Latent Community Adaptive Network Regression

Heather Mathews[1], Alexander Volfovsky[1]

[1]Department of Statistical Science, Duke University

Objective

Given a network, Y, and covariates, X, can we make inference about effects of X on connections in Y? We aim to efficiently estimate latent community structure that may be influencing effects of covariates. We condition on these structures since nodes belonging to different communities may be impacted by a particular covariate in different ways. Here, we allow for estimation of coefficient effects based on estimated latent community structure.

Introduction

Networks allow us to investigate connections between people and what drives those connections. For example, we might be interested in what impacts a students' ability to make friends. To adequately model impacts of covariates on a network, we must account for the dependencies in our data and potential latent structures (here, specifically latent community structure):

- Let's say we observe some covariate: Grades
- We suspect that there is latent community structure: High school cliques
- If we want to infer how grades impact connections, β, the effect of grades on connections might be different for each community; thus we would want to allow for each community to have a different coefficient estimate for that covariate

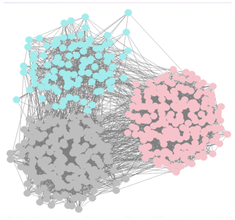

Figure 1. Example Network

How to Model Networks with Covariate Information: AMEN

First, we define the Additive and Multiplicative Effects Network Regression (AMEN) Model [1]:

$$Y_{i,j} = 1_{Z_{i,j}>0} \tag{1}$$

$$Z = \beta_0 11^T + \sum_{p=1}^{P}(X_{r,p}\beta_{r,p} + X_{c,p}\beta_{c,p}) + UV^T + a1^T + 1b^T + \epsilon \tag{2}$$

- X_r, X_c: Observed covariate information (r represents row covariates and c represents column covariates, P number of covariates)
- β: Coefficients of interest estimating covariate effects on connections
- a, b: Individual row (sociability) and column (popularity) random effects
- U, V: Latent factor matrices of rank R (latent multiplicative effects)
- Z: Response representing latent network structure
- $Y_{i,j}$: Observed binary response indicating whether nodes i and j are connected (Y is directed). Also, this can be extended to the fixed rank nomination likelihood. This means that individuals rank a fixed number of people as friends so Y is no longer binary.

Estimating Latent Community Structure

We alter the above model slightly. We focus on when we believe our latent multiplicative effects have an underlying community structure. As such, let's suppose we believe there exist K latent communities. We now define our multiplicative effects as $U\Lambda U^T$ (rather than UV^T). Now, U is an $n \times K$ membership matrix with each row, U_i, indicating the community a node belongs to and Λ being a $K \times K$ matrix describing the relationship between communities. To sample U and Λ, we do the following:

- Apply spectral clustering methods to Y in order to estimate the membership of each node
- Use these memberships to form an initial estimate of U
- Start our sampling at our initial estimate of U and then update memberships periodically using a Metropolis step

To sample Λ, we place a normal prior on it and sample using its full conditional. Note that this is much more computationally efficient as we are only updating a few nodes of U at each iteration and a $K \times K$ matrix rather than all entries for the original latent multiplicative effects from (2).

Estimating $\tilde{\beta}$: Allowing for community dependent β

If we have latent community structure influencing connections in our network, it is *intuitive* that the βs have community structure as well. In order to accommodate for this, we alter (2) such that:

$$Z = \beta_0 11^T + \sum_{p=1}^{P}(Diag(\tilde{\beta}_{r,p}U^T)X_{r,p} + X_{c,p}Diag(\tilde{\beta}_{c,p}U^T)) + U\Lambda U^T + a1^T + 1b^T + \epsilon \tag{3}$$

where $\tilde{\beta}_r$,$\tilde{\beta}_c$ are $P \times K$ matrices of coefficients representing row and column covariate estimates for each community. If we were estimating the coefficient for a covariate, and we have 3 latent communities, then we would have 3 β estimates associated with that single covariate.

Relationship between β in (2) and $\tilde{\beta}$ in (3)

Can we express the $P \times 1$ β as a function of the $P \times K$ matrix $\tilde{\beta}$? We consider a simplified version of (2) and (3) where U is known and $H_r = \{1_n \otimes [I_n \circ X_r]U\}$:

$$Z_{vec} = X_{vec}\beta_r + \epsilon_{vec} \text{ vs. } Z_{vec} = H_r\tilde{\beta}_r + \epsilon_{vec} \tag{4}$$

Theorem Let n_K be equal for all $k \in \{1, .., K\}$. Let the sample community variances be equal for each community. Then $\frac{\sum_{j=1}^{K}\hat{\tilde{\beta}}_{r,j}}{K} \overset{d}{=} \hat{\beta}_r$.

Proof. Using properties of ordinary least squares, we get the asymptotic distribution for $\hat{\beta}_r$ and $\hat{\tilde{\beta}}_r$:

$$\sqrt{n^2}(\hat{\beta}_r - \beta_r) \to N(0, n^2(X_{vec}^T X_{vec})^{-1}) \text{ and } \sqrt{n^2}(\hat{\tilde{\beta}}_r - \tilde{\beta}_r) \to N(0, n^2(H_r^T H_r)^{-1}) \tag{5}$$

By the continuous mapping theorem and Slutsky's theorem, $\frac{\sum_{j=1}^{K}\hat{\tilde{\beta}}_{r,j}}{K} \overset{d}{=} \hat{\beta}_r$. (i.e let $h(a_1, ..., a_K) = \frac{\sum_{i=1}^{K} a_i}{K}$ and compute the variance $\dot{h}(\tilde{\beta}_r)\Sigma\dot{h}(\tilde{\beta}_r)^T$).

Simulations/Results

We consider 2 simulations where we generate data from model (3). We fit various models on these simulated networks to compare posterior credible intervals for our covariate coefficients. We generate a network with $n = 150$, and there is an underlying community structure in which we have $K = 3$ communities.

1. Generate data such that we have one covariate with and without community dependence (binary likelihood, left panel)
2. Generate data from model (3) but with a fixed rank nomination structure (right panel).

β: Estimation using AMEN (not allowing for community dependence) (2)

- We consider AMEN models with no multiplicative effects ($R = 0$) as well as with multiplicative effects ($R = 3$ and MCMC Update). For $R = 3$, we allow for full estimation of UV^T. Note that our MH update of U yields estimates close to those of AMEN with $R = 3$
- For both simulations, we get estimates close to the mean of true community estimates; thus showing how ignoring community dependence can lead to poor inference

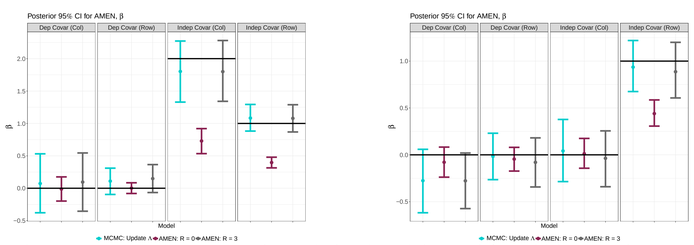

Figure 2: AMEN Estimation. MH update is indicative of the model in which we update Λ and U using a Metropolis update. The horizontal lines represent the true community means: $\frac{\sum_{i=1}^{K}\tilde{\beta}_{r,i}}{K}$, $\frac{\sum_{i=1}^{K}\tilde{\beta}_{c,i}}{K}$. (left: Binary network, right: FRN network)

Simulations/Results

Naive Approach: Partition data based on true community membership

- Here, we suppose that we know U. We partition the data by community and run a separate AMEN model for each. This leads to information loss, and our estimates are not quite accurate

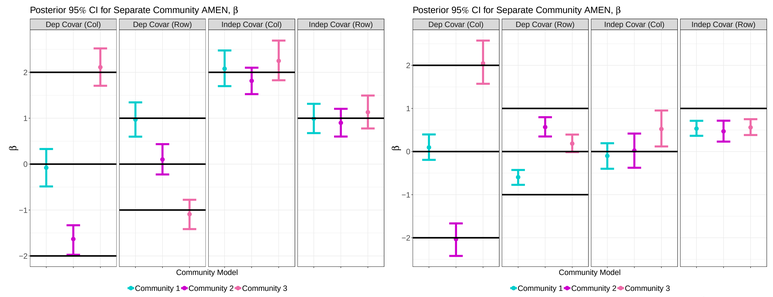

Figure 3. Run separate model on partitioned network. Horizontal black lines indicate true $\tilde{\beta}_k$ values (left: Binary network, right: FRN network)

$\tilde{\beta}$: Estimation of β dependent on latent communities (3)

- Now, we use our proposed model that allows for community dependent coefficient estimation
- We are able to estimate $\tilde{\beta}$ dependent on community membership and our true $\tilde{\beta}$ values are covered by our 95% credible intervals (figure 4). Also, this method allows for the most accurate inference on our effects of covariates. It also allows us to interpret the latent structure, U.

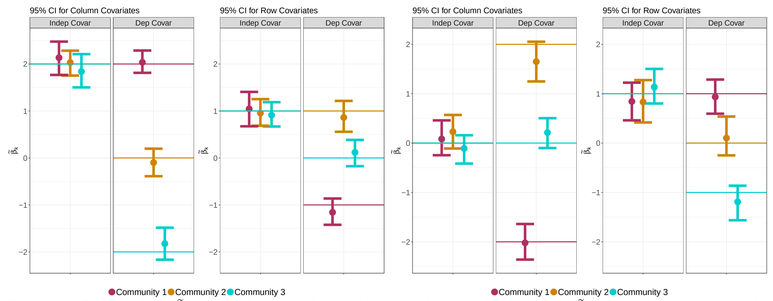

Figure 4: Dependent $\tilde{\beta}$ estimation. Horizontal lines indicate true value of $\tilde{\beta}_k$. (left: Binary network, right: FRN network)

Conclusions

- Allowing for community dependent covariate coefficient estimation leads to more interpretable and meaningful inference when community structure is present as we can determine specific effects of covariates on particular communities
- When we do not account for these dependencies, single estimation of β yields inaccurate estimates or estimates close to the community averages
- Our method allows for efficient estimation and updating of latent community structure as well as using these learned communities to improve inference on what we are interested in: effects of covariates on connections in a network

References

[1] Hoff, Fosdick, Volfovsky, and Stovel. Likelihoods for fixed rank nomination networks. *Network Science*, 1(03):253--277, 2013.

[2] P. Hoff. Additive and multiplicative effects network models. *arXiv preprint arXiv:1807.08038*, 2018.

Contact Info

heather.mathews@duke.edu

Duke UNIVERSITY