Tutorial: Inference in the Stochastic Block Model

MSc in Statistics for Smart Data – Introduction to graph analysis and modeling

Julien Chiquet, November the 7, 2017

Preliminaries

Goals.

1. Random graphs generation and corresponding Gaussian data: Erdös-Rényi, Community networks, star-network, scale-free
2. sparse inference methods for Gaussian Graphical models
3. Analysis of some real world data

Instructions. Each student *must* send an `R markdown` report generated via `R studio` to julien.chiquet@inra.fr at the end of the tutorial. This report should answer the questions by commentaries and codes generating appropriate graphical outputs. A cheat sheet of the markdown syntax can be found here.

Required packages. Check that the following packages are correctly available on your plateform:

```
library(huge)
library(glmnet)
library(sand)
```

You also need `Rstudio`, LaTeX and packages for markdown:

```
library(knitr)
library(rmarkdown)
```

1 Background

1.1 Notations

We let $\mathcal{P} = \{1, \dots, p\}$ be a set of nodes. Presence or absence of an edge between two nodes i and j is described by the random variable $X_{ij} = \mathbf{1}_{\{i \leftrightarrow j\}}$. We assume by convention that the nodes are not connected to themselves, that is, $X_{ii} = 0$ for all $i \in \mathcal{P}$.

1.2 Stochastic Block Model

This model has several representation. We adopt the one given by Daudin, Picard and Robin (2007), known as "mixture model for random graphs". This model spreads the nodes among a set of Q classes $\mathcal{Q} = \{1, \dots, Q\}$ with *a priori* distribution $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_Q)$. The hidden random indicator variables $(Z_{iq})_{i \in \mathcal{P}, q \in \mathcal{Q}}$ define the classes each node belongs to. Thus

$$\alpha_q = \mathbb{P}(Z_{iq} = 1) = \mathbb{P}(i \in q), \quad \text{such that} \sum_q \alpha_q = 1. \tag{1}$$

It is straightforward to see that $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iQ})$ has a multinomial distribution

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha}). \tag{2}$$

Finally, let $\pi_{q\ell}$ be the probability that a node in class q connects to a node in class ℓ[1]. The probability for having edge between nodes i and j is defined *conditionally on* the classes they belong to:

$$X_{ij} | \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q\ell}), \quad i \neq j. \tag{3}$$

To sum up, the parameters are

- $\mathbf{X} = (X_{ij})$, the $p \times p$ adjacency matrix of the graph,
- $\boldsymbol{\pi} = (\pi_{q\ell})$ the $Q \times Q$ connectivity matrix,
- $\boldsymbol{\alpha} = (\alpha_q)$, the size-$Q$ vector of class proportions.

1.2.1 Useful quantities in the variational EM

During this practical, you will implement the variational EM (VEM) algorithm studied during this morning lecture. Here are the expressions of the key quantities that you need to compute along the algorithm.

1.2.1.1 Variational lower bound. The variational lower bound of the loglikelihood maximized by the VEM is

$$J(\boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\alpha}) = \sum_{i,q} \tau_{iq} \log \alpha_q + \sum_{i<j,q,\ell} \tau_{iq}\tau_{j\ell} \log b(X_{ij}; \pi_{q\ell}) - \sum_{i,q} \tau_{iq} \log(\tau_{iq}),$$

where $b(x; \pi) = \pi^x (1 - \pi)^{1-x}$ is the probability density function of the Bernoulli distribution and τ_{iq} are the posterior probabilities for class belonging, aka the variational parameters.

[1] Since the network is undirected, the matrix $\mathbf{X}$ is symmetric and so $\pi_{q\ell} = \pi_{\ell q}$.

1.2.1.2 M step.

For fixed values of $\hat{\tau}_{iq}$ the estimators for α_q and $\pi_{q\ell}$ by maximizing the conditional expectation are

$$\hat{\alpha}_q = \frac{1}{n}\sum_i \hat{\tau}_{iq}, \quad \hat{\pi}_{q\ell} = \frac{\sum_{i\neq j} \hat{\tau}_{iq}\hat{\tau}_{j\ell}X_{ij}}{\sum_{i\neq j} \hat{\tau}_{iq}\hat{\tau}_{j\ell}}. \tag{4}$$

1.2.1.3 E step.

The variational parameters τ_{iq} verify the following fixed point relation:

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_\ell b(X_{ij}, \pi_{q\ell})^{\hat{\tau}_{j\ell}} \tag{5}$$

1.2.1.4 Integrated complete likelihood criterion (ICL).

The variational ICL used to compare models with different numbers of clusters is

$$\begin{aligned}\mathrm{vICL}(Q) = \sum_{i,q} \hat{Z}_{iq} \log \hat{\alpha}_q + \sum_{i<j,q,\ell} \hat{Z}_{iq}\hat{Z}_{j\ell} b(X_{ij}; \hat{\pi}_{q\ell}) \\ - \frac{1}{2}\left(\frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} + (Q-1)\log(n)\right)\end{aligned}$$

where $\hat{Z}_{iq}$ is the maximum a posteriori associated to the estimated probability $\hat{\tau}_{iq}$.

2 Introduction

This practical aims to provide a quick overview of sparse Gaussian Graphical Models (GGM) and their use in the context of network reconstruction for gene interaction networks.

To this end, we rely on the R-package **huge**, which implements some of the most popular sparse GGM methods and provides a set of basic tools for their handling and their analysis.

The first part focuses on an empirical analysis of the statistical models used for network reconstruction. The objective is to quickly study the range of applicability of these methods. It should also give you some insights about their limitations, especially toward the interpretability of the inferred network in terms of biology.

The second part applies these methods to one data sets: the first one consists in a transcriptomic data associated to a small regulatory network for which partial ground truth is available. The objective is to unravel the most striking interactions between differentially expressed genes.