Project Proposal

John Koo

Proposal

In this project, we will use Mean Field Variational Inference (MFVI) to estimate a simple two-community Bayesian Popularity Adjusted Block Model (PABM), which is a generalization of the Stochastic Block Model (SBM). A Bayesian version of the SBM can be described by the following generative model:

1. Define:
    - n, the number of vertices
    - K, the number of communities
    - $\pi_1, ..., \pi_K$, the community probabilities
    - a_1, b_1, a_2, b_2, priors for edge probabilities
2. Draw community memberships $Z_1, ..., Z_n \stackrel{iid}{\sim} Multinomial(\vec{\pi})$
3. Draw
    - Within-community edge probability $p \sim Beta(a_1, b_1)$
    - Between-community edge probability $q \sim Beta(a_2, b_2)$
4. For $i, j = 1, ..., n$ and $i < j$, draw the adjacency matrix:
    - If i, j in the same community, $A_{ij} \mid \vec{z}, p \sim Bernoulli(p)$
    - If i, j in different communities, $A_{ij} \mid \vec{z}, q \sim Bernoulli(q)$
    - $A_{ji} = A_{ij}$ and $A_{ii} = 0$
    - Alternatively, let $P \in [0, 1]^{n \times n}$ be the edge probability matrix. Then $P_{ij} = p$ if i, j are in the same community and $P_{ij} = q$ if i, j are in different communities. Then $A_{ij} \mid P_{ij}, \vec{z} \stackrel{indep}{\sim} Bernoulli(P_{ij})$ for $i < j$ (and then make A symmetric and hollow).

For additional details and a MFVI solution, see Zhang and Zhou [2]. For a slightly more complicated model and a Gibbs sampler for that model, see Koo [1].

For the PABM version, we would have to expand the edge probability matrix to a wider set of values than just two. More specifically, each vertex has K values assigned to it, each one representing its affinity toward each community. For instance, in the two-community case, vertex 5 will have two values, $\lambda_{5,1}$ (vertex 5's affinity toward community 1) and $\lambda_{5,2}$ (vertex 5's affinity toward community 2). Then if both vertex 5 and vertex 7 (which has corresponding values $\lambda_{7,1}, \lambda_{7,2}$) are in community 1, $P_{5,7} = \lambda_{5,1}\lambda_{7,1}$ and $A_{5,7} \sim Bernoulli(\lambda_{5,1}\lambda_{7,1})$. On the other hand, if vertex 8 is in community 2, then $P_{5,8} = \lambda_{5,2}\lambda_{8,1}$ and $A_{5,8} \sim Bernoulli(\lambda_{5,2}\lambda_{8,1})$. To modify the generative model for the SBM for the PABM, we would then have to say $\lambda_{ik} \sim Beta(a_{ik}, b_{ik})$ for each $i = 1, ..., n$ and $k = 1, ..., K$. To simplify this, we will limit the scope of this project to $K = 2$, and we will say $\lambda_{i,k} \sim Beta(a_1, b_1)$ if i belongs to community k and $\lambda_{i,k} \sim Beta(a_2, b_2)$ if i does not belong to community k. The MFVI approximation will be decomposing the full joint distribution:

$$p(\vec{z}, P|A) \propto p(A, \vec{z}, P) \approx q(\vec{z}, P) \approx q(\vec{z}) \prod_{i,k} q(\lambda_{ik})$$

References

[1] John Koo. Bayesian graph partitioning. https://github.com/johneverettkoo/stats-hw/tree/master/stats626, 2018.

[2] Anderson Y. Zhang and Harrison H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection, 2017.