Final Project Proposal

Nana Wang
Department of Statistics
University of California, Davis
Davis, CA 95616
nnawang@ucdavis.edu

1 Optimization problem

In networks, it is reasonable to have a block dependence model. Let $Y = (Y_{i,j})_{i,j=1,\cdots,N}$ be the adjacency matrix of a network. Assume there are K known blocks, and $p_{k,l}$ denotes the probability of observing an edge between a node in block k and a node in block l. The block model says $Y_{i,j} \sim \text{Bernoulli}(p_{z[i],z[j]})$ independently where $z[i] \in \{1, \cdots, K\}$ denotes the label of the neighborhood of node i. The dependence between blocks is modelled as $\eta_k = x_k^t \beta + \epsilon_k$, where $\eta_k = \log(\frac{p_{k,k}}{1-p_{k,k}})$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_k)^t \sim N(0, \Sigma)$. The p-dimensional vector covariates x_k, which is assumed to be known, describes some properties of neigborhood k and β is a p-dimensional parameter vector. Moreover, $p_{k,l}$ is independent of all the other $p_{s,t}$ for $(k,l) \neq (s,t)$.

One of the most important tasks is estimating the partial correlation between blocks, that is estimation of $D = \Sigma^{-1}$. Denote $\eta_k = \log(\frac{\hat{p}_{k,k}}{1-\hat{p}_{k,k}})$, where $\hat{p}_{k,k} = \frac{\sum_{z[i]=z[j]=k} Y_{i,j}}{n_k}$ and n_k is the number of all the possible edges in block k. We assume n iid observed networks $Y^{(i)}$ with unobserved iid $p^{(i)}$ (or iid $\eta^{(i)}$). The statistics $\hat{\eta}^{(i)}$ are thus also iid. By using the lasso method, we can get

$$\hat{\theta}^{a,\lambda} = \underset{\theta:\theta_a=0}{\arg\min} \left(n^{-1} \|\hat{\boldsymbol{\eta}}_a - \hat{\boldsymbol{\eta}}\theta\|_2^2 + \lambda \|\theta\|_1 \right). \tag{1}$$

Then based on $\{\hat{\theta}^{a,\lambda} : a \in \{1, \cdots, K\}\}$, we have the estimation of edge set E, denoted by $\hat{E}^{\lambda}$. In order to get a good estimator of E, it is important have a good estimator, $\hat{\theta}_a^{\lambda}$. In the final project, my optimization problem is solving the minimization equation (1).

2 Method

The 'lars' [1] and 'glmnet' are two well-known R packages for solving the lasso problem. One of the interesting thing is to implement the algorithms myself, then compare it with the two packages.

Right now, I still do not have a good data for the block dependent model. For the final project, I will first simulate the data from the model. Then apply my implementation and the two packages to the simulate data and compare the results.

3 Research plan

Before Nov.17, I will finish reading the material of the algorithms. Then I will complete the algorithm first in R before Nov.28. Before Dec. 5, I will finish writing my code in Julia.

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning (pp. 485-585). Springer New York.