Mixed Membership Stochastic Blockmodels

Journal of Machine Learning Research, 2008

by E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing
as interpreted by Ted Westling

STAT 572 Update Talk
May 8, 2014

Overview

1. Review
2. Variational Bayes: General Theory
3. Variational Bayes for the MMSB
4. Conclusion and next steps
5. Issues

Overview

1. Review

2. Variational Bayes: General Theory

3. Variational Bayes for the MMSB

4. Conclusion and next steps

5. Issues

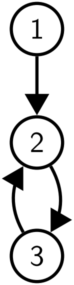
Network theory: notation

- Individuals $p, q \in \{1, \ldots, N\}$.
- We observe relations/interactions $R(p, q)$ on *pairs* of individuals.
- Here we assume $R(p, q) \in \{0, 1\}$, $R(p, p) = 0$, but do not assume $R(p, q) = R(q, p)$ (we deal with *directed* networks).

Network theory: data representations

p	q	R(p,q)
1	2	1
2	3	1
3	2	1

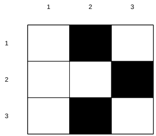
Table



Graph

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Adjacency matrix i, j element is $R(i, j)$

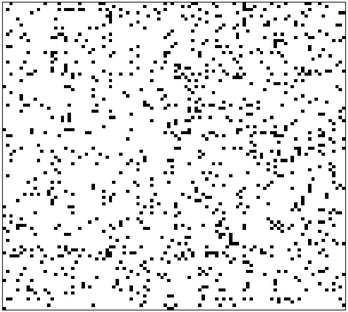


Adjacency matrix, black=1, white=0

The Problem: Scientific Motivation

- We believe the relations are a function of unobserved groupings among the individuals.
- We want to recover the groups so we can a) predict new relations or b) interpret the existing network structure.
- Example: Monk network.

The Problem: Pictures



The Problem: Pictures

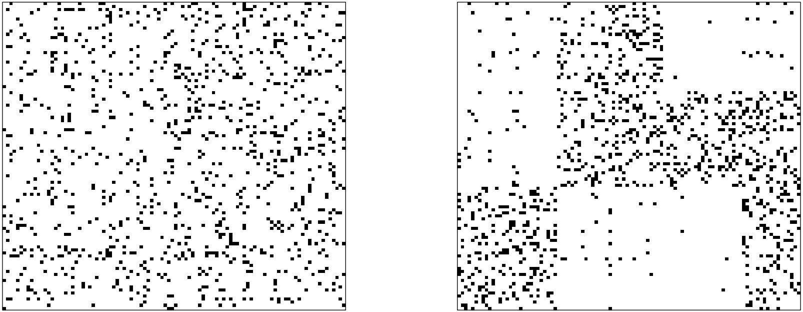


Figure: Two visualizations of the same binary adjacency matrix. Each filled-in square represents a directed edge. Left: ordered randomly. Right: ordered by group membership.

Brief blockmodel history

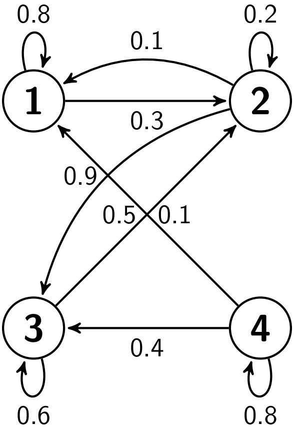
- 1975: CONCOR developed by Harrison White and colleagues
- 1983: Holland, Laskey & Leinhardt introduce *stochastic* blockmodel for blocks known *a priori*.
- 1987: Wasserman & Anderson extend to *a posteriori* estimation.
- 2004: Kemp, Griffiths & Tenenbaum allow unknown and unlimited number of blocks in the Infinite Relational Model.

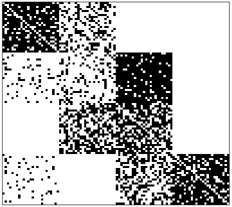
Infinite Relational Model

- Observe binary relations $R(p, q)$ between nodes $p, q \in \{1, \ldots, N\}$.
- Each node p is a member of exactly one block of K total blocks, $K \leq N$ unknown. Let z_p be an indicator vector of block membership for node p, i.e. $z_p = (0, 1, 0)$.
- B is a $K \times K$ matrix of block relationships. If p is in block g and q is in block h then the probability of observing an interaction from node p to node q is B_{gh}.
- $R(p, q) \sim \text{Bernoulli}(z_p^T B z_q)$.
- For example, if p is in block 3 and q is in block 2 then $P(R(p, q) = 1) = B_{32}$.

Block structure

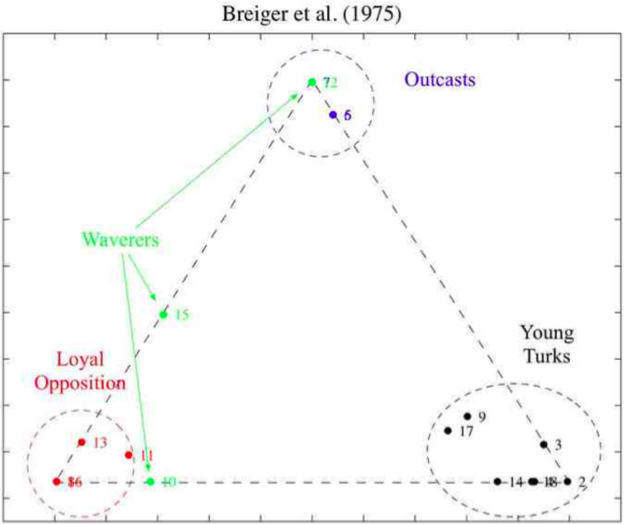
$$\begin{pmatrix} .8 & .3 & 0 & 0 \\ .1 & .2 & .9 & 0 \\ 0 & .5 & .6 & 0 \\ .1 & 0 & .4 & .8 \end{pmatrix}$$





The Mixed Membership Stochastic Blockmodel

- Previous models assume each node is assumed to belong to exactly one latent block - e.g. $z_p = (0, 1, 0, 0)$.
- Instead, in the MMB we assume each node has a distribution π_p over the latent blocks.
- For each interaction from p to q, both p and q draw a particular block to be a part of for the interaction: $z_{p \to q} \sim \text{Discrete}(\pi_p)$, $z_{p \leftarrow q} \sim \text{Discrete}(\pi_q)$.
- Then $R(p, q) \sim \text{Bernoulli}(z_{p \to q}^T B z_{p \leftarrow q})$.
- K chosen by BIC.



1 Ambrose
2 Boniface
3 Mark
4 Winfrid
5 Elias
6 Basil
7 Simplicius
8 Berthold
9 John Bosco
10 Victor
11 Bonaventure
12 Amand
13 Louis
14 Albert
15 Ramuald
16 Peter
17 Gregory
18 Hugh

Overview

Estimation Basics

- Strategy: treat $\{\pi, Z_{\rightarrow}, Z_{\leftarrow}\} \equiv \theta$ as random latent variables and obtain posterior distribution. Treat $\{\alpha, B\} \equiv \beta$ as fixed parameters to estimate via Empirical Bayes.
- The typical approach in this setting is to use the EM algorithm, which involves calculating the posterior distribution $p(\theta | Y, \beta)$.

Posterior Calculation

- Great! Write down the form of the posterior $p(\theta|Y,\beta)$:

$$\frac{p(Y|\theta,\beta)p(\theta|\beta)}{p(Y|\beta)}.$$

- The denominator requires calculating the integral

$$\int_{\Pi}\sum_{\Omega}\prod_{p,q}\left[P(Y(p,q)|z_{p\rightarrow q},z_{p\leftarrow q},B)p(z_{p\rightarrow q}|\pi_p)p(z_{p\leftarrow q}|\pi_q)\right]$$
$$\prod_{p}p(\pi_p|\alpha)d\pi_{1:N}$$

- No closed form solution

Variational Bayes

- Main idea: write down a simple parametric form $q(\theta|\Delta)$ for the posterior distribution that depends on free variational parameters Δ.
- At each E-step, minimize the KL divergence between q and the true posterior in terms of the free variational parameters.

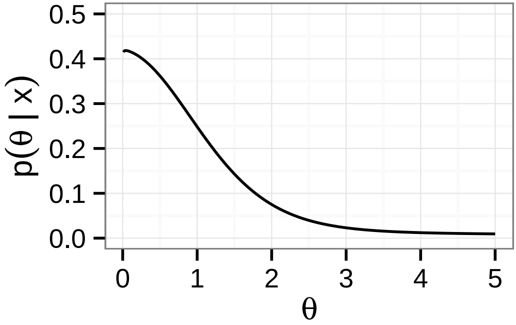
$$\begin{aligned}
K(q,p) &= \mathbb{E}_q\left[\log \frac{q(\theta|\Delta)}{p(\theta|Y,\beta)}\right] \\
&= \mathbb{E}_q\left[\log q(\theta|\Delta)\right] - \mathbb{E}_q\left[\log p(\theta|Y,\beta)\right] \\
&= \mathbb{E}_q\left[\log q(\theta|\Delta)\right] - \mathbb{E}_q\left[\log \frac{p(\theta, Y|\beta)}{p(Y|\beta)}\right] \\
&= \mathbb{E}_q\left[\log q(\theta|\Delta)\right] - \mathbb{E}_q\left[\log p(\theta, Y|\beta)\right] + \log p(Y|\beta).
\end{aligned}$$

This is equivalent to maximizing the *evidence lower bound* or ELBO

$$\mathcal{L}(\Delta|Y,\beta) = \mathbb{E}_q\left[\log p(\theta, Y|\beta)\right] - \mathbb{E}_q\left[\log q(\theta|\Delta)\right].$$

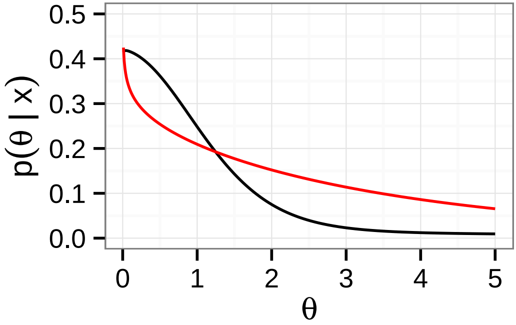
Variational Bayes: Example

Suppose we have a complicated posterior distribution p (the one below is a mix of lognormal and t - yuck). We use a variational posterior Gamma(α, β). We minimize the KL divergence between the true and variational posteriors in terms of α and β to get an approximate posterior.



Variational Bayes: Example

Suppose we have a complicated posterior distribution p (the one below is a mix of lognormal and t - yuck). We use a variational posterior Gamma(α, β). We minimize the KL divergence between the true and variational posteriors in terms of α and β to get an approximate posterior.



Variational Bayes: Mean-Field Approximation

- How do we know what sort of variational posterior q is easy to work with?
- Most popular strategy is called the mean field approximation: assume q factorizes into a term for each latent variable:

$$q(\theta_1, \ldots, \theta_r | \Delta) = \prod_{i=1}^{r} q_i(\theta_i | \Delta_i).$$

  The optimal form for q_i can be derived from the calculations of variations (hence the name variational inference).
- To minimize the KL divergence (maximize the ELBO) between p and and q, we use coordinate descent over the Δ_i.

Variational EM algorithm

1. Initialize $\beta^{(0)}$, $\Delta^{(0)}$.
2. E-step: Find the $\Delta_{1:r}^{(j)}$ that maximizes $\mathcal{L}(\Delta, \beta^{(j-1)})$ via coordinate ascent:
   (i) For $i = 1, \ldots, r$, maximize $\mathcal{L}(\Delta^{(jk)}, \beta^{(j-1)})$ with respect to Δ_i.
   (ii) If $\|\Delta^{(jk)} - \Delta^{(j(k+1))}\| > \epsilon$ has not converged, return to (i)
3. M-step: Find the $\beta^{(j)}$ that maximizes $\mathcal{L}(\Delta^{(j)}, \beta)$.
4. If $\|\Delta^{(j)} - \Delta^{(j-1)}\| > \epsilon$ or $\|\beta^{(j)} - \beta^{(j-1)}\| > \epsilon$, return to 2.

Overview

Mean-field approximation

- Recall $\theta = \{\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}\}$
- Recall the model distribution for θ:
  - $\pi_p \sim \text{Dirichlet}(\alpha)$
  - $z_{p\rightarrow q} \sim \text{Discrete}(\pi_p)$
  - $z_{p\leftarrow q} \sim \text{Discrete}(\pi_q)$
- Assume the posterior factors as:
  - $\pi_p \sim \text{Dirichlet}(\gamma_p)$
  - $z_{p\rightarrow q} \sim \text{Discrete}(\phi_{p\rightarrow q})$
  - $z_{p\leftarrow q} \sim \text{Discrete}(\phi_{p\leftarrow q})$
- Variation parameters $\Delta = \{\gamma_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}\}$
- Full approximate posterior:

$$q(\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \gamma_p, \Phi_{\rightarrow}, \Phi_{\leftarrow}) = \prod_p q_1(\pi_p|\gamma_p) \prod_{p,q} q_2(z_{p\rightarrow q}|\phi_{p\rightarrow q}) q_2(z_{p\leftarrow q}|\phi_{p\leftarrow q}).$$

Parameter updates

- Now we have to calculate the ELBO

$$
\begin{aligned}
\mathcal{L}(\Delta|Y,\beta) &= \mathbb{E}_q\left[\log p(\theta, Y|\beta)\right] - \mathbb{E}_q\left[\log q(\theta|\Delta)\right] \\
&= \mathbb{E}_q\left[\log p(\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}, Y|\alpha, B)\right] \\
&\qquad - \mathbb{E}_q\left[\log q(\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}|\gamma_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow})\right] \\
&= \cdots \\
&= \mathcal{L}(\gamma_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}|Y, \alpha, B)
\end{aligned}
$$

  (appendix calculations)
- Then differentiate with respect to γ_i, $\phi_{p\rightarrow q}, \phi_{p\leftarrow q}$, α_i, $B(g,h)$, set to 0, and solve.

Parameter updates

- Get closed form updates:

$$\hat{\phi}_{p\rightarrow q,g} \propto e^{\psi(\gamma_{p,g})-\psi(\Sigma_j \gamma_{p,j})} \prod_h \left(B(g,h)^{Y(p,q)} (1 - B(g,h))^{1-Y(p,q)} \right)^{\phi_{p\leftarrow q}}$$

$$\hat{\phi}_{p\leftarrow q,h} \propto e^{\psi(\gamma_{q,h})-\psi(\Sigma_j \gamma_{q,j})} \prod_g \left(B(g,h)^{Y(p,q)} (1 - B(g,h))^{1-Y(p,q)} \right)^{\phi_{p\rightarrow q}}$$

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_q \phi_{p\rightarrow q,k} + \sum_q \phi_{q\leftarrow p,k}$$

$$\hat{B}(g,h) = \frac{\sum_{p,q} \phi_{p\rightarrow q,g} \phi_{p\leftarrow q,h} Y(p,q)}{\sum_{p,q} \phi_{p\rightarrow q,g} \phi_{p\leftarrow q,h}}$$

- No closed form solution for α, so we use a Newton-Raphson algorithm to find maximum:

$$\alpha^{(t+1)} = \alpha^{(t)} - (H_\alpha(\mathcal{L}(\alpha^{(t)})))^{-1} \nabla_\alpha L(\alpha^{(t)})$$

where $\nabla_\alpha L(\alpha^{(t)})$ is the gradient (vector of first derivatives) of the ELBO $\mathcal{L}$ with respect to α and $H_\alpha(\mathcal{L}(\alpha^{(t)}))$ is the hessian (matrix of second derivatives) of the $\mathcal{L}$ with respect to α:

$$\frac{\partial}{\partial \alpha_k} \mathcal{L} = N \left[\psi(\Sigma_j \alpha_j) - \psi(\alpha_k)\right] + \sum_p (\psi(\gamma_{p,k}) - \psi(\Sigma_j \gamma_{p,j}))$$

$$\frac{\partial^2}{\partial \alpha_k \partial \alpha_l} \mathcal{L} = N \left[\psi'(\Sigma_j \alpha_j) - \psi'(\alpha_k) \mathbf{1}_{k=l}\right]$$

"Naive" algorithm for the MMSB

1. Initialize $B^{(0)}, \alpha^{(0)}, \gamma_{1:N}^{(0)}, \Phi_{\rightarrow}^{(0)}, \Phi_{\leftarrow}^{(0)}$.
2. E-step: Find the $\gamma_{1:N}^{(j)}, \Phi_{\rightarrow}^{(j)}, \Phi_{\leftarrow}^{(j)}$ that maximizes $\mathcal{L}(\cdot, \alpha^{(j-1)}, B^{(j-1)})$ via coordinate ascent:
   - (i) Update $\gamma_i^{(j)}$ for $i = 1, \ldots, N$
   - (ii) Update $\phi_{p\rightarrow q}, \phi_{p\leftarrow q}$ for all p, q
   - (iii) Until convergence
3. M-step: Update $B^{(j)}, \alpha^{(j)}$.
4. Until convergence.

"Nested" algorithm for the MMSB

1. Initialize $B^{(0)}, \alpha^{(0)}, \gamma_{1:N}^{(0)}, \Phi_{\rightarrow}^{(0)}, \Phi_{\leftarrow}^{(0)}$.
2. E-step:
   (a) for each p, q:
      (i) Update $\phi_{p\rightarrow q}, \phi_{p\leftarrow q}$ for all p, q
      (ii) Update γ_p, γ_q
      (iii) Update B.
   (b) Until convergence
3. M-step: Update $\alpha^{(j)}$.
4. Until convergence.

Overview

Conclusion and next steps

- Variational inference: approximate the posterior with an easier closed-form posterior; make the approximation as good as possible
- Pros: scales very well, get an exact posterior rather than samples. Cons: Posterior is approximate, and we don't know how good (or bad) the approximation is.
- Next steps: apply to data, simulations

Overview

Sparsity parameter

- The authors introduce a sparsity parameter ρ
- Instead of $Y(p,q) \sim z_{p\to q}^T B z_{p\leftarrow q}$, set $Y(p,q) \sim z_{p\to q}^T (1-\rho) B z_{p\leftarrow q}$.
- Identifiability?