Hierarchical Dirichlet Processes

AMS 241, Fall 2010

Vadim von Brzeski
vvonbrze@ucsc.edu

Reference

- *Hierarchical Dirichlet Processes*, Y. Teh, M. Jordan, M. Beal, D. Blei, Technical Report 653, Statistics, UC Berkeley, 2004.
  - Also published in NIPS 2004 : *Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes*
- Some figures and equations shown here are directly taken from the above references (indicated if so)

The HDP Prior

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0)$$

group index

G_0 is discrete : G_j DPs necessarily share atoms !

Source: Teh, 2004.

Stick Breaking Construction:

1st Level :

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H$$

$$\beta'_k \sim \mathrm{Beta}(1, \gamma) \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l)$$

2nd Level :

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}$$

Going back to original definition of DP, we can derive relationship between β and π :

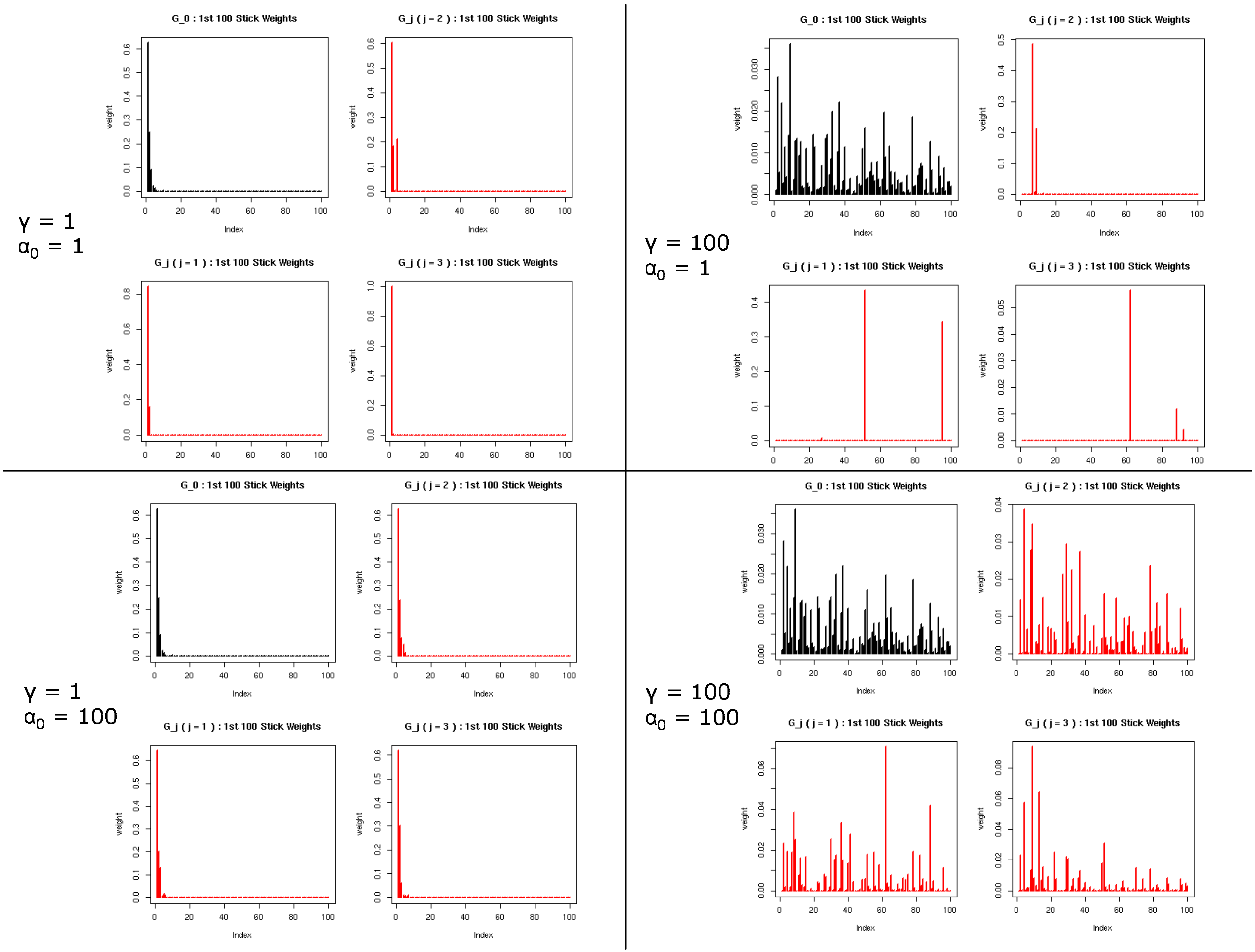
$$(G_j(A_1), ..., G_j(A_r)) \sim Dirichlet\,(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r))$$

$$\left(\sum_{k \in K_1} \pi_{jk}, ..., \sum_{k \in K_r} \pi_{jk} \right) \sim Dirichlet \left(\alpha_0 \sum_{k \in K_1} \beta_k, ..., \alpha_0 \sum_{k \in K_r} \beta_k \right) \qquad K_l = \{k : \theta_k \in A_l\}$$

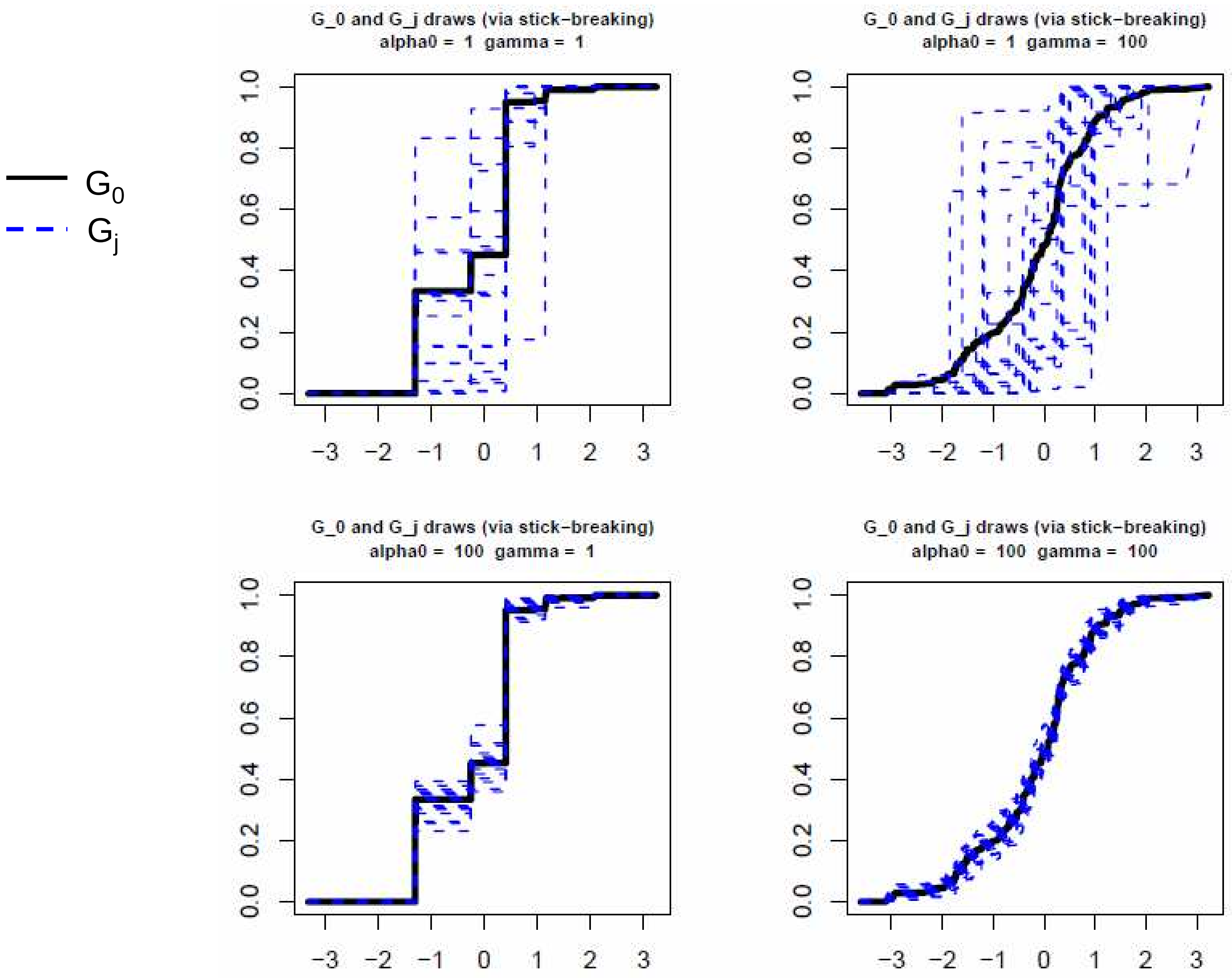
$$\Downarrow$$

$$\pi'_{jk} \sim \mathrm{Beta} \left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l \right) \right) \qquad \pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl})$$

Source: Teh, 2004.

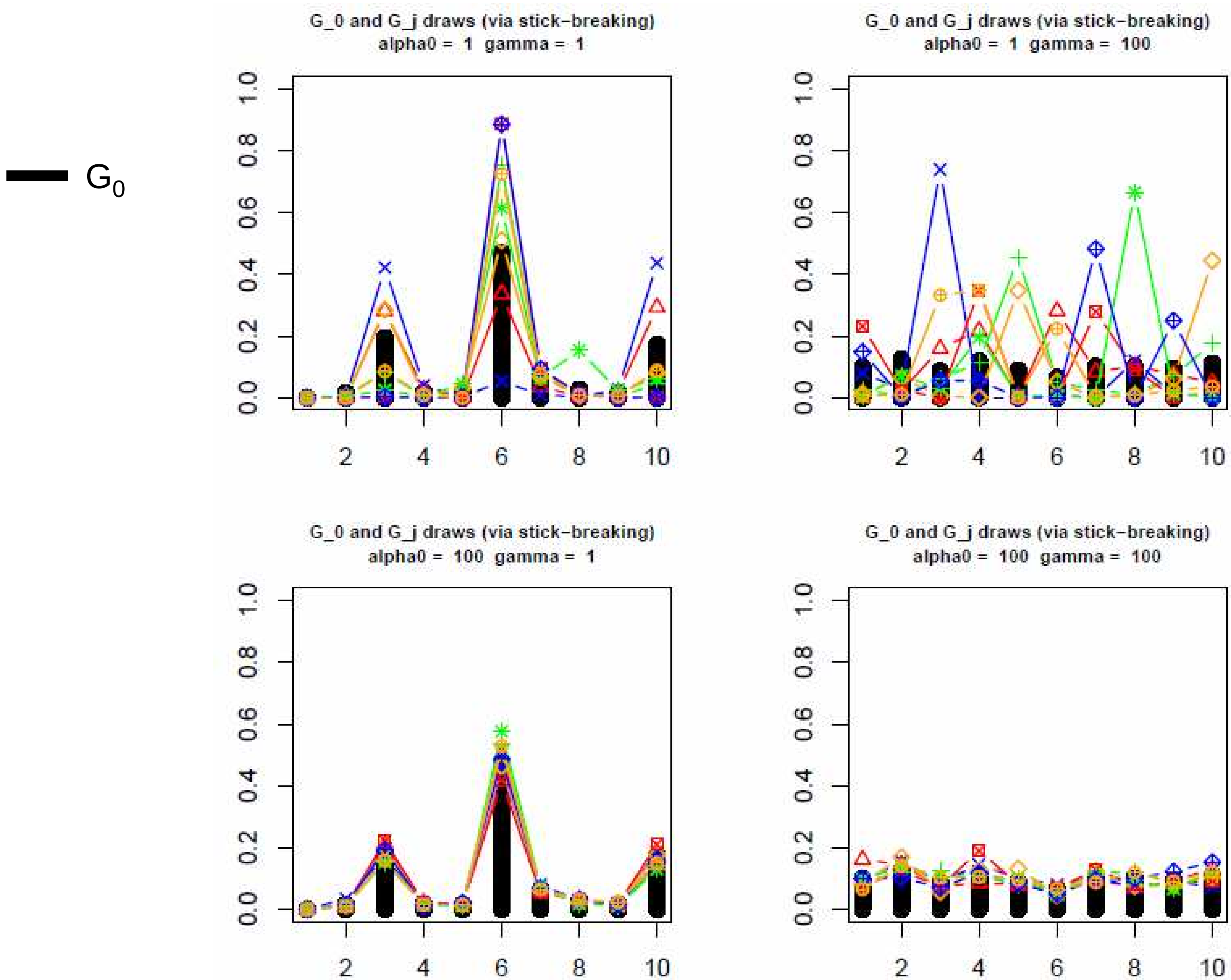

$\gamma = 1$
$\alpha_0 = 1$
G_0 : 1st 100 Stick Weights
G_j (j = 2) : 1st 100 Stick Weights
G_j (j = 1) : 1st 100 Stick Weights
G_j (j = 3) : 1st 100 Stick Weights
$\gamma = 100$
$\alpha_0 = 1$
G_0 : 1st 100 Stick Weights
G_j (j = 2) : 1st 100 Stick Weights
G_j (j = 1) : 1st 100 Stick Weights
G_j (j = 3) : 1st 100 Stick Weights
$\gamma = 1$
$\alpha_0 = 100$
G_0 : 1st 100 Stick Weights
G_j (j = 2) : 1st 100 Stick Weights
G_j (j = 1) : 1st 100 Stick Weights
G_j (j = 3) : 1st 100 Stick Weights
$\gamma = 100$
$\alpha_0 = 100$
G_0 : 1st 100 Stick Weights
G_j (j = 2) : 1st 100 Stick Weights
G_j (j = 1) : 1st 100 Stick Weights
G_j (j = 3) : 1st 100 Stick Weights
weight
Index

H : Normal(0,1)



G_0 and G_j draws (via stick-breaking)
alpha0 = 1 gamma = 1
G_0 and G_j draws (via stick-breaking)
alpha0 = 1 gamma = 100
G_0 and G_j draws (via stick-breaking)
alpha0 = 100 gamma = 1
G_0 and G_j draws (via stick-breaking)
alpha0 = 100 gamma = 100
0.0 0.2 0.4 0.6 0.8 1.0
−3 −2 −1 0 1 2 3
G_0
G_j

H : Dirichlet(0.1,0.1,…..,0.1), dim V=10



Prior and Data Model

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0)$$

$$\phi_{ji} \mid G_j \sim G_j$$

$$x_{ji} \mid \phi_j \sim F(\phi_{ji})$$

↑

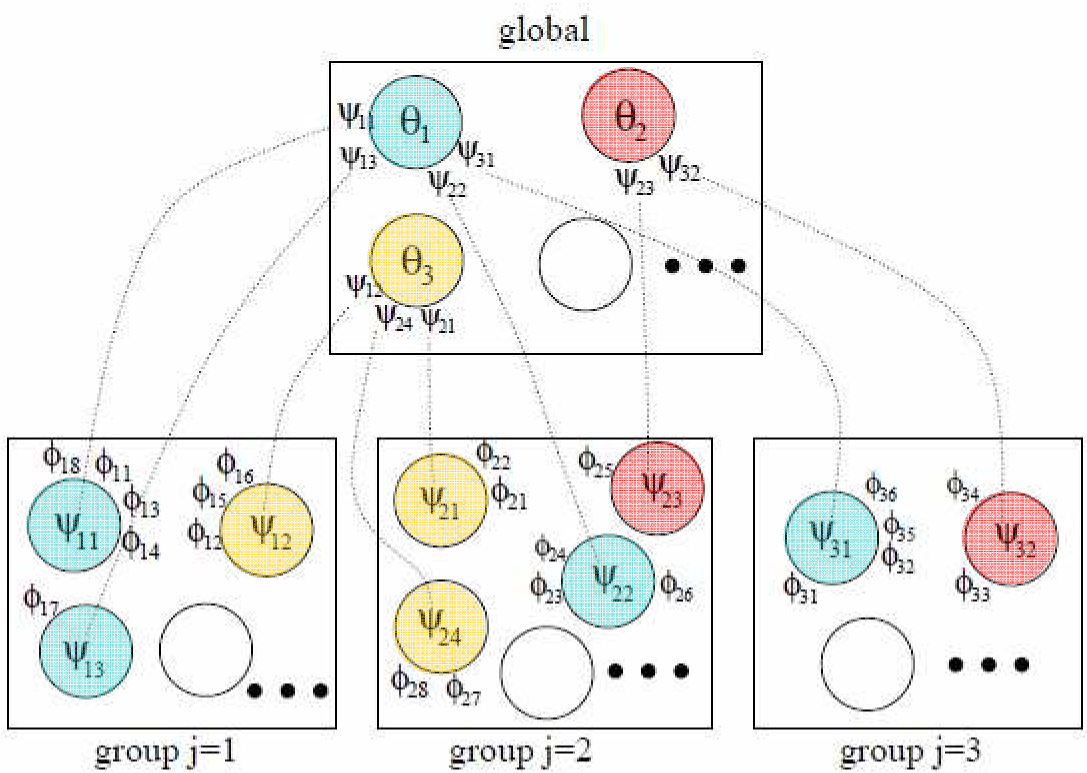
i^{th} datum in group j

Source: Teh, 2004.

Polya Urn Sampling via Chinese Restaurant (Process) Franchise :

$$\phi_{ji} \mid \phi_{j1}, \ldots, \phi_{j\,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{T_j} \overbrace{\frac{n_{jt}}{i-1+\alpha_0}}^{\text{existing table}} \delta_{\psi_{jt}} + \overbrace{\frac{\alpha_0}{i-1+\alpha_0}}^{\text{new table}} G_0$$

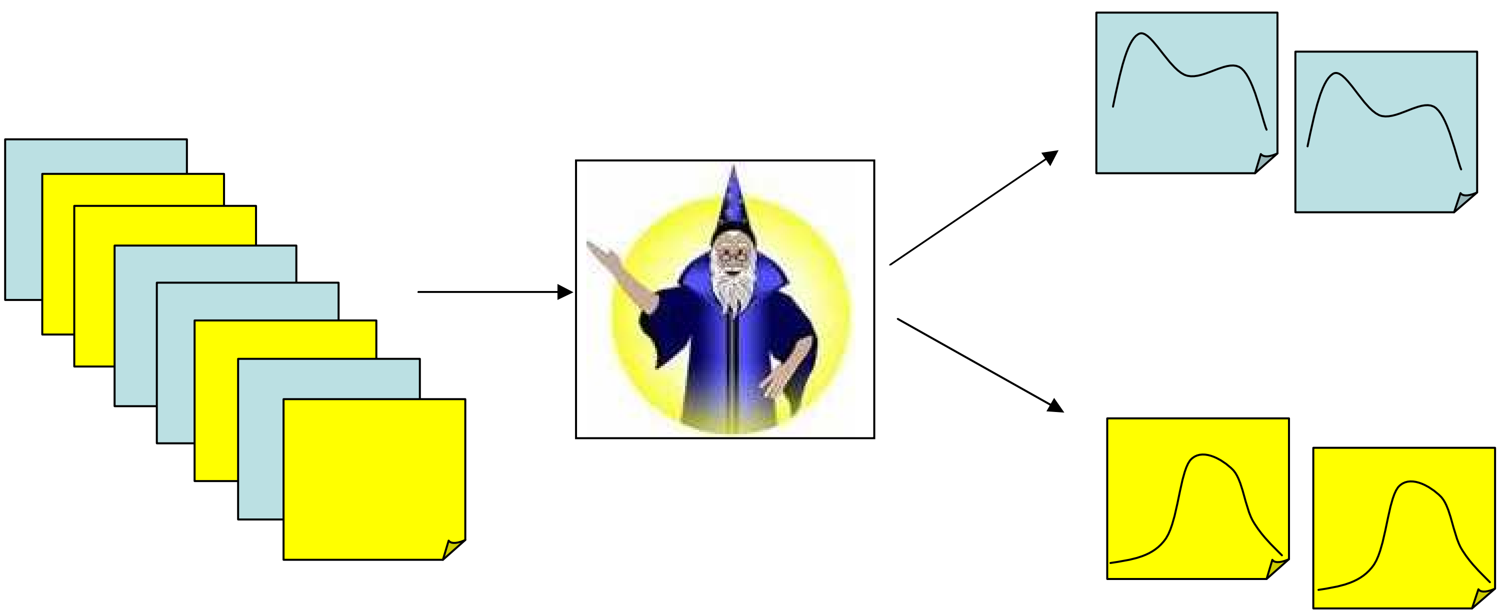
$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j\,t-1}, \gamma, H \sim \sum_{k=1}^{K} \overbrace{\frac{m_k}{\sum_k m_k + \gamma}}^{\text{existing component}} \delta_{\theta_k} + \overbrace{\frac{\gamma}{\sum_k m_k + \gamma}}^{\text{new component}} H$$



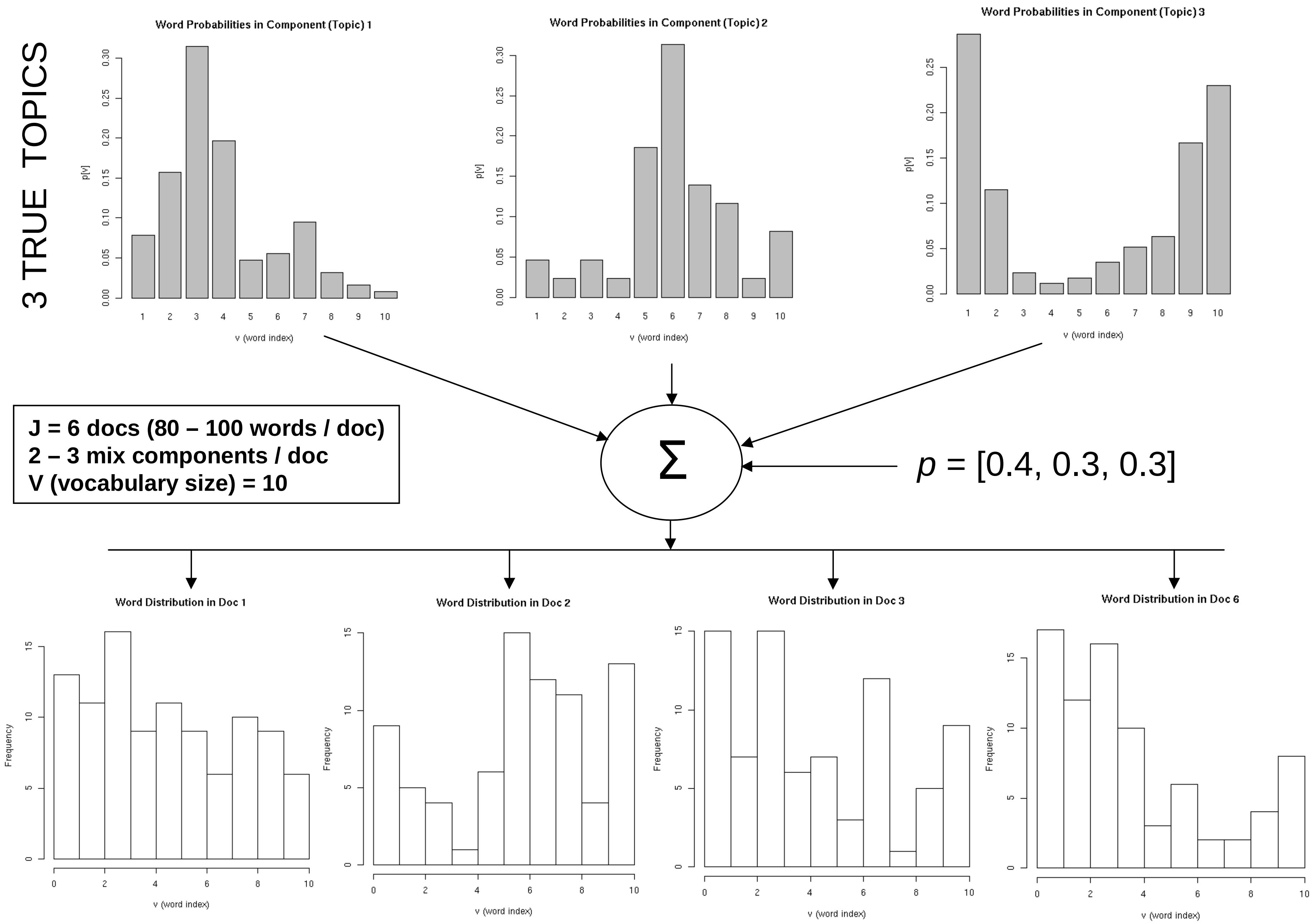
Source: Teh, 2004.

Application : Topic Modeling

- Topic = (multinomial) distribution over words
  - Fixed size vocabulary; p(word | topic)
  - F : Multinomial kernel, H : Dirichlet()
- Document = mixture of one or more topics
- Goal = recover latent topics; use topics for clustering, finding related documents, etc.



Study Model Inference Using Simulated Data



Inference via Gibbs Sampling

1. $$p(t_{ji}=t|\boldsymbol{t}^{-ji},\boldsymbol{k},\boldsymbol{\theta},\boldsymbol{x}) \propto \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt}}) & \text{if } t = t^{\text{new}}, \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & \text{if } t \text{ previously used.} \end{cases}$$

$$k_{jt^{\text{new}}} \mid \boldsymbol{k} \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma}\delta_k + \frac{\gamma}{\sum_k m_k + \gamma}\delta_{k^{\text{new}}}$$

$$\theta_{k^{\text{new}}} \sim H$$

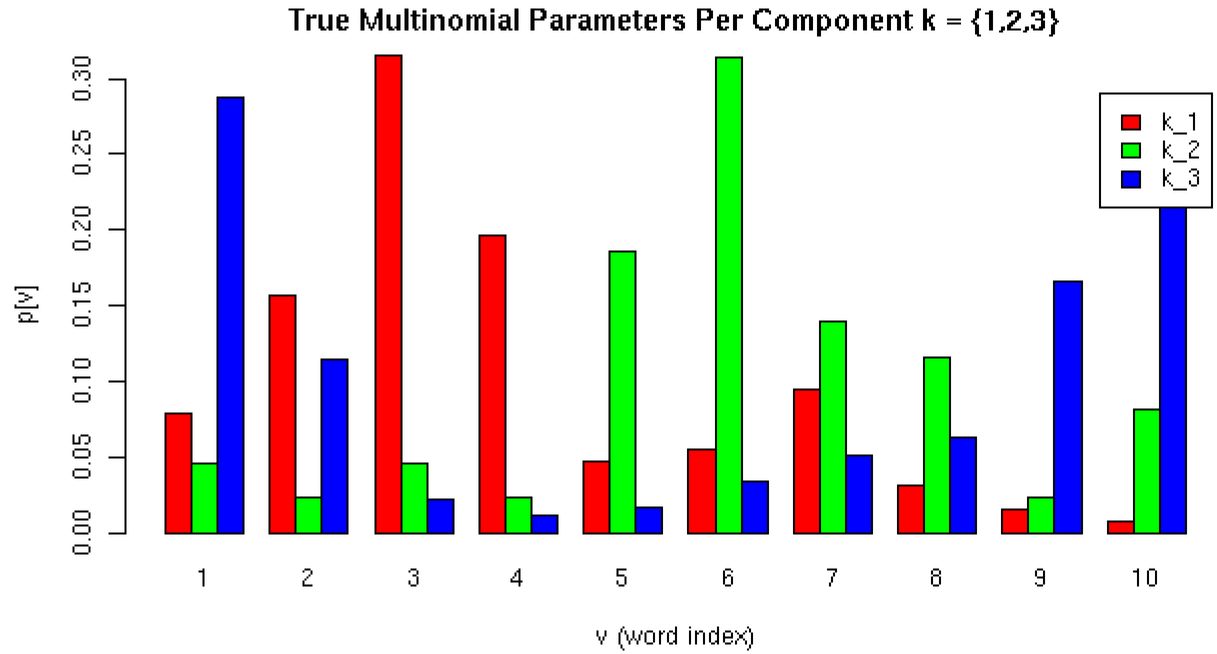
2. $$p(k_{jt}=k|\boldsymbol{t},\boldsymbol{k}^{-jt},\boldsymbol{\theta},\boldsymbol{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k = k^{\text{new}}, \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k \text{ is previously used.} \end{cases}$$

$$\theta_{k^{\text{new}}} \sim H$$

3. $$p(\theta_k|\boldsymbol{t},\boldsymbol{k},\boldsymbol{\theta}^{-k},\boldsymbol{x}) \propto h(\theta_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\theta_k)$$

Source: Teh, 2004.

TRUTH :

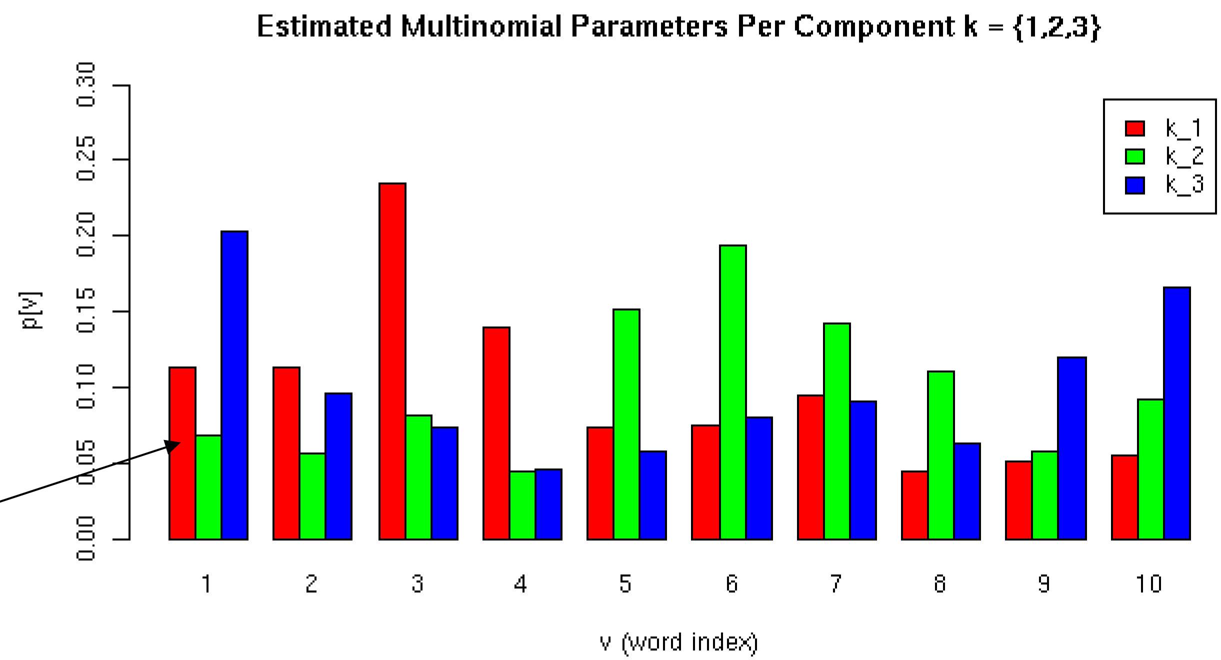


ESTIMATE :

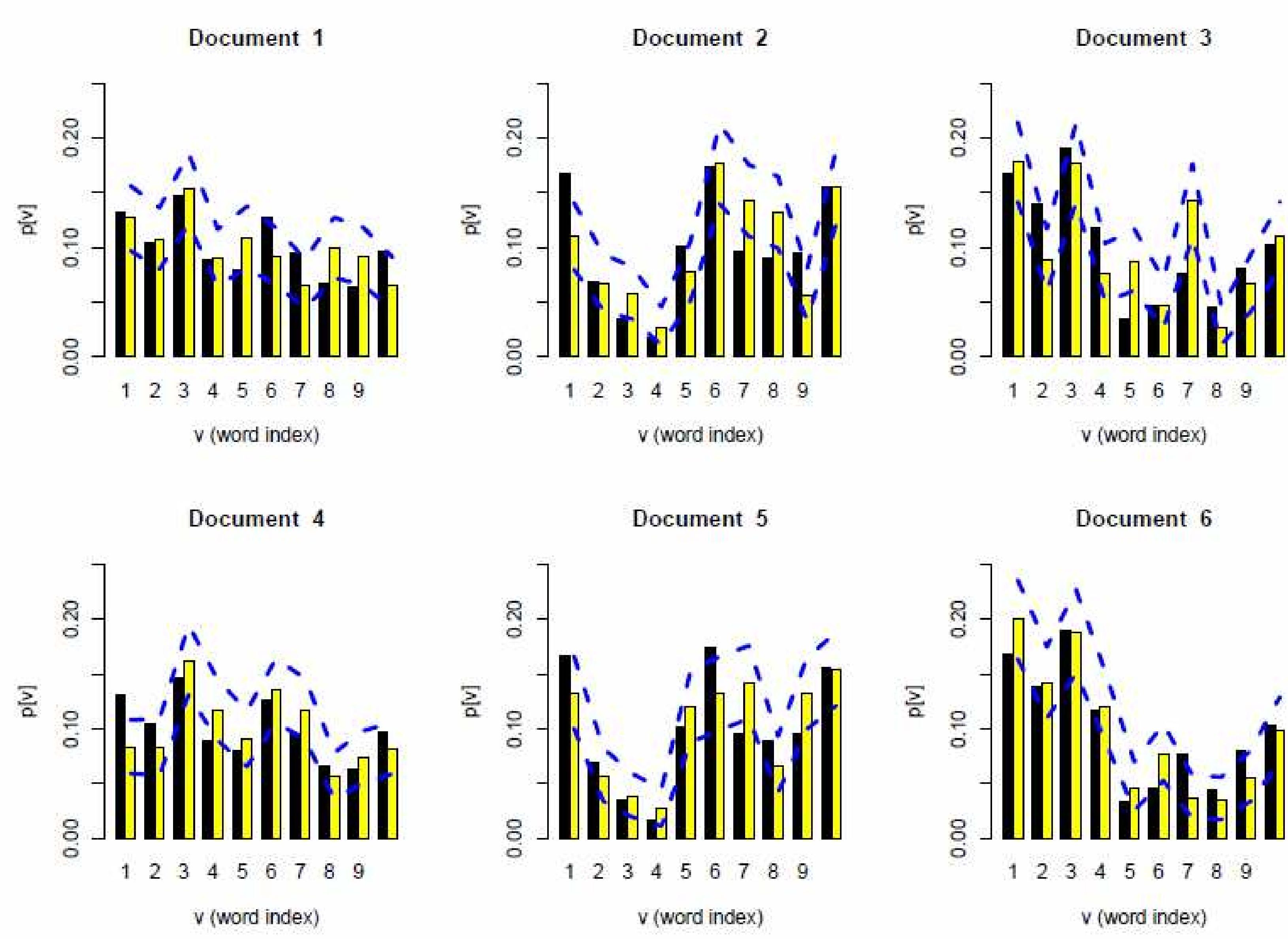
For each $\mathbf{x}_{ji}$ whose true component was k, we have B MCMC draws: $\{\theta_{ji}^{(1)}, \theta_{ji}^{(2)}, \ldots\ldots, \theta_{ji}^{(B)}\}$

$$\overline{\theta_{ji}^{(B)}} = \frac{1}{B} \sum_b \theta_{ji}^{(b)}$$

$$\overline{\theta}_k = \frac{1}{n_k} \sum \overline{\theta_{ji}^{(B)}}$$



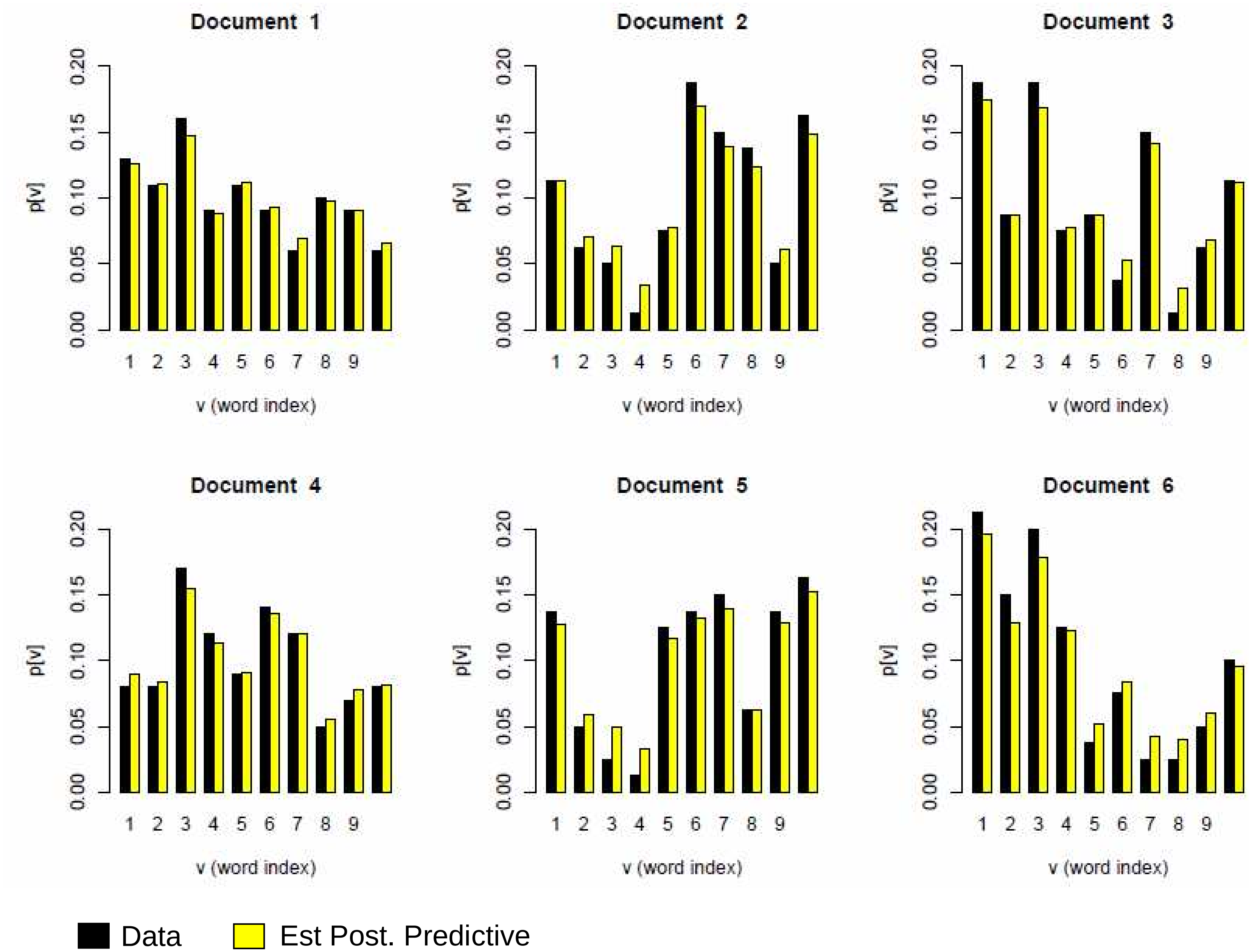
Truth vs. Posterior Point and 10/90 Interval Estimates for E[θ_j | data]



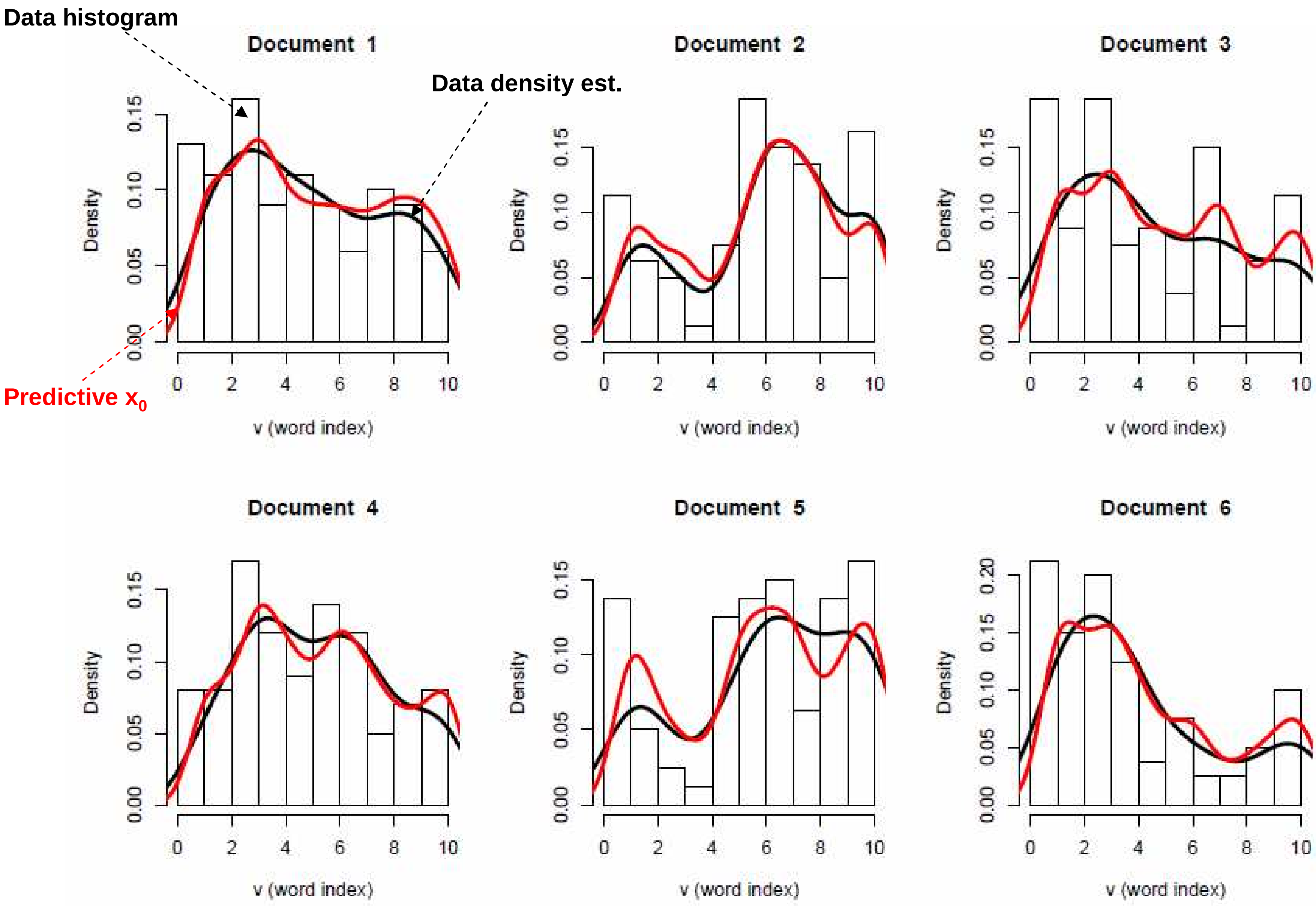
True θ_j Estimate

Simulated Data Histograms vs. Est. Posterior Predictive : E[θ_{j0} | data]

For each doc j : avg (over states b = 1..B) draws of $\theta_{j0}^{(b)}$ via CRP config @ state b.



Simulated Data Distributions vs. Est. Posterior Predictive for New Observation x_{j0}



R Code Available

- Works, but SLOOOOOOOOOOW….

http://www.numberjack.net/download/classes/ams241/project/R