A Flexible Latent Space Model for Multilayer Networks

Xuefei Zhang [1] **Songkai Xue** [1] **Ji Zhu** [1]

Abstract

Entities often interact with each other through multiple types of relations, which can be represented as multilayer networks. Multilayer networks among the same set of nodes usually share common structures, while each layer can also possess its distinct node connecting behaviors. This paper proposes a flexible latent space model for multilayer networks for the purpose of capturing such characteristics. Specifically, the proposed model embeds each node with a latent vector shared among layers and a layer-specific effect for each layer; both elements together with a layer-specific connectivity matrix determine edge formations. To fit the model, we develop a projected gradient descent algorithm for efficient parameter estimation. We also establish theoretical properties of the maximum likelihood estimators and show that the upper bound of the common latent structure's estimation error is inversely proportional to the number of layers under mild conditions. The superior performance of the proposed model is demonstrated through simulation studies and applications to two real-world data examples.

1. Introduction

Network data represent relationships among entities and are ubiquitous in various fields, such as social media, neuroscience, and computer science (Newman, 2010; Kolaczyk & Csárdi, 2014). In many applications, individuals often interact with each other through more than one type of relations. For example, people can be coworkers or friends (Lazega et al., 2001); or interactions among individuals can happen through social activities or money exchanges (Banerjee et al., 2013). Multiple types of relationships among entities naturally introduce multilayer networks, where different networks share matched node set, while each single network has a distinct edge type defined through a type of relationship. Tools designed for a single network can be naively used to deal with multilayer networks in two ways: either aggregating multiple layers into a single network or analyzing each single network separately. However, aggregating multilayer networks may lose the specific information contained in each layer, while analyzing each network separately does not leverage the information that may be shared across different relations. Therefore, it is of importance to design tailored tools for multilayer network data.

Real-world multilayer networks are often observed with both homogeneity shared between different layers and heterogeneity retained within each layer. For example, nodes usually have their own intrinsic traits that are consistent across different relations, and at the same time, specific activity levels of individual nodes and the overall network connecting characteristics, such as the edge density or homophily patterns, may vary across different layers. Figure 3 provides an example on multiple social networks among the same set of people and demonstrates the heterogeneous node individual behaviors in different social relations. In this paper, we propose a flexible model for multilayer networks which uses the latent space model for a single network (Hoff et al., 2002) as building blocks, with the goal of capturing aforementioned observed characteristics for multilayer networks. Specifically, we assume each node is represented by a common latent vector shared across layers such that the commonality among layers is kept. Moreover, we assume within each layer, nodes have layer-specific individual effects and connecting patterns, accommodating distinctions between layers. Model specifications and a scalable model fitting algorithm are introduced in Section 3. In Section 4, we establish theoretical properties of the maximum likelihood estimators of the proposed model. In particular, we study the relationship between the estimation of nodes' common latent representations and the number of layers. The theoretical properties are further supported by simulation studies in Section 5. We also demonstrate the performance of the proposed model in terms of latent variables estimation and link prediction on real-world examples in Section 6.

The main contributions of this paper include two aspects. The first one is on the model specification. Our proposed model is more flexible in comparison to existing models in the literature (summarized in Section 2) in the sense that it

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA. Correspondence to: Ji Zhu <jizhu@umich.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

allows layer-specific node individual effects, and such flexibility is shown to be necessary when fitting real-world multilayer networks. Introducing these additional layer-specific node individual parameters brings non-trivial challenges for studying theoretical properties of the model, because the mean structure of the resulting multilayer networks will go beyond the low-rank assumption. The second contribution of our work is on theory, in which we prove that the estimation error of the layer-shared node representations is inversely proportional to the number of layers. This result provides the insight that leveraging multilayer networks for joint estimation is more beneficial than separate estimation with single networks. To the best of our knowledge, this is the first theoretical guarantee on latent variable estimation for multilayer latent space models. Moreover, since our model contains several tensor factorization models (e.g., Nickel et al. (2011); Nickel & Tresp (2013)) as special cases, our results also provide theoretical support for these models, which is less studied in the literature.

2. Related Work

In recent years there has been a growth of statistical models for multilayer networks. The majority of the work extends the tools for modeling a single network to jointly modeling multiple networks, and examples include but are not limited to extensions of: the stochastic block model (SBM) (Han et al., 2015; Paul & Chen, 2015; 2020), the mixed membership SBM (De Bacco et al., 2017), the random dot-product graph model (Levin et al., 2017; Wang et al., 2017; Nielsen & Witten, 2018; Arroyo et al., 2019), and the latent space model (Gollini & Murphy, 2016; Salter-Townshend & McCormick, 2017; D'Angelo et al., 2018; DAngelo et al., 2019; Simpson et al., 2019; Wilson et al., 2020). We chose to build on the latent space model due to its flexibility in capturing commonly observed network characteristics, such as node degree heterogeneity, transitivity, homophily, etc.

Latent space models for a single network are first proposed in Hoff et al. (2002), and their variants are further developed in Hoff (2003; 2008) and Ma et al. (2020). Gollini & Murphy (2016) and DAngelo et al. (2019) proposed latent space models for multilayer networks, with the assumption that the latent representations for each node are the same across all layers and the variation between networks is captured through layer-specific parameters that control overall network characteristics, such as edge density or homophily patterns. The assumption of common node representations implicitly suggests that a node has consistent behaviors through all layers and the nodes that behave similarly in one layer should also behave similarly in other layers. This assumption is relatively strict, as it does not reflect the node-level differences across different layers. Salter-Townshend & McCormick (2017) allows each node to have a distinct latent representation in each layer. However, due to their specific model assumption, these latent representations are "conditional" and therefore not straightforward to interpret. D'Angelo et al. (2018) extends Gollini & Murphy (2016) and incorporates layer-specific node effects in each layer. This work is in spirit the closest to our proposed model. However, it considers a different family of latent space models and adopts Bayesian estimation for model fitting, which is computationally much more expensive, and further, there is no theoretical guarantee on model estimation. (Wilson et al., 2020) and (Simpson et al., 2019) allow more variations across layers by taking the layer-specific node covariates into account, though node covariates information are not generally avaiable in many applications.

Multilayer networks sometimes also refer to dynamic or time-evolving networks (Sewell & Chen, 2015; 2017; Gupta et al., 2018), in which connections among the same set of nodes are recorded at different timestamps. The focus is often on the dependency between different layers due to the time order, so the modeling framework is different from that of multiple types of relations. Lastly, multilayer networks can be viewed as a three-way tensor, where the first two dimensions are along the nodes and the third dimension is along the layers. Tensor factorization methods (Tucker, 1966; De Lathauwer et al., 2000; Nickel et al., 2011; Nickel & Tresp, 2013) have often been utilized for analyzing multi-relational data. Some special cases of our model reduce to existing tensor factorization models, such as the logistic RESCAL model (Nickel & Tresp, 2013). Therefore, the estimation approach and theoretical results we develop can be directly applied in these cases.

3. Proposed Model

Motivated by phenomena observed in real-world multilayer networks and limitations in existing work, we aim to propose a model that has the following properties. First, it should be able to capture the homogeneity and heterogeneity across multiple layers simultaneously. In particular, it should allow node individual effects to vary between layers. Secondly, model parameters should be straightforward to interpret. Lastly, scalable estimation approaches can be developed and theoretical guarantees can be established.

We start with introducing notations. Assuming the multilayer networks are composed of R different networks over a common set of n nodes, with each network representing one type of relation. For $r = 1, \ldots, R$, the rth layer network is represented by a binary adjacency matrix $A^{(r)} \in \{0,1\}^{n\times n}$, where $A^{(r)}_{ij} = A^{(r)}_{ji} = 1$ if node i and node j are connected in the rth relation and $A^{(r)}_{ij} = 0$ otherwise. Stacking the R adjacency matrices together, we obtain a three-way adjacency tensor, denoted by $A = [A^{(1)}; \ldots; A^{(R)}] \in \{0,1\}^{n\times n\times R}$. Note that our

development in this paper focus on adjacency tensor with binary entries, however, the proposed model and theoretical results can be naturally extended to edge types which are modeled by other exponentail family models, such as continuous or count edges.

3.1. Latent Space Model for Multilayer Networks

We extend the main idea in the latent space model for a single layer network, where the connecting probability between two nodes depends on their latent representations in an unobserved Euclidean space. Specifically, we assume that each node i is represented by a unique latent vector $U_i \in \mathbb{R}^k$. Given node latent vectors and layer-specific parameters, we assume connectivity between each pair of nodes i and j in all layers are conditionally independent Bernoulli random variables, i.e.,

$$A_{ij}^{(r)} \overset{ind}{\sim} \text{Bernoulli}\left(P_{ij}^{(r)}\right),$$

where

$$\text{logit}\left(P_{ij}^{(r)}\right) \coloneqq \Theta_{ij}^{(r)} = \alpha_i^{(r)} + \alpha_j^{(r)} + U_i^\top \Lambda^{(r)} U_j, \quad (1)$$

for $i, j = 1, \ldots, n$ and $r = 1, \ldots, R$. Note $\alpha^{(r)} = (\alpha_1^{(r)}, \ldots, \alpha_n^{(r)})^\top \in \mathbb{R}^n$ are node degree heterogeneity parameters for layer r. Specifically, when all other parameters are fixed, the larger $\alpha_i^{(r)}$, the more likely that node i connects with other nodes in the rth layer. The $\alpha^{(r)}$s are distinct across different layers, allowing nodes to have different degree heterogeneity in different types of relations. Moreover, the latent positions $U = [U_1, \ldots, U_n]^\top \in \mathbb{R}^{n\times k}$ are shared between all layers, which capture the common structure between multiple networks among the same set of nodes. The node latent variables U enter the model through a layer-specific connection matrix $\Lambda^{(r)} \in \mathbb{R}^{k\times k}$, $r = 1, \ldots, R$. In general $\Lambda^{(r)} \in \mathbb{R}^{k\times k}$ does not need to be diagonal. In the special case when $\Lambda^{(r)} = I_k$, model (1) for a single layer coincides with the inner-product model considered in Hoff (2003) and Ma et al. (2020). We propose to use non-diagonal $\Lambda^{(r)}$s as they allow not only different levels of homophily along different dimensions, but also general interactions between different dimensions of latent variables.

In summary, model (1) accommodates enough differences between layers, as it embeds each node through two components: layer-varying node individual effects $\{\alpha^{(r)}\}_{r=1}^R$ and layer-invariant latent positions U. Layer-specific connection matrices $\{\Lambda^{(r)}\}_{r=1}^R$ provide additional flexibility, allowing each layer to retain its own network-level characteristics. Therefore, information can be borrowed across different layers due to the shared latent structure, meanwhile each layer is also distinct in terms of its own node connecting behaviors.

Note in order for model (1) to be identifiable, additional constraints on parameters are necessary. The following proposition states the identifiability conditions, and its proof is provided in the Supplementary Material.

Proposition 1 (Identifiability conditions) *Suppose that two sets of parameters* $(\{\alpha^{(r)}\}_{r=1}^R, \{\Lambda^{(r)}\}_{r=1}^R, U)$ *and* $(\{\alpha_\dagger^{(r)}\}_{r=1}^R, \{\Lambda_\dagger^{(r)}\}_{r=1}^R, U_\dagger)$ *satisfy the following conditions:*

A1. $J_n U = U$, $J_n U_\dagger = U_\dagger$, *where* $J_n = I_n - \frac{1}{n} 1_n 1_n^\top$;
A2. $U^\top U = n I_k$ *and* $U_\dagger^\top U_\dagger = n I_k$;
A3. At least one of $\Lambda^{(r)}$*'s,* $r = 1, 2, \cdots, R$, *is full rank.*
Then

$$\begin{aligned}&\alpha^{(r)} 1_n^\top + 1_n \alpha^{(r)\top} + U \Lambda^{(r)} U^\top \\ =&\alpha_\dagger^{(r)} 1_n^\top + 1_n \alpha_\dagger^{(r)\top} + U_\dagger \Lambda_\dagger^{(r)} U_\dagger^\top\end{aligned}$$

for $r = 1, \ldots, R$ *implies that there exists an orthonormal matrix* $O \in \mathbb{R}^{k\times k}$ *where* $O^\top O = O O^\top = I_k$, *such that*

$$\alpha_\dagger^{(r)} = \alpha^{(r)}, U_\dagger = UO, \Lambda_\dagger^{(r)} = O^\top \Lambda^{(r)} O,$$

for $r = 1, \cdots, R$.

3.2. Parameter Estimation

We define the objective function as the negative conditional log-likelihood of A under model (1):

$$\begin{aligned}&L\left(U, \{\alpha^{(r)}\}_{r=1}^R, \{\Lambda^{(r)}\}_{r=1}^R\right) \\ =&-\log P\left(A \mid U, \{\alpha^{(r)}\}_{r=1}^R, \{\Lambda^{(r)}\}_{r=1}^R\right) \\ =&-\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^n \left\{A_{ij}^{(r)} \Theta_{ij}^{(r)} + \log\left(1 - \sigma(\Theta_{ij}^{(r)})\right)\right\},\end{aligned} \quad (2)$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid fucntion. The goal is to find estimates $\widehat{U}$, $\{\widehat{\alpha}^{(r)}\}_{r=1}^R$, and $\{\widehat{\Lambda}^{(r)}\}_{r=1}^R$ that minimize the objective function defined in (2). For the purpose of interpretation and estimation, we treat all the parameters including the node degree heterogeneity parameters $\{\alpha^{(r)}\}_{r=1}^R$ and latent positions U as fixed effects. This is different from the majority of existing work on single layer and multilayer network latent space models (Hoff et al., 2002; Hoff, 2003; Salter-Townshend & McCormick, 2017; D'Angelo et al., 2018; DAngelo et al., 2019), where latent vectors U and node effects (if considered) are treated as random effects and Bayesian approaches are adopted for estimation. Ma et al. (2020) is the first work that treated U as fixed latent representations and proposed a scalable projected gradient descent algorithm for estimating the single layer inner-product latent space model. In pursuit of computational efficiency, we adapt the projected gradient descent algorithm for estimating our multilayer network latent space model. Specifically, in each iteration, parameter estimates for U, $\{\alpha^{(r)}\}_{r=1}^R$ and $\{\Lambda^{(r)}\}_{r=1}^R$ are updated along the direction of their negative gradients of L and are further projected onto the set of parameter space that satisfies

the identifiability condition. The procedure is summarized in Algorithm 1. Note that the update of U leverages the network links of all types. Therefore, we expect a superior estimation of U, in comparison to the estimate when using a single network only.

4. Theoretical Results

In this section we present our main theoretical results on the maximum likelihood estimators of model (1). Note though each $\Theta^{(r)}$ under model (1) is of rank at most $k+2$, the tensor that represents the overall edge connection probabilities, $[\Theta^{(1)};\ldots;\Theta^{(R)}]$, is beyond the low rank structure due to the layer-specific parameters $\{\alpha^{(r)}\}_{r=1}^R$. This brings non-trivial challenges for studying theoretical properties of estimators for the model, since most tensor recovery methods rely on a low-rank assumption for the tensor's mean structure (Kolda & Bader, 2009; Wang & Song, 2017; Ghadermarzy et al., 2018; Wang & Li, 2018). We adopt recently developed tools in random tensor theory and tensor inequalities to establish an upper bound on the estimation error of $[\Theta^{(1)};\ldots;\Theta^{(R)}]$. Then using matrix perturbation theory, we further localize the overall error bound to a single network layer and apply the Davis-Kahan theorem to upper bound the estimation error of the common latent vectors U. Specifically, we first introduce the feasible parameter space as follows.

Definition 1 (Feasible parameter space). *For $n, R, k \in \mathbb{N}, \mu \in \mathbb{R}_+$, the feasible parameter space $\mathcal{F} = \mathcal{F}_{n,R,k}(\mu)$ is defined as*

$$\begin{aligned}\mathcal{F} &= \mathcal{F}_{n,R,k}(\mu)\\ &= \{\mathcal{T} = [\Theta^{(1)};\Theta^{(2)};\cdots;\Theta^{(R)}] \in \mathbb{R}^{n\times n\times R}:\\ &\quad \Theta^{(r)} = \alpha^{(r)}1_n^\top + 1_n\alpha^{(r)\top} + U\Lambda^{(r)}U^\top;\\ &\quad U\in\mathbb{R}^{n\times k}, U^\top U = nI_k, J_nU = U, \alpha^{(r)}\in\mathbb{R}^n,\\ &\quad \Lambda^{(r)}\in\mathbb{S}^{k\times k}, \|\Theta^{(r)}\|_{\max}\le\mu, r=1,2,\cdots,R\},\end{aligned} \tag{3}$$

where $J_n = I_n - \frac{1}{n}1_n1_n^\top$, $\mathbb{S}^{k\times k}$ includes all symmetric $k\times k$ matrices, and $\|\cdot\|_{\max}$ represents the maximum absolute value of entries in a matrix.

Suppose the estimator $\widehat{\mathcal{T}}$ is obtained by

$$\widehat{\mathcal{T}} = \arg\min_{\mathcal{T}\in\mathcal{F}} L(\mathcal{T}), \tag{4}$$

where L is defined in (2). Theorem 1 provides the result on the error bound for $\widehat{\mathcal{T}}$.

Theorem 1 *Given the true parameters $\mathcal{T}_\star \in \mathcal{F}$, there exist absolute constants c_1, c_2, such that with probability at least $1 - R\exp(-c_1 n) - \exp(-c_2(2n+R))$, we have*

$$\|\widehat{\mathcal{T}} - \mathcal{T}_\star\|_F^2 \le C_1 nR + C_2 k^2(2n+R), \tag{5}$$

where C_1 and C_2 only depend on μ.

The term C_1nR in (5) is induced by the layer specific parameters $\{\alpha^{(r)}\}_{r=1}^R$, and it grows linearly in the number of layers. The second term $C_2k^2(2n+R)$ is induced by $\{U\Lambda^{(r)}U^\top\}_{r=1}^R$, and due to the common latent variables U among layers, the order of this term would not grow as fast as the first term as R grows. In the next theorem, we specify the relationship between the upper bound on the estimation error of U and the number of layers.

Theorem 2 *Denote $\{\alpha_\star^{(r)}\}_{r=1}^R$, $\{\Lambda_\star^{(r)}\}_{r=1}^R$, and $U_\star$ as the true parameters that form $\mathcal{T}_\star \in \mathcal{F}$. Assume that $\Lambda_\star^{(1)}, \Lambda_\star^{(2)}, \cdots, \Lambda_\star^{(R)}$ are of full rank, i.e.*

$$\sigma_{\min}(\Lambda_\star^{(r)}) \ge \kappa \quad r = 1,2,\cdots,R \tag{6}$$

for some constant $\kappa > 0$. Assume there exists a constant $\delta > 0$ such that $R \le \delta n$, then with probability at least $1 - R\exp(-c_1 n) - \exp(-c_2(2n+R))$, we have

$$\min_{O:O^\top O = OO^\top = I_k} \left\{\|\widehat{U} - U_\star O\|_F^2\right\} \le 8\kappa^{-2}(C_1 + \widetilde{C}_2 k^2 R^{-1}), \tag{7}$$

where $\widetilde{C}_2 = C_2(2+\delta)$ and c_1, c_2, C_1, C_2 are the same constants as in Theorem 1.

Theorem 2 demonstrates that the upper bound of the estimation error of U decreases inversely as the number of layers R grows. The constant term C_1 is induced from the layer-specific terms $\{\alpha^{(r)}\}_{r=1}^R$. In Section 5, we will numerically further demonstrate that under the regime $R = \mathcal{O}(n)$, the upper bound in (7) is inversely proportional to R.

Remark 1 *Model (1) contains several tensor factorization models as special cases. For example, when $\alpha^{(r)} = 0$ for all $r = 1,\ldots,R$, it reduces to the logistic RESCAL model (Nickel & Tresp, 2013), for which theoretical properties were formerly unknown. The corollary below provides estimation property for latent factors under the logistic RESCAL model.*

Corollary 1 *Assume $\alpha^{(r)} = 0_n$, $r = 1,\ldots,R$ for all $\mathcal{T} \in \mathcal{F}$. Under the same assumptions as in Theorem 2, as $n \to \infty$, with probability going to 1 we have*

$$\min_{O:O^\top O = OO^\top = I_k} \left\{\|\widehat{U} - U_\star O\|_F^2\right\} \le C\kappa^{-2}k^2R^{-1}$$

for a constant C that only depends on μ. In other words, the upper bound of the estimation error of U decreases at the rate of $\mathcal{O}(R^{-1})$.

Proofs of Theorem 1, Theorem 2 and Corollary 1 are all provided in the Supplementary Materials.

5. Simulation Studies

In this section, we investigate empirical performance of the proposed method by simulation studies. Specifically, we

Algorithm 1 Projected Gradient Descent Algorithm for Parameter Estimation

Input: $A \in \mathbb{R}^{n\times n\times R}$; latent space dimension $k \geq 1$; initial estimates: $U_0, \{\alpha_0^{(r)}\}_{r=1}^R, \{\Lambda_0^{(r)}\}_{r=1}^R$; step sizes $\eta_u, \eta_\alpha, \eta_\lambda$; number of iterations T

Parameters:: $U, \{\alpha^{(r)}\}_{r=1}^R, \{\Lambda^{(r)}\}_{r=1}^R$

For $t = 0, 1, \ldots, T-1$

$U_{t+1} = U_t - \eta_u \nabla_U L = U_t + 2\eta_u \sum_{r=1}^R \left(A^{(r)} - \sigma(\Theta_t^{(r)})\right) U_t \Lambda^{(r)}$

$\alpha_{t+1}^{(r)} = \alpha_t^{(r)} - \eta_\alpha \nabla_{\alpha^{(r)}} L = \alpha_t^{(r)} + 2\eta_\alpha \left(A^{(r)} - \sigma(\Theta_t^{(r)})\right) 1_n, r = 1, \ldots, R$

$\Lambda_{t+1}^{(r)} = \Lambda_t^{(r)} - \eta_\lambda \nabla_{\Lambda^{(r)}} L = \Lambda_t^{(r)} + \eta_\lambda U_t^\top \left(A^{(r)} - \sigma(\Theta_t^{(r)})\right) U_t, r = 1, \ldots, R$

$U_{t+1} = J_n U_{t+1}, U_{t+1} = U_{t+1} W$ for $W \in \mathbb{R}^{k\times k}$ s.t. $U_{t+1}^\top U_{t+1} = nI_k$, $\Lambda_{t+1}^{(r)} = (W^{-1})^\top \Lambda_{t+1}^{(r)} W^{-1}$

Output: $\widehat{U} = U_T, \widehat{\alpha}^{(r)} = \alpha_T^{(r)}, \widehat{\Lambda}^{(r)} = \Lambda_T^{(r)}, r = 1, \ldots, R$

examine the estimation error of parameters with growing number of network layers. We also analyze the computational complexity of Algorithm 1.

We first study the relationship between the estimation error of the maximum likelihood estimators and the number of network layers. We set the true parameter values as follows.

- Generate $(U_\star)_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots k$; transform $U_\star$ by 1) centering $U_\star$ s.t. $J_n U_\star = U_\star$, 2) rotating $U_\star$ s.t. $U_\star^\top U_\star \propto I_k$, and 3) scaling $U_\star$ s.t. $U_\star^\top U_\star = nI_k$.
- Generate $(\alpha_\star^{(r)})_i \overset{iid}{\sim} \text{Uniform}(-2, -1)$ for $i = 1, \ldots, n$ and $r = 1, \ldots, R$.
- Generate $\Lambda_\star^{(r)} = \text{diag}(\lambda_{\star,1}^{(r)}, \cdots, \lambda_{\star,k}^{(r)})$ for $r = 1, \ldots, R$, where $\lambda_{\star,i}^{(r)} \overset{iid}{\sim} \text{Uniform}(-1, -0.5)$.

Note that though $\Lambda_\star^{(r)}$'s are set to be diagonal, we do not require $\widehat{\Lambda}^{(r)}$'s to be diagonal when fitting the model.

We set $n = 400$, $R = 100$, and $k = 2$. More simulation results with $(n, R, k) = (200, 50, 2)$ and $(400, 100, 4)$ are provided in the Supplementary Materials. We generate 30 independent copies of the adjacency tensor A based on model (1). The first R_0 out of R layers are used to fit the model. We examine how the estimation errors of $\widehat{U}$ and $\{\widehat{\Theta}^{(r)}\}_{r=1}^{R_0}$ change with R_0. The estimation error of $\{\widehat{\Theta}^{(r)}\}_{r=1}^{R_0}$ is evaluated by the relative error

$$\left(\sum_{r=1}^{R_0} \|\widehat{\Theta}^{(r)} - \Theta_\star^{(r)}\|_F^2\right) \Big/ \left(\sum_{r=1}^{R_0} \|\Theta_\star^{(r)}\|_F^2\right), \quad (8)$$

and the estimation error of $\widehat{U}$ is evaluated by

$$\min_{O: O^\top O = OO^\top = I_k} \left\{\|\widehat{U} - U_\star O\|_F^2\right\} / \|U_\star\|_F^2. \quad (9)$$

Finding the optimal O in (9) is known as the orthogonal Procrustes problem (Schönemann, 1966), which can be solved by singular value decomposition (SVD). In particular, denote the SVD of $\widehat{U}^\top U_\star$ be $S\Sigma V^\top$, then the optimal O is given by $VS^\top$.

Note Algorithm 1 requires initialization of U_0, $\{\alpha_0^{(r)}\}_{r=1}^{R_0}$ and $\{\Lambda_0^{(r)}\}_{r=1}^{R_0}$. When fitting the model, we initialize U_0 by first generating i.i.d. $\mathcal{N}(0,1)$ entries and then transforming it such that U_0 satisfies the identifiability condition. We initialize $\alpha_0^{(r)}$ as 0_n, i.e. the vector with all zeros, and we initialize $\Lambda_0^{(r)}$ as $\text{diag}(-1, \ldots, -1)$. The step sizes $\eta_\alpha, \eta_\lambda$ are chosen to be small and fixed, and η_u is proportional to R_0^{-1}.

Figure 1(a) shows the estimation error of $\{\widehat{\Theta}^{(r)}\}_{r=1}^{R_0}$ given in (8) versus R_0. We can see as the number of layers grows, the relative error of $\{\widehat{\Theta}^{(r)}\}_{r=1}^{R_0}$ decreases and is bounded below. By Theorem 1, we have

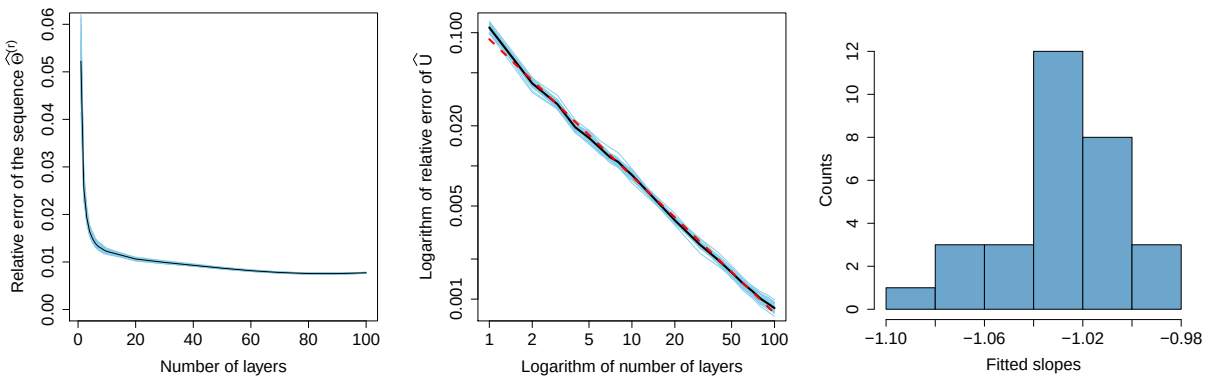
$$\sum_{r=1}^{R_0} \|\widehat{\Theta}^{(r)} - \Theta_\star^{(r)}\|_F^2 \leq C_1 nR_0 + C_2 k^2(2n + R_0) \quad (10)$$

with high probability. Also note that $\left(\sum_{r=1}^{R_0} \|\Theta_\star^{(r)}\|_F^2\right)$ is of order $\mathcal{O}(n^2 R_0)$ due to the constraints we put on the parameter space $\mathcal{F}$. Therefore, theoretically the bound of the relative error defined in (8) should be of order $\mathcal{O}\left(n^{-1} + n^{-1}R_0^{-1} + n^{-2}\right)$. For a fixed n, as R_0 increases, this term would decrease to some bound that depends on n^{-1}. The result in Figure 1(a) is then understandable as the "irreducible" estimation error of $\widehat{\Theta}^{(r)}$ comes from the first term in (10), i.e., the estimation error of layer-specific parameters $\widehat{\alpha}^{(r)}$, which does not decrease as the number of layers grows.

Figure 1(b) displays in log-log scale the estimation error of $\widehat{U}$ given by (9) against the number of network layers utilized for model fitting. For each replication, we fit a linear model to the result, i.e.,

$$\log(\text{relative error of } \widehat{U}) = a + b\log(R_0) + \epsilon.$$

Figure 1(c) shows the histogram of the fitted slopes. Note that all fitted slopes are close to -1, with the mean and standard deviation being -1.03 and 0.02 respectively. This agrees with the result in Theorem 2.



(a) Estimation error of $\{\widehat{\Theta}^{(r)}\}_{r=1}^{R_0}$ (b) Estimation error of $\widehat{U}$ (c) Histogram of fitted slopes

Figure 1. (a) and (b): Estimation error of parameters when $n = 400$, $R = 100$ and $k = 2$. Each light blue curve corresponds to one replication; the black curve corresponds to the average of all replications. The red dashed line in (b) corresponds to the line whose intercept and slope equal to the average of fitted intercepts and slopes respectively. (c): Histogram of all fitted slopes.

Since Algorithm 1 is a first-order method and aggregates gradient information of each layer in a linear manner, the running time should be proportional to R. Further, updating $\Theta_t^{(r)}$ in each iteration requires $\mathcal{O}(n^2k)$ operations. Therefore the computational complexity of Algorithm 1 is $\mathcal{O}(n^2Rk)$. To verify this, we examine the computing time under two settings: 1) fixing n, increasing R, and 2) fixing R, increasing n. As shown in Figure 2, the running time per iteration is linear in R and quadratic in n.

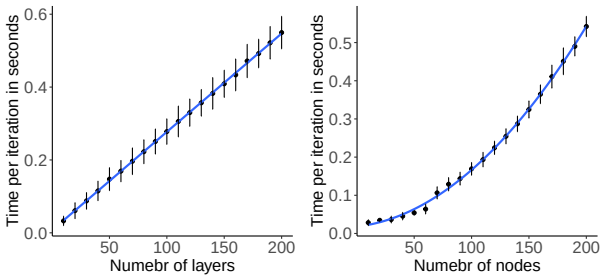


Figure 2. Average running time per iteration in seconds with one-standard-deviation error bars. Left: $n = 200$, and the R-square of a linear model is 0.998; Right: $R = 200$, and the R-square of a quadratic model is 0.998.

6. Real Data Applications

We apply the proposed model to two real-world examples. In practice, since there is no true value for the latent vectors, we can't evaluate the performance in terms of latent representation estimation. Instead, we consider two alternative approaches. First, though the latent vectors are not observed, there are usually observed node features which may be correlated with the latent vectors. Therefore, investigating the estimated latent representations against observed node features may provide insights on the estimation of latent vectors. Secondly, the estimated latent representations can often be used for downstream tasks, such as nodes classification, node clustering or link prediction. To examine the latent vector estimation, we demonstrate the performance of the proposed method on link prediction.

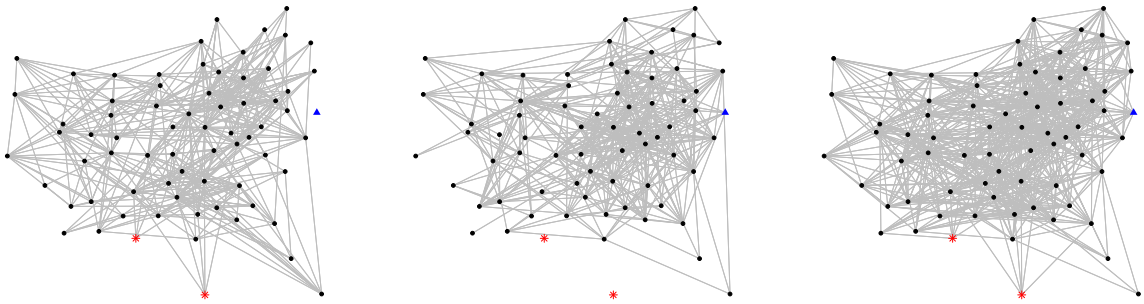
6.1. Lazega Lawyers Data

The Lazega Lawyers dataset records multiple connection relationships in a Northeastern US corporate law firm (Lazega et al., 2001). There are three types of networks between 71 lawyers, which are their co-worker network, advice network, and friendship network (Figure 3). The original network can be directed, for example, advice is often given in single direction and one nominates the other as a friend. We convert the direct networks to indirect ones by removing the directions. Besides the network relationships, multiple features of individuals are also recorded, including seniority, office, gender, law school attended, etc.

We fit both the multilayer version and single layer version of the proposed model (1). Initialization and stepsize choices are similar to what we do in Section 5. For comparison, we also fit model (1) without the layer-specific node individual effect terms, which reduces to the logistic RESCAL (L-RESCAL) model, as well as the COSIE model (Arroyo et al., 2019), which utilizes the random dot-product model for multilayer networks and assumes that

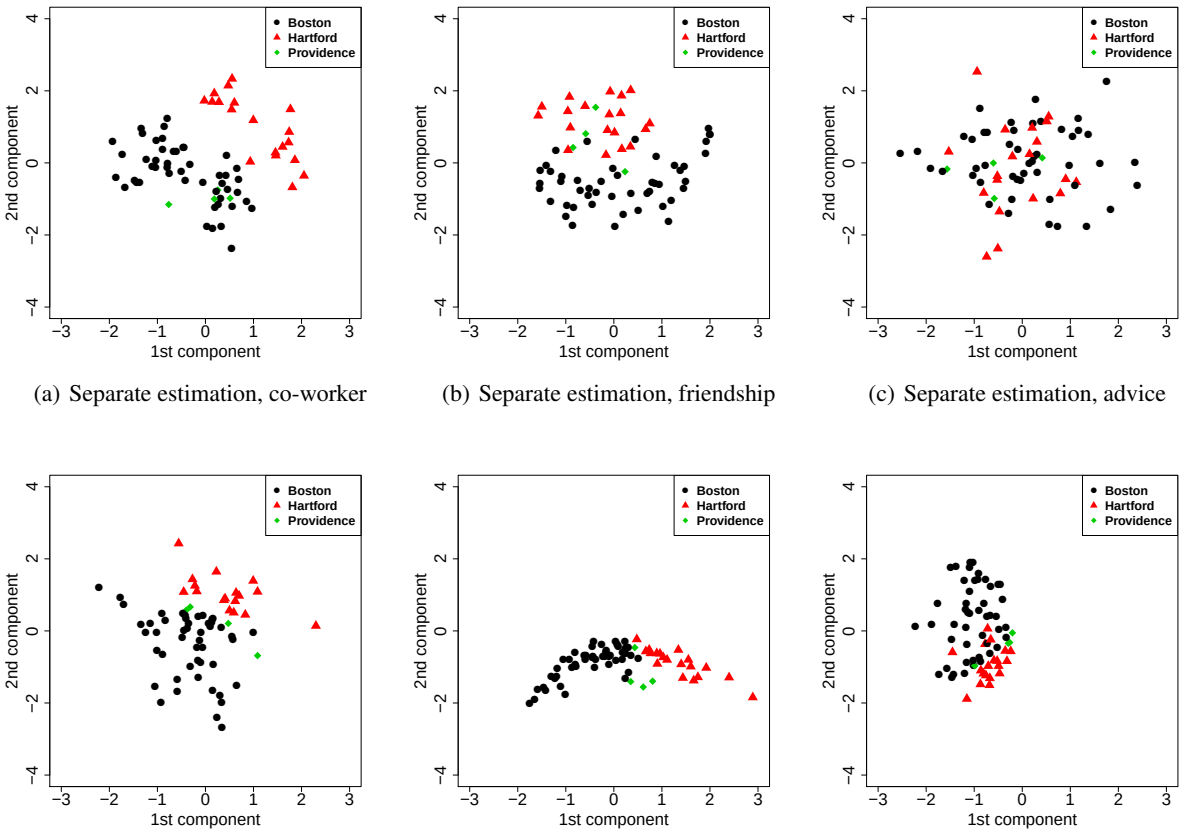
$$\mathbb{E}[A^{(r)}] = U\Lambda^{(r)}U^\top, \quad r = 1, \ldots, R.$$

We choose the dimension of the latent space to be $k = 2$ for the purpose of visualization. Figure 4 shows the estimated U from each model, and the colors are based on the



(a) Co-worker network (b) Friendship network (c) Advice network

Figure 3. Visualization of the Lazega lawyer data. Nodes in different layers exhibit different connecting patterns. For example, the two red nodes are isolated in the friendship network but have several links in other layers, while the blue node is not connected to other nodes in the co-worker network but is well-connected in the friendship and advice networks.



(a) Separate estimation, co-worker (b) Separate estimation, friendship (c) Separate estimation, advice

(d) Joint estimation, proposed model (e) Joint estimation, L-RESCAL model (f) Joint estimation, COSIE

Figure 4. Upper row: estimated U based on single networks. Lower row: jointly estimated U based on multilayer networks using different methods. Color represents the lawyer's office.

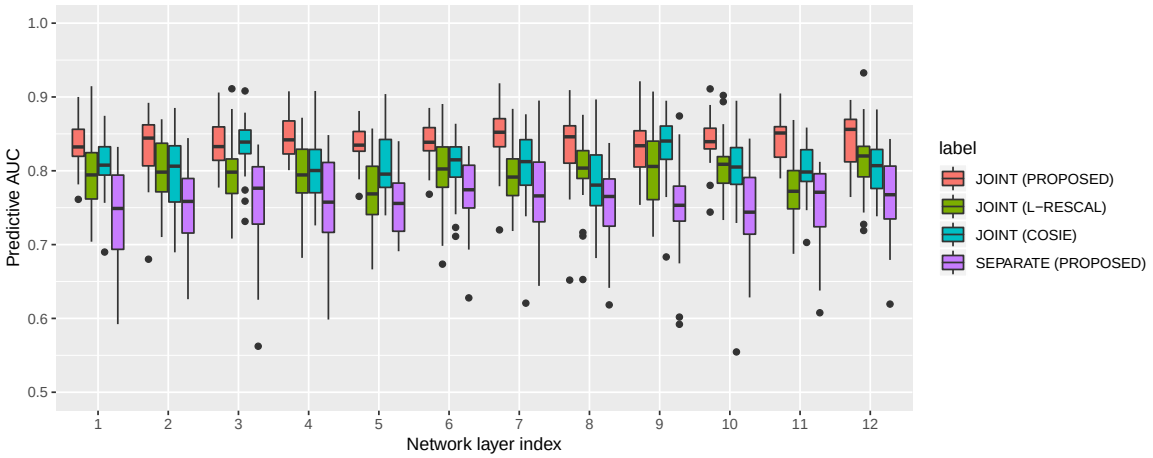


Figure 5. Link Prediction: AuROC on the test sets for 12 layers of the Karnataka Data.

lawyers' three offices: Boston, Hartford and Providence. From Figure 4, we can see more clear separation of people from different offices in the latent space based on the estimation from multilayer networks, in comparison to the estimation using single networks. Moreover, comparing the proposed method to the other two multilayer network models which do not incorporate node individual effects, the proposed method shows the most clear separation in terms of the lawyers' offices, which, among the available node features, is presumably the most correlated one with all three types of networks.

6.2. Karnataka Data

In practice, estimation of the latent space model is often an initial step, and the estimated model can be further used for downstream tasks on networks, for example, link prediction. Suppose we are interested in predicting missing links in a target network. With multilayer networks, we investigate whether connections in other layers would assist in the prediction of links in the target layer, due to the correlation in network structures between different layers.

Banerjee et al. (2013) provided multiple social networks in villages in rural southern Karnataka, India. Within each village, 12 types of social relations are recorded, including borrow money from, give advice to, help with a decision, borrow kerosene or rice from, lend kerosene or rice to, lend money to, obtain medical advice from, engage socially with, are related to, go to temple with, invite to one's home, and visit in another's home. Some of the relations are directed, and as in Section 6.1, the directed networks are converted to indirected ones based on the existence of any single directional edge between nodes. The networks are collected at both individual level and household level. We analyzed the data on the household level and selected one representative village with 99 nodes. For each type of relation, we randomly remove 20% entries of the adjacency matrix as missing. For comparison, we fit the single layer version of the proposed model, the multilayer latent space model with and without layer-specific node individual effect terms, and the COSIE model using the observed entries. Then we predict link probabilities on those missing entries using the fitted parameters and node latent representations. The dimension of the latent space is set to $k = 3$ for all methods. The experiments are replicated 30 times and we report AuROC for link prediction in Figure 5. As we can see, for each type of relation, using information from multiple networks has superior performances than using the layer itself only, which demonstrates that different layers share common structures and leveraging such information is beneficial. Moreover, the proposed method achieves the best performance in most layers, in comparison to the methods which do not take layer-specific node degree heterogeneity into account. This further supports the observed phenomenon that individual node behavior can vary from relation to relation, and modeling such flexibility is critical for capturing real-world network characteristics.

7. Conclusion and Discussion

In this paper, we have proposed a flexible and interpretable latent space model for multilayer networks. The proposed model is able to capture the common structure shared across different networks and meanwhile allows for heterogeneous layer-specific node connecting patterns. We have developed an efficient algorithm for parameter estimation. Moreover, theoretical guarantees on maximum likelihood estimators, in particular, improvements in the estimation of shared latent representations, are established. We have also demonstrated the proposed model on real-world data examples.

This work can be extended in several potential directions. Real-world networks are heterogeneous in the sense of not only multiple edge types, but also various node types (Sun & Han, 2013; Huang et al., 2018; Yu et al., 2018; Zhang & Chen, 2020; Zitnik et al., 2018). One interesting direction is to extend the proposed modeling framework to networks with both multiple edge types and multiple node types, with each node type embedded into a unique latent space.

Acknowledgements

This research was partially supported by NSF grant DMS-1821243.

References

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. Inference for multiple heterogeneous networks with a common invariant subspace. *arXiv preprint arXiv:1906.10026*, 2019.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. The Diffusion of Microfinance, 2013. URL https://doi.org/10.7910/DVN/U3BIHX.

D'Angelo, S., Alfò, M., and Murphy, T. B. Node-specific effects in latent space modelling of multidimensional networks. *arXiv preprint arXiv:1807.03874*, 2018.

De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, 2017.

De Lathauwer, L., De Moor, B., and Vandewalle, J. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

DAngelo, S., Murphy, T. B., Alfò, M., et al. Latent space modelling of multidimensional networks with application to the exchange of votes in eurovision song contest. *The Annals of Applied Statistics*, 13(2):900–930, 2019.

Ghadermarzy, N., Plan, Y., and Yilmaz, O. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2018.

Gollini, I. and Murphy, T. B. Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics*, 25(1):246–265, 2016.

Gupta, S., Sharma, G., and Dukkipati, A. Evolving latent space model for dynamic networks. *arXiv:1802.03725*, 2018.

Han, Q., Xu, K., and Airoldi, E. Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*, pp. 1511–1520, 2015.

Hoff, P. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pp. 657–664, 2008.

Hoff, P. D. *Random effects models for network data*. Technical report, 2003.

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

Huang, W., Liu, Y., and Chen, Y. Mixed membership stochastic blockmodels for heterogeneous networks. *Bayesian Analysis*, 2018.

Kolaczyk, E. D. and Csárdi, G. *Statistical Analysis of Network Data with R*, volume 65. Springer, 2014.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Lazega, E. et al. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand, 2001.

Levin, K., Athreya, A., Tang, M., Lyzinski, V., Park, Y., and Priebe, C. E. A central limit theorem for an omnibus embedding of random dot product graphs. *arXiv preprint arXiv:1705.09355*, 2017.

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

Newman, M. *Networks: An Introduction*. OUP Oxford, 2010.

Nickel, M. and Tresp, V. Logistic tensor factorization for multi-relational data. *arXiv preprint arXiv:1306.2084*, 2013.

Nickel, M., Tresp, V., and Kriegel, H.-P. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 809–816. Omnipress, 2011.

Nielsen, A. M. and Witten, D. The multiple random dot product graph model. *arXiv preprint arXiv:1811.12172*, 2018.

Paul, S. and Chen, Y. Community detection in multi-relational data with restricted multi-layer stochastic blockmodel. *arXiv preprint arXiv:1506.02699*, 2015.

Paul, S. and Chen, Y. Spectral and matrix factorization methods for consistent community detection in multilayer networks. *Annals of Statistics, in press*, 2020.

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

Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Sewell, D. K. and Chen, Y. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

Sewell, D. K. and Chen, Y. Latent space approaches to community detection in dynamic networks. *Bayesian Analysis*, 12(2):351–377, 2017.

Simpson, S. L., Bahrami, M., and Laurienti, P. J. A mixed-modeling framework for analyzing multitask whole-brain network data. *Network Neuroscience*, 3(2):307–324, 2019.

Sun, Y. and Han, J. Mining heterogeneous information networks: a structural analysis approach. *Acm Sigkdd Explorations Newsletter*, 14(2):20–28, 2013.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Wang, M. and Li, L. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*, 2018.

Wang, M. and Song, Y. Tensor decompositions via two-mode higher-order svd (hosvd). In *Artificial Intelligence and Statistics*, pp. 614–622, 2017.

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. Joint embedding of graphs. *arXiv preprint arXiv:1703.03862*, 2017.

Wilson, J. D., Cranmer, S., and Lu, Z.-L. A hierarchical latent space network model for population studies of functional connectivity. *Computational Brain & Behavior*, pp. 1–16, 2020.

Yu, L., Woodall, W. H., and Tsui, K.-L. Detecting node propensity changes in the dynamic degree corrected stochastic block model. *Social Networks*, 54:209–227, 2018.

Zhang, J. and Chen, Y. Modularity based community detection in heterogeneous networks. *Statistica Sinica, in press*, 2020.

Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):457466, 2018.