Community Detection in Stochastic Block Models

Carter Allen

12/04/2018

The Stochastic Block Model (SBM)

Brief Background

- Introduced in the social science literature (Holland et al., 1983) to model social networks
- Further developed by Nowicki & Snijers (2001)
- Authors sought to develop a less *ad hoc* way of modeling relational data
- Has been adopted in applications to network data in several domains
- Promising and flexible way to model gene networks, though not yet ubiquitous as a tool in biomedical applications

Community Detection vs. SBM

- Some authors draw a distinction between community detection and block modeling (McDaid 2012).
- While, SBM can be used to find communities, the results are often different than in pure community detection methods.
- Many community detection algorithms seek to maximize the intra-cluster edge density.
- SBMs can result in comparatively sparse clusters

Definitions

Original definition

A stochastic blockmodel is a special type of probability distribution over the space of adjacency arrays. - Holland (1983)

Modern definition

The stochastic block model (SBM) is a random graph model with planted clusters. It is widely employed as a canonical model to study the statistical and computational tradeoffs that arise in network and data sciences. - Abbe (2017)

Definitions

The SBM is encoded by a random adjacency matrix

$$\mathbf{A}_{n\times n} \sim SBM(n, \vec{Z}, \vec{\pi}, \mathbf{P})$$

- n: the number of nodes in the graph
- $\vec{Z}_{n\times 1} = (Z_1, ..., Z_n)^T$: The random community labels of each graph, where $Z_i \in \{1, ..., k\}$
- $\vec{\pi}_{k\times 1} = (\pi_1, ..., \pi_k)$: The probabilities governing community labeling, where $\pi_l = P(Z_i = l)$ for $i = 1, 2, ..., n$ and $l = 1, 2, ..k$
- $\mathbf{P}_{k\times k}$ The conditional probability matrix of edges. $A_{ij} \sim Bern.(P_{Z_i,Z_j})$, where $P_{Z_i,Z_j} = P(A_{ij} = 1|\vec{Z})$

Features of the SBM

- The graph defined by $\mathbf{A}_{n \times n}$ is *undirected* (i.e. $\mathbf{A}_{n \times n}$ is symmetric).
- Reflexive relations are not allowed (i.e. $A_{ij} = 0 \; \forall \; i = j$)
- There are $\frac{n(n-1)}{2}$ edges possible in a graph with n nodes.
- All nodes must belong to exactly one community. More advanced SBMs allow for mixed membership.

Features of the SBM

- The probability that two nodes are connected with an edge *depends only on the community membership of the two nodes*.
- In the case when $\mathbf{P}$ is constant ($P_{ij} = p \ \forall \ i, j$), the communities become meaningless.
- The ***planted partition*** model arised when P is compound symmetric. Here, the probability of an edge *within community* and the probability of an edge *between communities* is constant across communities.
- As $\mathbf{P} \to \mathbf{0}_{k \times k}$, $\mathbf{A}$ becomes sparse.

Features of the SBM

- Define ρ_n as the probability of an edge between *two randomly selected nodes* η_1 and η_2, members of communities a and b, respectively.

$$\rho_n = P(\{\eta_1 \in a\} \cap \{\eta_2 \in b\} \cap \{A_{\eta_1,\eta_2} = 1\})$$

$$= P(\{A_{\eta_1,\eta_2} = 1\} | \{\eta_1 \in a\} \cap \{\eta_2 \in b\})$$

$$= \sum_a \sum_b \pi_a \pi_b P_{ab}$$

- Note: ρ_n is a function of n, the number of nodes.
- Note: ρ_n depends on η_1 and η_2 only through their community memberships.

Features of the SBM

- Define λ_n as the expected degree of *one randomly selected node* η.

$$\lambda_n = \sum_{\eta' \neq \eta} E[A_{\eta',\eta}] = (n-1)\rho_n$$

- Define μ_n as the expected number of edges in a stochastic block model.

$$\mu_n = \sum_{i=1}^{n} \sum_{j>i}^{n} E[A_{i,j}] = \sum_{i=1}^{n} \frac{\lambda_n}{2} = \sum_{i=1}^{n} \frac{(n-1)\rho_n}{2}$$

$$= \frac{n(n-1)\rho_n}{2}$$

Visualizing SBMs

- Simple program for generating an observed graph from an underlying stochastic block model with two communities.
- This is an example of what data could be used as input to a SBM to recover communitiy memberships
- https://carter-allen.shinyapps.io/SBM2/

Defining the Likelihood

Derived from Bernoulli likelihoods

$$L(n, \vec{Z}, \vec{\pi}, \mathbf{P}|\mathbf{A}) = \prod_{i<j} (P_{Z_i,Z_j})^{A_{ij}} (1 - P_{Z_i,Z_j})^{1-A_{ij}} \prod_i \pi Z_i$$

$$\prod_{a \leq b} (P_{ab})^{O_{ab}(Z)} (1 - P_{ab})^{n_{ab}(Z) - O_{ab}(Z)} \prod_a \pi_a^{n_a(Z)}$$

Under a specific labeling Z, $O_{ab}(Z)$ is the number of edges between nodes labeled a and b, $n_{ab}(Z)$ is the number of possible edges between nodes labeled a and b, and $n_a(Z)$ is the number of nodes labeled a.

Bayesian Approach to SBMs

van der Pas & van der Vaart (2018)

- *Bayesian Community Detection* by van der Pas & van der Vaart (2018) extends the SBM literature by outlining how one can recover estimates of class labels in a Bayesian framework
- Main results of paper is presented in section 3.2
- Authors formally argue consistency of their Bayesian estimator
- Their method ***assumes k is known!***
- Redux: By placing priors on parameters in SBM and fixing k, obtain joint distribution $f(\mathbf{A}, \vec{Z}, \vec{pi}, \mathbf{P})$. Marginalize over $\vec{\pi}$ and $\mathbf{P}$ to obtain $f(\mathbf{A}, \vec{Z})$ and estimate $\vec{Z}$ from $f(\vec{Z}|\mathbf{A})$

Prior Structure

- The authors change notation from Z to e, reserving Z for the frequentist setting.

$$\pi \perp (P_{ab})$$

$$\pi \sim Dir(\alpha, ..., \alpha)$$

(often $\alpha = 1$)

$$P_{ab} \overset{iid}{\sim} Beta(\beta_1, \beta_2),\ 1 \leq i \leq j \leq k$$

$$e_i|\pi, P \overset{iid}{\sim} \pi, \quad 1 \leq i \leq n$$

$$A_{ij}|\pi, P, e \overset{indep.}{\sim} Bern.(P_{e_i}, P_{e_j}), \quad 1 \leq i \leq j \leq n$$

Posterior inference

The authors claim that

$$Q_B(e) = \frac{1}{n^2} \sum_{1 \leq a \leq b \leq K} logB(O_{ab}(e) + \beta_1, n_{ab}(e) - O_{ab}(e) + \beta_2)$$

$$+ \frac{1}{n^2} \sum_{a=1}^{K} log\Gamma(n_a(e) + \alpha) \propto p(e|\mathbf{A})$$

$$\Rightarrow \hat{e} = \underset{e}{\operatorname{argmax}} \; Q_B(e)$$

(i.e. Bayesian estimator is the posterior mode)

Computational Issues in SBMs

McDaid et al. (2013)

- Little detail is given in van der Pas & van der Vaart (2018) as to how computation is performed
- Authors refer the reader to *Improved Bayesian inference for the stochastic block model with application to large networks* by McDaid et al. (2013).
- An effecient algorithm in `C++` for estimating ***both*** the number of clusters and community membership

https://sites.google.com/site/aaronmcdaid/sbm

Computational issues

- Letting `K` be decided by the data introduces complexity
- MCMC is now concerned with estimating Z and K
- Searching over a space whose dimension depends on K can be challenging

Applying SBMs to Network Augmentation

Refresh on Network Augmentation

Several possible sources of information for learning about relationships between genes:

1) Manually curated database such as ***KEGG*** (Kyoto Encyclopedia of Genes and Genomes)
   - Reliable/validated baseline information
   - Difficult to scale
2) Literature mining database such as ***GAIL***
   - Easily scalable
   - Can suggest previously un-investigated relationships

Snowball method

```r
snowball <- function(core, n.iter = 5, crit.quantile = 0.99)
    {
        crit.val = quantile(cos_sims$cos,probs = crit.quantile)
        network = core
        for(i in 1:n.iter)
            {
                edgecounts = cos_sims %>%
                    filter(gene1 %in% network & gene2 %in% network == FALSE) %>%

                    mutate(edge = ifelse(cos > crit.val,1,0)) %>%
                    group_by(gene2) %>%
                    summarize(n_edges = sum(edge),avg_cos = mean(cos)) %>%
                    arrange(desc(n_edges))

                top_candidate = edgecounts %>%
                    top_n(1,wt = n_edges) %>%
                    pull(gene2) %>%
                    as.character() %>%
                    unname()

                network = c(network,top_candidate)
                print(paste("Iter:",i,"Added",top_candidate,"to network"))
            }

    }
```

Network Augmentation with SBMs

- One key issue: ***the stochastic block model does not propose new members***
- The SBM can estimate community structure ***based only on observed interconnectivity of edges in a network***
- One possible solution
  1) Fit SBM to core network (e.g. KEGG pathway)
  2) Use `snowball.R` to suggest new members via cosine similarity
  3) Refit SBM and observe community membership of new member

Network Augmentation with SBMs

- Another possible approach

  1) Generate list of potential members *a priori*
  2) Let $Z_i \in \{1, 2\} \; \forall \; i = 1, 2, ..., n$ (i.e. two possible classes)
  3) Place strong priors on $Z_i \; \forall \; i \in \mathbf{C}$, where $\mathbf{C}$ is core set.
  4) Place priors on remaining candidates proportional to their average connectivity to $\mathbf{C}$
  5) Observe adjacency matrix A after some number of interations of `snowball.R`
  6) Fit SBM to observed A under such prior structure

Weighted SBMs

- Work has been done by Christopher Aicher of University of Colorado, Boulder, and others, to incorporate edge weights in the stochastic block model

References on WSBMs:

1) *Adapting the Stochastic Block Model to Edge-Weighted Networks*, Aicher, C. et al. 2013.
2) *Learning Latent Block Structure in Weighted Networks*, Aicher, C. et al. Journal of Complex Networks. 2014.

These models might allow us to incorporate cosine similarity information as edge weights.

Summary

Summary

- SBMs provide a promising and flexible framework for modeling network data
- Some work will need to be done to develop a method suitable for GAIL data
- Next steps are to running models on test data and observing performance

References

- *Bayesian Community Detection*. S. L. van der Pas & A. W. van der Vaart. Bayesian Analysis. 2018.
- *Stochastic Block Models, First Steps*. Holland, P et al. Social Networks. 1983.
- *Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates*. Leger, J.-B. Journal of Statistical Software. 2016.
- *Adapting the Stochastic Block Model to Edge-Weighted Networks*, Aicher, C. et al. 2013.
- *Learning Latent Block Structure in Weighted Networks*, Aicher, C. et al. Journal of Complex Networks. 2014.