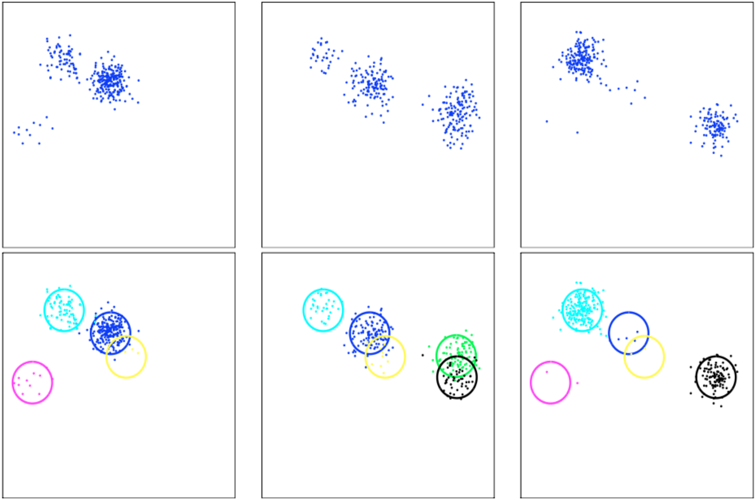
Bayesian Statistics

Debdeep Pati
Florida State University

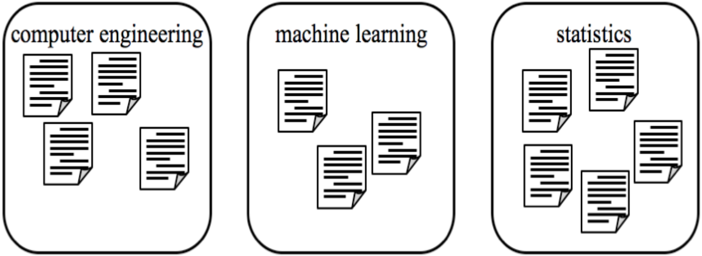
October 19, 2016

Application 3: Inference on Grouped data



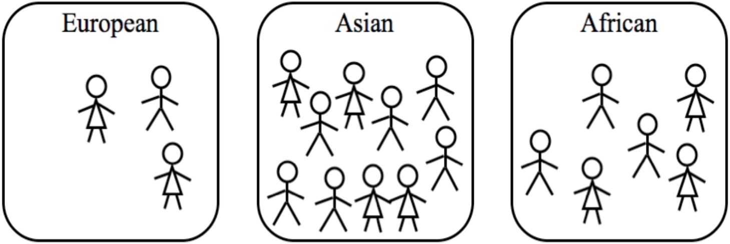
Example 1: Clustering of groups of documents sharing common topics

- Share topics across documents in a collection, and across different collections.
- More sharing within collections than across.
- Use DP mixture models as we do not know the number of topics a priori.



Example 2: Modeling populations sharing haplotypes

- Individuals inherit both ancient haplotypes dispersed across multiple populations, as well as more recent population-specific haplotypes.
- Sharing of haplotypes among individuals in a population, and across different populations.
- More sharing within populations than across.
- Use DP mixture models as we do not know the number of haplotypes.

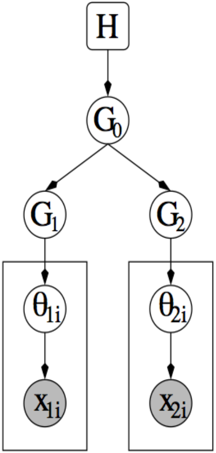


Hierarchical DP

- A hierarchical Dirichlet process:

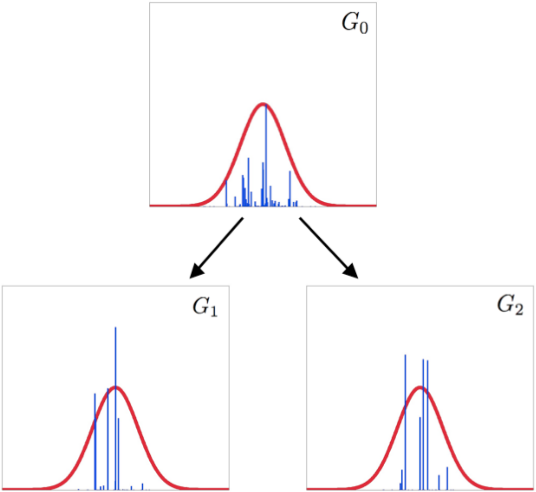
$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$$

- Extension to deeper hierarchies is straightforward.



Hierarchical DP

- Making G_0 discrete forces shared cluster between G_1 and G_2



Stick-breaking for HDP

- We shall assume the following HDP hierarchy

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

- The stick-breaking construction for the HDP is

$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \pi_{0k}\delta_{\phi_k}, \quad \phi_k \sim H \\
\pi_{0k} &= \beta_{0k}\prod_{l=1}^{k-1}(1-\beta_{0l}), \quad \beta_{0k} \sim \mathrm{Beta}(1,\gamma) \\
G_j &\sim \sum_{k=1}^{\infty} \pi_{jk}\delta_{\phi_k}, \quad \phi_k \sim H \\
\pi_{jk} &= \pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl}), \quad \pi'_{jk} \sim \mathrm{Beta}(\alpha\pi_{0k}, \alpha(1-\sum_{l=1}^{k}\pi_{0l}))
\end{aligned}
$$

Chinese Restaurant Franchise