

Bayesian community detection¹

Random graphs and network statistics

Yann McLatchie

¹<https://github.com/yannmclatchie/karate>

Aim

Pas and Vaart (2018) task to produce a consistent Bayesian estimator of the community structure for an SBM given a fixed number of communities.

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) , $i = 1, \dots, n$. One way to do this is to model the structure of the graph, and specifically model $A \stackrel{d}{=} \text{SBM}(z, P)$ with the link probability matrix P also unobserved.

Consistency

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

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

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

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

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 π_1, \dots, π_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

We wish to infer parameter $\theta \in \Theta$ over which we have prior information $p(\theta)$. We achieve a *posterior* belief $p(\theta \mid x_{1:n})$ by combining our prior with a likelihood $p(x_{1:n} \mid \theta)$ and by performing the belief update (Bernardo and Smith 2009),

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)p(\theta).$$

Prior choices

$$\pi \sim \text{Dirichlet}(\alpha, \dots, \alpha)$$

$$P_{i,j} \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(\beta_1, \beta_2)$$

$$e_i \mid \pi, P \sim \pi$$

$$A_{ij} \mid \pi, P, e \sim \text{Bernoulli}(P_{e_i, e_j})$$

(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

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

As such, we classify nodes into community based on the maximum *a posteriori* (MAP) estimate of e . The Bayesian modularity is connected to the *likelihood modularity* of Bickel and Chen (2009), in that the latter exists as a special case of the former.

Why be Bayesian?

- ▶ Computationally efficiency of Gibbs sampler compared to maximum likelihood for large n
- ▶ Complete posterior predictive distribution
 - ▶ Uncertainty quantification
 - ▶ Decision analysis
- ▶ Ability to encode prior beliefs
- ▶ “A Bayesian version will usually make things better.” (Gelman 2022)

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 (π has strictly positive coordinates, and rows of P are distinct) 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 $\mathbb{E}[\deg_G(i)] = (n-1)\rho_n$.

How much data is enough data?

$(n - 1)\rho_n \gg \log n$ is sufficient for weakly consistent community detection (Lei and Rinaldo 2015).

Bickel and Chen (2009) claim the likelihood modularity is strongly consistent for arbitrary K under $(n - 1)\rho_n \gg \log n$. **However**, this was shown under the assumption that the modularity is globally Lipschitz, which is not the case in general.

$(n - 1)\rho_n \gg (\log n)^2$ is sufficient for the likelihood (and thus also the Bayesian) modularity to be strongly consistent for arbitrary K . In the special case $K = 2$, $(n - 1)\rho_n \gg \log n$ is also sufficient.

An application in Stan: Zachary's karate club

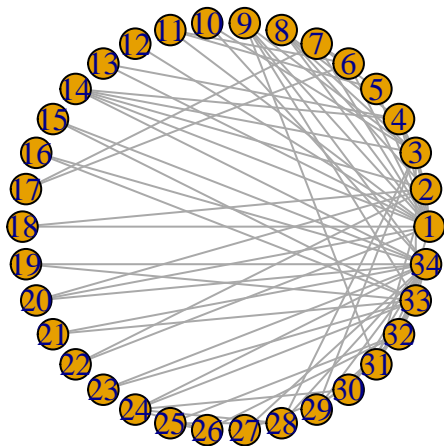


Figure 1: The karate club graph, with empirical average degree ~ 4.6 and $\log(n) \sim 3.5$.

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 community 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);  
    }  
  }  
}
```

The fitted model

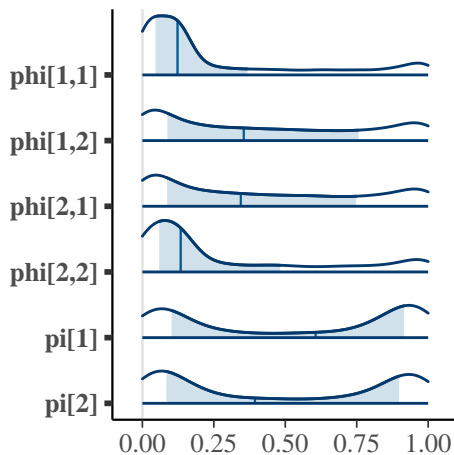


Figure 2: Posterior SBM parameters.

Decision analysis

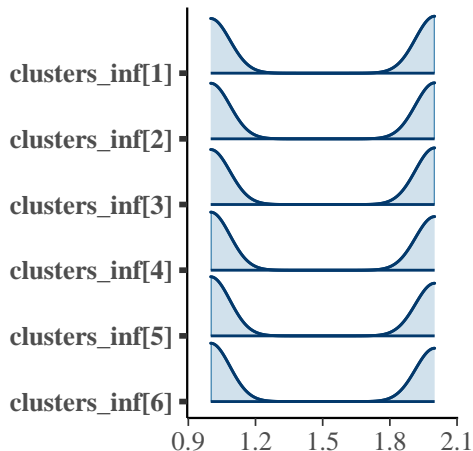


Figure 3: Posterior predictive distribution over six individual node communities.

Prediction

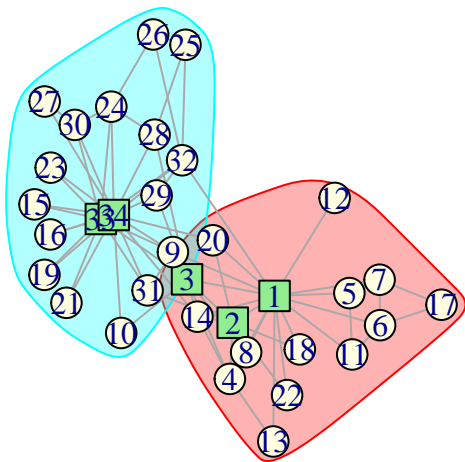


Figure 4: MAP cluster predictions from the Bayesian modularity. Shape and colour of node show predicted class, while coloured clouds indicate true communities.

Prior and likelihood sensitivity

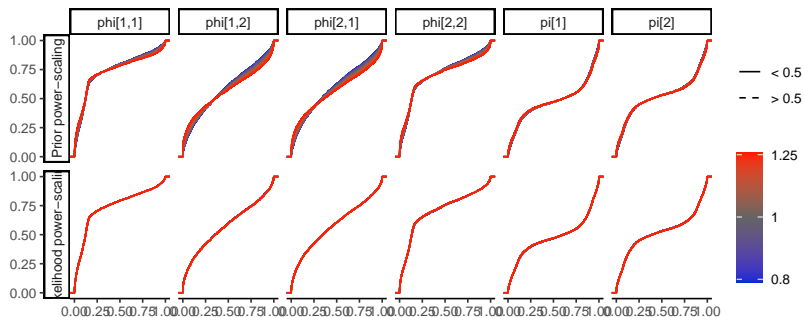


Figure 5: Prior and likelihood sensitivity of posterior.

The moral

1. Bayesian inference affords computational efficiency, the ability to encode priors, and general posterior decision analysis
2. The Bayesian modularity is closely related to the likelihood modularity, and both are strongly consistent under $(n - 1)\rho_n \gg (\log n)^2$
3. The karate club results do not reflect a failure of the inference method, rather a combination of small data and potentially misspecified likelihood.

References I

- Bernardo, J. M., and A. F. M. Smith. 2009. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley.
- Bickel, Peter J., and Aiyou Chen. 2009. "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities." *Proceedings of the National Academy of Sciences of the United States of America* 106 (50): 21068–73.
<http://www.jstor.org/stable/25593428>.
- Gelman, Andrew. 2022. "Andrew Gelman Quotes."
http://stat.columbia.edu/~gelman/book/gelman_quotes.pdf.
- Lei, Jing, and Alessandro Rinaldo. 2015. "Consistency of Spectral Clustering in Stochastic Block Models." *The Annals of Statistics* 43 (1). <https://doi.org/10.1214/14-aos1274>.
- 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>.

References II

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