Inférence du Modèle à Blocs Stochastiques en présence de données manquantes.

Timothée Tabouy [1] & Pierre Barbillon [1] & Julien Chiquet [1]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
prenom.nom@agroparistech.fr

Résumé. Le modèle à blocs stochastiques ou *Stochastic Block Model* (SBM) [8] est un modèle de graphe aléatoire généralisant le modèle d'Erdős-Reyni [4] à l'aide d'une structure latente sur les nœuds. L'utilisation de variables latentes dans le SBM permet de modéliser un large spectre de topologies de réseau, en particulier les graphes en affiliation, en étoile ou bipartite. L'inférence de ces modèles repose sur des modifications de l'algorithme EM (Expectation Maximization), comme par exemple l'approche EM variationnelle [1] ou Bayésienne variationnelle [7]. Dans ces approches, le réseau est toujours considéré comme parfaitement observé, alors que de nombreux cas d'application (en particulier en sociologie) suggèrent que son observation est partielle et guidée par une stratégie d'échantillonnage dépendant du réseau lui-même.

La motivation de ce travail vient du constat qu'un échantillonnage partiel du réseau peut induire un biais d'estimation dans le modèle SBM. Notre objectif est la modélisation de la stratégie d'échantillonnage utilisée et son intégration dans les procédure d'inférence. Dans cette optique, nous nous appuyons sur la théorie des données manquantes développée par D. Rubin [9] que nous adaptons au cadre du SBM. Nous proposons une typologie pour les stratégies d'échantillonnages dans le SBM pour lesquelles la prise en compte dans l'inférence varie. Les stratégies se regroupent essentiellement en deux classes : *i*) celles où la probabilité d'être échantillonné est indépendante de la valeur des données manquantes, dites "manquantes au hasard" (*Missing At Random* – MAR) et *ii*) leur contrepartie "non manquantes au hasard" (*Not Missing At Random* – NMAR). Dans le cas MAR, la stratégie d'échantillonnage ne perturbe pas l'inférence et il suffit de conduire l'inférence uniquement sur la partie observée du graphe. Au contraire, les stratégies NMAR nécessitent la prise en compte dans l'inférence de la stratégie d'échantillonnage employée pour récolter les données.

Pour toutes les stratégies MAR, nous avons adapté les algorithmes EM dans leur forme variationnelle pour l'inférence des paramètres du SBM binaire. Dans le cas NMAR, nous proposons une version stochastique de l'algorithme EM (SAEM) permettant de corriger les biais d'estimation. Nous présentons des simulations qui permettent de mettre en évidence la pertinence de ces approches.

Mots-clés. Modèle à blocs stochastiques · données manquantes · EM variationnel · EM stochastique

Abstract. The stochastic block model (SBM) [8] is a random graph model generalizing the Erdős-Reyni model [4] by means of a latent structure on the nodes. The use of latent variables in the SBM allows to model a broad variety of network topologies, in particular affiliation networks, star networks or bipartite networks. The inference of such models is based on modifications of the EM (Expectation Maximization) algorithm, such as the variational EM [1] or the variational Bayes algorithm [7]. In these approaches, the network is always considered to be perfectly observed, whereas many cases of application (particularly in sociology) suggest that its observation is partial and guided by a sampling strategy depending on the network itself.

The present work has been motivated by the fact that a partial sampling of the network may induce estimation biases in the SBM model. Our goal is to model the sampling strategy used in order to integrate this strategy in the inference procedure itself. In this perspective, we rely on the missing data theory developed by D. Rubin [9] that we adapt to the framework of the SBM. We propose a typology of the sampling strategies in the SBM, for which the integration in the inference strategy varies. The sampling strategies are grouped essentially in two classes: *i*) those where the probability of being sampled is independent of the value of the missing data (*Missing At Random* - MAR) and *ii*) their counterpart "Not missing at random" (*Not Missing At Random* - NMAR). In the MAR case, the sampling strategy does not disturb the inference and it suffices to conduct the inference only on the observed part of the graph. On the contrary, NMAR strategies require that the sampling strategy used to collect the data must be taken into account in the inference.

For all MAR strategies, we have adapted the EM algorithms in their variational form for the inference of the binary SBM. In the NMAR case, we propose a stochastic version of the EM algorithm (SAEM) to correct the estimation biases induced by the sampling. We present simulations to demonstrate the relevance of our approach.

Keywords. Stochastic block Model · missing data · variational EM · stochastic EM

1 Stochastic Block Model

The Stochastic Block Model (SBM) introduced in [8] is a mixture model for random graph [1] where the n vertices are distributed among a set $\mathcal{Q} = \{1, \dots, Q\}$ of hidden clusters that model the latent structure of the graph. The cluster of node i is described by the categorical variable $Z_i \in \mathcal{Q}$ with prior probabilities $\alpha = (\alpha_1, ..., \alpha_Q)$, so that $\mathbb{P}(Z_i = q) = \alpha_q$. Furthermore, the probability of an edge between any pair of nodes depends only on the clusters they belong to. Hence, let Y_{ij} be the binary variable which indicates the presence of an edge between i and j, then all $\{Y_{ij}, i, j = 1, \dots, n\}$ are independent conditionally on the latent clusters $\{Z_i, i = 1, \dots, n\}$. In other words,

$$Y_{ij} \mid Z_i = q, Z_j = \ell \sim^{\text{ind}} \mathcal{B}(\pi_{q\ell}), \quad \forall (i, j) \in \{1, \dots, n\}^2,$$

where $\mathcal{B}$ stands for the Bernoulli distribution. In the following, $Y = (Y_{ij})_{i,j=1,\dots,n}$ is the adjacency matrix of the random graph, $Z = (Z_1, \dots, Z_n)$ the vector of the latent clusters and $\theta = (\alpha, \pi)$ are the unknown parameters associated with the SBM.

2 SBM and missing data

Regarding SBM inference, a missing value corresponds to a missing entry in the adjacency matrix Y. We rely on the sampling matrix R to record the data sampled during this process:

$$(R_{ij}) = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The missing entries in Y – or equivalently the matrix R – results from the process that generates missing values, which we call *sampling design*. A sampling design is the description of a stochastic process that generates observed and missing data. It is fully characterized by the conditional distribution $h_\psi(R|Y)$. The parameters ψ associated with the sampling design (and then to R) is such that ψ and θ live in a product space $\Theta \times \Psi$.

We then follow the classical framework of [9] for missing data which, however, has to be adapted to the presence of latent variables: let $Y_{obs} = \{Y_{ij} : R_{ij} = 1\}$ and $Y_{mis} = \{Y_{ij} : R_{ij} = 0\}$ denote the sets of variables associated respectively with the observed and missing data. The joint probability density function of the observed data satisfies:

$$p_{\theta,\psi}(Y_{obs}, R) = \int\int p_\theta(Y_{obs}, Y_{mis}, Z) h_\psi(R|Y_{obs}, Y_{mis}, Z) dY_{mis} dZ. \tag{2}$$

Definition 1 (Missing At Random)**.** *A sampling design is called Missing At Random (MAR) if the sampling process described by R is such that*

$$R \perp\!\!\!\perp (Y_{mis}, Z) \mid Y_{obs}.$$

Proposition 1. *If the sampling design is MAR, then maximizing $p_{\theta,\psi}(Y_{obs}, R)$ is equivalent to maximize $p_\theta(Y_{obs})$.*

MAR example: random pair sampling Each pair of nodes has the same probability ρ to be observed:

$$\forall (i,j) \in \{1, \dots, n\}^2, \quad \mathbb{P}(R_{ij} = 1) = \rho\,.$$

NMAR example: double standard sampling. Let ρ_1 and ρ_0 be two probabilities. Double standard sampling consists in observing edges with probabilities :

$$\begin{cases} \mathbb{P}(R_{ij} = 1 | X_{ij} = 1) = \rho_1, \\ \mathbb{P}(R_{ij} = 1 | X_{ij} = 0) = \rho_0. \end{cases}$$

3 Statistical Inference

3.1 MAR case

As usual in the presence of latent variables, EM algorithm [3] is a natural choice. However, the E-step cannot be computed due to the dependence structure inherent to the SBM. The variational approach proposed in [1] overcomes this issue by maximizing the following lower bound of the likelihood:

$$\mathcal{J}(\mathcal{R}_{Y_{obs}}) = \log(p_\theta(Y_{obs})) - KL[\mathcal{R}_{Y_{obs}}(Z)||p_\theta(Z|Y_{obs})],$$

where KL denotes the Kullback-Leibler divergence, $p_\theta(Z|Y_{obs})$ the true conditional distribution of Z given Y, and $\mathcal{R}_{Y_{obs}}$ an approximation of this conditional distribution. The E-step becomes tractable when the approximate conditional distribution $\mathcal{R}_{Y_{obs}}$ factorizes as follows [5]:

$$\mathcal{R}_{Y_{obs}}(Z) = \left(\prod_i h(Z_i, \tau_i)\right)$$

where $\{\tau_i \in [0,1]^Q, \quad i = 1, \dots n\}$ are the variational parameters associated with the $\{Z_i, i = 1, \dots n\}$, and $h(\cdot, \tau_i)$ is the multinomial distribution.

Variational EM for MAR cases

1. Initialize $\{{\tau_i}^{(0)}\}$.

2. Iteratively update τ_i and θ:

$$\begin{array}{rcll} \left(\theta^{(h+1)}\right) & = & \arg\max_\theta \mathcal{J}\left(\mathcal{R}_{Y_{obs}};\ \tau_i^{(h)}, \theta\right), & \textbf{(M step)} \\ \{{\tau_i}^{(h+1)}\} & = & \arg\max_{\tau_i} \mathcal{J}\left(\mathcal{R}_{Y_{obs}};\ \tau_i, \theta^{(h+1)}\right). & \textbf{(VE step)} \end{array}$$

3. Repeat Step 2 until convergence (typically if the likelihood no longer significantly increases from successive steps).

3.2 NMAR case

In the NMAR case, we choose an SAEM (Stochastic Approximation EM) approach in the vein of [2] combined with Gibbs sampling as in [6]. Indeed, relying on a stochastic strategy allows to provide a general framework whatever the sampling strategy at play.

The general idea is to replace the E-step of the EM algorithm by a stochastic approximation of $Q_h(\theta, \psi) = \mathbb{E}[p(Y, R, Z; Y_{obs}, R, \hat{\theta}^{(h-1)}, \hat{\psi}^{(h-1)})]$ (see [3]) with simulated values of $Z^{(h+1)}$ and $Y_{mis}^{(h+1)}$ based on Gibbs sampling which generates an ergodic Markov chain

with $p(Y_{mis}, Z; \theta^{(h)})$ as unique stationary distribution. The stochastic approximation of the conditional expectation is :

$$Q_h(\theta, \psi) = Q_{h-1}(\theta, \psi) + \gamma_h \left(\log p_{\theta,\psi}(Y_{obs}, Y_{mis}^{(h)}, R, Z^{(h)}) - Q_{h-1}(\theta, \psi) \right), \tag{3}$$

using $(\gamma_h)_{h \geqslant 0}$ a sequence of positive numbers decreasing to 0, thus ensuring the convergence (see [6]).

Stochastic Approximation EM for NMAR cases

1. Initialize $\{\{\tau_i{}^{(0)}\}, \theta^{(0)}, \psi^{(0)}\}$.
2. Iteratively update Z_i, θ and ψ as follows:

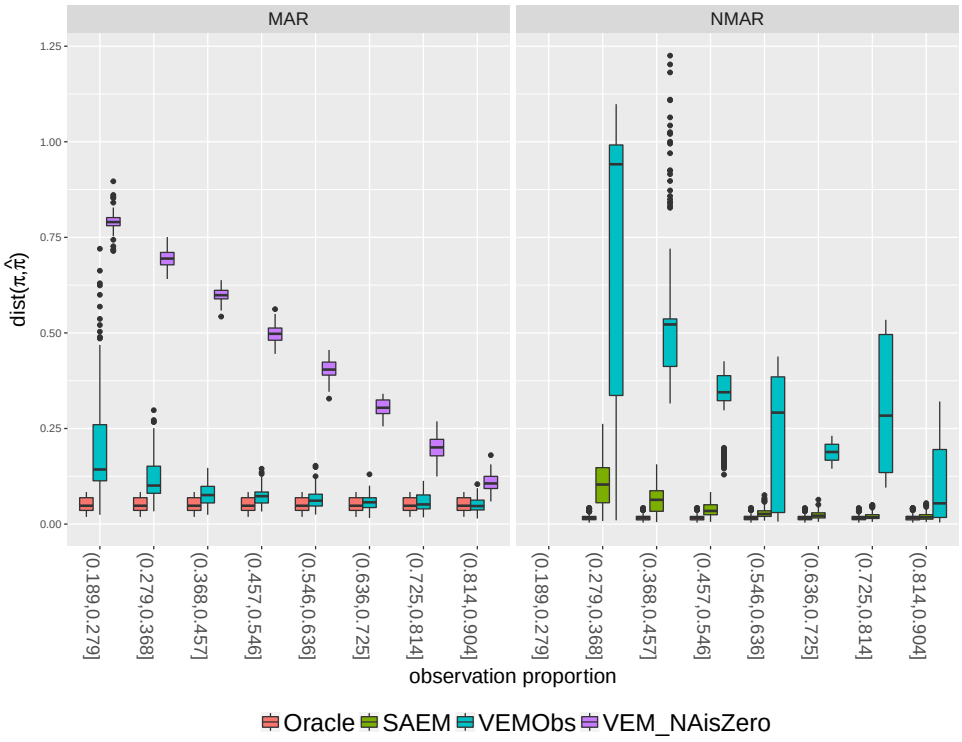
$$\begin{aligned} (Y_{mis}^{(h)}, Z^{(h)}) | (Y_{obs}, R) &\sim \hat{\mathbb{P}}_{\text{Gibbs}} && \textbf{(SE step)} \\ (\theta^{(h+1)}, \psi^{(h+1)}) &= \arg\max_{(\theta,\psi)} Q_h(\theta, \psi) && \textbf{(M step)} \end{aligned}$$

3. Repeat Step 2 until convergence.

4 Simulation

We illustrate the benefit of our approach on simulated random graphs following the SBM with different topologies (e.g., affiliation, star or bipartite networks). We assess the quality of the inference by computing the distance between estimated and true connectivity matrix π in terms of Frobenius norm. Concerning the estimation of the classification, we compute the adjusted Rand index between the true and the estimated classification.

An example of such a simulation is represented on the following figure for an affiliation network with $n = 100$ nodes. The left panel addresses the MAR case while the right panel concerns the NMAR case. On the left panel, the estimation error is represented as a function of the proportion of observed values in the adjacency matrix of the graph. We show that accounting for MAR strategy greatly improves the estimation of π compared with the commonly used strategy which consists in replacing NA value in the adjacency matrix by zeros. On the right panel, we consider the NMAR case with double standard sampling strategy. Our NMAR approach clearly outperforms the MAR algorithm, sticking to the oracle case where the adjacency matrix is considered as completely observed.



References

[1] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18:173–183, 2008.

[2] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27:94–128, 1999.

[3] A. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39(1):1–38, 1977.

[4] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[5] T. Jaakkola. *Advanced Mean Field Methods: Theory and Practice*, chapter : Tutorial on variational approximation methods. MIT Press, Cambridge, 2000.

[6] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[7] P. Latouche, E. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

[8] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[9] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.