Clustering in weighted networks using Binomial stochastic blockmodels

Abir El Haj [1,2] & Yousri Slaoui [1] & Pierre-Yves Louis [1] & Zaher Khraibani [2]

[1] *Laboratoire de Mathématiques et Applications, Université de Poitiers, 11 Boulevard Marie et Pierre Curie, Bâtiment H3 - TSA 61125, 86073 POITIERS CEDEX 9.*
abir.el.haj@math.univ-poitiers.fr, yousri.slaoui@math.univ-poitiers.fr, pierre-yves.louis@math.univ-poitiers.fr

[2] *Faculté de Sciences, Université Libanaise, Beyrouth, Liban.*
zaher.khraibani@gmail.com

Résumé. Le modèle à blocs stochastiques est un modèle de graphe aléatoire qui vise à partitionner les sommets d'un réseau en groupes appelés blocs, ou plus généralement clusters. Dans la plupart des réseaux du monde réel, les liens entre les noeuds sont affectés par des poids qui représentent la force des relations entre ces noeuds. Il est évidemment très intéressant de pouvoir modéliser et regrouper ces réseaux pondérés en utilisant la structure du réseau et la capacité de leurs liens. Cet article présente le modèle à blocs stochastiques binomial, qui est un modèle probabiliste pour les réseaux ayant les poids sur les arêtes distribués selon une loi binomiale. Un algorithme variationnel d'espérance-maximisation est proposé ici pour effectuer l'inférence. Enfin, nous démontrons l'efficacité de la méthode proposée en considérant un réseau de co-citation dans un contexte de text mining.

Mots-clés. Modèle à blocs stochastiques binomial, clustering, classes latentes, fouille de textes, inférence variationnelle, réseaux pondérés.

Abstract. The Stochastic Block Model is a random graph model which aims to partition the vertices of a network into groups called blocks, or more generally clusters. In most real-world networks, the ties among nodes have weights assigned to them which represent the strength of relationship between these nodes. It is obviously of strong interest to be able to model and cluster those weighted networks using the structure of network and the capacity of their ties. This paper introduces the binomial stochastic blockmodel, a probabilistic model for networks with binomial distributed edges weights. A variational expectation-maximization (`VEM`) algorithm is proposed here to perform inference. Finally, we demonstrate the effectiveness of the proposed method by considering a co-citation network in a text mining context.

Keywords. Binomial stochastic blockmodel, clustering, latent classes, text mining, variational inference, weighted networks.

1 Introduction

In this work, we are interested in the estimation of the parameters in a binomial stochastic block model for weighted networks as well as in detecting the community structure in these networks. The stochastic block model proposed by Anderson et al. (1992) and Holland et al. (1983) is a probabilistic random graph model which aims to produce clusters in networks. In this model, the nodes of the network are divided into disjoint blocks such that the nodes belonging to the same block have the same probability called inter-connection probability and the same connection probability with nodes belonging to others blocks. This probability is called intra-connection probability.

In most of the methods already treated in this context, the `SBM` is restricted to binary networks, in which edges are unweighted. Since the most of networks are weighted, we study here the case of weighted networks, where each edge is associated with an integer value representing the capacity of ties among nodes.

We proposed here the variational expectation maximization (`VEM`) algorithm developed by Daudin et al. (2008) and Jaakola (2000) which is an approximate method based on a variational approach. This approach is known to be consistent under the `SBM` model according to Celisse et al. (2012).

The proposed method allow us to estimate the parameters of the model for a fixed number of clusters in the network. We are interested in determining the one that will optimally fit the data. Daudin et al. (2008) proposed the Integrated Classification Likelihood (`ICL`) criterion to estimate the optimal number of clusters in a `SBM` model. This method is an approximation of the complete data likelihood.

2 The Model

A weighted undirected network is defined by its set of n nodes $[n] = \{1, \ldots, n\}$ for all $n \geq 1$ and by its edge-weighted symmetric matrix X of size n. Thus, $X_{ij} = m_{ij}$ if there is an edge joining the nodes i and j and is weighted by the value m_{ij} and 0 otherwise. We assume that the network has a fixed number of blocks equal to Q. The network is assumed to be generated as follows

- Each vertex i in the network belongs to an unobserved group q such as:

$$Z_i \sim \mathcal{M}(1, \alpha = (\alpha_1, \ldots, \alpha_Q)),$$

  where $Z = (Z_i)_{i \in \{1,\ldots,N\}}$ is a latent vector describing the belonging of the node i to cluster q when $Z_{iq} = 1$ and not when $Z_{iq} = 0$, $\mathcal{M}$ is the multinomial distribution and α is the vector of class proportion such as $\sum_q \alpha_q = 1$.

- Each observed edge X_{ij} joining node i, that belongs to group q, to node j , that belongs to group l, is sampled from a Binomial distribution such as:

$$X_{ij} | Z_{iq} Z_{jl} = 1 \sim \mathcal{B}(m, \pi_{ql}),$$

where the parameter m represents the maximal weight on the edges of the network and $\pi = (\pi_{ql})_{ql}$ represents the $Q \times Q$ matrix of connection probabilities between the latent groups of the network.

In the sequel, we are interested in the case of weighted undirected network without self-loops. However, we claim that all results can be extended to directed networks, with or without self-loops.

2.1 Variational Inference

The log-likelihood of the incomplete data can be obtained through the marginalization $\log \mathbb{P}_\theta(X) = \log \sum_Z \mathbb{P}_\theta(X, Z)$, where θ is the set of the parameters of the model $\theta = (\alpha, \pi)$ and $\mathbb{P}_\theta(X, Z)$ is the likelihood of the complete data defined by

$$\log \mathbb{P}_\theta(X, Z) = \sum_i \sum_q Z_{iq} \log(\alpha_q) + \sum_{i<j} \sum_{q,l} Z_{iq} Z_{jl} (\log C_m^{X_{ij}} + X_{ij} \log \pi_{ql} + (m - X_{ij}) \log(1 - \pi_{ql})).$$

This marginalization involves a summation over every possible matrix Z and thus may not be tractable except for small networks. So we use iterative method to tackle this issue. The expectation maximization (`EM`) algorithm is intractable in this context since the E-step require the computation of $\mathbb{P}_\theta(Z|X)$ which is impossible since the edges of the network are not independent. Following the work of Blei et al. (2003), we propose to rely on a variational decomposition. In the case of the `SBM` model, it leads to

$$\log \mathbb{P}_\theta(X) = J_\theta(R_X(.)) + \mathrm{KL}(R_X(.) \parallel \mathbb{P}_\theta(.|X)),$$

where $\mathbb{P}_\theta(Z|X)$ is the true conditional distribution of Z given Y, $R_X(Z)$ is an approximate distribution of $\mathbb{P}_\theta(Z|X)$, KL is the Kullback-Leibler divergence between $\mathbb{P}_\theta(Z|X)$ and $R_X(Z)$ defined by

$$\mathrm{KL}(R(.) \parallel \mathbb{P}_\theta(.|Z)) = -\sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(Z|X)}{R_X(Z)}$$

and $J_\theta(R_X(.))$ is the lower bound of the form

$$J_\theta(R_X(.)) \quad = \quad \sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(X, Z)}{R_X(Z)}.$$

Since $\log \mathbb{P}_\theta(X)$ does not depend on the distribution $R_X(.)$, maximizing the lower bound $J_\theta(R_X(.))$ with respect to $R_X(.)$ is equivalent to minimize the `KL` divergence. Following Blei et al. (2003), we assume that the distribution $R_X(Z)$ can be factorized over the latent variable Z as follows

$$R_X(Z) = \prod_{i=1}^n R_{X,i}(Z_i) = \prod_{i=1}^n h(Z_i; \tau_i),$$

where $\{\tau_i \in [0,1]^Q, i = 1, \ldots, n\}$ are the variational parameters associated with $\{Z_i, i = 1, \ldots, n\}$ such as $\sum_q \tau_{iq} = 1,\ \forall i \in \{1, \ldots, n\}$ and h is the multinomial distribution with parameters τ_i.

The comination of the equations above leads to

$$\begin{aligned} J_\theta(R_X) &= -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_{i<j} \sum_{q,l} \tau_{iq}\tau_{jl}(\log C_m^{X_{ij}} + X_{ij} \log \pi_{ql} \\ &\quad +(m - X_{ij}) \log(1 - \pi_{ql})). \end{aligned}$$

2.2 Optimization

In this section, we develop the steps of the `VEM` algorithm. These steps aims to estimate the parameters of the model by maximizing the lower bound $J_\theta(R_X)$. The `VEM` algorithm alternates between the optimization of τ and $\theta = (\alpha, \pi)$ until the convergence of the lower bound.

During the variational E-step, the parameters of the model are fixed. By maximizing the lower bound $J_\theta(R_X)$ with respect to τ, we obtain the estimate of τ by the following fixed point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_l \left(C_m^{X_{ij}} \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{m - X_{ij}} \right)^{\hat{\tau}_{jl}}. \tag{1}$$

The estimation of τ is obtained from (1) by iterating a fixed point algorithm until convergence.

Conversely, during the M-step, the parameter τ is fixed. By maximizing the lower bound $J_\theta(R_X)$ with respect to α and under the condition $\sum_q \alpha_q = 1$, we obtain the estimate of α_q

$$\hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}.$$

Moreover, by maximizing the lower bound $J_\theta(R_X)$ with respect to π, we obtain the estimate of π_{ql}

$$\hat{\pi}_{ql} = \frac{\sum_{i<j} \tau_{iq}\tau_{jl} X_{ij}}{m \sum_{i<j} \tau_{iq}\tau_{jl}}.$$

2.3 Model selection

In the sections above, we showed that the `SBM` model function requires number of latent groups Q as an input argument. We are interested here in estimating the optimal number of clusters $\hat{Q}$.

Daudin et al. (2008) proposed the Integrated Classification Likelihood (ICL) criterion to estimate Q in a `SBM` model. This method is an approximation of the complete data

likelihood. The ICL is of the form

$$\begin{aligned} ICL(Q) &= \sum_i \sum_q \hat{\tau}_{iq} \log \hat{\alpha}_q + \sum_{i<j} \sum_{q,l} \hat{\tau}_{iq}\hat{\tau}_{jl}(\log C_m^{X_{ij}} + X_{ij} \log \hat{\pi}_{ql} + (m - X_{ij}) \log(1 - \hat{\pi}_{ql})) \\ &\quad -\frac{1}{2}\left(\frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - (Q-1)\log n\right). \end{aligned}$$

The `VEM` algorithm is run for different values of Q and $\hat{Q}$ is chosen such that ICL is maximized.

3 Numerical experiments

The Reuters-21578 data set contains a collection of documents that appeared on Reuters newswire in 1987. For more explanation about this data, we refer the reader to (Lewis (1997)). We are interested in this example in 20 exemplary news articles from the Reuters-21578 data set of topic crude. The data is available in the package `tm` of the software `R` under the name of `crude` data where all documents belong to the topic crude dealing with crude oil see (Feinerer el al. (2008)). We build a term-by-document matrix of the corpus `crude` by doing a text mining treatment. We interpret a term as important according to a simple counting of frequencies, we chose the frequent terms that co-occur at least six times in the documents. Then, we compute the correlations between them in the term-by-document matrix and we chose those out higher than 0.5. The figure visualizing the correlation between these terms is available in (Feinerer el al. (2008)) .

We transform the term-by-document matrix into a one mode matrix which is the term-by-term matrix. The network associated to this matrix is an undirected network of 21 vertices and 97 edges, where each vertex is a term and there is an edge between a pair of terms if they co-occur together at least one time in the documents. The edge weights are represented in the obtained matrix where each cell indicates the number of documents where both the row and the column terms co-occur.

The graph associated with this network is visualized in Figure 1 using `Gephi` software with the layout algorithm Force Atlas.

We apply our algorithm. We obtain that the terms are grouped into four clusters as presented in Table 1. Table 1 shows the distribution of the network's terms into groups

Clusters	Vertices
1	oil opec prices
2	mln bpd month sources production saudi market
3	billion budget riyals government economics indonesia report
4	exchange nymex futures Kuwait

Table 1: Grouping the terms of the network of terms of the Reuters-21578 corpus into clusters using binomial `SBM`.

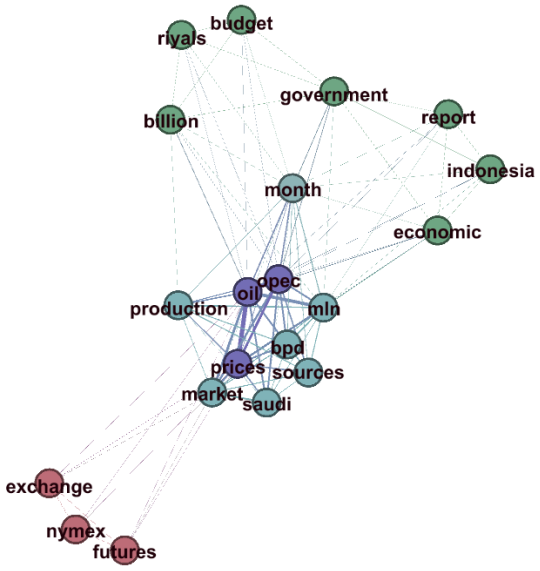

Figure 1: Network of terms of the the Reuters-21578 corpus visualization with `Gephi`.

which means that the terms of each group are frequently co-occuring together in the documents.

Bibliographie

Anderson, C. J., Wasserman, S., and Faust, K. (1992). Building stochastic blockmodels. *Social Networks*, 14, pp. 137–161.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation.*the Journal of machine Learning research*. 3, 993–1022.

Celisse, A., Daudin, J. J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6, 1847–1899.

Daudin, J., Picard, F., and Robin, S. (2008). A mixture model for random graph. *Statistics and Computing*, 18, 1–36.

Feinerer, I., Hornik, K. and Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25, 1–54.

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

Jaakkola. T. S. (2000). Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, 129–159.

Lewis, D. (1997). Reuters-21578 Text Categorization Collection Distribution 1.0. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html