Bayesian community detection[1]

Random graphs and network statistics

Yann McLatchie

[1]https://github.com/yannmclatchie/karate

Community detection

Suppose we observe an adjacency matrix $A = (A_{ij})$ of a graph, and task to infer the community memberships of each node (z_i), $i = 1, \ldots, n$. One way to do this is to model the structure of the graph, and specifically model $A \stackrel{d}{=} \mathsf{SBM}(z, P)$ with the link probability matrix P also unobserved.

Aim

We want to produce a Bayesian estimator (hopefully consistent) of the community structure for an SBM given a fixed number of communities.

Consistency

An estimator $\bar{X}_n$ of a random variable X is deemed *consistent* if it converges in probability to the true value of the variable X^*,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - X^*| > \epsilon) = 0, \; \forall \epsilon > 0.$$

An estimator $\bar{X}_n$ of a random variable X is deemed *strongly consistent* if it converges *almost surely* to the true value of the variable X^*,

$$\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = X^*) = 1.$$

Community structure in the SBM

We have an undirected random graph G on n nodes, each belonging to one of $K \in \mathbb{N}$ classes. Each node is randomly labelled according to i.i.d. $Z_1, \dots, Z_n$ random variables with probability $\pi_1, \dots, \pi_K$. Given this set of labels, edges between nodes are independently sampled from a Bernoulli random variable dependent on the label, $\mathbb{P}(A_{ij} = 1 \mid Z) = P_{Z_i, Z_j}$.

The likelihood of our SBM is then defined as

$$\prod_{1 \leq i < j \leq n} P_{Z_i, Z_j}^{A_{ij}} (1 - P_{Z_i, Z_j})^{1 - A_{ij}} \prod_{1 \leq i \leq n} \pi_{Z_i}.$$

Bayesian inference

What is Bayesian inference? Why Bayesian inference?

Prior choices

$$
\begin{aligned}
\pi &\sim \text{Dirichlet}(\alpha, \dots, \alpha) \\
P_{i,j} &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\beta_1, \beta_2) \\
e_i \mid \pi, P &\sim \pi \\
A_{ij} \mid \pi, P, e &\sim \text{Bernoulli}(P_{e_i, e_j})
\end{aligned}
$$

(Hyper-priors over α, β_1, β_2 also available, and not very sensitive, can use for instance $\alpha = 0.5$ and $\beta_1 = \beta_2 = 0.5$).

The posterior

Pas and Vaart (2018) call the posterior class distribution $p(e \mid A)$ the *Bayesian modularity*, denoted $Q_B(e)$, and we then assign class labels according to

$$\hat{e} = \arg\max_{e} Q_B(e).$$

A classification $\hat{e}$ is said to be weakly consistent if the fraction of misclassified nodes tends to zero, and strongly consistent if the probability of misclassifying any of the nodes tends to zero in the limit of the number of nodes (Pas and Vaart 2018).

The main result

Theorem

Denote $\rho_n = \sum_{i,j} \pi_i \pi_j P_{i,j}$, then:
1. If (P, π) is fixed and identifiable then the MAP estimator $\hat{e}$ is strongly consistent;
2. If $P = \rho_n S$ with (S, π) is fixed and identifiable then the MAP estimator $\hat{e}$ is strongly consistent if $(n-1)\rho_n \gg (\log n)^2$, where $\mathbb{E}[deg_G(i)] = (n-1)\rho_n$.

An application in Stan: Zachary's karate club

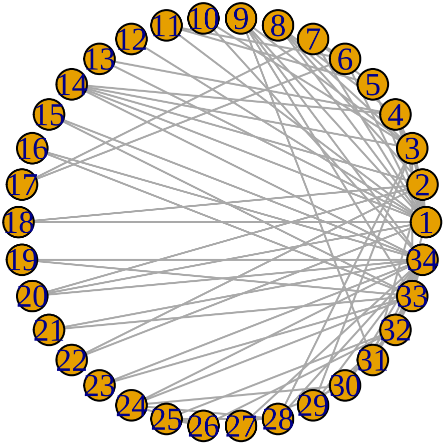

Figure 1: The karate club graph

The model in Stan

This code is adapted from Sarkar (2018).

```
model {
  for(i in 1:K){
    for(j in 1:K){
      // prior over kernel matrix
      phi[i][j] ~ beta(beta[1], beta[2]);
    }
  }
  // prior over mixture distribution
  pi ~ dirichlet(alpha);
  for(i in 1:N){
    for(j in i+1:N){ //symmetry and ignore diagonals
      // likelihood
      graph[i][j] ~ bernoulli(pi' * phi * pi);
    }
  }
}
```

Fitting the model

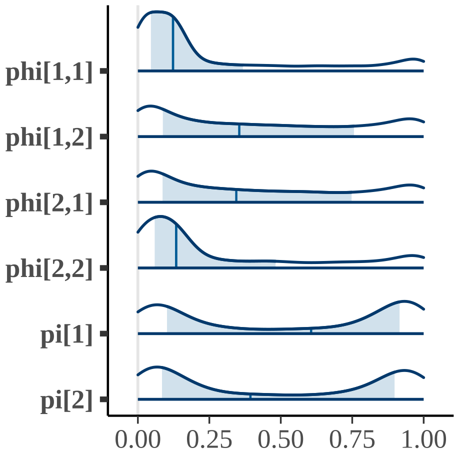

Figure 2: Posterior parameters

Prediction

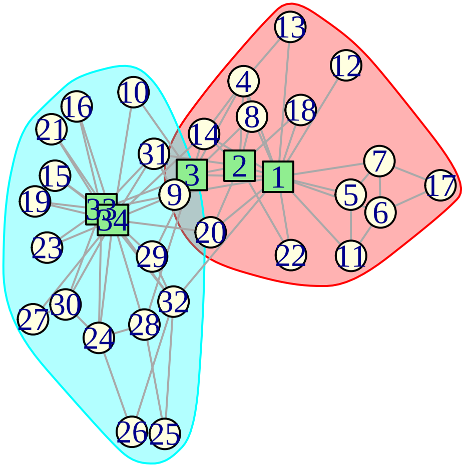

Figure 3: Cluster predictions.

Prior and likelihood sensitivity

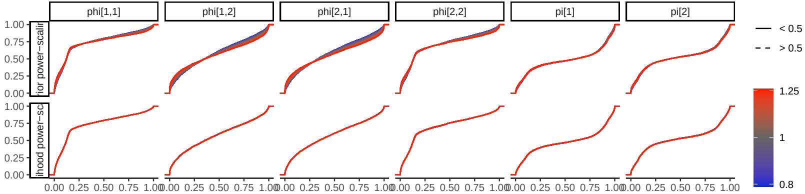

Figure 4: Prior and likelihood sensitivity of posterior.

References

Pas, S. L. van der, and A. W. van der Vaart. 2018. "Bayesian Community Detection." *Bayesian Analysis* 13 (3): 767–96. https://doi.org/10.1214/17-BA1078.

Sarkar, Arindam. 2018. "Extensions of Powerlawgraph." *GitHub Repository*. https://github.com/arindamsarkar93/powerlawgraph-ex; GitHub.