MULTI-LEVEL STOCHASTIC BLOCKMODELS FOR MULTIPLEX NETWORKS

Maria Francesca Marino [1], Matteo Sani [1] and Monia Lupparelli [1]

[1] Department of Statistics, Computer Science, Applications "G.Parenti", University of Florence, (e-mail: mariafrancesca.marino@unifi.it, monia.lupparelli@unifi.it, matteo.sani@stud.unifi.it)

ABSTRACT: Multiplex arises when the network for the same set of nodes is repetitively observed on different layers that can represent, for instance, different statistical units or different criteria to connect the nodes. A multi-level Stochastic Blockmodel for multiplexes is introduced to provide a joint clustering of layers and nodes. This is achieved by considering two different sets of discrete latent variables. A former set allows us identifying groups of layers sharing similar connectivity patterns. A letter set of discrete latent variables, nested within the former, allows us identifying groups of nodes sharing similar relational features. A variational Expectation-Maximization algorithm is derived for estimation purposes.

KEYWORDS: network data, model-based clustering, finite mixtures, EM algorithm, variational inference.

1 Introduction

Uncover patterns underlying relations between nodes of a network is a complex task, especially when the network is repeatedly observed on a number of statistical units, or when different criteria to connect the nodes are available. For instance, connections between brain regions may be observed on a number of individuals, or imports/exports between countries may entail different types of products. In such cases, data provide a multilevel structure and multiplexes can be effectively used to describe, analyze, and model interactions between nodes (Barbillon *et al.*, 2017).

Stochastic blockmodels (SBMs - Daudin *et al.*, 2008) represent a valuable approach for identifying clusters of nodes sharing common relational features. These are identified by including in the model specification a set of node-specific, discrete, latent variables inducing nodes' partitioning. When multiplexes are available, one can decide to apply a SBM to each layer of the data structure, thus obtaining a separate clustering of nodes for each layer. As an alternative, the multivariate nature of dyadic relations may be properly taken into consideration and nodes' clustering may be defined by fully exploiting the richness of the data at hand (Barbillon *et al.*, 2017).

We introduce a specification of the SBM for multiplexes that allows us to obtain a clustering of both layers and nodes. In detail, we introduce a multilevel SBM where layer-specific, discrete, latent variables allow us to cluster layers (i.e., the statistical units) sharing similar connectivity patterns. Within each of such clusters, nodes characterized by similar relational features are clustered by means of a further set of node-specific, discrete, latent variables. As typical of SBMs, Maximum Likelihood (ML) parameter estimates cannot be computed due to the intractability of the likelihood function. This makes infeasible the use of an Expectation-Maximization (EM - Dempster *et al.*, 1977) algorithm, as the posterior distribution of the random variables to compute at the E-step of the algorithm still requires the derivation of the likelihood function. To overcome the issue, we employ an extended variational EM algorithm, where the true, intractable, posterior distributions are substituted by their approximate versions, having a tractable form; see e.g., Blei *et al.*, 2017 for a thorough treatment of the topic.

2 Model definition

Let $\mathcal{G} = \{\mathcal{G}^k\}_{k\in(1,\dots,K)}$ denote a multiplex characterized by K layers. Each graph $\mathcal{G}^k = (\mathcal{N}, \mathcal{E}^k) \in \mathcal{G}$ is defined by the same node set $\mathcal{N} = \{1,\dots,n\}$ and the layer-specific edge set $\mathcal{E}^k$, with $k = 1,\dots,K$. Equivalently, the multiplex $\mathcal{G}$ may be defined in terms of the *adjacency array* $\mathcal{Y} = \{Y^k\}_{k\in(1,\dots,K)}$, with Y^k being the adjacency matrix associated to the k-th layer. Its generic element is

$$Y_{ij}^k = \begin{cases} 1 & \text{if the pair } (i,j) \in \mathcal{E}^k, \\ 0 & \text{else.} \end{cases}$$

That is, $Y_{ij}^k = 1$ iff nodes i and j are joined by an edge in the network associated to the k-th layer. For simplicity, we focus on the case of undirected networks, even though the extension to the directed case is straightforward.

Let $\{U_k\}_{k=(1,\dots,K)}$ denote layer-specific, independent and identically distributed, latent variables defined over the support $\{1,\dots,s\}$ and let $\eta_v = \Pr(U_k = v)$, for all $k \in 1,\dots,K$. Furthermore, let $Z_i^k, i = 1,\dots,n$, be a node-level latent variable, nested with respect to $U_k, k = 1,\dots,K$, defined over the support $\{1,\dots,m\}$ and let $\alpha_{qv} = \Pr(Z_i^k = q \mid U_k = v)$.

We assume that, conditional on the latent variables U_k, Z_i^k, and Z_j^k, the random variables Y_{ij}^k are independent each other and follow a Bernoulli distribution with tie probability only depending on the block membership of layers and nodes involved in the relation. That is,

$$Y_{ij}^k \mid Z_i^k = q, Z_j^k = l, U_k = v \overset{iid}{\sim} \mathcal{B}e(\pi_{qlv}).$$

Based on the above assumptions and denoting with θ the set of all free model parameters, the log-likelihood function can be written as

$$\begin{aligned} \ell(\theta) &= \log p(y) = \log \sum_u \sum_z p(y \mid u, z) p(z \mid u) p(u) \qquad (1) \\ &= \log \sum_u \sum_z \left\{ \left[\prod_{k=1}^K \prod_{i=1}^n \prod_{j>i} \mathcal{B}e(\pi_{z_i^k, z_j^k, u_k}) \right] \left[\prod_{k=1}^K \prod_{i=1}^n \alpha_{z_i^k, u_k} \right] \left[\prod_{k=1}^K \eta_{u_k} \right] \right\}, \end{aligned}$$

where y is a realization of $\mathcal{Y}$, and $\sum_u$ and $\sum_z$ are shorthands for $\sum_{u_1} \dots \sum_{u_K}$ and $\sum_{z_1^1} \sum_{z_1^2} \dots \sum_{z_n^{K-1}} \sum_{z_n^K}$, respectively.

As evident, deriving parameter estimates by either a direct or an indirect maximization of equation (1) is impractical. Indeed, this would require the computation of multiple summations, which is infeasible from a computational standpoint, even for networks of very limited size. To overcome the issue, an EM algorithm based on a variational approximation of the likelihood function may be employed as an effective alternative, as detailed in the following section.

3 Parameter estimation and inference

To derive parameter estimates, we extend the variational approach firstly introduced by Daudin *et al.*, 2008 in the SBM framework. Accordingly, starting from the likelihood function detailed in equation (1), estimates are derived by maximizing the following lower bound

$$\mathcal{F}(q(z,y), \theta) = \ell(\theta) - KL[q(z,u) \mid\mid p(z,u \mid \mathcal{Y}, \theta)], \qquad (2)$$

where $KL[\cdot \mid\mid \cdot]$ denotes the Kullback-Leibler divergence between the true, intractable, posterior distribution of the latent variables $p(z,u \mid y)$ and the corresponding approximating function $q(z,u)$. As we are not able to let KL vanish due to intractability of the likelihood, we look for the best approximation $q(z,u)$ in the class of completely factorized distributions

$$q(z,u) = q(u)q(z) = \prod_{k=1}^K \mathcal{M}ult(1, \tau_k) \prod_{i=1}^n \mathcal{M}ult(1, \phi_i).$$

The variational EM (VEM) algorithm alternates between two separate steps until convergence: (*i*) a VE-step, in which we maximize equation (2) with respect to the variational parameters τ_k and ϕ_i; (*i*) a VM-step, in which maximize (2) with respect to model parameters θ. Different works in the literature show the effectiveness of the variational approach in recovering the true value of model parameters in θ both with finite samples (see e.g., Mariadassou *et al.*, 2010) and asymptotically (see e.g., Celisse & Pierre, 2012).

To select the optimal number of blocks s and m, we may rely on an Integrated Classification Likelihood criterion (ICL - Biernacki *et al.*, 2000), as typically done in the SBM framework. Once the optimal model is selected, layer and node memberships are determined on the base of the parameter estimates $\hat{\tau}_k$ and $\hat{\phi}_i$, obtained at convergence of the estimation algorithm.

References

BARBILLON, PIERRE, DONNET, SOPHIE, LAZEGA, EMMANUEL, & BAR-HEN, AVNER. 2017. Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **180**, 295–314.

BIERNACKI, CHRISTOPHE, CELEUX, GILLES, & GOVAERT, GÉRARD. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725.

BLEI, DAVID M, KUCUKELBIR, ALP, & MCAULIFFE, JON D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, **112**, 859–877.

CELISSE, ALAIN, & PIERRE, LAURENT. 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, **6**, 1847–1899.

DAUDIN, J-J, PICARD, FRANCK, & ROBIN, STÉPHANE. 2008. A mixture model for random graphs. *Statistics and computing*, **18**(2), 173–183.

DEMPSTER, ARTHUR P, LAIRD, NAN M, & RUBIN, DONALD B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**, 1–22.

MARIADASSOU, MAHENDRA, ROBIN, STEPHANE, & VACHER, CORINNE. 2010. Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics*, **4**(2), 715–742.