Inférence de graphe avec contrôle du taux de faux positifs

Tabea Rebafka [1], Etienne Roquain [1]& Fanny Villers [1]

[1] *Sorbonne Université, Université de Paris, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, Paris, France.*
tabea.rebafka@upmc.fr, etienne.roquain@upmc.fr, fanny.villers@upmc.fr

Résumé. Une nouvelle procédure de test multiple est proposée pour inférer un graphe à partir d'une observation bruitée de ce graphe. Dans ce but, nous introduisons d'abord un modèle à blocs stochastiques bruité (NSBM) et développons un algorithme EM variationnel pour estimer les paramètres et calculer un clustering des nœuds. La procédure de test pour l'inférence du graphe que nous proposons exploite la topologie d'un NSBM. Nous montrons que la version oracle de notre procédure contrôle le taux de faux positifs, et maximise le taux de vrais positifs. Des résultats numériques illustrent les propriétés de notre procédure et montrent qu'elle est plus performante que des procédures de test classiques.

Mots-clés. Inférence de graphe, modèle à blocs stochastiques, taux de faux positifs, algorithme EM variationnel.

Abstract. A new multiple testing procedure is proposed for the problem of inferring a graph from a noisy observation of that graph. To this end the so-called noisy stochastic block model (NSBM) is introduced and a variational expectation-maximization algorithm is developed that provides parameter estimates and a clustering of the nodes. A new test procedure for graph inference is proposed that exploits the graph topology of the NSBM. We show that the oracle version of our procedure controls the false discovery rate and, in some sense, maximizes the true discovery rate. Numerical experiments illustrate the performance of our test procedure and show that it outperforms classical methods.

Keywords. Graph inference, stochastic block model, false discovery rate, variational expectation-maximization algorithm.

1 Introduction

In many applications of network analysis, an essential task is to infer a reliable version of the network. Often, when a dense network is observed, this one is considered to be a perturbation of some underlying less dense graph, and we would like to remove the edges that are only due to noise. Once the graph is inferred, in many applications a deeper analysis is carried out to describe the communication structure of the network, for

instance by community detection, that is, clustering of the nodes in groups with similar connecting behavior.

Graph inference is a task with a long history, especially in the context of estimating marginal or partial correlations between the nodes of a graph. In this context, often a Gaussian graphical model (Lauritzen, 1996) is used for the correlation or the precision matrix. In the literature, this task is classically done by some graphical lasso-type approach (Meinshausen and Bühlmann, 2006; Friedman et al., 2007; Banerjee et al., 2008; Ravikumar et al., 2011).

A popular random graph model for clustering the nodes is the stochastic bloc model (SBM) (Holland et al., 1983), that models network heterogeneity by varying connecting behavior of different groups of nodes. More precisely, each node is supposed to belong to exactly one group and the edge probability of a pair of nodes depends entirely on the group membership of these two nodes. Thus, clustering becomes the problem of estimating the group memberships in the stochastic block model.

In this work we propose a variant of the SBM which is suitable for simultaneously inferring the clustering and the graph: the so-called noisy stochastic block model (NSBM). In this model, we do not observe the graph of interest, which is itself a latent structure, but only a noisy version of it, with the following blurring mechanism: in place of missing edges, pure random noise is observed, and in place of present edges, we observe an effect, whose intensity may depend on the block memberships of the nodes in the latent graph. Based on the NSBM, we adapt procedures of the multiple testing literature for mixture models, namely a q-value approach (Storey, 2003; Castillo and Roquain, 2018). Thus, our approach leads to a new procedure for simultaneously inferring a clustering of the nodes and the graph itself, with a clear interpretation in terms of false positives: among the edges discovered by the procedure, there are, on average, at most 5% (say) of errors.

2 Gaussian noisy stochastic block model

Consider an undirected graph with n nodes. Denote $\mathcal{A} = \{(i,j) : 1 \leq i < j \leq n\}$ the set of all possible edges. We observe a symmetric, real-valued matrix $X = (X_{i,j})_{1 \leq i,j \leq n} \in \mathbb{R}^{n^2}$ representing the observed interactions between all node pairs (i,j).

To model X we introduce the *noisy stochastic block model* (NSBM) defined as a perturbation of a standard binary stochastic block model (SBM) . More precisely, we suppose that X is a noisy version of a binary adjacency matrix $A = (A_{i,j})_{1 \leq i,j \leq n} \in \{0,1\}^{n^2}$, where $A_{i,j} = 1$ if and only if there is an edge between node i and node j. We assume that A is a binary SBM and our aim is to derive A from the observation X.

Let Q be the number of latent node blocks. The NSBM is defined by the following random layers. First, generate a vector $Z = (Z_1, \ldots, Z_n)$ of block memberships of the nodes such that Z_i, $1 \leq i \leq n$ are i.i.d. taking their values in $\{1, 2, \ldots, Q\}$ with probabilities $\pi_q = \mathbb{P}(Z_1 = q)$ for $q \in \{1, \ldots, Q\}$ with parameter $\pi = (\pi_q)_{q \in \{1,\ldots,Q\}} \in [0,1]^Q$ such

that $\sum_{q=1}^{Q} \pi_q = 1$. Then, conditionally on Z, the variables $A_{i,j}$, $(i,j) \in \mathcal{A}$ are independent Bernoulli variables with parameter w_{Z_i,Z_j}, that is,

$$A_{i,j} \mid Z \sim \mathcal{B}(w_{Z_i,Z_j}),$$

for some parameter $w = (w_{q,\ell})_{q,\ell \in \{1,\dots,Q\}} \in [0,1]^{Q^2}$, where w is symmetric, since the graph is undirected, that is, $w_{q,\ell} = w_{\ell,q}$ for all $q, \ell \in \{1,\dots,Q\}$. Note that only $A_{i,j}$, $(i,j) \in \mathcal{A}$ are sampled randomly and we set $A_{j,i} = A_{i,j}$ for all $(i,j) \in \mathcal{A}$ and $A_{i,i} = 0$ for $i \in \{1,\dots,n\}$. Finally, conditionally on (Z,A), the observed variables $X_{i,j}$, $(i,j) \in \mathcal{A}$ are independent Gaussian variables and every $X_{i,j}$ has the distribution given by

$$X_{i,j} \mid (Z,A) \sim (1 - A_{i,j})\mathcal{N}(0,\sigma_0^2) + A_{i,j}\mathcal{N}(\mu_{Z_i,Z_j}, \sigma^2_{Z_i,Z_j}),$$

with parameters $\sigma_0^2 > 0, \mu_{q,\ell} \in \mathbb{R}$ and $\sigma^2_{q,\ell} > 0$.

The rationale behind this model is that, in place of missing edges ($A_{i,j} = 0$), we observe pure random noise modeled by the null distribution $\mathcal{N}(0,\sigma_0^2)$, and in place of present edges ($A_{i,j} = 1$), we observe an effect, whose distribution $\mathcal{N}(\mu_{Z_i,Z_j}, \sigma^2_{Z_i,Z_j})$ depends on the block membership of the interacting nodes in the underlying SBM. Instead of the normal distribution, one may use other parametric families like the exponential or Poisson distribution.

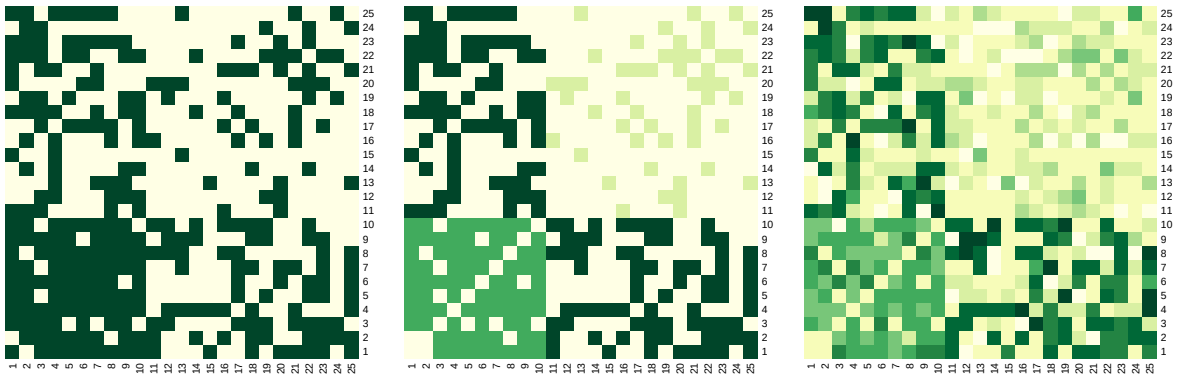
The unknown global model parameter is $\theta = (\pi, w, \sigma_0^2, \mu, \sigma)$ with $\mu = (\mu_{q,\ell})_{q \le \ell}$ and $\sigma^2 = (\sigma^2_{q,\ell})_{q \le \ell}$. The observation is X, while both Z and A are unobserved, that is, (Z,A) are the latent variables of this model.

An illustration is provided in Figure 1, where two latent blocks are considered: one is a community, the other has few connections among nodes in the same block and more connexions with nodes of the other block, see (a). In (b) the associated gaussian means are illustrated: 0 if there is no edge, i.e. $A_{i,j} = 0$, and some non zero mean $\mu_{q,\ell}$ otherwise, depending on the group memberships of the interacting nodes. Finally, (c) displays the observations $X_{i,j}$ that are random perturbations of the means of the matrix in (b). Recall that our aim is to recover A from X. The fundamental idea of our method is that learning the block memberships helps in the decision about the presence or absence of edges. From (c) it is clear that there is few ambiguity when $X_{i,j}$ is associated with two nodes in the first group, while for values associated with two nodes belonging to the second block is a much more difficult task. It follows that a light green value of $X_{i,j}$ should be interpreted differently depending on the group memberships of the interacting nodes.

Our results

It can be shown that the NSBM is **identifiable** up to label swapping under classical assumptions (Rebafka et al., 2019).

Furthermore, the maximum likelihood estimator of θ can be approached by a **variational EM-algorithm**. The development of this algorithm is involved, but similar to



(a) Latent graph A. (b) Gaussian means $(\mu_{Z_i,Z_j})_{i,j}$. (c) Observations X.

Figure 1: Gaussian NSBM with two groups. (a) Latent adjacency matrix A (b) Matrix of gaussian means according to presence/absence of edge $A_{i,j}$ and group membership (c) observed random perturbation of the matrix in (b).

other SBM-type models. As a byproduct, the variational EM-algorithm also provides a clustering $\hat{Z}$ of the nodes into Q classes. See Rebafka et al. (2019) for details.

3 Test procedure

For graph inference, the goal is to recover the adjacency matrix A from the observation X. In the multiple testing paradigm, the aim is to make a simultaneous test of

$$H_{0,i,j} : A_{i,j} = 0 \quad \text{against} \quad H_{1,i,j} : A_{i,j} = 1, \qquad \text{for } (i,j) \in \mathcal{A},$$

which corresponds to test "there is no edge between i and j in the latent graph" against "there is an edge between i and j in the latent graph". A multiple testing procedure is any measurable function $\varphi(X) \in \{0,1\}^{\mathcal{A}}$, with the convention that $\varphi_{i,j}(X) = 1$ if and only if $H_{0,(i,j)}$ is rejected.

The false discovery rate (FDR) of a given multiple testing procedure $\varphi(X)$ is the average proportion of errors among the discovered edges. The power of test or true discovery rate (TDR) is the average proportion of discovered edges in the underlying latent graph. A good testing procedure detects a maximum number of significant edges, without making too many false detections. In this sense, for a given level $\alpha \in (0,1)$, we aim at finding a testing procedure $\varphi = \varphi_\alpha$ such that for all θ,

$$\mathrm{FDR}(\theta, \varphi) \leq \alpha, \quad \text{with } \mathrm{TDR}(\theta, \varphi) \text{ as large as possible.}$$

That is, the FDR is controlled at level α, while many true edges are discovered.

3.1 Oracle procedure

To infer the latent graph A in the NSBM the best classification rule is the Bayes rule based on the posterior distribution of A, that is on the distribution of A given X. However, this distribution is intractable like in numerous other latent variable models.

So, to start with, we assume that the latent clustering Z is known. Then it is natural to consider as test statistics the posterior probabilities of $A_{i,j}$ given X and Z, that is

$$\begin{aligned}\ell_{i,j}(X, Z, \theta) &= \mathbb{P}_\theta(A_{i,j} = 0 \mid X, Z) \\ &= \frac{(1 - w_{Z_i,Z_j}) f_{\mathcal{N}(0,\sigma_0^2)}(X_{i,j})}{(1 - w_{Z_i,Z_j}) f_{\mathcal{N}(0,\sigma_0^2)}(X_{i,j}) + w_{Z_i,Z_j} f_{\mathcal{N}(\mu_{Z_i,Z_j},\sigma^2_{Z_i,Z_j})}(X_{i,j})}.\end{aligned}$$

We refer to the $\ell_{i,j}(X, Z; \theta)$ as the *ℓ-values*. A convenient multiple testing procedure rejects $H_{0,i,j}$ provided that $\ell_{i,j}(X, Z; \theta) \leq t$ for some threshold t. This threshold t should be chosen such that the FDR is lower than or equal to α. According to the approach of q-values, we define

$$q_{i,j}(X, z; \theta) = \frac{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{A}} (1 - A_{i,j}) \mathbb{1}\{\ell_{i,j}(X, Z, \theta) \leq \ell_{i,j}(X, z; \theta)\}\right]}{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\{\ell_{i,j}(X, Z, \theta) \leq \ell_{i,j}(X, z; \theta)\}\right]},$$

and reject $H_{0,i,j}$ whenever $q_{i,j}(X, z; \theta) \leq \alpha$. The quantities $q_{i,j}(X, z; \theta)$ are referred to as the *q-values*, a term that goes back to Storey (2003). Thus, we define the *oracle multiple testing procedure* as

$$\varphi^*_{i,j} = \mathbb{1}\{q_{i,j}(X, Z; \theta_0) \leq \alpha\}, \ \ (i, j) \in \mathcal{A},$$

where X follows the NSBM with true parameter $\theta_0 \in \Theta$ and latent clustering Z.

Our results

One can show that the oracle procedure maximizes the TDR among all procedures controlling the marginal FDR at level α (under appropriate assumptions). Proofs are quite technical (Rebafka et al., 2019).

3.2 New test procedure

Obviously, the oracle procedure is unknown, but it can be approximated using the estimator $\widehat{\theta} = (\widehat{\pi}, \widehat{w}, \widehat{\sigma}_0, \hat{\mu}, \hat{\sigma}^2)$ of θ_0 and the estimated clustering $\widehat{Z}$ obtained by the VEM algorithm for the NSBM. Thus, we define the following feasible multiple testing procedure

$$\varphi^{\text{VEM}}_{i,j} = \mathbb{1}\{q_{i,j}(X, \widehat{Z}; \widehat{\theta}) \leq \alpha\}, \quad (i, j) \in \mathcal{A}.$$

Our results

Numerous simulation results illustrate the performance of the new test procedure φ^{VEM}, namely the FDR is close to the nominal value α. Compared to various other test procedures, our procedure φ^{VEM} achieves the highest TDR in a large variety of scenarios. See Rebafka et al. (2019) for details.

Acknowledgments

This work has been supported by the grants ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS) and ANR-18-CE02-0010-01(EcoNet) of the French National Research Agency ANR.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516.

Castillo, I. and Roquain, E. (2018). On spike and slab empirical Bayes multiple testing. *arXiv e-prints*, page arXiv:1808.09748.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

Rebafka, T., Roquain, E., and Villers, F. (2019). Graph inference with clustering and false discovery rate control.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035.