Problem_Set_3

Bayesian community detection in assortative Stochastic Block Models

Pierpaolo De Blasi

deadline 2025-06-03 11.59pm

Introduction

A key task in network analysis is community detection, which typically consists in dividing the nodes into groups such that nodes within a group are strongly connected, while connections between groups are relatively scarcer. We speak of *assortative* communities of nodes. A generative model well-suited for the formation of such communities is the *assortative* Stochastic Block Model (SBM), which prescribes a higher probability of a connection between nodes belonging to the same block rather than to different blocks. The number of blocks k is of pivotal importance, as in more traditional clustering applications, in that it determines the intrinsic dimension of the model. We keep k fixed here, and we are interested to recover communities via Bayesian methods, in particular in the *over-fitted case*, that is when the number of communities (if any) in the observed network is smaller than k. The task is to study, tthrough a Monte Carlo study on synthetic networks, the effects of enforcing assortativity through the Bayesian modeling on community detection. See Section "Monte Carlo Study" below for details.

Bayesian model

Consider an undirected graph with no self-loops made of n nodes, labeled $1, \dots, n$, described by a symmetric $n \times n$ adjacency matrix A, with entry $A_{ij} = 1$ if node i and node j are connected and $A_{ij} = 0$ otherwise. Each node belongs to one of k mutually exclusive groups, called blocks, labelled $1, \dots, k$, that share a similar connectivity patterns according to

$$p(A_{ij} = 1) = P_{z_i z_j}$$

independent across $i \neq j$. Here $P = (P_{ab})$ is the connectivity matrix, a $k \times k$ symmetric matrix with $P_{ab} \in (0, 1)$ being the probability of an edge between nodes of blocks labelled a and b. Also $z_i \in \{1, \dots, k\}$ is the block assignment, or label, of the ith node. We will refer to $z = (z_1, \dots, z_n)$ as the vector of labels or block assignments. Finally, let $\pi = (\pi_1, \dots, \pi_k)$ be probabilities of community assignments, so that $p(z_i = a) = \pi_a$ for $a = 1, \dots, k$ independently across $i = 1, \dots, n$. In summary,

$$\begin{aligned} z_i|\pi &\overset{\text{iid}}{\sim} \text{Multinomial}(1, \pi), \quad i = 1, \dots, n \\ A_{ij}|z, P &\overset{\text{ind}}{\sim} \text{Bernoulli}(P_{z_i, z_j}), \quad i < j \end{aligned}$$

We do observe A only, so the z_i are treated as latent variables like in cluster analysis. The *augmented* likelihood is given by

$$p(A, z|P, \pi) = \prod_{i<j} P_{z_i, z_j}^{A_{ij}} (1 - P_{z_i, z_j})^{1 - A_{ij}} \prod_i \pi_{z_i} = \prod_{a \leq b} P_{ab}^{O_{ab}(z)} (1 - P_{ab})^{n_{ab}(z) - O_{ab}(z)} \prod_a \pi_a^{n_a(z)}$$

where $n_a(z) = \sum_{i=1}^n \mathbb{1}(z_i = a)$ is the number of nodes labelled a,

$$O_{ab}(z) = \sum_{i \neq j} \mathbb{1}(z_i = a, z_j = b) A_{ij}, \quad n_{ab}(z) = \sum_{i \neq j} \mathbb{1}(z_i = a, z_j = b),$$

are the number of edges between nodes labelled a and b by the labelling z, and the maximum number of edges that can be created between nodes labeled a and b, respectively. As for the latter, we have

$$n_{ab}(z) = \begin{cases} n_a(z) n_b(z), & a \neq b \\ \binom{n_a(z)}{2} = \frac{1}{2} n_a(z)(n_a(z) - 1), & a = b. \end{cases}$$

We complete the Bayesian model by specifying a prior on π and P. The prior on π is symmetric Dirichlet

$$\pi \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \tag{1}$$

for $\gamma > 0$. The *default* prior on P is

$$P_{ab} \overset{\text{iid}}{\sim} \text{beta}(\alpha, \beta), \quad 0 \leq a \leq b \leq k \tag{2}$$

for $\alpha, \beta > 0$ and with π and P_{ab} independent. We sample from the posterior distribution of z and P by integrating out the probabilities π of assignment. According to (1), the conjugacy between the Bernoulli and Dirichlet distribution gives the marginal probability of label assignments

$$p(z) = \frac{\Gamma(\gamma k)}{\Gamma(\gamma)^k \Gamma(n + \gamma k)} \prod_a \Gamma(n_a(z) + \gamma)$$

from which one obtains the *prior predictive* (assignment) probabilities

$$p(z_i = a | z_{-i}) = \frac{n_a(z_{-i}) + \gamma}{n - 1 + \gamma k} \tag{3}$$

where z_{-i} is the vector of $n - 1$ assignments but the ith one. The full conditional of z_i is

$$\begin{aligned} p(z_i = a | z_{-i}, A, P) &= \frac{p(z_i = a, A | z_{-i}, P)}{p(A | z_{-i}, P)} = p(z_i = a | z_{-i}) \frac{p(A | z_i = a, z_{-i}, P)}{p(A | z_{-i}, P)} \\ &= p(z_i = a | z_{-i}) \frac{p(A | z_i = a, z_{-i}, P)}{p(A_{-i} | z_{-i}, P) p(A_i | A_{-i}, z_{-i}, P)} \\ &\propto p(z_i = a | z_{-i}) \frac{p(A | z_i = a, z_{-i}, P)}{p(A_{-i} | z_{-i}, P)} \end{aligned} \tag{4}$$

that is proportional to the product of the prior predictive probability $p(z_i = a | z_{-i})$, and the *likelihood contribution to* $z_i = a$ which simplifies to

$$\frac{p(A | z_i = a, z_{-i}, P)}{p(A_{-i} | z_{-i}, P)} = \prod_{j > i} P_{a\, z_j}^{A_{ij}} (1 - P_{a\, z_j})^{(1 - A_{ij})} \prod_{k < i} P_{z_k\, a}^{A_{ki}} (1 - P_{z_k\, a})^{(1 - A_{ki})} = \prod_{j \neq i} P_{a\, z_j}^{A_{ij}} (1 - P_{a\, z_j})^{(1 - A_{ij})} \tag{5}$$

Here A_i refers to the connections involving the ith node and A_{-i} to the $(n-1) \times (n-1)$ adjacency matrix upon removal of the ith node. As for the full conditional of P, by exploiting the standard binomial-beta update

$$P_{ab} | A, z \overset{\text{ind}}{\sim} \text{beta}(O_{ab}(z) + \alpha, n_{ab}(z) - O_{ab}(z) + \beta). \tag{6}$$

We report the Gibbs sampling algorithm next.

Algorithm 1 Gibbs sampler for SBM, k fixed

procedure SBM-K

Require $n \times n$ adjacency matrix A, number of communities k, number of iterations B, prior hyperparameters γ, α, β.

Initialize $z = (z_1, ..., z_n)$

for each iter $b = 1$ to B **do**

- update $P = (P_{ab})$ conditional on z and A from (6)

for each iter $i = 1$ to n **do**

- update z_i conditional on z_{-i}, P and A from (4) via (3) and (5)

end for

end for

end procedure

We move on to assortative SBM, which prescribes

$$\max_{a \neq b} P_{ab} < \min_{a} P_{aa}, \tag{7}$$

We discuss first how to induce a stochastic order $q < p$ between two random probabilities $p, q \in (0, 1)$. We introduce an auxiliary variable $\epsilon \in (0, 1)$, and set both p and q conditional on ϵ as truncated on $(\epsilon, 1)$ and $(0, \epsilon)$, respectively. We will refer to ϵ as the *cutoff*. For illustration purposes we confine to ϵ uniformly distributed on $(0, 1)$, $\epsilon \sim \mathrm{U}(0, 1)$, extensions to beta$(\alpha, \beta)$ distribution can be worked out similarly. We have

$$\epsilon \sim U(0, 1),$$
$$p|\epsilon \sim U(\epsilon, 1), \quad q|\epsilon \sim U(0, \epsilon).$$

The marginal and joint densities are then found to be:

$$f(p) = -\log(1 - p), \quad f(q) = -\log q$$
$$f(p, q) = \int_q^p x^{-1}(1 - x)^{-1} \mathrm{d}x = \log \frac{p(1 - q)}{q(1 - p)} \mathbb{1}_{\{q<p\}}$$

Figure 1 visually represents these densities.

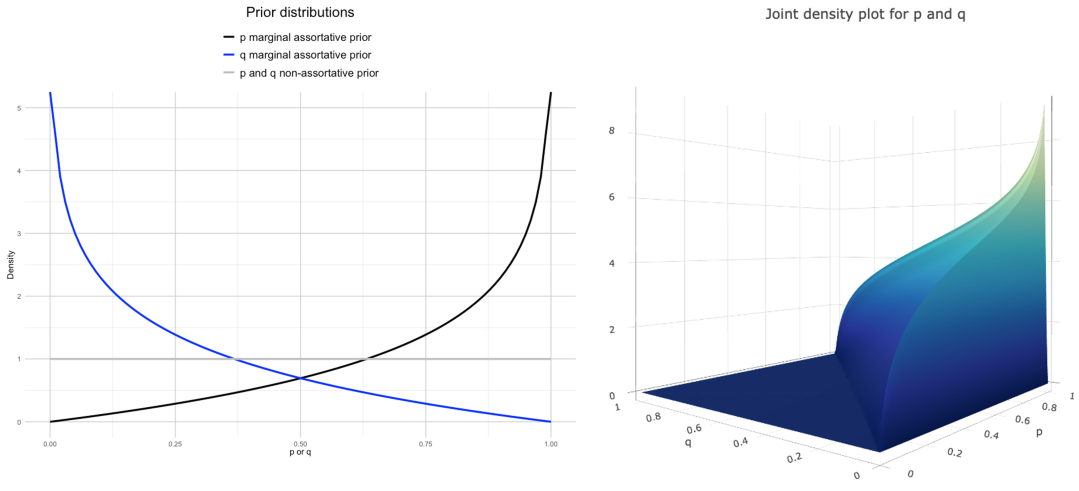

Figure 1: The left panel illustrates the marginal of p and q, the grey horizontal line marks the uniform density. The right panel shows the joint density which is supported on $q < p$.

We now have in place of (2)

$$\epsilon \sim U(0,1), \tag{8a}$$

$$P_{aa}|\epsilon \overset{\text{iid}}{\sim} \mathrm{U}(\epsilon,1), \quad a=1,\ldots,k, \tag{8b}$$

$$P_{ab}|\epsilon \overset{\text{iid}}{\sim} \mathrm{U}(0,\epsilon), \quad 1\le a<b\le k. \tag{8c}$$

As for the conditional distribution given A and z, we have to differently treat the diagonal elements P_{aa}, the off-diagonal elements P_{ab} and the cutoff ϵ. The full conditional of P and ϵ is found to be

$$\{P_{aa}\}_a, \{P_{ab}\}_{a<b}, \epsilon | A, z \propto \epsilon^{-k(k-1)/2}(1-\epsilon)^{-k} \prod_a P_{aa}^{O_{aa}(z)}(1-P_{aa})^{n_{aa}(z)-O_{aa}(z)}\mathbb{1}_{\{P_{aa}>\epsilon\}} \prod_{a<b} P_{ab}^{O_{ab}(z)}(1-P_{ab})^{n_{ab}(z)-O_{ab}(z)}\mathbb{1}_{\{P_{ab}<\epsilon\}}.$$

which leads to the following update for the connection probabilities

$$\begin{aligned} P_{aa}|\epsilon, A, z &\sim \mathrm{beta}_{(\epsilon,\ 1)}(O_{aa}(z)+1, n_{aa}(z)-O_{aa}(z)+1), \quad 1\le a\le k, \\ P_{ab}|\epsilon, A, z &\sim \mathrm{beta}_{(0,\ \epsilon)}(O_{ab}(z)+1, n_{ab}(z)-O_{ab}(z)+1), \quad 1\le a<b\le k. \end{aligned} \tag{9}$$

Here $\mathrm{beta}_{(\epsilon,\ 1)}$ and $\mathrm{beta}_{(0,\ \epsilon)}$ are beta distributions truncated over $(\epsilon,\ 1)$ and $(0,\ \epsilon)$, respectively. As for the cut off ϵ, we have

$$\epsilon | P, A, z \propto \epsilon^{-k(k-1)/2}(1-\epsilon)^{-k}\mathbb{1}_{q<\epsilon<p\}}, \quad p=\min_a P_{aa},\ q=\max_{a<b} P_{ab}. \tag{10}$$

We can sample from Equations (10) by introducing a latent variable y which has joint density with ϵ given by

$$\epsilon, y | P \propto \epsilon^{-k(k-1)/2}\mathbb{1}_{\{y<(1-\epsilon)^{-k},\ q<\epsilon<p\}}$$

The full conditional for y is given by the uniform density on the interval $(0,(1-\epsilon)^{-k})$. The full conditional for ϵ is

$$\epsilon | y, P \propto \epsilon^{-k(k-1)/2}\mathbb{1}_{\{\max\{q,1-y^{-1/k}\}<\epsilon<p\}}$$

This is easily sampled using the inverse cdf technique.

We report the Gibbs sampling algorithm for assortative-SBM next.

Algorithm 2 Gibbs sampler for assortative-SBM, k fixed

procedure SBM-K
Require $n\times n$ adjacency matrix A, number of communities k, number of iterations B, prior hyperparameters γ, α, β.
Initialize $z=(z_1,...,z_n)$ and ϵ
for each iter $b=1$ to B **do**
- update $P=(P_{ab})$ conditional on ϵ, z and A from (9)
- update ϵ conditional on P, z and A from (10)
for each iter $i=1$ to n **do**
- update z_i conditional on z_{-i}, P and A from (4) via (3) and (5)
end for
end for
end procedure

Monte Carlo study

We consider a synthetic network with a core/periphery structure according to an assortative SBM and intentionally designed to highlight the effect of enforcing the assortative constraint in Bayesian community detection. The network is made of 3 communities, a core, that is a central, larger community (60 nodes) with dense internal connections and links to two smaller external communities, or peripheries, of 20 nodes each. The two peripheries are each sparsely connected internally and with limited connections between them. To achieve so, connections are to be generated according to the following connectivity matrix

$$P = \begin{pmatrix} 0.30 & 0.08 & 0.08 \\ 0.08 & 0.10 & 0.02 \\ 0.08 & 0.02 & 0.10 \end{pmatrix}. \tag{11}$$

Use Algorithms 1 and 2 with $k = 3$ and $k = 4$ initializing the node labels at random. For consistency, use uniform prior on $(0, 1)$ on connection probabilities in the standard case, that is $\alpha = \beta = 1$ in (2). Also use unit shape parameter $\gamma = 1$ for the symmetric Dirichlet prior on assignment probabilities, cf. (1). Run each algorithm $B = 2000$ times after a burn-in of 500 iterations. Repeat each run 100 times, each time for a fresh new set of connections generated according to (11). For each run, estimate Maximum At Posteriori (MAP) community assignments based on N posterior samples of label vectors z using the R package `salso` using default loss function. Report histograms (4 in total) over the 100 runs of the estimated number of communities, as implied by the MAP assignments, together with boxplots of the Rand index of the MAP assignments against the true assignments. Any other summary statistics, or plots, are welcome in investigating the effect of assortative constrained SBM against standard SBM. Provide a report written in Markdown together with a separated script file with the R code. Set clearly the seed in the script file for reproducible random number generation.