STA6800 - Statistical Analysis of Network Stochastic Block Models

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

Introduction

- Blockmodeling: To partition the vertex set into subsets called blocks in such a way that the block structure and the pattern of edges between the blocks capture the main structural features of the graph.
- Priori blockmodeling: The blocks are known.
- Posteriori blockmodeling: The blocks have to be inferred from the edge pattern.
- We extended the concept of blockmodeling to stochastic version.

Stochastic Blockmodel

1. Adjacency matrix Y
   - $y_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices i and j} \\ 0 & \text{otherwise} \end{cases}$
   - $y_{ii} = 0$ for all i
   - $y_{ij} = y_{ji}$

2. Color vector (or block structure) $\boldsymbol{x}$
   - $x_i = k$ if vertex i has color k
   - If i and h belong to the same block, then they relate to the other vertices in the same way

Stochastic Blockmodel

- Under stochastic blockmodel, the probability that an edge is present between two vertices depends only on the colors of the vertices
- The random colors are iid with $Pr(X_i = k) = \theta_k$
- Conditional on the vertex colors, edges are independent, with $Y_{ij} \sim Bernoulli(\eta_{X_i, X_j})$
- $P(y, x; \theta, \eta) = \theta_1^{n_1} \cdots \theta_m^{n_m} \prod_{1 \le k \le l \le m} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}}$
- Edges often refer to a relation which is more frequent within blocks than between blocks ($\eta_{kk} > \eta_{kl}$). (But there exist exceptions : $\eta_{kl} > \eta_{kk}$)

Stochastic Blockmodel

- We consider a structure (Y,x) assuming that only the relational structure Y can be observed; that is, the color vector x is unobserved (latent).
- Given this a posteriori blockmodel, we wish to estimate the parameters and predict the unobserved coloring x.
- The probability of observing edge pattern y can be written as $Pr(y|\theta,\eta) = \sum_x Pr(y,x|\theta,\eta) = \sum_x Pr(y|x,\theta,\eta) \times Pr(x)$.

Identifiability Problems and Invariant Parameters

- A model is identifiable if different values of the parameters generate different probability distributions for **Y**.
- SBM is **not identifiable** because the partition of vertices does not change in appropriate ways when the parameters are transformed in a compatible way.
- This kind of non-identifiability occurs in all finite mixture and latent class models, especially using a Bayesian estimation.
- A nonuniform prior distribution can identify the class labels and thereby circumvent the problems caused by nonidentifiability.

Identifiability Problems and Invariant Parameters

- If there is no identifying prior information, consider the posterior distributions of functions of $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{X})$ that are invariant w.r.t relabeling the classes.
- Examples of functions of $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{X})$
  - $\theta_{X_i} = \sum_{k=1}^{c} \theta_k I\{X_i = k\}$
  - $\eta(X_i, X_j) = \sum_{1 \le k,h \le c} \eta(k,h) I\{X_i = k, X_j = h\}$
  - Matrix of the posterior predictive probabilities, $\{Pr(X_i = X_j \mid \boldsymbol{y})\}_{1 \le i \ne j \le n}$

Maximum Likelihood Estimation

- Because of the intractable form of the likelihood function, explicit formulae for the maximum likelihood estimators cannot be obtained.

- We must use numerical methods for maximizing the likelihood function.
  1. The direct maximization
  2. The EM algorithm

Maximum Likelihood Estimation

- The EM algorithm(for calculation of mle with missing data)
  - E: Compute $Q(\theta, \eta|\theta^{(p)}, \eta^{(p)})$
  - M: Choose $(\theta^{(p+1)}, \eta^{(p+1)})$ that maximizes $Q(\theta, \eta|\theta^{(p)}, \eta^{(p)})$
  - Iteration scheme converges to the maximum likelihood estimate
  - EM algorithm is considerably faster than the direct maximization

Bayesian Estimation

- Bayesian estimators and the posterior standard deviations of the parameters can approximate the maximum likelihood estimators for θ and η and their standard errors.
- Gibbs sampling
  1. Sample $\boldsymbol{\theta}^{(p+1)}, \boldsymbol{\eta}^{(p+1)}$ from $Pr(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{X}^{(p)}, \boldsymbol{y})$
  2. Sample $X_i^{(p+1)}$ from

$$Pr(X_i \mid \boldsymbol{\theta}^{(p+1)}, \boldsymbol{\eta}^{(p+1)}, \boldsymbol{y}, X_1^{(p+1)}, ..., X_{i-1}^{(p+1)}, X_{i+1}^{(p)}, ..., X_n^{(p)})$$

- Unlike MLE, bayesian estimation is also practical for larger graphs.

Asymptotic Recovery of Colors

- One of the main goals in posterior blockmodeling is to recover (or predict) the colors x_i from the observation of the edge pattern y.
- It turns out that asymptotically for $n \to \infty$, it is possible, under certain weak conditions, to recover the colors x_i correctly with probability tending to 1.
- This property will be called *the asymptotically correct distinction of vertex colors*.
- This finding is important because it implies that statistical inference in a posteriori blockmodeling can be almost as good as inference in a priori block modeling.

Asymptotic Recovery of Colors

Procedure 1.

- A latent two-blockmodel situation
- $y_{(i)+}$: the ordered degrees.
- I: the index i maximize $y_{(i+1)+} - y_{(i)+}$ for $1 \leq i \leq n-1$.
- $D = y_{(I)+}$
- $F_i(y) = \begin{cases} 1 & \text{if } y_{i+} \leq D; \\ 2 & \text{if } y_{i+} > D. \end{cases}$
- Let $n \to \infty$, and assume that $n_2/n \to \theta \in (0,1)$,

$$\theta\eta_{12} + (1-\theta)\eta_{11} < \theta\eta_{22} + (1-\theta)\eta_{12}$$

.

- Then, $P(X_i = F_i(Y) \text{ for } i = 1, ..., n \mid X = x) \to 1$.

Asymptotic Recovery of Colors

Procedure 2.

- Define matrix C, where $C_{ij} = \sum_{h \neq i,j} Y_{hi} Y_{hj}$, dyad-wise partnership.
- $\boldsymbol{s}$: the ordered vertex numbers, $\boldsymbol{d}$: an ordering vector of C_{ij}.
- Suppose that we obtained $s_1, ..., s_i$ and $d_1, ..., d_i$.
- To determine s_{i+1}, calculate
  $C^*_{ij} = \min(C_{hj} \mid h \in \{s_1, ..., s_i\})$, for $j \notin \{s_1, ..., s_i\}$.
- Choose r maximizing C^*_{ir}, then, $s_{i+1} = r$, $d_{i+1} = C^*_{ir}$.
- Split at t maximizing $d_{t+1} - d_t$.
- This second procedure works better in practice than the first procedure.

Color Prediction for Finite n

- The procedures of the previous section are not necessarily satisfactory for small and intermediate values of n.
- Profile predictive likelihood
  - $L_p(\boldsymbol{x} \mid \boldsymbol{y}) = \sup_{\theta,\eta} P(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\eta})$
  - Iterating 2 steps,
    1. Find $\hat{\theta}^{mle}$, $\hat{\boldsymbol{\eta}}^{mle}$.
    2. Find set of X maximizing $P(\boldsymbol{y}, \boldsymbol{x}; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$.

Color Prediction for Finite n

- Conditional predictive likelihood
  - $L_c(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta})}{P(t(\boldsymbol{y},\boldsymbol{x});\boldsymbol{\theta},\boldsymbol{\eta})}$.
  - The minimal sufficient statistic $t(\boldsymbol{y}, \boldsymbol{x}) = (N_2, E_{11}, E_{22}, E_{12})$.
- Bayesian approach
  - The relative frequencies of $\{X_i^{(p)} = k\}$ for $k = 1, 2$ are Monte Carlo estimates of the Bayesian posterior predictive probabilities for $X_i = k$.
  - A block structure $\boldsymbol{x}$ yielded from the profile predictive likelihood method, is same with the result of Gibbs sampler .

Example: Hansell's Student Data

- Data
  - Sociomatrix of friendship among 13 male and 14 female sixth-graders.
  - Assume the presence of friendship between two students whenever at least one of the two students expressed liking for the other. (Symmetric)
  - The adjacency matrix presents the data blocked according to sex.
- Investigate whether the partitioning of the class into two subgroups based on **friendship ties** differs from the partitioning based exclusively on **gender**.

Example: Hansell's Student Data

- Applying the procedure 2 of slide 13, obtain the block structure,

$$\boldsymbol{x}^{(1)} = (11112211111112222222122122 11).$$

- Using $\boldsymbol{x}^{(1)}$ as the starting point of the Gibbs sampler led to the posterior means:

$$\hat{\theta} = 0.479,\ \hat{\eta}_{11} = 0.326,\ \hat{\eta}_{12} = 0.248,\ \hat{\eta}_{22} = 0.763.$$

- Standard errors of the estimated parameters:

$$S.E.(\hat{\theta}) = 0.122, \quad S.E.(\hat{\eta}_{11}) = 0.062,$$
$$S.E(\hat{\eta}_{12}) = 0.073, \quad S.E.(\hat{\eta}_{22}) = 0.080.$$

Example: Hansell's Student Data

Result

- Students 5, 6, 14-19, and 21-25 belong to block 2.
- 1-4, 7-13, 20, and 26-27 belong to block 1.
- The posterior blocking does not entirely follow gender lines because 5, 6, 20, 26, and 27 are the exceptions.
- The adjacency matrix shows that these students indeed have different friendship patterns from the others.
- 5,6 have relatively many female friends.
- 20 has more male than female friends.
- 26, 27 have few friends of either gender.

Discussion

- This model assumes that
  - the probability distribution of the relation between two vertices depends only on the latent classes to which the vertices belong
  - the relations are independent conditionally on these classes.
- Extend the model with observed covariates.
- Adopt more complex conditional dependence assumptions for entries in **Y** conditional on the latent classes **X**.
- Drop the assumption that the colors X_i are iid, and replace this by a more complex model for **X**, expressing ”emergent” social roes such as leader/follower or center/periphery.

Introduction

- One of the ways to analyze relational data is *Clustering*
  - grouping the objects to uncover a structure based on the observed patterns of interactions
- the latent stochastic blockmodel (Wang and Wong, 1987; Snijders and Nowicki, 1997)
  - Each object belongs to a cluster and the relationships between objects are governed by the corresponding pair of clusters.
  - limitation
    : each object can only belong to one cluster, or in other words, play a single latent role

Introduction

- However, many relational data sets are multi-facet.
  - (EX) When a social actor interacts with different partners, different social contexts may apply and thus the actor may be acting according to different latent roles they can possible play.
- In this paper, develop a *mixed membership model* for relational data
  - Relax the assumption of single-latent-role for actors
  - Associate each unit of observation with multiple clusters via a membership probability-like vector
  - Assume that the ensemble of mixed membership vectors help govern the relationships of each object

Setting

- Observed relational data : graph $G = (\mathcal{N}, Y)$
- $Y(p, q)$: pairs of nodes to values $Y(p, q) \in \{0, 1\}$
- Directed (binary) edges: Positive responses to survey questions about a specific sociometric relation
- N: # of monks in the monastery
- K: # of distinct groups a monk can belong to

Setting

- A randomly drawn vector $\vec{\pi}_i$
  - $\pi_{i,g}$: the probability of monk i belonging to group g
- A matrix of Bernoulli rates $B_{K\times K}$: the probabilities of interactions between different groups
  - $B(g,h)$: the probability of having a link between a monk from group g and a monk from group h
- The indicator vector $\vec{z}_{p\to q}$: The group membership of monk p when he responds to survey questions about monk q
- The indicator vector $\vec{z}_{p\leftarrow q}$: The group membership of monk q when he responds to survey questions about monk p
- $\{\vec{z}_{p\to q} : p,q \in \mathcal{N}\} =: Z_{\to}$ and $\{\vec{z}_{p\leftarrow q} : p,q \in \mathcal{N}\} =: Z_{\leftarrow}$

Mixed Membership Stochastic Blockmodel

1. For each node $p \in \mathcal{N}$:
   - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim \textit{Dirichlet}(\vec{\alpha})$
2. For each pair of nodes $(p, q) \in \mathcal{N} \times \mathcal{N}$:
   - Draw membership indicator for the initiator $\vec{z}_{p \to q} \sim \text{Multinom}(\vec{\pi}_p)$
   - Draw membership indicator for the receiver $\vec{z}_{p \leftarrow q} \sim \text{Multinom}(\vec{\pi}_q)$
   - Sample the value of their interaction,

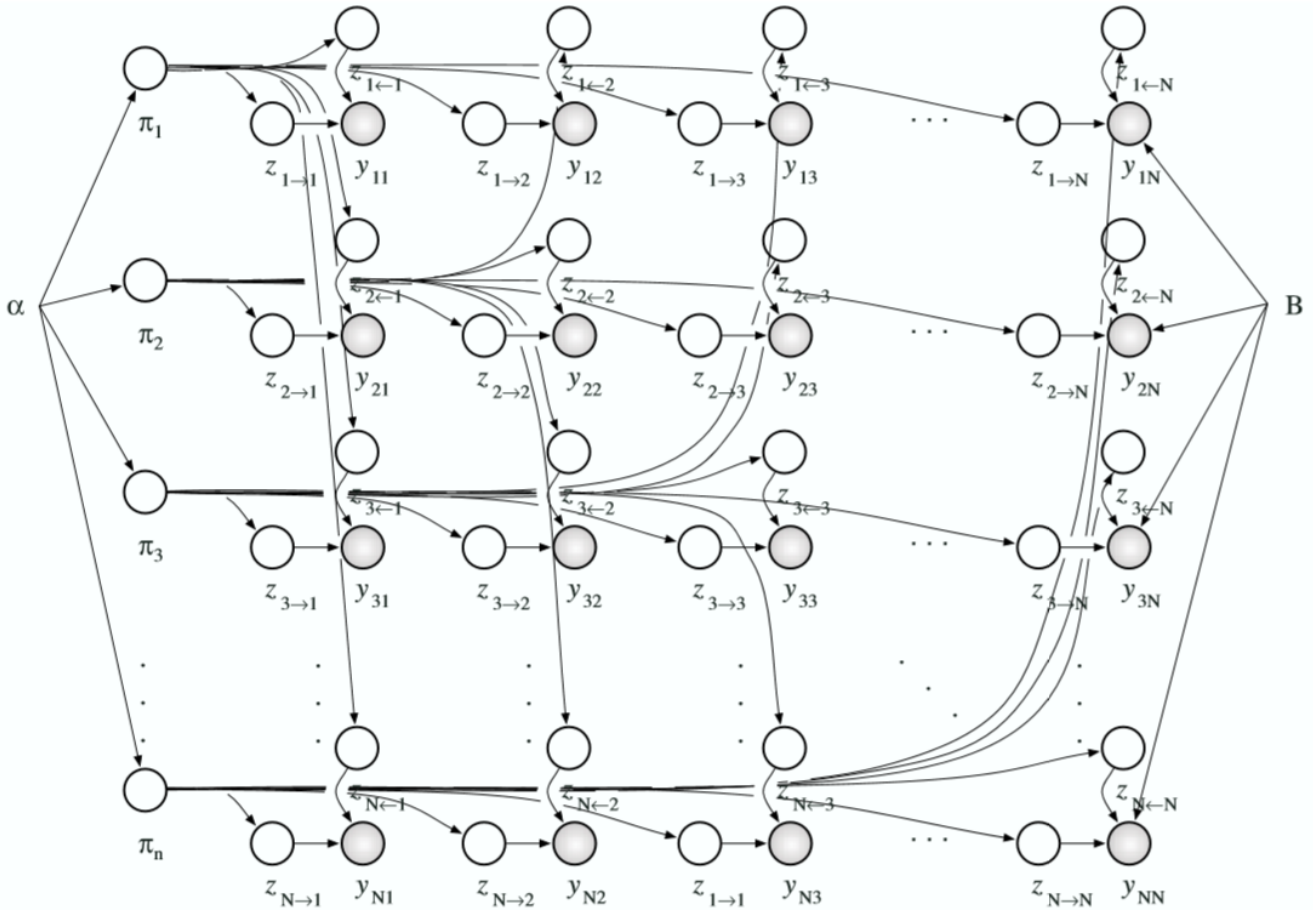
$$Y(p, q) \sim \text{Bernoulli}(\vec{z}_{p \to q}^T B \vec{z}_{p \leftarrow q})$$

Mixed Membership Stochastic Blockmodel

- The **joint probability** of the data and the latent variables

$$p(Y, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \vec{\alpha}, B) = \prod_{p,q} \Big\{ P(Y(p,q) | \vec{z}_{p \rightarrow q}, \vec{z}_{p \leftarrow q}, B) \times P(\vec{z}_{p \rightarrow q} | \vec{\pi}_p) P(\vec{z}_{p \leftarrow q} | \vec{\pi}_q) \Big\} \prod_p P(\vec{\pi}_p | \vec{\alpha})$$

Graphical Model



Context Dependent

- The group membership of each node is *context dependent.*
- Each node may assume different membership when interacting to or being interacted by different peers

Modeling Sparsity

- Adjacency matrices encoding binary pairwise measurements are often **sparse**. They contain many zeros or non-interactions
- Distinguish two sources of non-interaction
  1. The rarity of interactions in general
  2. The pair of relevant blocks rarely interact

Modeling Sparsity

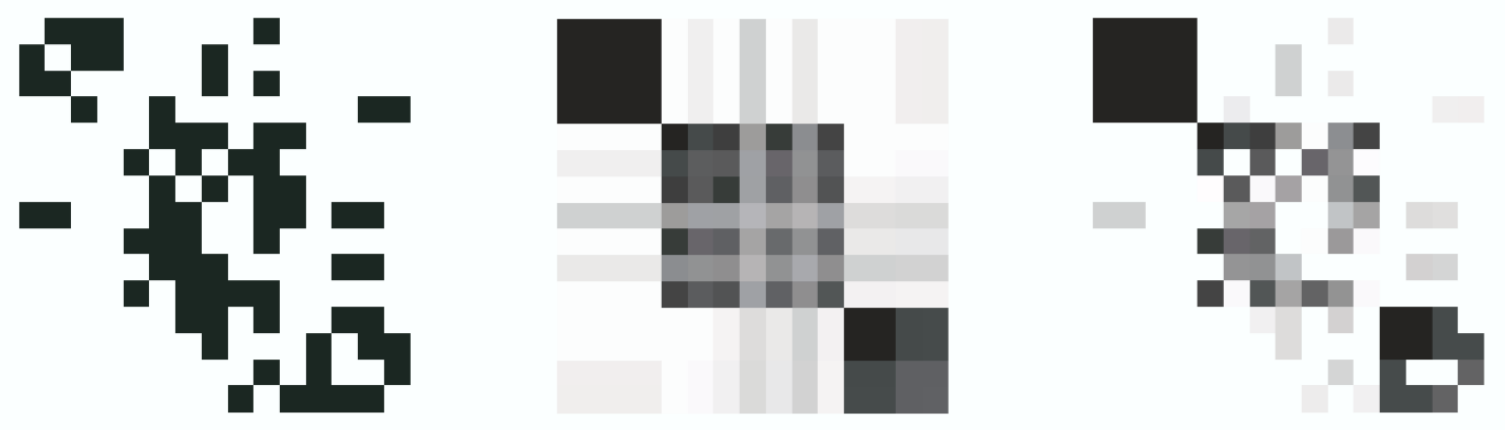
- Introduce a sparsity parameter $\rho \in [0, 1]$ in the MMSB to characterize the source of non-interaction
- Assume that the probability of a non-interaction comes from a mixture, $1 - \sigma_{pq} = (1 - \rho) \cdot \vec{z}^T_{p \to q}(1 - B)\vec{z}_{p \leftarrow q} + \rho$
- The weight ρ capture the portion zeros that should not be explained by the blockmodel B
- Down-weight the probability of successful interaction

$$\vec{z}^T_{p \to q} B \vec{z}_{p \leftarrow q} \Rightarrow (1 - \rho) \cdot \vec{z}^T_{p \to q} B \vec{z}_{p \leftarrow q}$$

Two Types of Data Analysis

- Distinguish two types of data analysis that can be performed with the MMSB
  1. Summarize the data, Y, in terms of the global blockmodel, B, and the node-specific mixed memberships, Πs
     $\Rightarrow$ the amount of relational information that is captured by in $\hat{\alpha}, \hat{B}$, and $\mathbb{E}[\vec{\pi}|Y]$ leads to a coarse reconstruction of the original sociomatrix.
  2. de-noise the data Y, in terms of the global blokmodel, B, and interaction-specific single memberships, Zs
     $\Rightarrow$ the amount of relational information that is revealed by $\hat{\alpha}, \hat{B}, \mathbb{E}[\vec{\pi}|Y]$, and $\mathbb{E}[Z_{\rightarrow}, Z_{\leftarrow}|Y]$ leads to a finer reconstruction of the original sociomatrix, Y.

Monk Data



Brief Introduction of VB

- Goal: To compute the posterior distribution of the latent variables given a collection of observations
- Why?: The variational distribution is simpler than the true posterior so that the optimization can be approximately solved.

1. Posit a distribution of the latent variables with free parameters.
2. Fit those parameters such that the distribution is close in Kullback-Leibler divergence to the true posterior.
3. Iterate until the convergence occurs.

Naive vs. Nested

- *Naive* variational algorithm often converged only after too many iterations. *Nested* variational algorithm can improve convergence.
- Dependence between $\vec{\gamma}_{1:N}$ & B should be considered.
- Nested algorithm can maintain the dependence by keeping the block of free parameters optimized given the other variational parameters.

Pseudo-code

1. initialize $\vec{\gamma}^0_{pk} = \frac{2N}{K}$ for all p, k
2. **repeat**
3. **for** $p = 1$ to N
4. **for** $q = 1$ to N
5. get **variational** $\vec{\phi}^{t+1}_{p \to q}$ and $\vec{\phi}^{t+1}_{p \leftarrow q} = f\,(\,Y(p,q), \vec{\gamma}^t_p, \vec{\gamma}^t_q, B^t\,)$
6. partially update $\vec{\gamma}^{t+1}_p$, $\vec{\gamma}^{t+1}_q$ and B^{t+1}
7. **until** convergence

5.1. initialize $\phi^0_{p \to q,g} = \phi^0_{p \leftarrow q,h} = \frac{1}{K}$ for all g, h
5.2. **repeat**
5.3. **for** $g = 1$ to K
5.4. update $\phi^{s+1}_{p \to q} \propto f_1\,(\,\vec{\phi}^s_{p \leftarrow q}, \vec{\gamma}_p, B\,)$
5.5. normalize $\vec{\phi}^{s+1}_{p \to q}$ to sum to 1
5.6. **for** $h = 1$ to K
5.7. update $\phi^{s+1}_{p \leftarrow q} \propto f_2\,(\,\vec{\phi}^s_{p \to q}, \vec{\gamma}_q, B\,)$
5.8. normalize $\vec{\phi}^{s+1}_{p \leftarrow q}$ to sum to 1
5.9. **until** convergence

Variational EM Algorithm

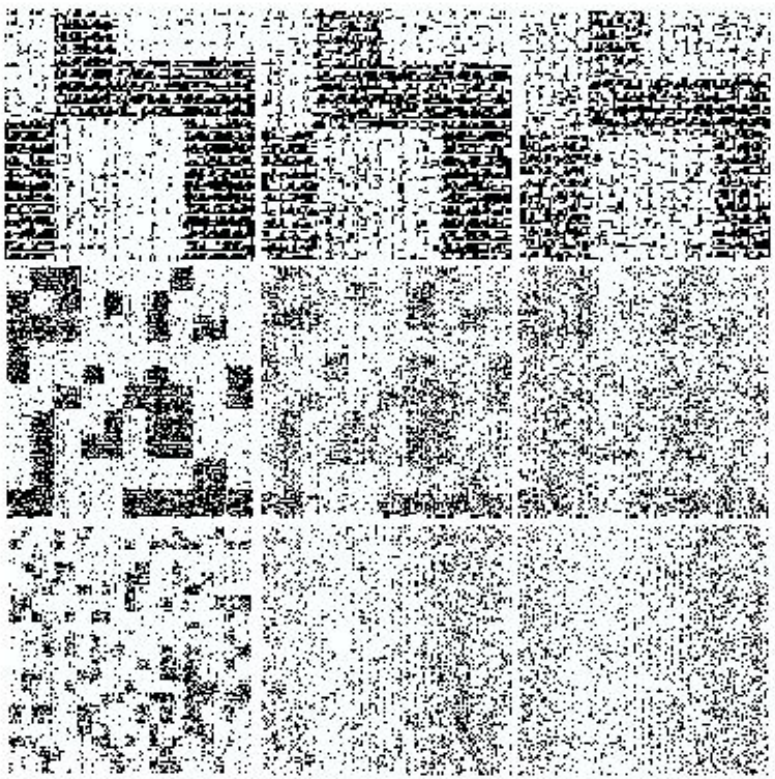
- Goal: To find parameters α,B,ρ corresponding to local optimum of the bound.
- Variational EM uses the lower bound as a surrogate likelihood.
- Procedure:
  1. E step: Fit the variational distribution q to approximate posterior
  2. M step: Maximize the corresponding bound with respect to the parameters
- A closed form solution for the $\hat{\alpha}$ does not exist. We use Newton-Raphson method to estimate this parameter.

Selecting the Optimal K

- 2 strategies for choosing the optimal K:
  1. Large networks: Select K corresponding to the highest averaged held-out likelihood in cross-validation
  2. Small networks: Select K using approximation to BIC

$$BIC = 2 \cdot \log p(Y) \approx 2 \cdot \log p(Y|\hat{\vec{\pi}}, \hat{\vec{Z}}, \hat{\vec{\alpha}}, \hat{\vec{B}}) - |\vec{\alpha}, B| \cdot \log|Y|$$

Simulation Example



- Sampled graphs of 100,300,600 nodes from blockmodels with 4,10,20 clusters (α = 0.05, 0.1, 0.25)

Simulation Example

- Results:
  1. K=10 was selected by CV when nodes # = 300
  2. Variational EM successfully recovers both B & $\vec{\pi}_{1:N}$
  3. As α increases, each node is likely to belong to more clusters
  4. Nested algorithm converges faster to its peak than the naive algorithm

Friendship Network

- A questionnaire was administered to a sample of students who were allowed to nominate up to 10 friends
- Measured the student's grade of membership, and the mixed membership vectors provide a mapping between grades & blocks.

Friendship Network

Blockmodel

$$\hat{B} = \begin{bmatrix} 0.3235 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3614 & 0.0002 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2607 & 0.0 & 0.0 & 0.0002 \\ 0.0 & 0.0 & 0.0 & 0.3751 & 0.0009 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0002 & 0.3795 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3719 \end{bmatrix}$$

Posterior Mixed Membership Scores

1. Better goodness-of-fit than other previous models (via correspondence - grade & block)
2. Extra-flexibility of MMSB reduces bias in the prediction of membership of students

Protein Data

- Goal: To reveal the proteins' diverse functional roles by analyzing their local and global patterns of interaction
- Considered physical interactions among 871 proteins in yeast (MIPS protein data)
- Mapped the 871 proteins in our collections to the high-level functions of the MIPS annotation tree

Protein Data: Direct Correspondence

- Goal: Explore the direct correspondence between protein-specific mixed memberships to blocks ($\vec{\pi}_{1:871}$) & functional annotations ($\vec{b}_{1:871}$)
- Estimate a permutation of the components of $\vec{\pi}_n$ to find out specific functional annotations $\vec{b}_n$ (using training set / K=15)

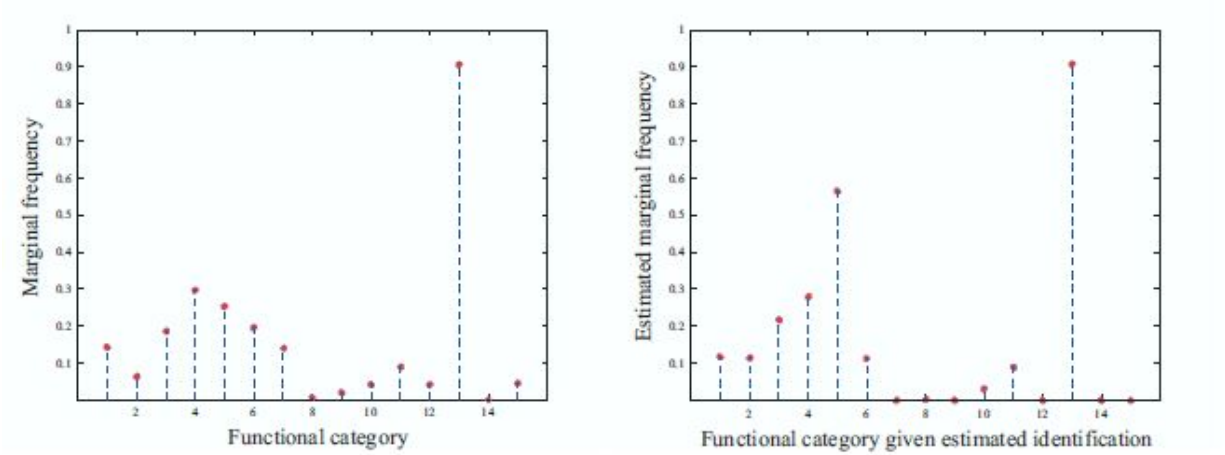
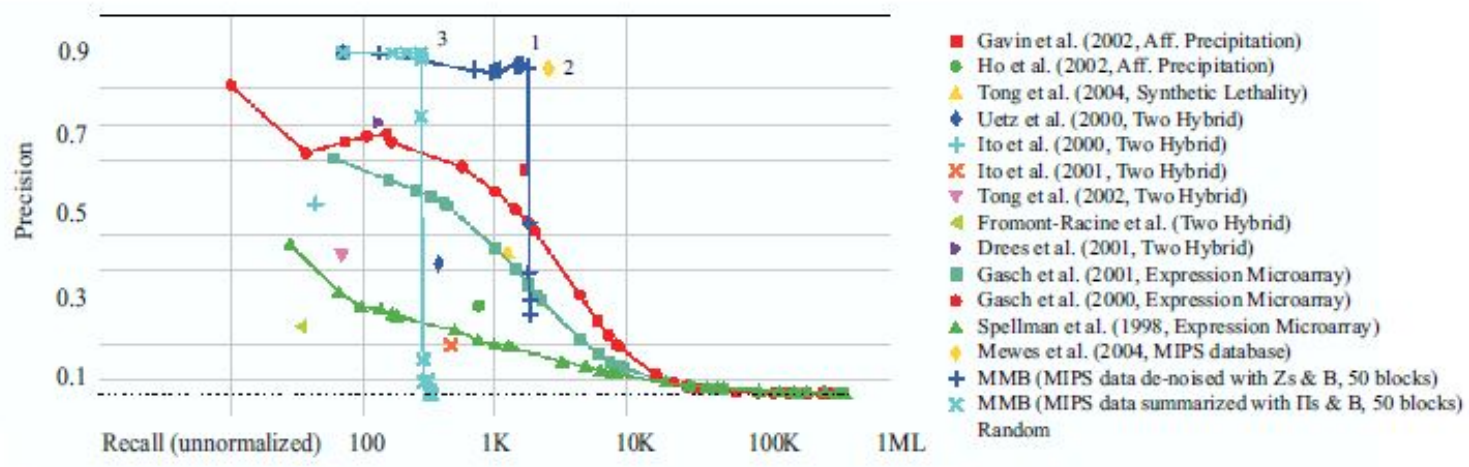


Figure: Accuracy of the predicted annotations

Protein Data: Indirect Correspondence

- Selected the mixed membership blockmodel best for predicting out-of-sample interactions (K=50 via CV)
- MMSB successfully reduces the dimensionality of the data while revealing substantive information about functionality



Discussion

- MMSB vs. LSM:
  1. Distribution over the latent vectors - Dirichlet vs. Gaussian
  2. VB vs. MCMC
- Limitation: In a simulation setting, the model does not readily generate *hubs*
- MMSB generalizes to 2 cases:
  1. Multiple data collections on the same objects can be generated by the same latent vectors
  2. B can be a matrix that parameterizes any kind of distribution

Conclusion

- Can assume MULTIPLE memberships per node
- 2 Analysis: Summarize & De-noise
- Variational methods for estimation
- Can explain the real world network better than previous models