Lesson 5: Non parametric Bayes and application to relational model

Richard Yi Da Xu

School of Computing & Communication, UTS

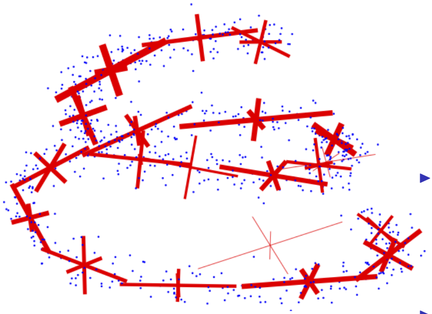
June 10, 2015

Non parametric Bayes - Some background knowledge

- Getting Harder: introduciton to Dirichlet Process - A machine learning research topic
- Also see MCMC used in practice
- Get the chance to study a real modelling application - Relational Model

Dirichlet Process: A diagrammatic representation

Rasmussen, Infinite Gaussian Mixture Model (1999):



- For a mixture model:
  Let $\mathbf{X} = x_1, \dots, x_N$:

$$P(\mathbf{X}|\theta_1, \dots \theta_K, w_1, \dots w_K) = \sum_{l=1}^{K} w_l f(\mathbf{X}|\theta_l)$$

where $\sum_{l=1}^{K} w_l = 1$

- If we allow K to also vary, what happens if you want to:

$$\arg\max_{\theta_1,\dots\theta_K, w_1,\dots,w_K, K} P(\mathbf{X}|\theta_1, \dots \theta_K, w_1, \dots w_K, K)?$$

- K = N for Gaussian case. Of course it's not desirable!

Dirichlet Process: Motivation

- For data $x_1, \dots, x_N$, each x_i is associating with a parameter θ_i
- We need to a good prior for $\Pr(\theta_1 \dots \theta_N)$:
- You also want K potentially be infinite
- A "clustering" property, controllable through a single parameter α
- Let's define it using Hierarchical prior, its marginal is:

$$p(\theta_1, \dots \theta_n) = \int_G \Pr(\theta_1, \dots, \theta_n | G)\mathbf{p}(\mathbf{G})$$

So, we are interested in the property of G:

- G needs to be **discrete** random distribution.
- Perhaps it should also some resemblence with some basic distribution H.

Dirichlet Process Definition

We say G is a Dirichlet process, distributed with base distribution H and concentration parameter α:

$$G \sim DP(\alpha, H), \text{if}$$
$$(G(A1), ..., G(Ar)) \sim \text{Dir}(\alpha H(A1), ..., \alpha H(Ar))$$

for every finite measurable partition $A_1, ..., Ar$ of Θ.
What does this all mean? Let's visualise it!

Beta Process

- G is a Beta process, with base distribution H and concentration parameter α:

$$G \sim BP(\alpha H), \text{if}$$
$$G(A_k) \sim \text{Beta}(\alpha H(A_k), \alpha(1 - H(A_k)))$$

- Given an infinitesimal partition $(A_1, \dots, A_K)$ with $K \to \infty$ and $H(A_k) \to 0$ the samples correspond to the density function:

$$G = \sum \pi_i \delta_{\theta_i}$$
$$\text{where} \qquad \pi_i \sim \text{Beta}(0, \alpha)$$
$$\theta_i \sim H$$

- Beta process is a Completely Random Measure with Levy measure on product space $[0, 1] \times \Omega$ with Levy measure:

$$\nu(\text{d}\pi d\theta) = \alpha\pi^{-1}(1 - \pi)^{\alpha-1}\text{d}\pi H(\text{d}\theta).$$

Gamma Process

- Γ is a Gamma process, distributed with base distribution H and concentration parameter α:
- Given a partition $(A_1, \ldots, A_K), A_i \in \Omega \implies G(A_i) \sim \text{Gamma}(H(A_i), 1/\alpha)$
- Let $\Gamma = \{(\pi_i, \theta_i)\}_{i=1}^{\infty}$ be a realization of a Gamma process in product space $\mathbb{R}^+ \times \Theta$:

$$
\begin{aligned}
& \Gamma \sim GaP(\alpha, H) \\
& \quad = \sum \pi_i \delta_{\theta_i} \\
\text{where} \qquad & \pi_i \sim \text{Gamma}(0, \alpha) \\
& \theta_i \sim H
\end{aligned}
$$

- Gamma process is a Completely Random Measure with Levy measure:

$$
\nu(\mathrm{d}\pi \mathrm{d}\theta) = \pi^{-1} \exp^{-\alpha\pi} \mathrm{d}\pi H(\mathrm{d}\theta)
$$

- $G \sim DP(\alpha, H)$ and N data points, the probability of K is:

$$\Pr(K = k|N, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}|s(N,k)|\alpha^k, \qquad k = 0, 1, \dots, N$$

- This means that,

$$\sum_{k=1}^{N} |s(N,k)|\alpha^k = \frac{\Gamma(N+\alpha)}{\Gamma(\alpha)}$$

- it can be sampled as $k = \sum_{n=1}^{N} b_n, \qquad b_n \sim \text{Bernoulli}\left(\frac{\alpha}{n-1+\alpha}\right)$

Negative Binomial Process

- X is a Negative Binomial Process with base measure H and another measure $\mathcal{P}$:
- Let $X = \{(n_i, \theta_i)\}_{i=1}^{\infty}$ be a realization of a Gamma process in the product space $\mathbb{Z}^+ \times \Theta$:

$$
\begin{aligned}
X &\sim \text{NBP}(\mathcal{P}, H) \\
&= \sum n_i \delta_{\theta_i}
\end{aligned}
$$

(From probability notes) Relationship between Multinomial distribution and Poisson

$$\text{Poisson}(x|\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda) \qquad \text{Mult}(n_1, \dots, n_k | p_1, \dots p_k) = \frac{(\sum n_i)!}{n_1! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

suppose:

- $x_1 \sim \text{Poisson}(x|\lambda_1), \dots, x_k \sim \text{Poisson}(x|\lambda_k) \implies$
- The above generated two random variables:

1st random variable: $\left(n = \sum_{i=1}^{k} x_i \right) \sim \text{Poisson}(\lambda_1 + \lambda_2 + \dots + \lambda_k)$

2nd random variable: $\mathbf{x} = (x_1, \dots, x_k)|n \sim \text{Mult}(n, p_1, \dots p_k)$ where $p_i = \dfrac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}$

Extend this Relationship to **Process**

- Grouped data $x_1, \dots x_J$ for any measurable disjoint partition $A_1, \dots A_Q$ of Ω,
- Jointly model the count random variables $\{X_j(A_q)\}$.
- Poisson process $X_j \sim \text{PP}(G)$, with a shared Completely Random Measure G on $\Omega : X_j(A) \sim Pois(G(A))$
- $X_j \sim \text{PP}(G)$

  $\equiv X_j \sim \text{MP}(X_j(\Omega), \tilde{G}), \qquad X_j(\Omega) \sim Pois(G(\Omega)) \qquad \text{where } \tilde{G} = \frac{G}{G(\Omega)}$

$$
\begin{aligned}
X_j &\sim \text{NBP}\left(G_0, \frac{1}{c+1}\right) = \int_G \text{PP}(X_j|G)\text{GaP}(c, G_0)\text{d}G \\
&\sim \text{NBP}\left(G_0, p\right) = \int_G \text{PP}(X_j|G)\text{GaP}\left(\frac{J(1-p)}{p}, G_0\right)\text{d}G
\end{aligned}
$$

$$
X = \left(\sum_{j=1}^{J} X_j\right) \sim \text{NBP}(G_0, p) \qquad X_j(A) \sim \text{NBP}(G_0(A), p)
$$

Negative Binomial Process

- $L \sim \text{CRTP}(X, G_0)$ as CRT process:

$$\text{for each } A \in \Omega: \qquad L(A) = \sum_{\omega \in \Omega} L(\omega), \qquad L(\omega) \sim \text{CRT}(X(\omega), G_0(\omega))$$

- $X(A)$ customer count and $L(A)$ table count. Each $A \in \Omega$. Number of tables:

$$L(A) \sim \text{Pois}(-G_0(A) \ln(1-p))$$

- assign $\log(p)$ customers to each table, with $X(A)$ total number of customers.
- $X(A) \sim \text{NB}(G_0(A), p)$ customers and assign them into $L(A) \sim \sum_{\omega \in A} \text{CRT}(X(\omega), G_0(\omega))$ tables:

$$\frac{X \sim \sum_{t=1}^{L} \log(p), \qquad L \sim \text{PP}(-G_0 \ln(1-p))}{\text{is equivalent: } L \sim \text{CRTP}(X, G_0), \qquad X \sim \text{NBP}(G_0, p)}$$

Negative Binomial Process (2)

$$
\begin{aligned}
(\gamma_0 = G_0(\Omega)) &\sim \text{Gamma}\left(e_0, \frac{1}{f_0}\right) \\
p &\sim \text{Beta}\left(a_0, \frac{1}{b_0}\right) \\
G|X, p, G_0 &\sim \text{GaP}\left(\frac{J}{p}, G_0 + X\right) \\
p|X, G &\sim \text{Beta}(a_0 + X(\Omega), b_0 + \gamma_0) \\
L|X, G_0 &\sim \text{CRTP}(X, G_0) \\
\gamma_0|L, p &\sim \text{Gamma}\left(e_0 + L(\Omega), \frac{1}{f_0 - \ln(1p)}\right)
\end{aligned}
$$

You need both the posterior and predictive distribution of Multinomial-Dirichlet:

Posterior

$$
\begin{aligned}
&P(p_1, \dots, p_k | n_1, \dots, n_k) \\
&\propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}}_{\text{Dir}(p_1,\dots,p_k|\alpha_1,\dots,\alpha_k)} \underbrace{\frac{n!}{n_1! \dots n_k!} \prod_{i=1}^k p_i^{n_i}}_{\text{Mult}(n_1,\dots,n_k|p_1,\dots p_k)} \\
&\propto \prod_{i=1}^k p_i^{\alpha_i - 1} \prod_{i=1}^k p_i^{n_i} = \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \text{Dir}(p_1, \dots p_k | \alpha_i + n_i, \dots \alpha_k + n_k)
\end{aligned}
$$

Marginal

$$
\begin{aligned}
p(n_1, \dots n_k) &= \int_{p_1,\dots,p_k} P(p_1, \dots, p_k, n_1, \dots, n_k) \\
&= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{n!}{n_1! \dots n_k!} \int_{p_1,\dots,p_k} \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \frac{N!}{n_1! \dots n_k!} \times \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^k \Gamma(\alpha_i + n_i)}{\Gamma\left(N + \sum_{i=1}^k \alpha_i\right)}
\end{aligned}
$$

Expectation

for any measurable set $A_i \in \Theta$: we have $E[G(A_i)] = H(A_i)$, why?
For a dirichlet distribution:

$$f(x_1, \ldots, x_{K-1} | \alpha_1, \ldots, \alpha_K) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

The expectation: $E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$
Therefore,
$E[G(A_i)] = \frac{\alpha H(A_i)}{\sum_i \alpha H(A_i)} = \frac{\alpha H(A_i)}{\alpha \sum_i H(A_i)} = H(A_i)$

Variances

Variances for Dirichlet Distribution: $Var[X_i] = \frac{\alpha_i\left(\left(\sum_i^K \alpha_{i=1}\right) - \alpha_i\right)}{(\sum_i^K \alpha_{i=1})^2(\sum_i^K \alpha_{i=1}+1)}$

Therefore:

$$
\begin{aligned}
Var(G(A_i)) &= \frac{\alpha H(A_i)\left(\alpha - \alpha H(A_i)\right)}{\alpha^2\left(\alpha + 1\right)} \\
&= \frac{H(A_i)\left(1 - H(A_i)\right)}{(\alpha + 1)}
\end{aligned}
$$

Posterior

From well-known multinomial-dirichlet conjugacy, we have:

$$G' = G(A1), \dots, G(Ar)|\theta_1, \dots, \theta_n \sim Dir(\alpha H(A_1) + n_1, \dots, \alpha H(A_k) + n_k)$$

This is equivalently of saying,

$$G' \sim \text{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right) \text{ or,}$$
$$G' \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right)$$

DP provides a conjugate family of priors over distributions that is **closed under posterior updates given observations**

Predictive

Let $P(\theta_{n+1} \in A|G) = G(A)$:

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, \dots, \theta_n) &= \int_G P(\theta_{n+1} \in A|G)P(G|\theta_1, \dots, \theta_n)dG \\ &= E(G(A)|\theta_1, \dots, \theta_n) \\ &= E(G'(A)) \end{aligned}$$

We know that $E(G(A)) = H(A) \implies E(G'(A)) = \frac{\alpha}{\alpha+n}H(A) + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}$

Stick-Breaking construction

- $\beta_k \sim \text{Beta}(1, \alpha)$
- $\theta^k \sim H$
- $\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$
- $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

Predictive

Let $\alpha_i = \frac{\alpha}{k}$: compute the density of i^{th} data belonging to existing component m.

$$
\begin{aligned}
\Pr(z_i = m|\mathbf{z}_{-1}) &= \int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_k)P(p_1,\dots,p_k|n_{1,-i},\dots,n_{k,-i}) \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K)P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)}{P(n_{1,-i},\dots,n_{k,-i})} \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K)P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)}{\int_{p_1,\dots,p_K} P(n_1^{-i},\dots,n_K^{-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)} \\
&= \frac{\Gamma(\frac{\alpha}{k} + n_{m,-i} + 1)\prod_{l=1,l\neq m}^{k}\Gamma(\frac{\alpha}{k} + n_{l,-i})}{\Gamma(N+\alpha)} \times \frac{\Gamma(N-1+\alpha)}{\prod_{l=1}^{k}\Gamma(\frac{\alpha}{k} + n_{l,-1})} \\
&= \frac{\frac{\alpha}{k} + n_{m,-i}}{N+\alpha-1} \qquad \text{Let } k \to \infty = \frac{n_{m,-i}}{N+\alpha-1}
\end{aligned}
\tag{1}
$$

$\Pr(z_i = \text{new}) = \frac{\alpha}{N+\alpha-1}$.

Some Extensions to DP

- Hierarchical Dirichlet Process (HDP)
- HDP-Hidden Marko Model
- Indian Buffet Process

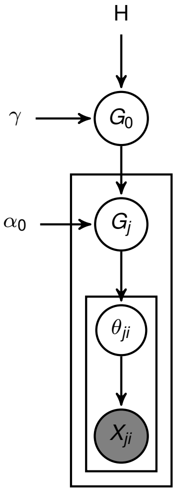
Hierarchical Dirichlet Process (HDP)

Generative model

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_j &\sim \mathrm{DP}(\alpha_0, G_0) \\
\theta_{ji} &\sim G_j \\
X_{ji} &\sim F(x|\theta_{ij})
\end{aligned}
$$

Drawing $G_0 \sim \mathrm{DP}(.)$ is difficult. Therefore, we need some "construction" method:

Graphical model



HDP - Stick breaking construction

Generative model

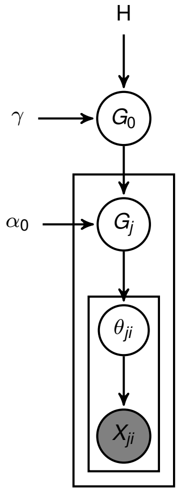
$$\boldsymbol{\beta} \sim GEM(\gamma) \qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

$$\boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}) \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

$$z_{ji} \sim \pi_j \qquad \phi_k \sim H \qquad X_{ji} \sim F(x|\phi_{z_{ji}})$$

- See in the next slide how to sample π_j DIRECTLY from stick-breaking process from β, i.e., NO NEED to perform stick-breaking.
- Using β as a base, discrete distribution define on range $\{0 \ldots \infty\}$. (takes advantage that partition of space is given)

Graphical model



Sample π_j DIRECTLY from stick-breaking process from β:

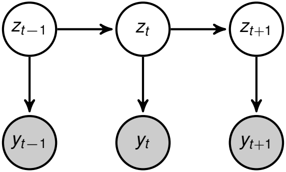
Suppose $\beta|\gamma \sim \text{GEM}(\gamma)$ and $\pi|\alpha, \beta \sim \text{DP}(\alpha, \beta)$. Notice that the support is $\{1, \dots, k, \dots, \infty\}$:

$$
\begin{aligned}
&(G_j(A_1), \dots, G_j(A_r)) \sim \text{Dir}\left(\alpha G_0(A_1), \dots, \alpha G_0(A_r)\right) \\
\implies &\left(\sum_{k\in K_1} \pi_k, \dots, \sum_{k\in K_r} \pi_k\right) \sim \text{Dir}\left(\alpha \sum_{k\in K_1} \beta_k, \dots, \alpha \sum_{k\in K_r} \beta_k\right) \\
\implies &\left(\sum_{l=1}^{k-1} \pi_l, \pi_k, \sum_{l=k+1}^{\infty} \pi_l\right) \sim \text{Dir}\left(\alpha \sum_{l\in 1}^{k-1} \beta_l, \alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right) \\
\implies &\left(\frac{\pi_k}{1 - \sum_{l=1}^{k-1} \pi_l}, \frac{\sum_{l=k+1}^{\infty} \pi_l}{1 - \sum_{l=1}^{k-1} \pi_l}\right) \sim \text{Dir}\left(\alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right) \\
\implies &\left(\frac{\pi_k}{1 - \sum_{l=1}^{k-1} \pi_l}, \frac{1 - \sum_{l=1}^{k} \pi_l}{1 - \sum_{l=1}^{k-1} \pi_l}\right) \sim \text{Dir}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right) \\
\implies &\frac{\pi_k}{1 - \sum_{l=1}^{k-1} \pi_l} \sim \text{Beta}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right) \text{ or } \pi' \sim \text{Beta}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right)
\end{aligned}
$$

Traditional HMM

Under normal HMM, you have a transition matrix A, let the j^{th} row of A to be π_i, then:

$$A = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \dots \\ \pi_K \end{bmatrix} = \begin{bmatrix} p(z_{t+1} = 1 | z_t = 1) & p(z_{t+1} = 2 | z_t = 1) & \dots & p(z_{t+1} = K | z_t = 1) \\ p(z_{t+1} = 1 | z_t = 2) & p(z_{t+1} = 2 | z_t = 2) & \dots & p(z_{t+1} = K | z_t = 2) \\ \dots & \dots & \dots & \dots \\ p(z_{t+1} = 1 | z_t = K) & p(z_{t+1} = 2 | z_t = K) & \dots & p(z_{t+1} = K | z_t = K) \end{bmatrix}$$



To obtain the current latent state, we need to sample $z_t \sim \pi_{z_{t-1}}$.

well, the same idea has been extended to non-parametric bayes, to allow π_j to have infinite many components. Therefore, matrix A has size $\infty \times \infty$. But the "recovered" number of states are finite, so you only "jumping around" in the upper-left corner of matrix A.

$$\begin{bmatrix} p(z_{t+1} = 1 | z_t = 1) & p(z_{t+1} = 2 | z_t = 1) & \dots & p(z_{t+1} = \infty | z_t = 1) \\ p(z_{t+1} = 1 | z_t = 2) & p(z_{t+1} = 2 | z_t = 2) & \dots & p(z_{t+1} = \infty | z_t = 2) \\ \dots & \dots & \dots & \dots \\ p(z_{t+1} = 1 | z_t = \infty) & p(z_{t+1} = 2 | z_t = \infty) & \dots & p(z_{t+1} = \infty | z_t = \infty) \end{bmatrix}$$

Generative model

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma)$$

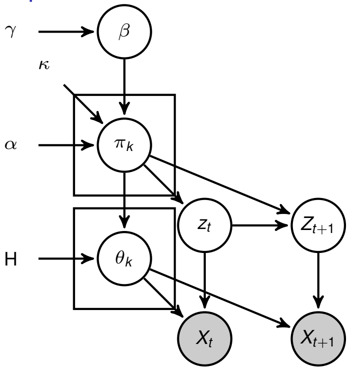
$$\boldsymbol{\pi}_j \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_j}{\alpha + \kappa}\right)$$

$$z_{t+1} \sim \pi_{z_t}$$

$$\theta_k \sim H$$

$$X_t \sim F(x|\theta_{z_t})$$

Graphical model



DP

- $\Pr(z_1 \dots z_N)$, where $z_i \in (1 \dots K)$ indicate category.
- You also want K potentially be infinite
- A "clustering" property, controllable through a single parameter α
- Can also be thought as a special $N \times K$ Z matrix, where there is only one "1" in each row.

IBP

- More general than DP: z_i can take multiple values $\in (1, \dots K)$
- This is equivelently of saying that, z_i is a binary vector of K elements.
- Given N such data, we have a binary matrix of size $N \times K$
- A "clustering" property, controllable through a single parameter α, a column with more 1, results it to have more 1s.

The big Z matrix

An example of Z matrix:

1	0	1	1	0	...	1
0	1	0	0	0	...	0
...	...	...	...	...	...	0
1	1	0	0	0	...	0

For each column: $Pr(z_{ik} = 1) \sim \text{Ber}(\mu_k)$ independently.
Each $u_k \sim \text{Beta}\left(\frac{\alpha}{k}, 1\right)$ is also distributed independently.
The marginal distribution:

Bernoulli- Beta vs Multinomial-Dirichlet: Posterior

Multinomial-Dirichlet

$$
\begin{aligned}
&P(p_1, \ldots, p_k | n_1, \ldots, n_k) \\
&\propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}}_{\mathrm{Dir}(p_1, \ldots, p_k | \alpha_1, \ldots, \alpha_k)} \underbrace{\frac{n!}{n_1! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i}}_{\mathrm{Mult}(n_1, \ldots, n_k | p_1, \ldots p_k)} \\
&\propto \prod_{i=1}^{k} p_i^{\alpha_i - 1} \prod_{i=1}^{k} p_i^{n_i} = \prod_{i=1}^{k} p_i^{\alpha_i - 1 + n_i} \\
&= \mathrm{Dir}(p_1, \ldots p_k | \alpha_i + n_i, \ldots \alpha_k + n_k)
\end{aligned}
$$

Bernoulli-Binomial

$$
\begin{aligned}
&P(p | n_1 = m) \\
&\propto \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1} (1 - p)^{\beta - 1}}_{\mathrm{Beta}(p | \alpha, \beta)} \underbrace{\frac{N!}{m!(N - m)!} p^k (1 - p)^{N - k}}_{\mathrm{Binomial}(n_1, n_2 | p)} \\
&\propto p^{\alpha - 1} (1 - p)^{\beta - 1} p^k (1 - p)^{N - k} \\
&= p^{\alpha - 1 + k} (1 - p)^{\beta - 1 + N - k} \\
&= \mathrm{Beta}(p | \alpha_i + k, \beta + N - k)
\end{aligned}
$$

Bernoulli- Beta vs Multinomial-Dirichlet: Marginal

Multinomial-Dirichlet

$$\int_{p_1,\dots,p_k} P(p_1,\dots,p_k,n_1,\dots,n_k)$$

$$= \frac{N!}{n_1!\dots n_k!}\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\frac{\prod_{i=1}^k \Gamma(\alpha_i+n_i)}{\Gamma\left(N+\sum_{i=1}^k \alpha_i\right)}$$

Bernoulli-Beta

$$\int_p P(p,n_1,n_2)$$

$$= \frac{N!}{k!(N-k)!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+k)\Gamma(\beta+N-k)}{\Gamma(N+\alpha+\beta)}$$

$\mu_k \sim \text{Beta}\left(\frac{\alpha}{k}, 1\right) \qquad \Pr(z_{ik} = 1) \sim \text{Ber}(\mu_k)$.
$n_{k,-i}$ is the number of 1s of k^{th} column, above row i.
Let $\alpha_i = \frac{\alpha}{k}$: compute the density of i^{th} data belonging to existing component m.

$$\Pr(z_{ik} = 1|\mathbf{z}_{-i,k}) = \int_p \Pr(z_{ik} = 1|p)P(p|\underbrace{n_{-i,k}}_{n_1}, \underbrace{i - 1 - n_{-i,k}}_{n_2})$$

$$= \frac{\int_p \Pr(z_{ik} = 1|p)\Pr(n_1, n_2|p)P(p)}{\Pr(n_1, n_2)} = \frac{\int_p \Pr(z_{ik} = 1|p)\Pr(n_1, n_2|p)P(p)}{\int_p \Pr(n_{-i,k}, i - 1 - n_{-i,k}|p)P(p)}$$

$$= \frac{\Gamma(\frac{\alpha}{k} + n_{-i,k} + 1)\Gamma(1 + i - 1 - n_{-i,k})}{\Gamma\left(i + \frac{\alpha}{k} + 1\right)} \frac{\Gamma\left(i - 1 + \frac{\alpha}{k} + 1\right)}{\Gamma(\frac{\alpha}{k} + n_{-i,k})\Gamma(1 + i - 1 - n_{-i,k})} = \frac{\frac{\alpha}{k} + n_{-i,k}}{i + \frac{\alpha}{k}}$$

One more factor: relationship between Binomial and Poisson

$$\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$$

Let $\lambda = np$:

$$\begin{aligned}
\text{Binomial}(x|n,p) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} \frac{\lambda}{n}^x (1 - \frac{\lambda}{n})^{n-x} \\
&= \underbrace{\frac{\lambda^x}{x!}}_{\text{constant}} \underbrace{\frac{n!}{(n-x)!} \frac{1}{n^x}} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \frac{\overbrace{n(n-1),\ldots(n-x+1)}^{n \text{ terms}}}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \frac{n}{n} \frac{n-1}{n} \ldots \frac{n-x+1}{n} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} 1 \left(1 - \frac{1}{n}\right) \ldots \left(1 - \frac{x+1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}
\end{aligned}$$

$$\lim_{n\to\infty} \text{Binomial}(x|n,p) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(1 - \frac{1}{n}\right) \ldots \lim_{n\to\infty} \left(1 - \frac{x+1}{n}\right) \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = \frac{\lambda^x}{x!} \exp(-\lambda)$$

Taking limit $k \to \infty$

$$\lim_{k\to\infty} \Pr(z_{ik}) = \lim_{k\to\infty} \frac{\frac{\alpha}{k} + n_{-i,k}}{i + \frac{\alpha}{k}} = \frac{n_{-i,k}}{i}$$

$$\lim_{n\to\infty} \text{Binomial}(\frac{\lambda}{n}, n) = \text{Poisson}(\lambda)$$

$$\text{Let } k \to \infty : \qquad = \frac{n_{-i,k}}{i}$$

For "new" dishes, i.e., $n_{-i,k} = 0$, then, $\Pr(z_{ik} = 1) = \text{Bernoulli}\left(\frac{\frac{\alpha}{K}}{i+\frac{\alpha}{K}}\right)$

i.e., how many new dishes across all columns would be: $\text{Binomial}\left(\frac{\frac{\alpha}{K}}{i+\frac{\alpha}{K}}, K\right)$

Since $\frac{\frac{\alpha}{k}}{i+\frac{\alpha}{k}} \times k = \frac{\alpha}{i+\frac{\alpha}{k}}$, we have:

$$\lim_{K\to\infty} \text{Binomial}\left(\frac{\frac{\alpha}{K}}{i + \frac{\alpha}{K}}, K\right) = \text{Poisson}\left(\frac{\alpha}{i}\right)$$

So, how many K^+ columns there are?

Let $n_i \sim \text{Poisson}\left(\frac{\alpha}{i}\right)$ $\left(\sum_{i=1}^N n_i\right) \sim \text{Poisson}\left(\sum_{i=1}^N \frac{\alpha}{i}\right)$

What is Factor Analysis? There are $N = 1000$ students, each having $(p = 10)$ scores. Therefore:

$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1N} \\ y_{21} & y_{22} & \dots & y_{2N} \\ \dots & \dots & \dots & \dots \\ y_{p1} & y_{p2} & \dots & y_{pN} \end{bmatrix} = \begin{bmatrix} g_{11} & \dots & g_{1k} \\ g_{21} & \dots & g_{2k} \\ \dots & \dots & \dots \\ g_{p1} & \dots & g_{pk} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kN} \end{bmatrix} + \mathbf{E}$$

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1N} \\ e_{21} & e_{22} & \dots & e_{2N} \\ \dots & \dots & \dots & \dots \\ e_{p1} & e_{p2} & \dots & e_{pN} \end{bmatrix} \text{ and } k << p$$

Or in a matrix form: $\mathbf{Y} = \mathbf{GX} + \mathbf{E}$.

Factor analysis cont.

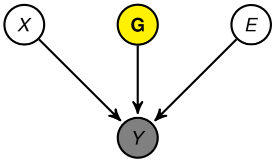
What this means is that a person's i's raw mark is interpretted as:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ \dots \\ y_{pi} \end{bmatrix} = x_{1i} \begin{bmatrix} g_{11} \\ g_{21} \\ \dots \\ g_{p1} \end{bmatrix} + x_{2i} \begin{bmatrix} g_{11} \\ g_{21} \\ \dots \\ g_{p1} \end{bmatrix} + \dots x_{ki} \begin{bmatrix} g_{1k} \\ g_{2k} \\ \dots \\ g_{pk} \end{bmatrix} + \begin{bmatrix} e_{1i} \\ e_{2i} \\ \dots \\ e_{pi} \end{bmatrix}$$

- Given a set of k loading factors (vectors) each with dimension p: $\{\mathbf{g}_{:,i}\}_{i=1}^{k}$, the $x_{:,i}$ can be thought as the latent linear weights.
- Of course, you are only given data matrix Y, one has to infer the latent structure. **G**, **X** and **E**. Ths is not as silly as it seems, as DoF is much reduced.

The Bayesian Treatment:

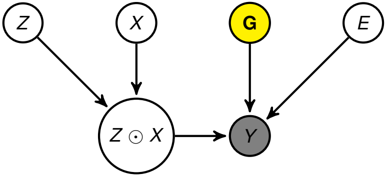
$$e_i \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) \qquad \sigma_e^2 \sim \mathcal{IG}(a, b)$$
$$g_k \sim \mathcal{N}(0, \sigma_G^2) \qquad \sigma_G^2 \sim \mathcal{IG}(c, d)$$
$$x_{ki} \sim \mathcal{N}(0, 1) \qquad y_i = \mathbf{G}x_i + e_i$$



- Knowles, d and Ghahramani, Z, Innite Sparse Factor Analysis
- K should known beforehand. What about making K a variable?
- Although $[x_{1,i}, \dots x_{k,i}]^T$ has a reduced dimension, it can still cause "overfitting".
- We need to introcuce variable number of latent factors K, at the same time, have **sparsity**!

How?

$$
\begin{aligned}
e_i &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) & \sigma_e^2 &\sim \mathcal{IG}(a, b) \\
g_k &\sim \mathcal{N}(0, \sigma_G^2) & \sigma_G^2 &\sim \mathcal{IG}(c, d) \\
Z &\sim \mathcal{IBP}(\alpha) & \alpha &\sim \mathcal{G}(e, f) \\
x_{ki} &\sim \mathcal{N}(0, 1) & y_i &= \mathbf{G}(x_i \odot z_i) + e_i
\end{aligned}
$$

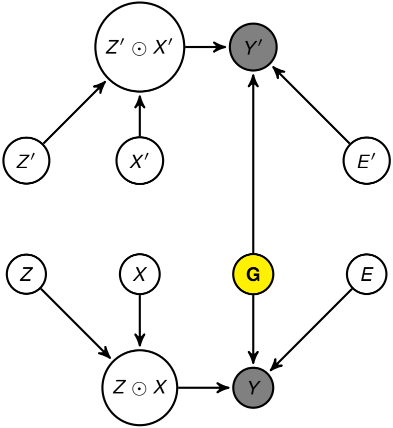


A proposed work

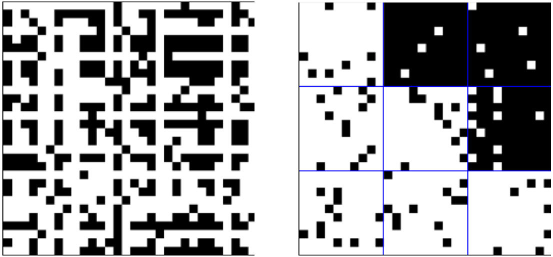
- What about if there are two sets of data matrix $\mathbf{Y}$ and $\mathbf{Y}'$, each having different number of entries. They share the same loading vectors $\mathbf{G}$, but with different level of **sparsities**.

$$
\begin{aligned}
e_i &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) & \sigma_e^2 &\sim \mathcal{IG}(a, b) \\
g_k &\sim \mathcal{N}(0, \sigma_G^2) & \sigma_G^2 &\sim \mathcal{IG}(c, d) \\
Z &\sim \mathcal{IBP}(\alpha) & \alpha &\sim \mathcal{G}(e, f) \\
x_{ki} &\sim \mathcal{N}(0, 1) & y_i &= \mathbf{G}(x_i \odot z_i) + e_i
\end{aligned}
$$

$$
\begin{aligned}
e'_i &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) & \sigma_e^2 &\sim \mathcal{IG}(a', b') \\
g_k &\sim \mathcal{N}(0, \sigma_G^2) & \sigma_G^2 &\sim \mathcal{IG}(c, d) \\
Z' &\sim \mathcal{IBP}(\alpha') & \alpha' &\sim \mathcal{G}(e', f') \\
x'_{ki} &\sim \mathcal{N}(0, 1) & y'_i &= \mathbf{G}(x'_i \odot z'_i) + e'_i
\end{aligned}
$$



- Community learning is an emerging topic applicable to many social networking problems and “hot” in machine learning.
- Partition a network of nodes into different groups based on their pairwise and directional binary observations.
- Many models were proposed in the last few years and they become increasing sophisticated.
- Data is **directional** i.e., I like you doesn’t mean you like me.



Simple Stochastic Block Model assumption: Fixed K communities

The model:

- There is a hidden "compatibility" matrix $\mathbf{B}$, size $K \times K$, each element $\mathbf{B}_{kl} \sim Beta(\lambda_1, \lambda_2)$, a realization example:

0.5	0.2	0.1	0.1	0	...	0.1
0.3	0.91	0.2	0.4	0.2	...	0.5
...	...	...	...	...	...	0.2
0.32	0.2	0.96	0.4	0.7	...	0.9

- Suppose that person i is in latent community 2, i.e., $z_i = 2$ and person j is in latent community 3, i.e., $z_j = 3$.
- Then $\boldsymbol{e}_{ij} \sim Bernoulli(\mathbf{B}_{z_i=2, z_j=3}) = Bernoulli(0.2)$.
- $z_i \sim \pi$: some weights of communities

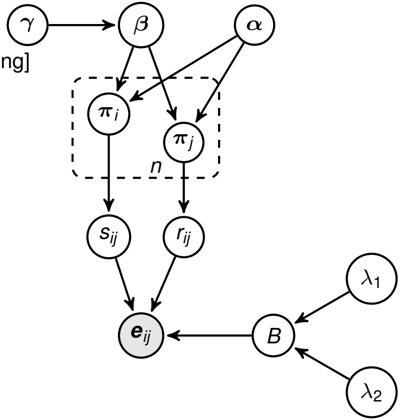
Inference:

- Then, our task is to perform posterior inference on: $\Pr(z_1, \dots z_n, \pi, \mathbf{B} | \{\boldsymbol{e}_{ij}\})$

Early work assumes a fixed number of K communities exist a node i can potentially belong to. However, in many applications, an accurate guess of K can be impractical. **Infinite Relational Model** (Kemp 2006)

- Infinite Relational Model was incorporated to address this problem, where K can be inferred from the data itself, and potentially be ∞.
- That's where Non-parametric Bayes comes in!
- Still a drawback: assumes each node i must belong to only a single community k (i.e., $z_i = k$).

Mixed-Membership Stocastic Blockmodel (MMSB)
[Airoldi et al.(2008)Airoldi, Blei, Fienberg, and Xing]



- mixed-membership concept: each node i may belong to multiple communities. Having individual distribution π_i
- e_{ij} no longer dependant only on each pair of community indicators z_i and z_j. Instead, they are sampled from pairs of interactions between nodes i and j. (s_{ij}, r_{ji}).

Generative Model

1. $\boldsymbol{\beta} \sim GEM(\gamma)$
2. $\{\pi_i\}_{i=1}^{n} \sim DP(\alpha \cdot \boldsymbol{\beta})$
3. $s_{ij} = \pi_i, \; r_{ij} = \pi_j$
4. $B_{k,l} \sim Beta(\lambda_1, \lambda_2), \forall k, l;$
5. $\boldsymbol{e}_{ij} \sim Bernoulli(B_{s_{ij}, r_{ij}})$.

The priors:

- Each element $\mathbf{B}_{kl} \sim Beta(\lambda_1, \lambda_2)$ still: a realization example:

0.5	0.2	0.1	0.1	0	...	0.1
0.3	0.91	0.2	0.4	0.2	...	0.5
...	...	...	...	...	...	0.2
0.32	0.2	0.96	0.4	0.7	...	0.9

- Suppose that interaction *i* **sent to** *j* is of latent community 2, i.e., $s_{ij} = 2$,
- Interaction *j* **received from** *i* is in latent community 3, i.e., $r_{ji} = 3$.
- Note that s_{ij} do not generally equal r_{ji}.
- Then $\boldsymbol{e}_{ij} \sim Bernoulli(\mathbf{B}_{s_{ij}=2, r_{ji}=3}) = Bernoulli(0.2)$.
- $\{s_{i,k}, r_{i,k}\} \sim \pi_i$: There are altogether N πs

The posterior

- Then, our task is to perform posterior inference on:
  $\Pr(\{s_{i,j}, r_{j,i}\}_{\forall, 1 \leq i,j \leq N}, \mathbf{B}, \pi_1, \ldots \pi_N | \{\boldsymbol{e}_{ij}\})$

Recent variants of MMSB

A few variants were subsequently proposed from MMSB, examples include:

- [**?**, Xing et al.(2010)Xing, Fu, and Song] extends the mixture-membership model with a dynamic setting;
- [Koutsourelakis and Eliassi-Rad(2008)] extends the MMSB into the infinite case; and
- [Kim et al.(2012)Kim, Hughes, and Sudderth] incorporates the node's metadata information into MMSB.

Incorporating meta data information work

Sudderth's method:

The model

$$v_{:i} \sim \mathcal{N}(\eta^T \phi_{:i}, \lambda_{V_i}^{-1})$$

Instead of $v_i \sim \beta(1, \alpha)$:

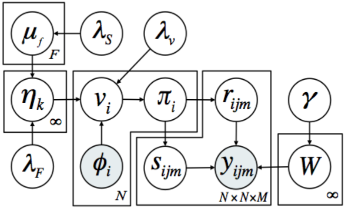
$$\pi_{ki} = \psi(v_{ki}) \prod_{l=1}^{k-1} \psi(-v_{li})$$

$$\text{where } \psi(v_{ki}) = \frac{1}{1 + \exp(-v_{ki})}$$

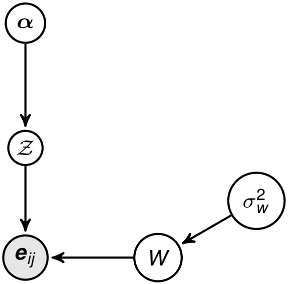
For each community k:

$$\eta_{:k} \sim \mathcal{N}(\mu, \lambda_F^{-1} I_F)$$

Graphic Model



$$\mathcal{Z} \sim \mathcal{IBP}(\alpha)$$
$$e_{ij} \sim Z_i W Z_j^T$$
$$W_{k,k'} \sim \mathcal{N}(0, \sigma_W)$$



Our work: Copula Mixed-Membership Stochastic Blockmodel with Subgroup Correlation

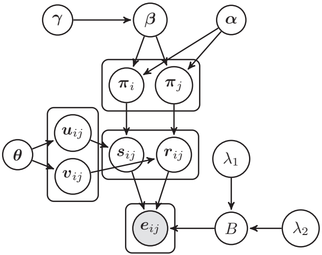
- Despite MMSB's powerful representations,it assumes that the distributions of relational membership indicators between the two nodes are independent.
- Under many social network settings, possible that certain known subgroups of people may have higher correlations in terms of their membership categories towards each other
- We introduce a new framework where individual Copula function is to be employed to model jointly the membership pairs of those nodes within the subgroup of interest.
- Various Copula functions may be used to suit the scenario, while maintaining the membership's marginal distribution, as needed for modeling membership indicators with other nodes outside of the subgroup of interest.
- Experimental results shows a superior performance when comparing with the exisiting models on both the synthetic and real world datasets.

The model

Generative Model

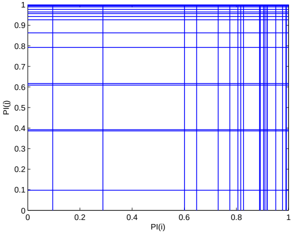
1. $\boldsymbol{\beta} \sim GEM(\gamma)$
2. $\{\pi_i\}_{i=1}^{n} \sim DP(\alpha \cdot \boldsymbol{\beta})$
3. $\begin{cases} (u_{ij}, v_{ij}) \sim Copula(\theta), & g_{ij} = 1; \\ u_{ij}, v_{ij} \sim U(0,1), & g_{ij} = 0. \end{cases}$
4. $s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij})$
5. $B_{k,l} \sim Beta(\lambda_1, \lambda_2), \forall k, l;$
6. $\boldsymbol{e}_{ij} \sim Bernoulli(B_{s_{ij}, r_{ij}})$.

Graphical model

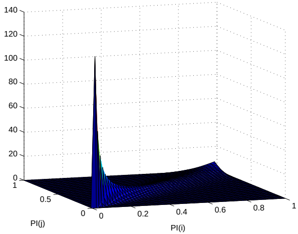


Diagrammatic Representation

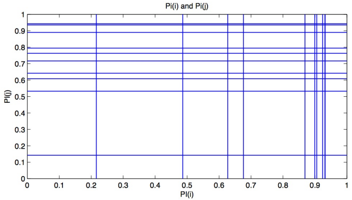
1. $\pi_i, \pi_j sim DP(\alpha \cdot \boldsymbol{\beta})$

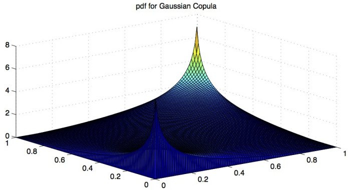


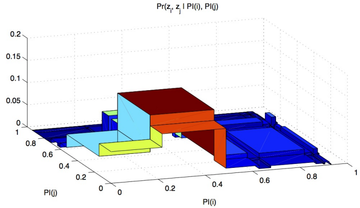
1. $(u_{ij}, v_{ij}) \sim Copula(\theta), g_{ij} = 1$



Diagrammatic Representation 2







So what is Copula

- A bivariate copula function $C(u, v)$ is a Cumulative Distribution Function over the interval $[0, 1] \times [0, 1]$ with uniform marginal distribution.
- *Sklar's theorem:* Let X and Y be random variables with distribution functions F and G respectively and joint distribution function H. Then there exists a Copula C such that for all $(x, y) \in R \times R$:

$$H(x, y) = C(F(x), G(y))$$

- C is unique if F and G are continuous, then the joint probability density function is:

$$h(x, y) = c(F(x), G(y)) \cdot f(x)g(y) \tag{2}$$

Here $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ is notes for copula density function.

So what is Copula Cont.

- Sklar's theorem ensures the uniqueness of copula function $C(F(x), G(y))$
- Change Copula function does not change the marginal distributions. This is what we want!
- Copula is popular! Many are availability: Commonly used copula functions includes, Gaussian Copula (Gaussian, t), Archimedean Copula (Clayton, Gumbel, Frank, etc.), Empirical Copula.

The ideal case

$$\Pr(s_{ij}, r_{ij}) = \int_{\pi_{i,1},\ldots,\pi_{i,K+1}} \int_{\pi_{j,1},\ldots\pi_{j,K+1}} \int_{(u_{ij},v_{ij})} \cdot \mathbf{1}\left(s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij})\right) \cdot dC(u_{ij}, v_{ij}) dF(\pi_{i1}, \cdots, \pi_{iK+1}) dF(\pi_{j1}, \cdots, \pi_{jK+1}) \tag{3}$$

Unfortunately, we cannot get it to an analytical form without any integrals present.

Marginal conditional on π only ($cMMSB^{\pi}$)

$$
\begin{aligned}
p_{ij}^{kl}(\pi_i, \pi_j) \equiv \Pr(s_{ij} = k, r_{ij} = l | \pi_i, \pi_j, \theta_d) \\
= \int_{\hat{\pi}_i^{k-1}}^{\hat{\pi}_i^{k}} \int_{\hat{\pi}_j^{l-1}}^{\hat{\pi}_j^{l}} dC(u, v|\theta_d) \\
= C(\hat{\pi}_i^k, \hat{\pi}_j^l) + C(\hat{\pi}_i^{k-1}, \hat{\pi}_j^{l-1}) - C(\hat{\pi}_i^k, \hat{\pi}_j^{l-1}) - C(\hat{\pi}_i^{k-1}, \hat{\pi}_j^l)
\end{aligned}
$$

$$
\hat{\pi}_i^k = \begin{cases} 0, & k = 0; \\ \sum_{q=1}^{k} \pi_{iq}, & k > 0 \end{cases}
$$

- Easily calculate this “rectangular” area.
- When no correlations, $p_{ij}^{kl}(\pi_i, \pi_j) = \pi_{ik}\pi_{jl}$

Properties of a Copula function: marginal of $\Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta_d)$ remain π_i and π_j respectively:

$$
\begin{aligned}
\sum_{l=1}^{K+1} \Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta_d) = \pi_{ik}; \\
\sum_{k=1}^{K+1} \Pr(s_{ij} = k, r_{ij} = l|\pi_i, \pi_j, \theta_d) = \pi_{jl}.
\end{aligned}
\tag{4}
$$

Marginal conditional on u, v only ($cMMSB^{uv}$)

- Integrate over $\{\pi_i\}_{i=1}^{n}$ given $\{(u_{ij}, v_{ij})\}_{i,j}$
- Given $\{(u_{ij}, v_{ij})\}_{i,j}$, $\Pr(s_{ij} = k)$ and $\Pr(r_{ij} = l)$ are independent.
- Copula function leaves marginal distributions of s_{ij} and r_{ij} invariant, which remains the same as the classical posterior of π_i in *MMSB*:

$$\pi_i|\alpha, \beta, \{N_{ik}^{-ij}\}_{k=1}^{K} \sim Dir(\alpha\beta_1 + N_{i1}^{-ij}, \cdots, \alpha\beta_K + N_{iK}^{-ij}, \alpha\beta_{K+1})$$

- $\Pr(s_{ij} = k)$ is equal to computing the probability of u_{ij} falling in π_i's k^{th} interval:

$$\Pr\left(\sum_{d=1}^{k-1} \pi_{id} \le u_{ij} < \sum_{d=1}^{k} \pi_{id}\right)$$

- The fact that the set $\{u_{ij} \in [0, 1]| \sum_{d=1}^{k-1} \pi_{id} \le u_{ij}\}$ can be decomposed into two ***disjoint*** sets:

$$\begin{aligned}
&\{u_{ij} \in [0, 1]| \sum_{d=1}^{k-1} \pi_{id} \le u_{ij}\} \\
=&\{u_{ij} \in [0, 1]| \sum_{d=1}^{k-1} \pi_{id} \le u_{ij} < \sum_{d=1}^{k} \pi_{id}\} \cup \{u_{ij} \in [0, 1]| \sum_{d=1}^{k} \pi_{id} \le u_{ij}\}
\end{aligned} \tag{5}$$

where $\sum_{d=1}^{k} \pi_{id} \sim Beta(\sum_{d=1}^{k} \alpha\beta_d + N_{id}, \sum_{d=k+1}^{K+1} \alpha\beta_d + N_{id})$.

Marginal conditional on u, v only ($cMMSB^{uv}$) cont.

We have:

$$\begin{aligned}
&\Pr(\sum_{d=1}^{k-1} \pi_{id} \leq u_{ij} < \sum_{d=1}^{k} \pi_{id}) \\
=&\Pr(\sum_{d=1}^{k-1} \pi_{id} \leq u_{ij}) - \Pr(\sum_{d=1}^{k} \pi_{id} \leq u_{ij}) \\
=&I_{u_{ij}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}}(h_i^k, \hat{h}_i^k)
\end{aligned}$$

$$h_i^k = \sum_{d=1}^{k} \alpha\beta_d + N_{id} \qquad \hat{h}_i^k = \sum_{d=k+1}^{K+1} \alpha\beta_d + N_{id}$$

$I_u(a, b)$ is Beta c.d.f. of u with parameter a, b

Non-negativity is guaranteed by the fact that
$\{u_{ij} \in [0, 1] | \sum_{d=1}^{k} \pi_{id} \leq u_{ij}\} \subseteq \{u_{ij} \in [0, 1] | \sum_{d=1}^{k-1} \pi_{id} \leq u_{ij}\}$ on the same π_i.

Inference

- In $cMMSB^{\pi}$), variables of interest are $\{\pi_i\}, \{s_{ij}, r_{ij}\}, \boldsymbol{\beta}$.
- In $cMMSB^{uv}$), variables of interest include $\{u_{ij}, v_{ij}\}, \{s_{ij}, r_{ij}\}, \boldsymbol{\beta}$, and an auxiliary variable $\boldsymbol{m}$.

Inference $cMMSB^{\pi}$ - Sampling π_i

- When a Copula is introduced, $p(\pi_i)$ and $\Pr(s_{ij}|\pi_i)$ are no longer a conjugate pair.
- Therefore, resort to the use of Metropolis-Hastings (M-H) Sampling in each (τ)-th Gibbs iteration

For each node i, posterior distribution of π_i is:

$$
\begin{aligned}
& p(\pi_i|\alpha, \boldsymbol{\beta}, \{s_{ij}, r_{ij}\}_{i,j}) \\
\propto & \prod_{k=1}^{K+1} \pi_{ik}^{\alpha\boldsymbol{\beta}_k - 1} \cdot \prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i) \right]
\end{aligned}
$$

Corresponding proposal of π_i:

$$
q(\pi_i^*|\alpha, \boldsymbol{\beta}, \{s_{ij}, r_{ij}\}_{i,j}) \propto \prod_{k=1}^{K+1} [\pi_{ik}^*]^{\alpha\boldsymbol{\beta}_k + N_{ik} - 1}
$$

Acceptance ratio becomes:

$$
A(\pi_i^*, \pi_i^{(\tau)}) = \min(1, a) \tag{6}
$$

$$
a = \frac{\prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i^*, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i^*) \right]}{\prod_{j=1}^{n} \left[p_{ij}^{s_{ij}r_{ij}}(\pi_i^{(\tau)}, \pi_j) p_{ji}^{s_{ji}r_{ji}}(\pi_j, \pi_i^{(\tau)}) \right]} \cdot \frac{\prod_{k=1}^{K+1} \left[\pi_{ik}^{(\tau)} \right]^{N_{ik}}}{\prod_{k=1}^{K+1} \left[\pi_{ik}^* \right]^{N_{ik}}} \tag{7}
$$

Inference $cMMSB^{\pi}$ - Sampling $\{s_{ij}, r_{ij}\}$

$$
\begin{aligned}
&\Pr(s_{ij}, r_{ij} | e_{ij}, \lambda_1, \lambda_2, \theta_d, \pi_i, \pi_j, \{(s_{ij}, r_{ij})\}_{i,j}) \\
\propto & p_{ij}^{s_{ij}, r_{ij}}(\pi_i, \pi_j) \cdot p(e_{ij} | \lambda_1, \lambda_2, \{(s_{ij}, r_{ij})\}_{i,j})
\end{aligned}
$$

$$
p(e_{ij} | \lambda_1, \lambda_2, \{(s_{ij}, r_{ij})\}_{i,j}) = \begin{cases} n_{s_{ij}, r_{ij}}^{1, -e_{ij}} + \lambda_1, & e_{ij} = 1; \\ n_{s_{ij}, r_{ij}}^{0, -e_{ij}} + \lambda_2, & e_{ij} = 0. \end{cases}
$$

- obvious choice of M-H proposal of β its prior $p(\beta|\gamma) = GEM(\gamma)$.
- this proposal can be non-informative, which results in a low acceptance rate.
- We sample β^* conditioned on an auxiliary variable $\boldsymbol{m}$: $(\boldsymbol{\beta}_1^*, \cdots, \boldsymbol{\beta}_K^*, \boldsymbol{\beta}_{K+1}^*) \sim Dir(\boldsymbol{m}_1, \cdots, \boldsymbol{m}_K, \gamma)$, in order to increase the M-H's acceptance rate
- instead of sampling β directly from $\boldsymbol{m}$ as in [Teh et al.(2006)Teh, Jordan, Beal, and Blei], we only use it for our proposal distribution, as we have explicitly sampled $\{\pi_i\}_{i=1}^n$. The acceptance ratio is hence:

$$A(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(\tau)}) = \min(1, a)$$

$$a = \frac{\prod_{i=1}^n \left[\prod_{d=1}^{K+1} \Gamma(\alpha\boldsymbol{\beta}_d^{(\tau)}) \cdot \pi_{id}^{\alpha\boldsymbol{\beta}_d^*}\right]}{\prod_{i=1}^n \left[\prod_{d=1}^{K+1} \Gamma(\alpha\boldsymbol{\beta}_d^*) \cdot \pi_{id}^{\alpha\boldsymbol{\beta}_d^{(\tau)}}\right]} \cdot \frac{\prod_{d=1}^K \left[\boldsymbol{\beta}_d^{(\tau)}\right]^{\boldsymbol{m}_d - \gamma}}{\prod_{d=1}^K \left[\boldsymbol{\beta}_d^*\right]^{\boldsymbol{m}_d - \gamma}} \tag{8}$$

Inference $cMMSB^{u,v}$ - Sampling (u_{ij}, v_{ij})

The Copula function is used as its proposal, and therefore, its corresponding acceptance ratio becomes that of:

$$A\left((u_{ij}^{(\tau)}, v_{ij}^{(\tau)}), (u_{ij}^{*}, v_{ij}^{*})\right) = \min(1, a)$$

$$a = \frac{I_{u_{ij}^{*}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}^{*}}(h_i^{k}, \hat{h}_i^{k})}{I_{u_{ij}^{(\tau)}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}^{(\tau)}}(h_i^{k}, \hat{h}_i^{k})} \cdot \frac{I_{v_{ij}^{*}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}^{*}}(h_j^{l}, \hat{h}_j^{l})}{I_{v_{ij}^{(\tau)}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}^{(\tau)}}(h_j^{l}, \hat{h}_j^{l})}$$

Here $h_i^k, \hat{h}_i^k$'s definitions are the same as in Eq. (7) in the paper; assuming $s_{ij} = k, r_{ij} = l$.

$$
\begin{aligned}
&\Pr(s_{ij} = k, r_{ij} = l | e_{ij}, \lambda_1, \lambda_2, n_{kl}, u_{ij}, v_{ij}, \{h_i^k\}_k, \{\hat{h}_i^k\}_k, \{h_j^k\}_k, \{\hat{h}_j^k\}_k) \\
\propto &\Pr(s_{ij} = k | u_{ij}, , \{h_i^k\}_k, \{\hat{h}_i^k\}_k) \cdot \Pr(r_{ij} = l | v_{ij}, \{h_j^k\}_k, \{\hat{h}_j^k\}_k) \cdot \Pr(e_{ij} | \lambda_1, \lambda_2, n_{kl}) \\
\propto &(I_{u_{ij}}(h_i^{k-1}, \hat{h}_i^{k-1}) - I_{u_{ij}}(h_i^k, \hat{h}_i^k)) \cdot (I_{v_{ij}}(h_j^{l-1}, \hat{h}_j^{l-1}) - I_{v_{ij}}(h_j^l, \hat{h}_j^l)) \cdot \Pr(e_{ij} | \lambda_1, \lambda_2, n_{kl})
\end{aligned}
$$

The likelihood is:

$$
\begin{aligned}
&\Pr(e_{ij} | s_{ij} = k, r_{ij} = l, \lambda_1, \lambda_2, n_{kl}^{-e_{ij}}) \\
\propto &P(e_{ij}, \boldsymbol{e} \backslash \{e_{ij}\}, s_{ij} = k, r_{ij} = l, , n_{kl}^{-e_{ij}}, \lambda_1, \lambda_2) \\
= &\frac{\Gamma(e_{ij} + n_{k,l}^1 + \lambda_1)\Gamma(1 - e_{ij} + n_{k,l}^0 + \lambda_2)}{\Gamma(1 + n_{k,l} + \lambda_1 + \lambda_2)} \\
&P(e_{ij} | \boldsymbol{e} \backslash \{e_{ij}\}, s_{ij} = k, r_{ij} = l, \boldsymbol{Z} \backslash \{s_{ij}, r_{ij}\}, \lambda_1, \lambda_2) \\
= &\begin{cases} \frac{n_{k,l}^1 + \lambda_1}{n_{k,l} + \lambda_1 + \lambda_2} & \text{if } e_{ij} = 1; \\ \frac{n_{k,l}^0 + \lambda_2}{n_{k,l} + \lambda_1 + \lambda_2} & \text{if } e_{ij} = 0. \end{cases}
\end{aligned}
$$

Here $n_{k,l} = \sum_{i'j'} \mathbf{1}(s_{i'j'} = k, r_{i'j'} = l)$, $n_{k,l}^1 = \sum_{s_{i'j'} = k, r_{i'j'} = l} e_{i'j'}$, and $n_{k,l}^0 = n_{k,l} - n_{k,l}^1$.

- Selected and report the results on 3 datasets. NIPS Co-authorship dataset, the lazega-lawfirm dataset and the MIT Reality Mining dataset
- ten-folds cross-validation to complete this task, where we randomly select one out of ten for each node's link data as test data and the others as training data
- he criteria in evaluating this predict ability includes the train error ($0-1$ loss), the test error ($0-1$ loss), the test log likelihood and the AUC (Area Under the roc Curve) score $cMMSB^{uv}$) obtain either comparable or better performance with other state-of-the-art.

Experiment - Result

Table: Different models' performance (Mean $\mp$ Standard Deviation) on Real world datasets

dataset		*Train error*	*Test error*	*Test log likelihood*	*AUC*
NIPS co-author	*IRM*	0.0317 ∓ 0.0004	0.0423 ∓ 0.0014	-135.0467 ∓ 7.3816	0.8901 ∓ 0.0162
	LFRM	0.0482 ∓ 0.0794	0.0239 ∓ 0.0735	-105.2166 ∓ 179.5505	0.9348 ∓ 0.1667
	MMSB	0.0132 ∓ 0.0042	0.0301 ∓ 0.0064	-86.2134 ∓ 10.1258	0.9524 ∓ 0.0215
	iMMM	$\mathbf{0.0061 \mp 0.0019}$	0.0253 ∓ 0.0035	-83.4264 ∓ 9.4293	0.9574 ∓ 0.0155
	$cMMSB^{\pi}$	0.0066 ∓ 0.0038	$\mathbf{0.0231 \mp 0.0043}$	-83.4261 ∓ 9.4280	0.9569 ∓ 0.0159
	$cMMSB^{uv}$	0.0097 ∓ 0.0047	0.0240 ∓ 0.0065	$\mathbf{-83.4257 \mp 9.4292}$	$\mathbf{0.9581 \mp 0.0153}$
MIT realtiy	*IRM*	0.0627 ∓ 0.0002	0.0665 ∓ 0.0004	-133.8037 ∓ 1.1269	0.8261 ∓ 0.0047
	LFRM	0.0397 ∓ 0.0017	0.0629 ∓ 0.0037	-143.6067 ∓ 10.0592	0.8529 ∓ 0.0179
	MMSB	0.0243 ∓ 0.0105	0.0716 ∓ 0.0043	-129.4354 ∓ 7.6549	0.8561 ∓ 0.0176
	iMMM	0.0297 ∓ 0.0055	0.0625 ∓ 0.0015	-126.7876 ∓ 3.4774	0.8617 ∓ 0.0124
	$cMMSB^{\pi}$	0.0246 ∓ 0.0016	0.0489 ∓ 0.0016	-125.3876 ∓ 3.2689	0.8794 ∓ 0.0159
	$cMMSB^{uv}$	0.0283 ∓ 0.0035	0.0438 ∓ 0.0015	-123.3876 ∓ 3.1254	0.8738 ∓ 0.0364
Lazega lawfirm	*IRM*	0.0987 ∓ 0.0003	0.1046 ∓ 0.0012	-201.7912 ∓ 3.3500	0.7056 ∓ 0.0167
	LFRM	0.0566 ∓ 0.0024	0.1051 ∓ 0.0064	-222.5924 ∓ 16.1985	0.7970 ∓ 0.0197
	MMSB	0.0391 ∓ 0.0071	0.0913 ∓ 0.0030	-212.1256 ∓ 3.2145	0.7789 ∓ 0.0102
	iMMM	0.0487 ∓ 0.0068	0.1096 ∓ 0.0026	-202.7148 ∓ 5.3076	0.7874 ∓ 0.0141
	$cMMSB^{\pi}$	0.0246 ∓ 0.0050	0.1023 ∓ 0.0056	-201.0154 ∓ 5.2167	0.8273 ∓ 0.0148
	$cMMSB^{uv}$	0.0276 ∓ 0.0043	0.1143 ∓ 0.0019	-204.0289 ∓ 9.5460	0.8215 ∓ 0.0167

The second work: Dynamic Infinite Mixed-Membership Stocastic Blockmodel

Summary of its advantage:

- allows the infinite number of communities;
- it allows mixed-membership for each node
- the model extends to the dynamic settings.
- it is apparent that in many social networking applications, a node's membership may become persistent over consecutive times, for example, a person's opinion of his peer is more likely to be consistent in two consecutive times.

Continue, ...

- $E = \{e_{ij}^t\}_{n \times n}^{1:T}$: entire set observations: if i has a relationship to node j at time t, it implies $e_{ij}^t = 1$. Otherwise, $e_{ij}^t = 0$.
- Each e_{ij}^t is specific to each pair of communities membership indicators (s_{ij}^t, r_{ij}^t).
- For each node pair i and j, at t, s_{ij}^t refers to the sender's community membership indicator. r_{ij}^t is for the receiver's community membership indicator.
- Z to denote all the hidden labels $\{s_{ij}^t, r_{ij}^t\}$.

Notations (2)

- Each node i at time t, has a mixed-membership distribution, $\boldsymbol{\pi}_i^t$ having infinite components, and the k^{th} component of $\boldsymbol{\pi}_i^t$: $\boldsymbol{\pi}_{ik}^t$ represents the "significance" of community k for node i.
- Role-compatibility matrix W: its $(k, l)^{\text{th}}$ entry, i.e., $W_{k,l}$ represents compatibilities between communities k and l. dimension of W can potentially be $\infty \times \infty$. Each $W_{k,l}$ is i.i.d from $Beta(\lambda_1, \lambda_2)$ which gives conjugacy to the Bernoulli distribution used to generate e_{ij}^t

Notations (3)

- $n^t_{k,l}$ denote the number of links from communities k to l, i.e., the number of times in which $s^t_{ij} = k$ and $r^t_{ij} = l$ simultaneously. $n^t_{k,l} = n^{t,1}_{k,l} + n^{t,0}_{k,l}$. **scalar**
- $n^{t,1}_{k,l}$ denotes the part of $n^t_{k,l}$ where the corresponding $e^t_{ij} = 1$.
- The number of times that a node i has participated in community k (both as a sending and receiving) at time t is represented by N^t_{ik} **vector**

Mixture Time Variant (MTV) and Mixture Time Invariant (MTI) Models

- To address the **phenomenon** that one's social community's memberships may change over times: allow each node's mixed-membership indicators to change cross times.
- Additionally, imperative that these indicators should have some persistence with its past values which reflects the reality of social behaviour.
- The persistence is achieved in two ways:
  - **"Mixture Time Variant (MTV)"** Mixed-membership distributions itself to change over times. Membership indicator of a node at time t is dependent on the "statistics" of all membership indicators of the same node at $t-1$ and $t+1$.
  - **"Mixture Time Invariant (MTI)"** mixed-membership distributions to stay invariant over times. Membership indicator at time t is dependent and more likely to have the same value as it was in $t-1$.

Mixture Time Variant (MTV)

Generative model

(1) Across all times $1:T$:

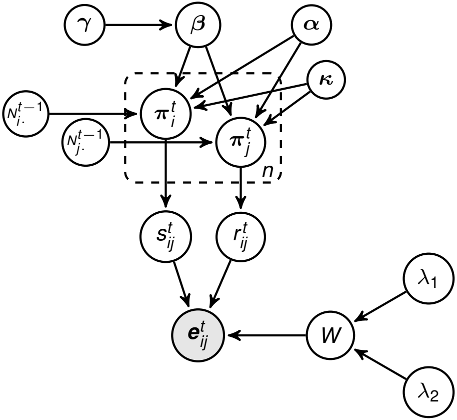
- $\boldsymbol{\beta} \sim GEM(\gamma)$
- $W_{k,l} \sim Beta(\lambda_1, \lambda_2)\ \forall k, l$

(2) Membership distributions:

- $\boldsymbol{\pi}_i^t \sim DP\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \frac{\kappa}{2n} \cdot \sum_k N_{ik}^{t-1} \boldsymbol{\delta}_k}{\alpha + \kappa}\right)$
  node i's mixed membership distribution at t.

$N_{ik}^{t-1} = \sum_{l=1}^{N} \mathbf{1}(s_{il}^{t-1} = k) + \sum_{l=1}^{N} \mathbf{1}(r_{li}^{t-1} = k)$, count number of nodes associated with a community k at time $t-1$.

Graphical model



Generative model

(3) Relationship Sampling: For each pair of $i, j \in \{1, \cdots, n\}, t \in \{1, \cdots, T\}$

- $s_{ij}^t \sim Multi(\boldsymbol{\pi}_i^t)$: sending community's indicator;
- $r_{ij}^t \sim Multi(\boldsymbol{\pi}_j^t)$: receiving community's indicator;
- $e_{ij}^t \sim Bernoulli(W_{s_{ij}^t, r_{ij}^t})$

Graphical model only shows for t, and omit the other times, where the structure is identical.

Graphical model

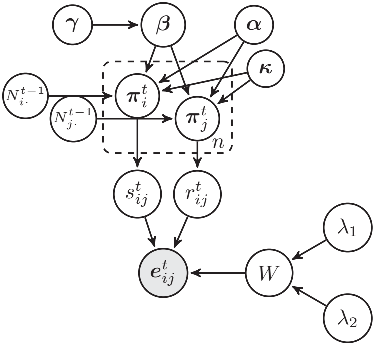


Figure: The *MTV-DIM3* Model

Mixture Time Variant (MTV): more explaination

- β "global" representing the "significance" of all existing communities at all times, while W is the communities' compatibility matrix.
- The prior $P(W)$ is element-wise *Beta* distributed, which is conjugate to the Bernoulli distribution $P(e^t_{i,j}|.)$: can obtain a marginal distribution of $P(e^t_{i,j}) = \int_W p(e^t_{i,j}|W)p(W)d(W)$ analytically. Do not explicitly sample values of W.
- Mixed-membership distribution $\{\boldsymbol{\pi}^t_i\}^{1:T}_{1:n}$ is sampled from the Dirichlet Process with a concentration parameter $(\alpha + \kappa)$ and a base measure $\frac{\alpha\boldsymbol{\beta} + \frac{\kappa}{2n}\sum_k N^{t-1}_{ik}\boldsymbol{\delta}_k}{\alpha+\kappa}$. There will be $N \times T$ of these distributions. They jointly describe each node's activities. It should be noted that each $\boldsymbol{\pi}^t_i$ is responsible to generate both the senders' label $\{s^t_{ij}\}^n_{j=1}$ from node i and receivers' label $\{r^t_{ji}\}^n_{j=1}$ to node i.

Mixture Time Variant (MTV): persistency

- Sticky parameter κ stands for each node's time influence on its mixed-membership distribution.
- In another words, we assume that each node's mixed-membership distribution at time t will be largely influenced by its activities at time $t-1$.
- This is reflected in the hidden label's multinomial distribution that the previous explicit activities will occupy a fixed proportion $\frac{\kappa}{\alpha+\kappa}$ to the current distribution. The larger the value of κ, the more weight that the activities at $t-1$ is going to play at time t.

Generative model

1. Across all times $1:T$.
   - $\boldsymbol{\beta} \sim GEM(\gamma)$
   - $W_{k,l} \sim Beta(\lambda_1, \lambda_2), \forall k, l$
2. Membership distribution
   - $\boldsymbol{\pi}_i^{(k)} \sim DP\left(\alpha_i + \kappa, \frac{\alpha_i \boldsymbol{\beta} + \kappa \boldsymbol{\delta}_k}{\alpha_i + \kappa}\right)$
3. Relationship Sampling, For $i, j \in \{1, \cdots, n\}, t \in \{1, \cdots, T\}$
   - $s_{ij}^t \sim Multi(\boldsymbol{\pi}_i^{(s_{ij}^{t-1})})$
   - $r_{ij}^t \sim Multi(\boldsymbol{\pi}_j^{(r_{ij}^{t-1})})$
   - $e_{ij}^t \sim Bernoulli(W_{s_{ij}^t, r_{ij}^t})$ relation from nodes i to j at time t.

Graphical model

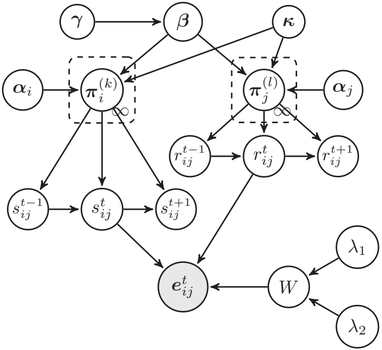


Figure: The *MTI-DIM3* Model

Mixture Time Invariant (MTI) Model: More explanation

- Each node has a variable number of membership distributions associated with, potentially infinite. At each time t, its membership indicator s_{ij}^t is generated from $\pi_{s_{ij}^{t-1}}$. In order to encourage persistence, each π_{ik} was generated from a corresponding β, where κ was added to β's k^{th} component [**?**].
- $\boldsymbol{\beta}$ and W's generation is the same MTV The set of membership indicators $\{s_{ij}^t, r_{ji}^t | j = 1, \cdots, n, t = 1, \cdots, T\}$ will be sampled from the time-invariant mixed-membership distribution set, $\{\boldsymbol{\pi}_i^{(k)}\}_{k=1}^{\infty}$, where each member is independently distributed from a Dirichlet Process with a concentration parameter $(\alpha + \kappa)$ and a base measure $\frac{\alpha\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k}{\alpha+\kappa}$.
- At time t, a membership indicator s_{ij}^t(or r_{ji}^t) is sampled from the distribution $\boldsymbol{\pi}_i^{(s_{ij}^{t-1})}$(or $\boldsymbol{\pi}_i^{(r_{ji}^{t-1})}$)$\forall i \in \{1, \cdots, n\}$.

Inference

Two sampling schemes are implemented for *MTV-DIM3*:

- Standard Gibbs sampling
- Slice-Efficient sampling

MTI-DIM3 sampling is very similar, so will not duplicate in these slides.

Two sampling schemes are implemented for *MTV-DIM3*:

- Standard Gibbs sampling
- Slice-Efficient sampling

- Largely based on [Teh et al.(2006)Teh, Jordan, Beal, and Blei]).
- Variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables eating dish k without counting the tables that generated from the sticky portion, i.e., κN_{ik}^{t-1}.
- We do not sample $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$, as it gets integrated out.

Sampling $\boldsymbol{\beta}$:
$\boldsymbol{\beta}$ is the prior for all $\{\boldsymbol{\pi}_i^t\}s$, can be thought as the ratios between the community components for all communities. Posterior distribution is obtained through auxiliary variable $\hat{\boldsymbol{m}}$:

$$(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K, \boldsymbol{\beta}_\mu) \sim Dir(\hat{\boldsymbol{m}}_{.1}, \cdots, \hat{\boldsymbol{m}}_{.K}, \gamma)$$

Gibbs Sampling details

Posterior $P(s_{ij}^t = k, r_{ij}^t = l | Z \backslash \{s_{ij}^t, r_{ij}^t\}, e, \boldsymbol{\beta}, \alpha, \lambda_1, \lambda_2, \kappa) \propto T_1 \times T_2 \times T_3$:

$$T_1 = \frac{\Gamma(\alpha\beta_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa)}{\Gamma(\alpha\beta_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t})} \frac{\Gamma(\alpha\beta_k + \kappa N_{ik}^{t,-s_{ij}^t})}{\Gamma(\alpha\beta_k + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa)} \times$$

$$\begin{cases} \alpha\beta_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t} & k \in \{1, \dots, K\} \\ \alpha\beta_u & k = K + 1 \end{cases}$$

$$T_2 = \frac{\Gamma(\alpha\beta_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa)}{\Gamma(\alpha\beta_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t})} \frac{\Gamma(\alpha\beta_l + \kappa N_{jl}^{t,-r_{ij}^t})}{\Gamma(\alpha\beta_l + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa)} \times$$

$$\begin{cases} \alpha\beta_l + \kappa N_{jl}^{t-1} + N_{jl}^{t,-r_{ij}^t}, & l \in \{1, \dots, l\} \\ \alpha\beta_u, & l = l + 1 \end{cases}$$

$$T_3 = \begin{cases} \frac{n^{t,1,-e_{ij}^t} + \lambda_1}{n^{t,-e_{ij}^t} + \lambda_1 + \lambda_2}, & \text{if } e_{ij}^t = 1 \\ \frac{n^{t,0,-e_{ij}^t} + \lambda_2}{n^{t,-e_{ij}^t} + \lambda_1 + \lambda_2}, & \text{if } e_{ij}^t = 0 \end{cases}$$

Using the restaurant-table-dish analogy, we denote $\boldsymbol{m}^t_{ik}$ as the number of tables eating dish k $\forall i, k, t$. This is related to the variable $\hat{\boldsymbol{m}}$ used in sampling $\boldsymbol{\beta}$, but also including the counts of the "un-sticky" portion, i.e., $\alpha\boldsymbol{\beta}_k$.
The sampling of $\boldsymbol{m}^t_{ik}$ is to incorporate a similar strategy as [Teh et al.(2006)Teh, Jordan, Beal, and Blei, **?**], which is independently distributed from:

$$\Pr(\boldsymbol{m}^t_{ik} = m | \alpha, \boldsymbol{\beta}_k, N^{t-1}_{ik}, \kappa)$$
$$\propto S(N^t_{ik}, m)(\alpha\boldsymbol{\beta}_k + \kappa N^{t-1}_{ik})^m$$

Here $S(\cdot, \cdot)$ is the Stirling number of first kind.

Sampling $\hat{\boldsymbol{m}}$ cont.

For each node, the ratio of generating new tables can result from two factors: (1) Dirichlet prior with parameter $\{\alpha, \boldsymbol{\beta}\}$ and (2) the sticky configuration from membership indicators at $t-1$, i.e., κN_{ik}^{t-1}.
To sample $\boldsymbol{\beta}$, we need to only include tables generated from the "un-sticky" portion, i.e., $\hat{\boldsymbol{m}}$, where each $\hat{\boldsymbol{m}}_{ik}^{t}$ can be obtained from a single Binomial draw:

$$\hat{\boldsymbol{m}}_{ik}^{t} \sim Binomial(\boldsymbol{m}_{ik}^{t}, \frac{\alpha\boldsymbol{\beta}_k}{\frac{\kappa}{2n}N_{ik}^{t-1} + \alpha\boldsymbol{\beta}_k}).$$

Adapted Slice-Efficient Sampling

- Incorporate the slice-efficient sampling [**?**][**?**]. The original sampling scheme was designed to sample the Dirichlet Process Mixture model.
- We use auxiliary variables $U = \{u^t_{ij,s}, u^t_{ij,r}\}$ for each of the latent membership pair $\{s^t_{ij}, r^t_{ij}\}$. Having the Us, we are able to limit the number of components in which $\boldsymbol{\pi}_i$ needs to be considered, which is infinite otherwise.
- Under the slice-efficient sampling framework, the variables of interest are now extended to: $\boldsymbol{\pi}^t_i, \{u^t_{ij,r}, u^t_{ij,s}\}, \{s^t_{ij}, r^t_{ij}\}, \boldsymbol{\beta}, \boldsymbol{m}$:

Adapted Slice-Efficient Sampling

- Incorporate the slice-efficient sampling [**?**][**?**]. The original sampling scheme was designed to sample the Dirichlet Process Mixture model.
- We use auxiliary variables $U = \{u^t_{ij,s}, u^t_{ij,r}\}$ for each of the latent membership pair $\{s^t_{ij}, r^t_{ij}\}$. Having the Us, we are able to limit the number of components in which $\boldsymbol{\pi}_i$ needs to be considered, which is infinite otherwise.
- Under the slice-efficient sampling framework, the variables of interest are now extended to: $\boldsymbol{\pi}^t_i, \{u^t_{ij,r}, u^t_{ij,s}\}, \{s^t_{ij}, r^t_{ij}\}, \boldsymbol{\beta}, \boldsymbol{m}$:

For each node $i = 1, \cdots, N$: we generate $\boldsymbol{\pi}'^{t}_{i}$ using sticky-breaking process [**?**], where each k^{th} component is generated using:
$\boldsymbol{\pi}'^{t}_{ik} \sim$beta$(\boldsymbol{\pi}'^{t}_{ik}; a^t_{ik}, b^t_{ik})$, where

$$
\begin{aligned}
a^t_{ik} &= \alpha\boldsymbol{\beta}_k + N^t_{ik} + \kappa N^{t-1}_{ik} \\
b^t_{ik} &= \alpha(1 - \sum_{l=1}^{k} \boldsymbol{\beta}_l) + N^t_{i,k_0>k} + \kappa N^{t-1}_{i,k_0>k}
\end{aligned}
$$

Here $\boldsymbol{\pi}_k = \boldsymbol{\pi}'_k \prod_{i=1}^{k-1}(1 - \boldsymbol{\pi}'_i)$.

Sampling $u^t_{ij,s}, u^t_{ij,r}, s^t_{ij}, r^t_{ij}$

We use $u^t_{ij,s} \sim U(0, \boldsymbol{\pi}^t_{is^t_{ij}})$, $u^t_{ij,r} \sim U(0, \boldsymbol{\pi}^t_{jr^t_{ij}})$. Then the obtained hidden label is independently sampled from the finite candidates:

$$
\begin{aligned}
& P(s^t_{ij} = k, r^t_{ij} = l | Z, e^t_{ij}, \boldsymbol{\beta}, \alpha, \kappa, N, \boldsymbol{\pi}, u^t_{ij,s}, u^t_{ij,r})) \\
\propto & \mathbf{1}(k : \boldsymbol{\pi}^t_{ik} > u^t_{ij,s}) \cdot \mathbf{1}(l : \boldsymbol{\pi}^t_{jl} > u^t_{ij,r}) \\
& \cdot \prod_{l=1}^{2n} P(z^{t+1}_{il} | z^t_{i\cdot} / s^t_{ij}, s^t_{ij} = k, \boldsymbol{\beta}, \alpha, \kappa, N^{t+1}_i) \\
& \cdot \prod_{l=1}^{2n} P(z^{t+1}_{jl} | z^t_{j\cdot} / r^t_{ij}, r^t_{ij} = l, \boldsymbol{\beta}, \alpha, \kappa, N^{t+1}_j) \\
& \cdot P(e^t_{ij} | E \backslash \{e^t_{ij}\}, s^t_{ij} = k, r^t_{ij} = l, \boldsymbol{Z} \backslash \{s^t_{ij}, r^t_{ij}\}, \lambda_1, \lambda_2)
\end{aligned}
$$

Sampling β, $\boldsymbol{m}$

This is the same as the Gibbs sampling.

Hyper-parameter Sampling

- Hyper-parameters are γ, α, κ. It is impossible to compute the posterior individually. Therefore, three prior are used:
  - $\mathcal{G}(1,1)$ is placed on both γ and $(\alpha + \kappa)$
  - A beta prior is placed on the ratio $\frac{\kappa}{\alpha+\kappa}$.
- Sample γ value, since $\log(\gamma)$'s posterior is log-concave (not concave!), use Adaptive Rejection Sampling (ARS) method [Rasmussen(2000)].
- Sample $(\alpha + \kappa)$, use auxiliary variable $\boldsymbol{m}$ as proposed in [Teh et al.(2006)Teh, Jordan, Beal, and Blei].
- Sample $\frac{\kappa}{\alpha+\kappa}$, place $\mathcal{B}(1,1)$ on it, with a likelihood of $\{\boldsymbol{m}_{ik}^{t} - \hat{\boldsymbol{m}}_{ik}^{t}, \forall i, k, t > 1\}$, the posterior is in an analytical and samplable form, due to conjugacy.

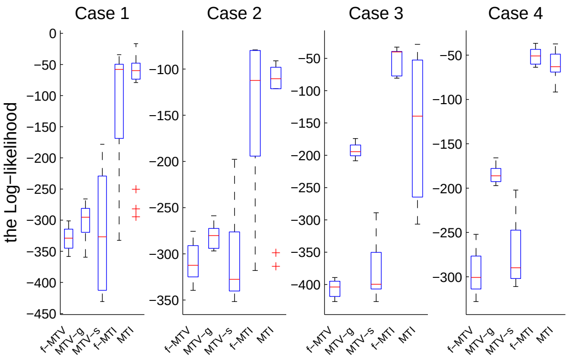
:

- Gibbs Sampling integrates out the mixed-membership distribution $\{\pi_i^t\}$. It is the "marginal approach" [Papaspiliopoulos and Roberts(2008)]. The property of community exchangeability makes it simple to implement. However, theoretically, it mix slowly as the sampling of each label is dependent on other labels.
- The Slice-Efficient Sampling is one "conditional approach" [**?**] while the membership indicators are independently sampled from $\{\pi_i^t\}$. In each iteration, given $\{\pi_i^t\}$, we can parallelize the process of sampling membership indicators, which may help to improve the computation, especially when the number of nodes, i.e., N becomes larger, and the number of communities, i.e., k becomes smaller.

Experiments

- *DIM3* model run on synthetic datasets. Test against finite-communities case as baseline, namely *f-MTV* & *f-MTI*.
- For synthetic data generation, variables generated following [**?**]: $N = 20, T = 3 \implies E$ is a $20 \times 20 \times 3$ asymmetric and binary matrix. The parameters are set up such that, 20 nodes are equally partitioned into 4 groups. The ground-truth of the mixed-membership distribution for each of the groups are: $[0.8, 0.2, 0; 0, 0.8, 0.2; 0.1, 0.05, 0.85; 0.4, 0.4, 0.2]$.
- Consider 4 different test case role-compatibility matrix:
  - **Case 1**: large diagonal values and small non-diagonal values
  - **Case 2**: large diagonal values and mediate non-diagonal values
  - **Case 3**: large non-diagonal values and small diagonal values
  - **Case 4**: small diagonal values and mediate non-diagonal values

Results: Log-likelihood Performance



From the log-likelihood comparison, we can see that the *MTI* model performs better than the *MTV* model.

Further results

The average l_2 distance between the mixed-membership distributions and its ground-truth; and the one between the posterior role-compatibility matrix and its ground-truth:

Table: Average l_2 Distance to the Ground-truth

Cases	Role-Compatibility Matrix					Mixed-Memberships				
	f-MTV	*MTV-g*	*MTV-s*	*f-MTI*	*MTI*	*f-MTV*	*MTV-g*	*MTV-s*	*f-MTI*	*MTI*
1	0.529	0.625	0.848	0.114	**0.086**	0.366	0.384	0.403	0.199	**0.191**
2	0.439	0.225	0.339	**0.195**	0.204	0.355	0.355	0.319	**0.207**	0.227
3	0.134	0.201	0.513	0.117	**0.087**	0.278	0.289	0.589	0.208	**0.187**
4	**0.195**	0.214	0.267	0.220	0.219	0.258	0.285	0.277	0.192	**0.182**

On the average l_2 distance to the ground-truth performance, the *MTI* model also performs better.

Current work

- MCMC diagnostics
- Test real relational data
- Vision applications?

- Airoldi, E., D. Blei, S. Fienberg, and E. Xing, 2008: Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, **9**, 1981–2014.
- Kim, D., M. Hughes, and E. Sudderth, 2012: The nonparametric metadata dependent relational model. *Proceedings of the 29th Annual International Conference on Machine Learning*, ACM.
- Koutsourelakis, P. and T. Eliassi-Rad, 2008: Finding mixed-memberships in social networks. *Proceedings of the 2008 AAAI spring symposium on social information processing*.
- Papaspiliopoulos, O. and G. Roberts, 2008: Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, **95 (1)**, 169–186.
- Rasmussen, C., 2000: The infinite gaussian mixture model. *Advances in neural information processing systems*, **12 (5.2)**, 2.
- Teh, Y., M. Jordan, M. Beal, and D. Blei, 2006: Hierarchical dirichlet processes. *Journal of the American Statistical Association*, **101 (476)**, 1566–1581.
- Xing, E., W. Fu, and L. Song, 2010: A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, **4 (2)**, 535–566.