Bayesian Degree-Corrected Stochastic Blockmodels for Community Detection

Lijun Peng and Luis Carvalho
Electronic Journal of Statistics, 2016

Carter Allen
December 02, 2019

Introduction

Peng and Carvalho propose an extension of the **parametric** Bayesian stochastic block model (SBM).

Peng and Carvalho extend their work by

1. Allowing for degree correction and explicitly characterizing community behavior via priors.

2. Addressing label switching through a remapping of community labels that forces an identifiable likelihood.

3. Utilizing Polya-Gamma data augmentation to allow for Gibbs sampling.

Background

Community detection has been a topic of interest in many fields such a social science, physics, and statistics.

Traditional methods for community detection include graph partitioning, hierarchical clustering, and spectral clustering.

These methods rely on **heuristics** which seek to identify bottlenecks of connectivity in terms of edge densities between groups.

Algorithms to maximize **modularity** have been widely used, but they tend to have high resolution limits.

Background

Peng and Carvalho propose a **parametric statistical** approach for community detection using the SBM.

This allows us to:

1. Infer community structure
2. Assess how likely a proposed community labeling is according to the assumed model

The Bayesian parametric SBM first proposed by Nowicki and Snijders (2001) allows for further flexibility.

Bayesian SBM

The authors propose a **parametric** Bayesian SBM: i.e., number of communities K is fixed *a priori*.

Data take the form of $\mathbf{A}$, an $n \times n$ adjacency matrix, with A_{ij} corresponding to the observed relationship between nodes (actors) i and j.

Nodes are mapped to communities via $\sigma : \{1,\ldots,n\} \mapsto \{1,\ldots,K\}$.

Thus, $\sigma_i = k$ denotes that node i belongs to community k.

Bayesian SBM

The SBM assumes the probability of an edge between nodes i and j depends only on σ_i and σ_j. A simple model for A_{ij} is

$$A_{ij} | \boldsymbol{\sigma}, \boldsymbol{\theta} \sim \text{Bern}(\theta_{\sigma_i, \sigma_j}) \text{ for } i, j = 1, \ldots, n; i < j$$

$$\sigma_i \sim \text{Multinom}(1, \boldsymbol{\pi}) \text{ for } i = 1, \ldots, n,$$

where $\boldsymbol{\theta}$ is a ragged array encoding the edge probability.

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{\sigma_1,\sigma_1} & & & \\ \theta_{\sigma_2,\sigma_1} & \theta_{\sigma_2,\sigma_2} & & \\ \vdots & \vdots & \ddots & \\ \theta_{\sigma_n,\sigma_1} & \theta_{\sigma_n,\sigma_2} & \cdots & \theta_{\sigma_n,\sigma_n} \end{bmatrix}$$

Bayesian SBM

Peng and Carvalho adopt a GLM approach:

$$A_{ij} | \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\eta} \sim \text{Bern}(\theta_{\sigma_i,\sigma_j}) \text{ for } i, j = 1,\ldots,n; i < j$$

$$\text{logit}(\theta_{\sigma_i,\sigma_j}) = \gamma_{\sigma_i,\sigma_j} + \eta_i + \eta_j$$

Here η_i and η_j are node-specific intercepts that measure the expected degree of nodes i and j, and

$\gamma_{\sigma_1,\sigma_2}$ captures between and within community association.

Identifiability

Non-Identifiability is a problem that arises in mixture models, where the likelihood is invariant to permutations of cluster labels.

In practice, this causes **label switching** in MCMC draws, i.e., when the cluster labels are permuted from one iteration to the next.

A common approach is to fix an arbitrary ordering of parameters, but this can lead to imperfect parameter estimation if true values are close together.

To prevent label switching, Peng and Carvalho propose a novel **canonical projection** of the community labels.

Identifiability

Let $\mathbf{L} = \{1,\ldots,K\}$ and $\mathscr{L} = \{\sigma \in \mathbf{L}^n : N_k(\sigma) > 1, k = 1,\ldots,K\}$.

Let $\mathrm{ord}(\sigma)$ return a vector with the order in which each $k = \{1,\ldots,K\}$ appears in σ.

Peng and Carvalho restrict community assignments to

$$\mathbf{Q} = \{\boldsymbol{\sigma} : \mathrm{ord}(\sigma) = \mathbf{L}\}$$

See section 4.1 on pg. 2753.

Identifiability

Illustrative example: $n = 8$; $K = 4$

$$\boldsymbol{\sigma} = (2,\ 2,\ 3,\ 1,\ 3,\ 4,\ 2,\ 1)$$

$$\mathrm{ord}(\sigma) = (\underset{1}{2},\ \underset{2}{3},\ \underset{3}{1},\ \underset{4}{4})$$

$$\rho(\boldsymbol{\sigma}) = (1,\ 1,\ 2,\ 3,\ 2,\ 4,\ 1,\ 3)$$

Identifiability

The remapping ρ guarantees correspondence between labels and order of appearance of a community in the label configuration.

Node $i = 1$ always belongs to community 1, with the next differently labeled node always belonging to community 2, and so on.

In MCMC sampling, ρ is also applied to corresponding parameters $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$.

This procedure forces identifiability and avoids the need for *post-hoc* relabeling algorithms (!).

Bayesian Inference

Peng and Carvalho adopt conjugate Multinomial-Dirichlet framework for $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$.

They also utilize Polya-Gamma data augmentation and assume Normal priors for all regression coefficients to allow for closed for Normal full conditionals.

Initial values are chosen via an efficient greedy optimization routine, which makes Gibbs sampling more efficient.

Posterior Estimators

1. **MAP Estimator** (maximum *a posteriori*):

$$\hat{\boldsymbol{\sigma}}_M = \operatorname{argmax}_{\tilde{\boldsymbol{\sigma}} \in \{1,\ldots,K\}^n} \mathrm{P}(\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}} \,|\, \mathbf{A})$$

2. **Binder Estimator**: $\hat{\boldsymbol{\sigma}}_B = \operatorname{argmin}_{\tilde{\boldsymbol{\sigma}} \in \{1,\ldots,K\}^n} \mathrm{E}_{\boldsymbol{\sigma}|\mathbf{A}}(B(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}))$, where

$$B(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = \sum_{i<j} I(\tilde{\sigma}_i \neq \tilde{\sigma}_j) I(\sigma_i = \sigma_j) + I(\tilde{\sigma}_i = \tilde{\sigma}_j) I(\sigma_i \neq \sigma_j)$$

3. **Centroid estimator**: $\hat{\boldsymbol{\sigma}}_B = \operatorname{argmin}_{\tilde{\boldsymbol{\sigma}} \in \{1,\ldots,K\}^n} \mathrm{E}_{\boldsymbol{\sigma}|\mathbf{A}}(H(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}))$, where

$$H(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} I(\tilde{\sigma}_i \neq \sigma_i)$$

Posterior Estimators

The authors prefer a refined version of Hamming loss that accounts for single and double label permutations.

I.e., for a permutation ϕ, $H(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = H(\phi(\tilde{\boldsymbol{\sigma}}), \phi(\boldsymbol{\sigma}))$, but $H(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = H(\tilde{\boldsymbol{\sigma}}, \phi(\boldsymbol{\sigma}))$ or $H(\tilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = H(\phi(\tilde{\boldsymbol{\sigma}}), \boldsymbol{\sigma})$ are not necessarily true.

The authors refine the centroid estimator as

$$\hat{\boldsymbol{\sigma}}_C = \rho\left(\text{argmin}_{\tilde{\boldsymbol{\sigma}} \in \{1,\ldots,K\}^n} \mathrm{E}_{\boldsymbol{\sigma}|\mathbf{A}}\left[H(\tilde{\boldsymbol{\sigma}}, \rho(\boldsymbol{\sigma}))\right]\right)$$

Posterior Estimators

In practice, we use the N posterior samples of $\boldsymbol{\sigma}$ to estimate $\mathrm{P}(\tilde{\boldsymbol{\sigma}} \mid \mathbf{A})$.

$$\hat{\mathrm{P}}(\sigma_i = k \mid \mathbf{A}) = \frac{1}{N} \sum_{t=1}^{N} I(\sigma_i^{(t)} = k)$$

Software

Peng and Carvalho implement their method in the `R` package `sbmlogit` provide in the supplementary materials of the 2016 paper.

Requires installation of the `igraph` `C` library (instructions here).

Still working through the learning this package.

Summary

In summary, Peng and Carvalho make important contributions:

1. A GLM approach is used to model edge placement in the SBM and implement degree correction. This establishes a generalizable and interpretable framework for including other covariates.

2. Polya-Gamma data augmentation is used in the logistic regression edge models to allow for closed form full conditionals of all model parameters.

3. A canonical mapping of community labels is proposed to avoid label switching.

4. A refined posterior estimator is proposed to go along with (3).