A Short Course on Bayesian Nonparametrics

Lecture 4 - Nonparametric methods under partial exchangeability

Abel Rodriguez - UC, Santa cruz

Universidade Federal Do Rio de Janeiro
March, 2011

Exchangeability

Definition

A sequence $y_1, y_2, \ldots$ is exchangeable if for any I

$$P(y_1, \ldots, y_I) = P(y_{\pi(1)}, \ldots, y_{\pi(I)})$$

for all permutations π of the set $\{1, \ldots, I\}$.

Exchangeability

Definition

$y_1, y_2, \ldots$ is exchangeable if there is a measure Q defined on $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ such that

$$P(y_1, \ldots, y_I) = \int_{\mathcal{F}} \left\{ \prod_{i=1}^{I} F(y_i) \right\} dQ(F)$$

or, in a more familiar hierarchical formulation,

$$y_i | F \sim_{iid} F \qquad\qquad F \sim Q$$

Partial exchangeability

- Exchangeability involves an assumption about complete symmetry among all the observables.
- In many situations this assumption is too restrictive, but judgements about partial exchangeability might be justified.
  - Unrestricted exchangeability of exchangeable sequences.
  - Separate exchangeability.
  - Joint exchangeability.

Unrestricted exchangeability of exchangeable sequences

- Consider collecting data on academic performance for eleventh graders from different schools, so that $y_{i,j}$ corresponds to the grade of the j-th students in the i-th school in the sample, with $i = 1, 2, \ldots$ and $j = 1, 2, \ldots$.
- We might not be willing to assume that the sequence $y_{1,1}, y_{1,2}, \ldots, y_{2,1}, y_{2,2}, \ldots$ is completely exchangeable, because that would imply that schools have no effect on academic performance.
- However, we might be willing to assume that students are exchangeable within each school, and that schools are exchangeable among themselves.

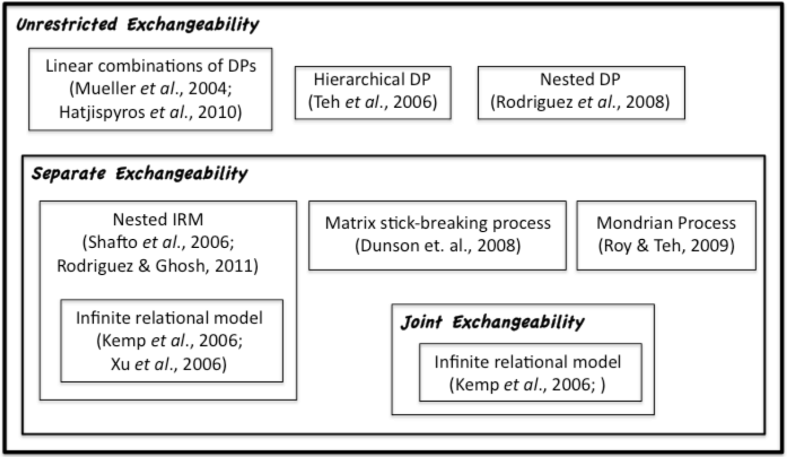
Separate exchangeability

- Consider modeling the voting record of U.S. senators. In this case $y_{i,j}$ represents the vote of senator i on bill j, with $y_{i,j} = 1$ if the vote was positive and $y_{i,j} = 0$ otherwise.
- Both exchangeability and unrestricted exchangeability would be difficult to defend in this example; the first would imply that the probability of a positive vote is the same for every senator and every bill, the second would imply that the probability of a positive vote would depend on the senator, but not on the content of the bill under consideration.
- However, it might seem reasonable to assume that, for any ordering of the bills, the model must be invariant to permutations in the ordering of the senators, and viceversa.

Joint exchangeability

- Consider modeling the number of email exchanges among Facebook users during one day, so that $y_{i,j}$ is the number of emails sent by user i to user j.
- Again, none of the partial exchangeability notions discussed so far is suitable for this problem, as they would imply that the rate at which individuals interact does not depend on the identity of the subjects.
- However, it would be reasonable to assume that the probability model should be invariant when the same permutation is applied two both rows and columns.

Some models BNP models for partial exchangeability



Unrestricted Exchangeability

Linear combinations of DPs (Mueller *et al.*, 2004; Hatjispyros *et al.*, 2010)

Hierarchical DP (Teh *et al.*, 2006)

Nested DP (Rodriguez *et al.*, 2008)

Separate Exchangeability

Nested IRM (Shafto *et al.*, 2006; Rodriguez & Ghosh, 2011)

Infinite relational model (Kemp *et al.*, 2006; Xu *et al.*, 2006)

Matrix stick-breaking process (Dunson et. al., 2008)

Mondrian Process (Roy & Teh, 2009)

Joint Exchangeability

Infinite relational model (Kemp *et al.*, 2006;)

Unrestricted exchangeability

Definition

$y_{i,1}, y_{i,2}, \ldots$ for $i = 1, 2, \ldots$ are unrestrictedly exchangeable if, for any finite integers I and $J_1, \ldots, J_I$, we have

$$\begin{aligned} P(\{y_{i,j} : j = 1, \ldots, J_i, i = 1, \ldots, I\}) \\ = P(\{y_{\sigma(i),\pi_{\sigma(i)}(j)} : j = 1, \ldots, J_i, i = 1, \ldots, I\}) \end{aligned}$$

for any permutation σ of the set $\{1, \ldots, I\}$ and π_i of the set $\{1, \ldots, J_i\}$ with $i = 1, \ldots, I$.

Unrestricted exchangeability

Definition

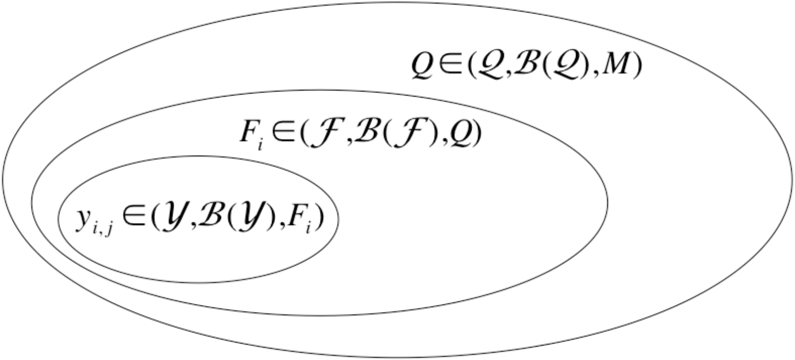
Unrestrictedly exchangeable sequences admit a representation of the form

$$P(\{y_{i,j} : j = 1, \ldots, J_i, i = 1, \ldots, I\}) \\ = \int_{\mathcal{Q}} \left(\prod_{i=1}^{I} \int_{\mathcal{F}} \left\{ \prod_{j=1}^{J_i} F_i(y_{i,j}) \right\} dQ(F_i) \right) dM(Q)$$

Or, in a more familiar hierarchical formulation,

$$y_{i,j} | F_i \sim_{iid} F_i \qquad F_i | Q \sim_{iid} Q \qquad Q \sim M$$

Structure of unrestricted exchangeable models


$Q \in (\mathcal{Q}, \mathcal{B}(\mathcal{Q}), M)$
$F_i \in (\mathcal{F}, \mathcal{B}(\mathcal{F}), Q)$
$y_{i,j} \in (\mathcal{Y}, \mathcal{B}(\mathcal{Y}), F_i)$

Unrestricted exchangeability

- We work with hierarchical mixtures,

$$y_{i,j}|\theta_{i,j} \sim \psi(y_{i,j}|\theta_{i,j}) \quad \theta_{i,j}|G_i \sim G_i \quad G_i|Q \sim Q \quad Q \sim M$$

- Independence and complete dependence among the G_is are special cases.
- Complete dependence ($G_i = G$ for all i) implies full exchangeability.

Linear combinations of DPs - Definition

- Introduced by Mueller et al, 2004.

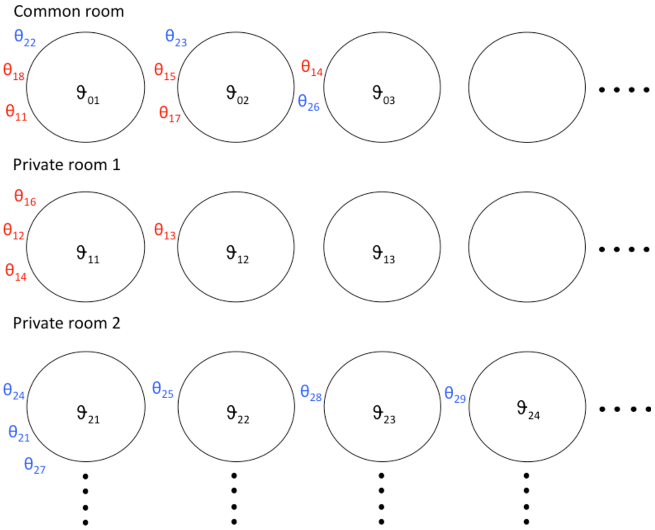
  $$G_i = \epsilon G_i^* + (1-\epsilon) G_0^* \quad G_i^* \sim \mathsf{DP}(\alpha, H) \quad G_0^* \sim \mathsf{DP}(\beta, H) \quad \epsilon \sim \mathsf{beta}(a, b)$$

  with $\epsilon \sim \mathsf{beta}(a, b)$.

- G_0^* is the "common" distribution and G_i^* is the "idiosyncratic" distribution associated with population i.
- $\epsilon = 0$ leads to complete exchangeability, while $\epsilon = 1$ leads to a set of independent random measures.
- Correlation structure induced by the model

$$\mathsf{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \frac{\epsilon^2}{(1+\alpha)} \mathsf{Var}_H(\vartheta) + \frac{(1-\epsilon)^2}{(1+\beta)} \mathsf{Var}_H(\vartheta) & i = i', j \neq j' \\ \frac{(1-\epsilon)^2}{(1+\beta)} \mathsf{Var}_H(\vartheta) & i \neq i' \end{cases}$$

Linear combinations of DPs - Joint distribution



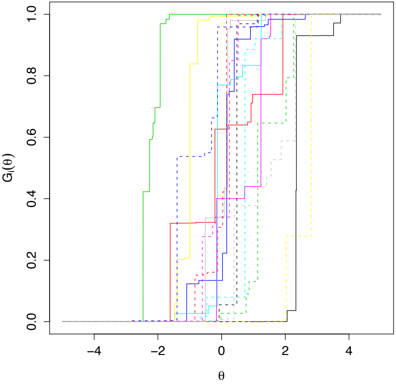
Dependence through random baseline measures

Just having conditionally independent DPs with a random parametric baseline measure is not flexible enough.

$$y_{i,j} | \theta_{i,j} \sim \mathrm{N}(\theta_{i,j}, 1) \quad \theta_{i,j} | G_i \sim G_i$$

$$G_i | \mu \sim \mathrm{DP}\left\{\alpha, \mathrm{N}(\mu, 1)\right\} \quad \mu \sim \mathrm{N}(0, 1)$$

This is not quite the same as an MDP!!!



Hierarchical Dirichlet processes

- Introduced by Teh et al, 2006.

$$G_i \sim \mathrm{DP}(\alpha, G_0) \qquad\qquad G_0 \sim \mathrm{DP}(\beta, H)$$

  where H is a parametric baseline measure.

- Since $G_0(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\vartheta_k}(\cdot) s$ is a.s. discrete,

$$G_i(\cdot) = \sum_{k=1}^{\infty} \varpi_{i,k} \delta_{\vartheta_k}(\cdot),$$

  with $(\varpi_{i,1}, \varpi_{i,2}, \ldots) \sim \mathrm{Dir}(w_1, w_2, \ldots)$.

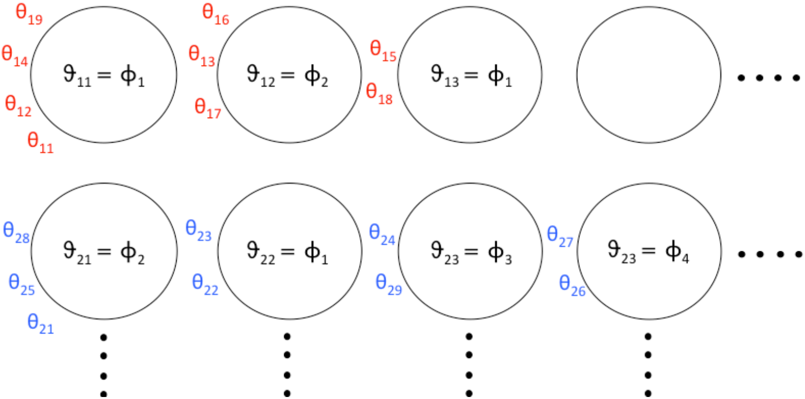
- If $\alpha \to \infty$ then we have $G_i = G_0$ for all i.
- If $\beta \to \infty$, then the G_is are independent from each other

Hierarchical Dirichlet processes

- Covariance structure

$$\mathrm{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \left[\frac{1}{(1+\beta)} + \frac{\beta}{(1+\beta)}\frac{1}{(1+\alpha)}\right]\mathrm{Var}_H(\vartheta) & i = i', j \neq j' \\ \frac{1}{(1+\beta)}\mathrm{Var}_H(\vartheta) & i \neq i' \end{cases}$$

Chinese restaurant franchise (CRPf)



Nested Dirichlet processes

- Previous models cluster observations but assume G_is are all different.
- Introduced in Rodriguez et al (2008),

$$G_i | Q \sim Q = \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*} \qquad G_k^* \sim_{iid} \mathrm{DP}(\beta, H)$$

  where $\pi_k = v_k \prod_{s<k}(1 - v_s)$ and $v_k \sim_{iid} \mathrm{beta}(1, \alpha)$

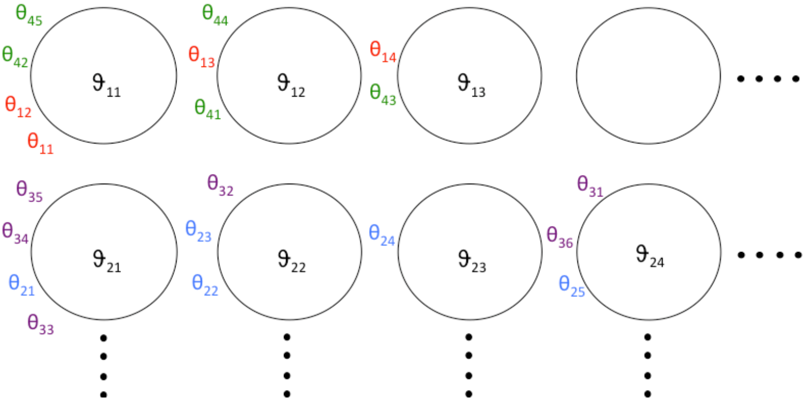
- This construction implies $Q \sim \mathrm{DP}\{\alpha, \mathrm{DP}(\beta, H)\}$.
- Cluster populations, and subjects within groups of populations.

Nested Dirichlet processes

- If $\alpha \to 0$ then we $G_i = G_j$ for all i and j.
- If $\alpha \to \infty$ then the G_is are independent.
- If $\beta \to 0$ then parametric clustering.
- The marginal correlation structure generated by the NDP is such that

$$\mathrm{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i = i', j \neq j' \\ \frac{1}{(1+\alpha)} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i \neq i' \end{cases}$$

Nested Dirichlet processes



Separate exchangeability

Definition

(Aldous, 1981) An infinite two-dimensional array $Y = \{y_{i,j} : i = 1, 2, \ldots, , j = 1, 2, \ldots\}$ of random variables is separately exchangeable if for any I and J

$$
\begin{aligned}
P(\{y_{i,j} : i = 1, \ldots, I, j = 1, \ldots, J\}) \\
= P(\{y_{\sigma(i),\pi(j)}, i = 1, \ldots, I, j = 1, \ldots, J\})
\end{aligned}
$$

where σ and π are permutations of the sets $\{1, \ldots, I\}$ and $\{1, \ldots, J\}$.

Separate exchangeability

Definition

An array $Y = [y_{ij}]$ is separately exchangeable if and only if

$$y_{i,j} = g(\eta, \lambda_i, \gamma_j, \varsigma_{i,j})$$

for some function g and independent random variables η, $\{\lambda_i\}$, $\{\gamma_j\}$ and $\{\varsigma_{i,j}\}$.

An example of separate exchangeability

- $y_{i,j}$ represents the vote of senator i on bill j, with $y_{i,j} = 1$ if the vote was positive and $y_{i,j} = 0$ otherwise.
- Assume that

$$y_{i,j} \sim \mathrm{Ber}(\theta_{i,j}) \quad \Phi^{-1}(\theta_{i,j}) = \mu + \phi_i + \varphi_j \quad \phi_i \sim \mathrm{N}(0,1) \quad \varphi_j \sim \mathrm{N}(0,1)$$

with Φ cdf of normal.

- Note that

$$y_{i,j} = g(\lambda_i, \gamma_j, \varsigma_{i,j}) = \begin{cases} 1 & \Phi^{-1}(\varsigma_{i,j}) < \mu + \Phi^{-1}(\lambda_i) + \Phi^{-1}(\gamma_j) \\ 0 & \text{otherwise} \end{cases}$$

where λ_i, γ_i and $\varsigma_{i,j}$ are indep. uniform random variables.

Infinite relational models (IRMs)

- Introduced in Kemp et al (2006) and Xu et al (2006)

$$y_{i,j}|\theta_{i,j} \sim \psi(y_{i,j}|\theta_{i,j}), \qquad \theta_{i,j} = \vartheta_{\zeta_i,\xi_j}$$

and

$$\zeta_i|\{w_k\} \sim \sum_{k=1}^{\infty} \omega_k \delta_k \qquad \xi_j|\{\varpi_l\} \sim \sum_{l=1}^{\infty} \varpi_l \delta_l \qquad \vartheta_{k,l} \sim H$$

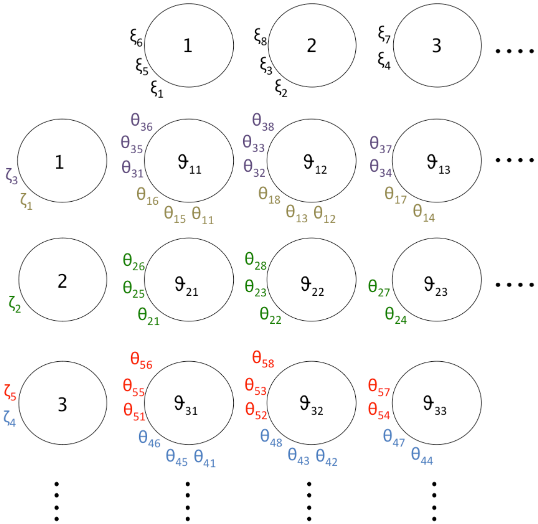
- $\omega_k = u_k \prod_{s<k}\{1 - u_s\}$ with $u_k \sim_{iid}$ beta$(1, \alpha)$
- $\varpi_l = v_l \prod_{s<l}\{1 - v_s\}$ with $v_l \sim_{iid}$ beta$(1, \beta)$

Correlation structure for the IRM

- Given the structure of the model, the marginal correlation structure generated by the IRM can be easily obtained, yielding,

$$\mathrm{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i = i', j \neq j' \\ \frac{1}{(1+\alpha)} \mathrm{Var}_H(\vartheta) & i \neq i', j = j' \\ \frac{1}{(1+\alpha)} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i \neq i', j \neq j' \end{cases} .$$

A Pólya urn for the IRM



Nested infinite relational models (NIRM)

- Discussed in Shaftko et al (2006) and Rodriguez & Ghosh, (2011).

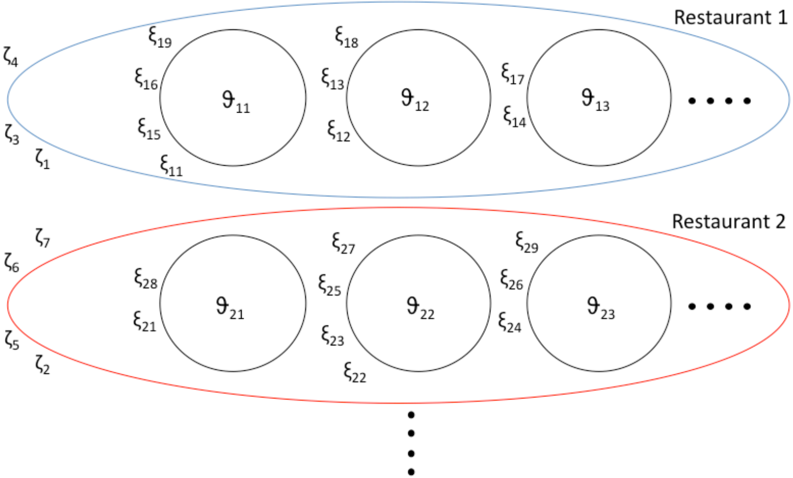
$$y_{i,j}|\theta_{i,j} \sim \psi(y_{i,j}|\theta_{i,j}) \qquad\qquad \theta_{i,j} = \vartheta_{\zeta_i,\xi_{\zeta_i,j}}$$

with

$$\zeta_i|\{w_k\} \sim \sum_{k=1}^{\infty} \omega_k \delta_k \quad \xi_{k,j}|\{\varpi_{k,l}\} \sim \sum_{l=1}^{\infty} \varpi_{k,l}\delta_l \quad \vartheta_{k,l} \sim H$$

- $\omega_k = u_k \prod_{s<k}\{1 - u_s\}$ with $u_k \sim \text{beta}(1,\alpha)$.
- $\varpi_{k,l} = v_{k,l} \prod_{s<l}\{1 - v_{k,s}\}$ with $v_{k,l} \sim \text{beta}(1,\beta)$.

A Pólya urn for the NIRM



Some properties of the NIRM

- Prior covariance structure is similar to the IRM.

$$\mathrm{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i = i', j \neq j' \\ \frac{1}{(1+\alpha)} \mathrm{Var}_H(\vartheta) & i \neq i', j = j' \\ \frac{1}{(1+\alpha)} \frac{1}{(1+\beta)} \mathrm{Var}_H(\vartheta) & i \neq i', j \neq j' \end{cases} .$$

- Expected number of distinct dishes is also the same

$$\left\{ \sum_{i=1}^{I} \frac{\alpha}{\alpha + i - 1} \right\} \left\{ \sum_{j=1}^{J} \frac{\beta}{\beta + j - 1} \right\} = o\left(\{\log I\}\{\log J\} \right)$$

- The main advantage of the NIRM is that it provides additional flexibility and interpretability by allowing the clusters of columns to vary across clusters of rows.

Hierarchical IRMs and NIRMs

- Share dishes across restaurants as in the CRP franchise. This reduces the number of unique dishes and allows for more complex (but still interpretable) clustering patterns.
- Obtained by setting $\vartheta_{k,l} | G_0 \sim G_0$ and $G_0 \sim \mathrm{DP}(\epsilon, H)$ in either the IRM or the NIRM.

Matrix stick-breaking processes

- Introduced in Dunson et al, (2008).

$$y_{i,j} \sim \psi(y_{i,j}|\theta_{i,j}) \qquad \theta_{i,j}|G_{i,j} \sim G_{i,j} \qquad G_{i,j} = \sum_{k=1}^{\infty} \omega_{i,j,k}\delta_{\vartheta_k}$$

- $\omega_{i,j,k} = u_{i,k} v_{j,k} \prod_{l<k} \{1 - u_{i,j,l} v_{i,j,l}\}$
- $u_{i,k} \sim_{iid}$ beta$(1,\alpha)$ and $v_{j,k} \sim_{iid}$ beta$(1,\beta)$.
- One distribution per cell!!!
- Covariance structure

$$\text{Cov}(\theta_{i,j}, \theta_{i',j'}) = \begin{cases} \frac{1}{(\beta+2)(\alpha+1)-1} & i = i', j \neq j' \\ \frac{1}{(\alpha+2)(\beta+1)-1} & i \neq i', j = j' \\ \frac{(\alpha+1)(\beta+1)}{2(\alpha+1)(\beta+1)-1} & i \neq i', j \neq j \end{cases}$$

Joint exchangeability

Definition

An infinite array $Y = [y_{ij}]$ of random variables is said to be jointly exchangeable if for any I

$$P(\{y_{i,j} : i = 1, \ldots, I, j = 1, \ldots, I\}) = \\ P(\{y_{\sigma(i),\sigma(j)}, i = 1, \ldots, I, j = 1, \ldots, I\})$$

where σ is a permutation of $\{1, \ldots, I\}$.

Joint exchangeability

Definition

An array is jointly exchangeable if and only if

$$y_{i,j} = g(\eta, \lambda_i, \lambda_j, \varsigma_{i,j})$$

for some function g and independent random variables η, $\{\lambda_i\}$, and $\{\varsigma_{i,j}\}$.

Note the difference with separate exchangeability, where rows and columns have their own distinct random variables.

Joint exchangeability

- The literature on nonparametric methods for jointly exchangeable arrays is limited.
- The best known model is an extension of the IRM

$$y_{i,j}|\theta_{i,j} \sim \psi(y_{i,j}|\theta_{i,j}) \quad \theta_{i,j} = \vartheta_{\zeta_i,\zeta_j} \quad \zeta_i|\{w_k\} \sim \sum_{k=1}^{\infty} \omega_k \delta_k \quad \vartheta_{k,l} \sim H$$

- $\omega_k = u_k \prod_{s<k}\{1-u_s\}$ and $u_k \sim \text{beta}(1,\alpha)$.
- This is an infinite dimensional version of the stochastic blockmodel $\Rightarrow$ ζ_is cluster subject into factions.

Text modeling

- Latent Dirichlet allocation (LDA) model (Blei et al, 2003).
- Very simple models: Words are assumed to be exchangeable.
- $y_{i,j} = d$ if the j-th word in the i-th document of the corpus equals the d-th word in the dictionary,

$$\Pr(y_{i,j} = d|\theta_i) = \theta_{i,d} \qquad \theta_i|G \sim G \qquad G \sim \mathrm{DP}\{\alpha, \mathrm{Dir}(\eta)\}$$

- All words within a document are drawn from a common θ_i, which is a probability a distribution over the words in the dictionary (topic).
- Documents are clustered according to the topic they cover.

Text modeling

- Hierarchical DP version:

$$\begin{aligned} \Pr(y_{i,j} = d | \theta_{i,j}) &= \theta_{i,j,d} \\ \theta_{i,j} | G_i &\sim G_i \\ G_i | G_0 &\sim \mathrm{DP}(\alpha, G_0) \\ G_0 &\sim \mathrm{DP}\{\alpha, \mathrm{Dir}(\eta)\} \end{aligned}$$

- Each document is made of a different mixture of “simple” topics.
- The topics are common among all documents (words are clustered).
- No clustering of documents.

Text modeling

- A model that allows for multiple topics per document as well as document clustering combines the HDP and the NDP

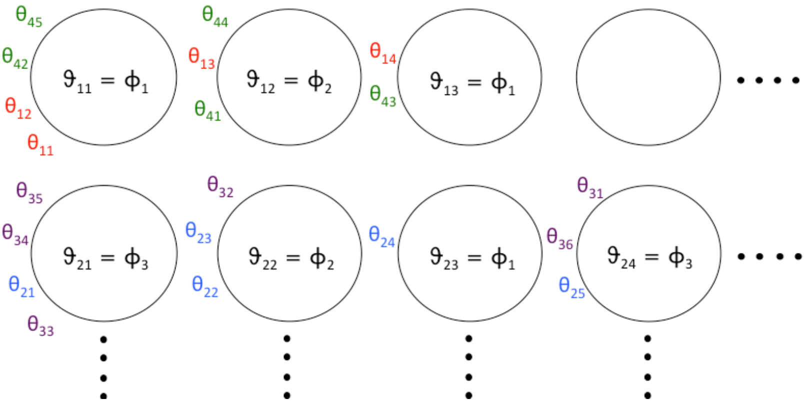
$$\Pr(y_{i,j} = d | \theta_{i,j}) = \theta_{i,j,d} \quad \theta_{i,j} | G_i \sim G_i \quad G_i | Q \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*}$$

where $\pi \sim \mathrm{SB}(\alpha)$

$$G_k^* | G_0 \sim \mathrm{DP}(\beta, G_0) \qquad G_0 \sim \mathrm{DP}\{\epsilon, \mathrm{Dir}(\eta)\}$$

- Sampling using a blocked Gibbs sampler that truncates all the distributions involved.

HNDP


θ_{45}
θ_{42}
θ_{12}
θ_{11}
$\vartheta_{11} = \phi_1$
θ_{44}
θ_{13}
θ_{41}
$\vartheta_{12} = \phi_2$
θ_{14}
θ_{43}
$\vartheta_{13} = \phi_1$
θ_{35}
θ_{34}
θ_{21}
θ_{33}
$\vartheta_{21} = \phi_3$
θ_{32}
θ_{23}
θ_{22}
$\vartheta_{22} = \phi_2$
θ_{24}
$\vartheta_{23} = \phi_1$
θ_{31}
θ_{36}
θ_{25}
$\vartheta_{24} = \phi_3$

Modeling multiple networks

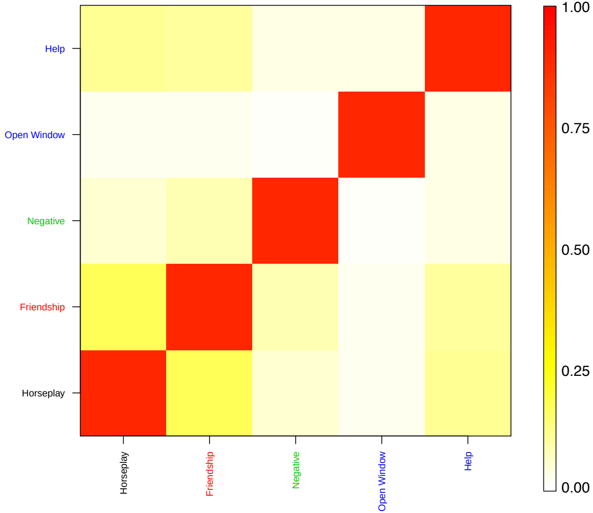
- Data on five different types of interaction (friendship, help, horseplay, negative comments, and open window) among a group of 14 people.
- Let $Y_r = [y_{i,j,r}]$ be the r-th matrix of interactions, where $y_{i,j,r} = 1$ if subjects i and j interact under relationship r.

$$y_{i,j,r} | \{\vartheta_{k,l,r}\}, \zeta_r, \xi_i, \xi_j \sim \mathsf{Ber}(\vartheta_{\xi_{\zeta_r,i}, \xi_{\zeta_r,j}, r}) \qquad \vartheta_{k,l,r} \sim \mathsf{beta}(1,1)$$

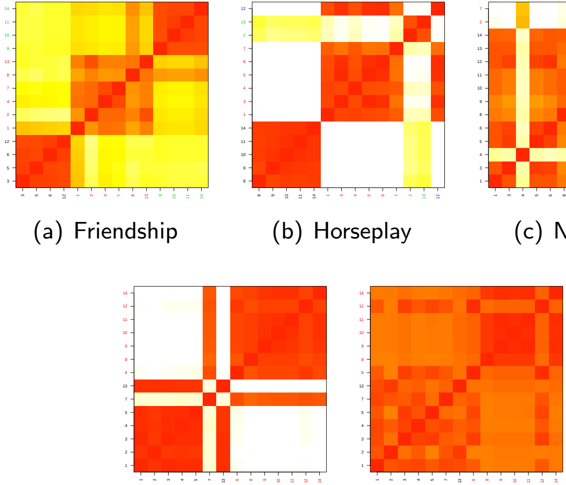
$$\zeta_r | \{\omega_k\} \sim \sum_{k=1}^{\infty} \omega_k \delta_k \qquad \xi_{k,j} | \{\varpi_k\} \sim \sum_{l=1}^{\infty} \varpi_{k,l} \delta_l$$

- $\{\xi_{k,i}\}$ cluster subjects into “factions” with similar patterns of interactions
- $\{\zeta_r\}$ cluster relationships that present similar structure in their factions

Multiple networks



Multiple networks



(a) Friendship

(b) Horseplay

(c) Negative

(d) Open window

(e) Help