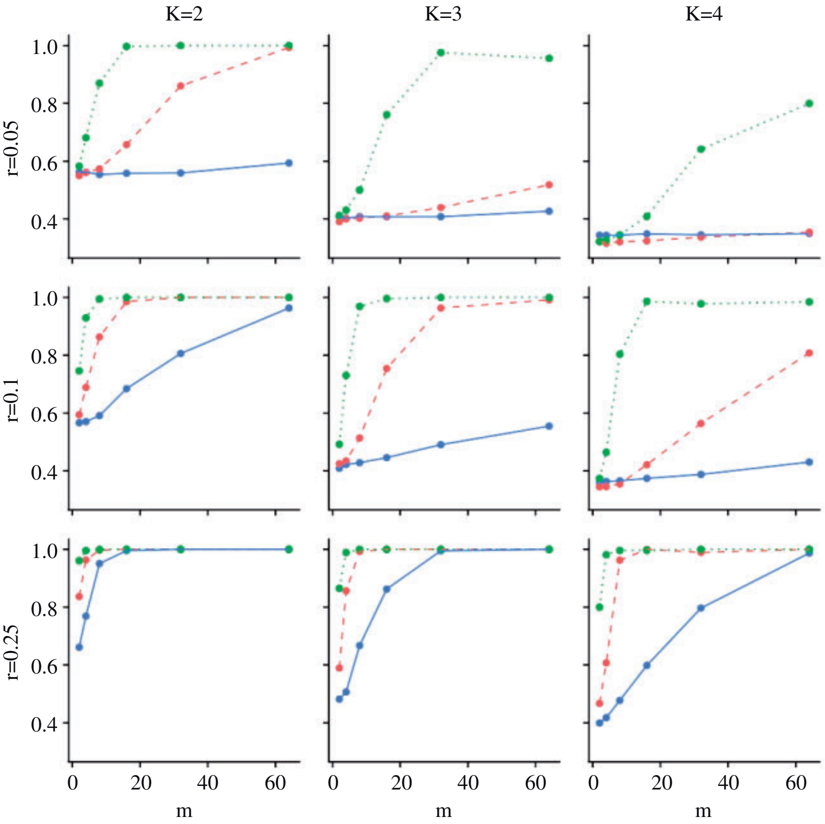

Fig. 3. Simulation I: Proportion of nodes correctly assigned by the multi-layer method, for $m \in \{2, 4, 8, 16, 32, 64\}$ and $r \in \{0.05, 0.1, 0.25, 0.5, 1\}$. The blue solid line, red dashed line and green dotted line correspond to $n = 50$, 100 and 200, respectively.

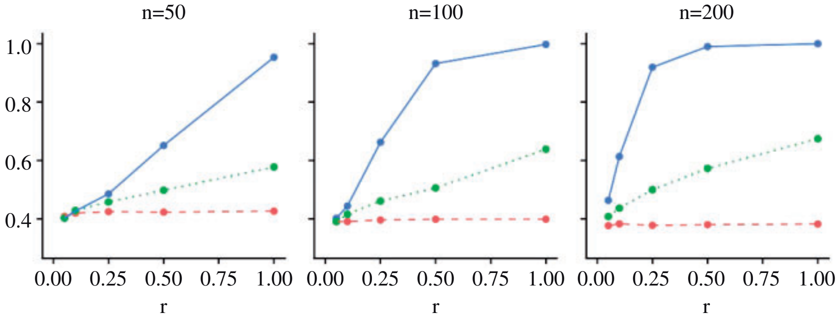

Fig. 4. Simulation II: Proportion of nodes correctly assigned for $r \in \{0.05, 0, 1, 0.25, 0.5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer method applied to the average of the layers, and the green dotted line is for the single-layer method applied to the first layer.

indicate the number of embryonic days of age. For the postnatal period, they consider a five-layer network corresponding to five layers within the medial prefrontal cortex, labelled L2 to L6. Studies of the medial prefrontal cortex have been used to understand developmental brain disorders, and Liu et al. (2018) make special note of sets of genes that are significantly enriched for neural projection guidance, which has been shown to be related to autism spectrum disorder. These genes are marked in red in their Figs. 5 and S10. We focus on the set of neural projection

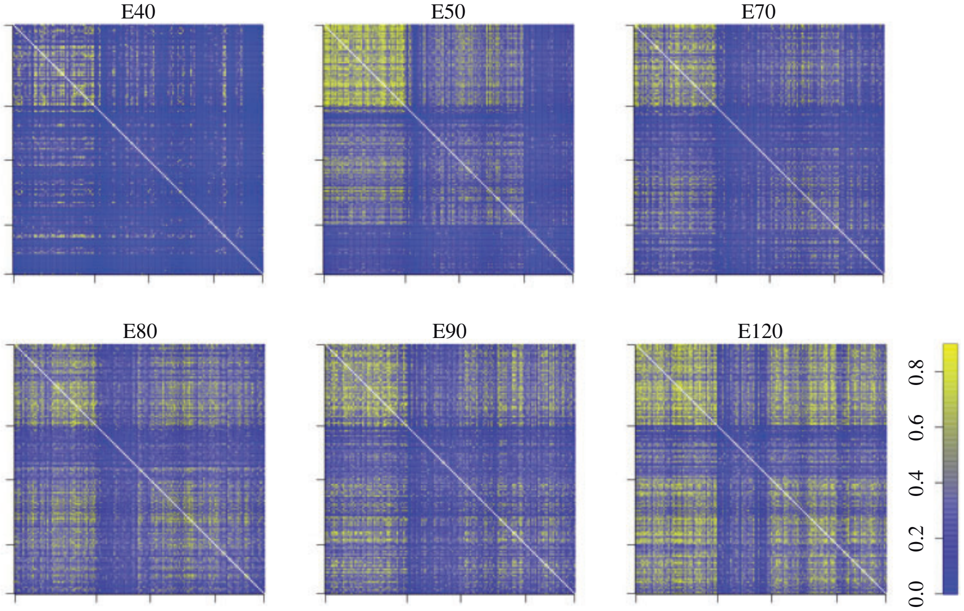

Fig. 5. The connectivity matrices for each layer of the prenatal data, with genes ordered by the clusters. Tick marks denote the boundaries between the clusters.

guidance-enriched genes, which results in $n = 154$ nodes for the prenatal network and $n = 117$ nodes for the postnatal network. The networks were constructed by soft thresholding the sample correlation computed from 423 samples from several groups, and we refer to Liu et al. (2018) for details of the calculation.

Our clustering results are visualized for the prenatal and postnatal data in Figs. 5 and 6, respectively, in which each layer's connectivity matrix is ordered according to the results of the multi-layer clustering. Here, $K = 4$ clusters are used based on visual inspection of the communities of red genes in Figures 5 and S10 of Liu et al. (2018). In both the prenatal and postnatal networks, the individual layers show the four clusters grouped in different ways, giving partial views of the overall clustering and revealing the advantage of our method in finding a cohesive portrait of the communities. In Fig. 5, the connectivity matrix for the first layer E40 seems to differentiate two groups of genes, including the first cluster and the combined next three clusters. The second layer E50 shows three groups, including the first cluster, the second two clusters, and the fourth cluster. In the third layer E70, clusters 1 and 3 appear distinct, while clusters 2 and 4 look like they could be grouped. In E80, E90 and E120, the last two clusters seem to be grouped, the first cluster seems to be either distinct or grouped with the last two, and the second cluster is distinct. The postnatal network analysis, shown in Fig. 6, reveals similar phenomena. Each of the layers gives partial or weak information about the overall clustering, and combining them gives a stronger signal that captures the structure of the gene clustering over all layers.

5. Discussion

The greedy algorithm works very well in numerical experiments, but the rigorous theoretical analysis of the approximate solutions is still an open question, even in single-layer network data analysis. We believe this would be an interesting and challenging topic for future study.

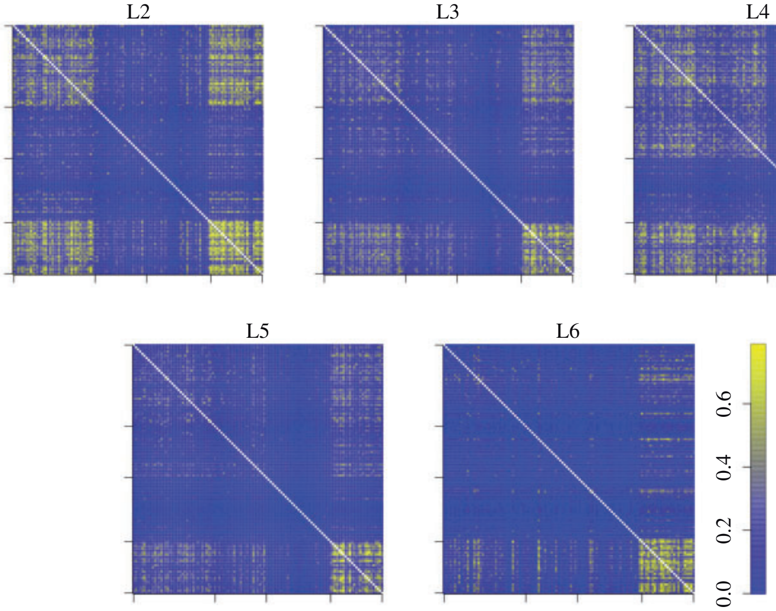

Fig. 6. The connectivity matrices for each layer of the postnatal data, with genes ordered by the clusters. Tick marks denote the boundaries between the clusters.

As pointed out in the introduction, there are also various other estimators for multi-layer network data proposed in the literature, especially tensor spectral methods. The theoretical analysis of these methods, and subsequent comparisons, are also of interest.

Our model can also be considered through the perspective of modelling non-binary interactions, extending the traditional single-layer network that records binary interactions among nodes. There are other types of pairwise interactions considered in the literature, such as the categorical interaction in Lelarge et al. (2015) and the continuous-valued interaction in Xu et al. (2018). It would be interesting to investigate and understand these models in a unified framework.

Acknowledgement

The authors are grateful for the comments the reviewers and editors provided to improve the paper. The authors thank Dr. Kathryn Roeder and Dr. Fuchen Liu for providing the gene network data and for helpful discussion. This research was partially supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online contains proofs of the lemmas and main theorem.

References

Abbe, E., Bandeira, A. S. & Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Info. Theory* **62**, 471–87.

Amini, A. A., Chen, A., Bickel, P. J. & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41**, 2097–122.

ANAGNOSTOPOULOS, A., DASGUPTA, A. & KUMAR, R. (2008). Approximation algorithms for co-clustering. In *Proc. Principles of Database Systems*, pp. 201–10.

ARTHUR, D. & VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proc. 18th Ann. ACM-SIAM Symp. on Discrete Algorithms*. Society for Industrial and Applied Mathematics.

BHATTACHARYYA, S. & CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv:*1805.10594.

BICKEL, P. J. & CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106**, 21068–73.

BORGS, C., CHAYES, J. & SMITH, A. (2015). Private graphon estimation for sparse graphs. In *Proc. 28th Int. Conf. on Neural Information Processing Systems*, vol. 1, pp. 1369–77.

CHEN, K. & LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Am. Statist. Assoc.* **113**, 241–51.

CHEN, P.-Y. & HERO, A. O. (2017). Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms. *IEEE Trans. Sig. Info. Proces. Networks* **3**, 553–67.

CHIN, P., RAO, A. & VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proc. Conf. on Learning Theory*, pp. 391–423.

CHOI, D. S., WOLFE, P. J. & AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–84.

DONG, X., FROSSARD, P., VANDERGHEYNST, P. & NEFEDOV, N. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Trans. Sig. Proces.* **60**, 5820–31.

FEIGE, U. & OFEK, E. (2005). Spectral techniques applied to sparse random graphs. *Random Struct. Algor.* **27**, 251–75.

GAO, C., LU, Y. & ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43**, 2624–52.

GHASEMIAN, A., ZHANG, P., CLAUSET, A., MOORE, C. & PEEL, L. (2016). Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Phys. Rev. X* **6**, 031005.

HAN, Q., XU, K. & AIROLDI, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *Proc. Int. Conf. on Machine Learning*, pp. 1511–20.

HOLLAND, P. W., LASKEY, K. B. & LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5**, 109–37.

KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. & PORTER, M. A. (2014). Multilayer networks. *J. Complex Networks* **2**, 203–71.

LE, C. M., LEVINA, E. & VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Struct. Algor.* **51**, 538–61.

LEI, J. & RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43**, 215–37.

LELARGE, M., MASSOULIÉ, L. & XU, J. (2015). Reconstruction in the labelled stochastic block model. *IEEE Trans. Networks Sci. Eng.* **2**, 152–63.

LIU, F., CHOI, D., XIE, L. & ROEDER, K. (2018). Global spectral clustering in dynamic networks. *Proc. Nat. Acad. Sci.* **115**, 927–32.

MATIAS, C. & MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Statist. Soc.* B **79**, 1119–41.

PAUL, S. & CHEN, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Statist.* **10**, 3807–70.

PAUL, S. & CHEN, Y. (2018). Consistency of community detection in multi-layer networks using spectral and matrix factorization methods. *arXiv:*1704.07353.

PENSKY, M., & ZHANG, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electron. J. Statist.* **13**, 678–709.

STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Statist. Soc.* B **62**, 795–809.

TANG, W., LU, Z. & DHILLON, I. S. (2009). Clustering with multiple graphs. In *Proc. Int. Conf. on Data Mining (ICDM)*, pp. 1016–21. IEEE.

TAYLOR, D., SHAI, S., STANLEY, N. & MUCHA, P. J. (2016). Enhanced detectability of community structure in multilayer networks through layer aggregation. *Phys. Rev. Lett.* **116**, 228301.

XU, K. S. & HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Selec. Topics Sig. Proces.* **8**, 552–62.

XU, M., JOG, V. & LOH, P.-L. (2018). Optimal rates for community estimation in the weighted stochastic block model. *arXiv:*1706.01175v2.

ZHANG, J. & CAO, J. (2017). Finding common modules in a time-varying network with application to the *Drosophila melanogaster* gene regulation network. *J. Am. Statist. Assoc.* **112**, 994–1008.

ZHAO, Y., LEVINA, E. & ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40**, 2266–92.

[*Received on* 2 *October* 2018. *Editorial decision on* 28 *May* 2019]