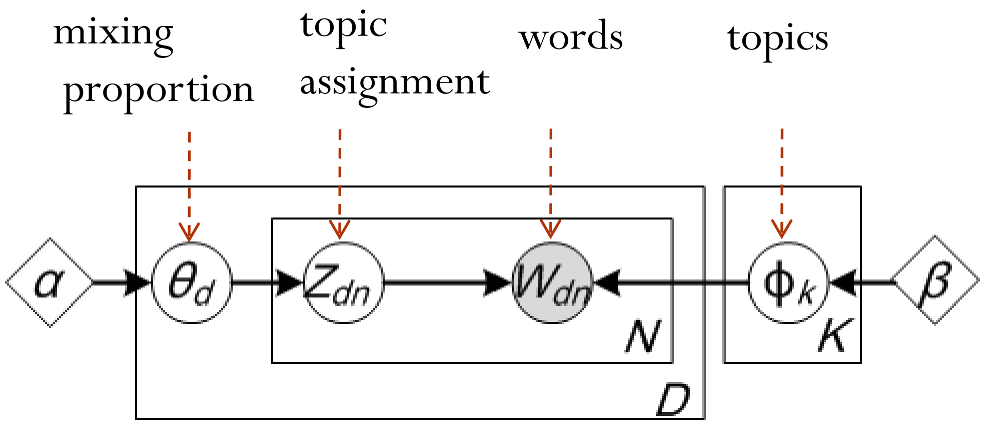
Nonparametric Bayesian Methods (Dirichlet Process Mixtures)

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

May 12, 2015

Recap. of LDA



$$p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta) = \prod_{k=1}^{K} p(\Phi_k | \beta) \prod_{d=1}^{D} p(\theta_d | \alpha) \Big(\prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \Phi) \Big)$$

- Given a set of documents, infer the posterior distribution

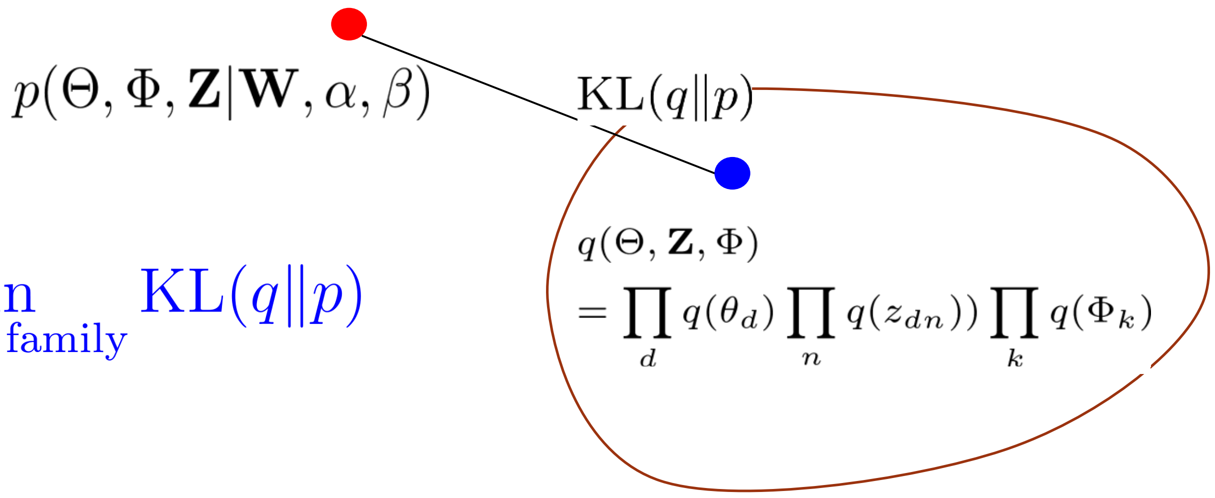
$$p(\Theta, \Phi, \mathbf{Z} | \mathbf{W}, \alpha, \beta) = \frac{p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta)}{p(\mathbf{W} | \alpha, \beta)}$$

OR

$$p(\mathbf{Z} | \mathbf{W}, \alpha, \beta) = \frac{\int_{\Theta, \Phi} p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta)}{p(\mathbf{W} | \alpha, \beta)}$$

Dealing with the Intractability of Inference

- Variational Inference (Blei et al., 2003; Teh et al., 2006)



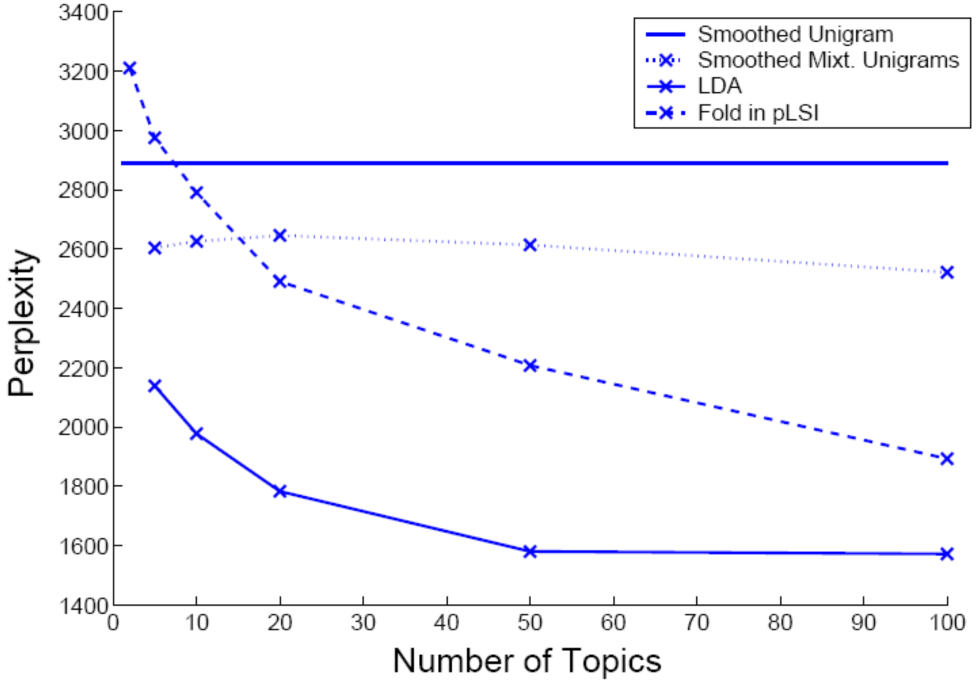
$$q^* = \min_{q \in \text{some family}} \mathrm{KL}(q \| p)$$

- Monte Carlo Markov Chains (Griffiths & Steyvers, 2004)
  - Collapsed Gibbs samplers iteratively draw samples from the local conditionals

$$p(z_{dn}^k = 1 | Z_{\neg})$$

Problem with K

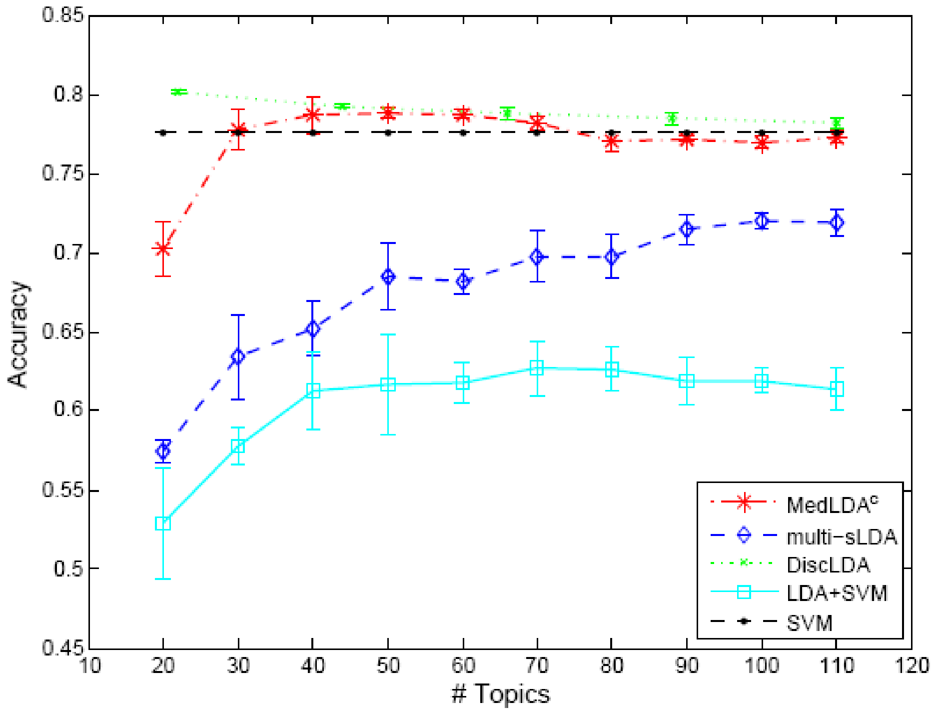
- K represents the model complexity
- It matters a lot in practice



[Blei et al., JMLR 2003]

Problem with K

- K represents the model complexity
- It matters a lot in practice



[Zhu et al., JMLR 2012]

- Today, we will discuss nonparametric Bayesian methods

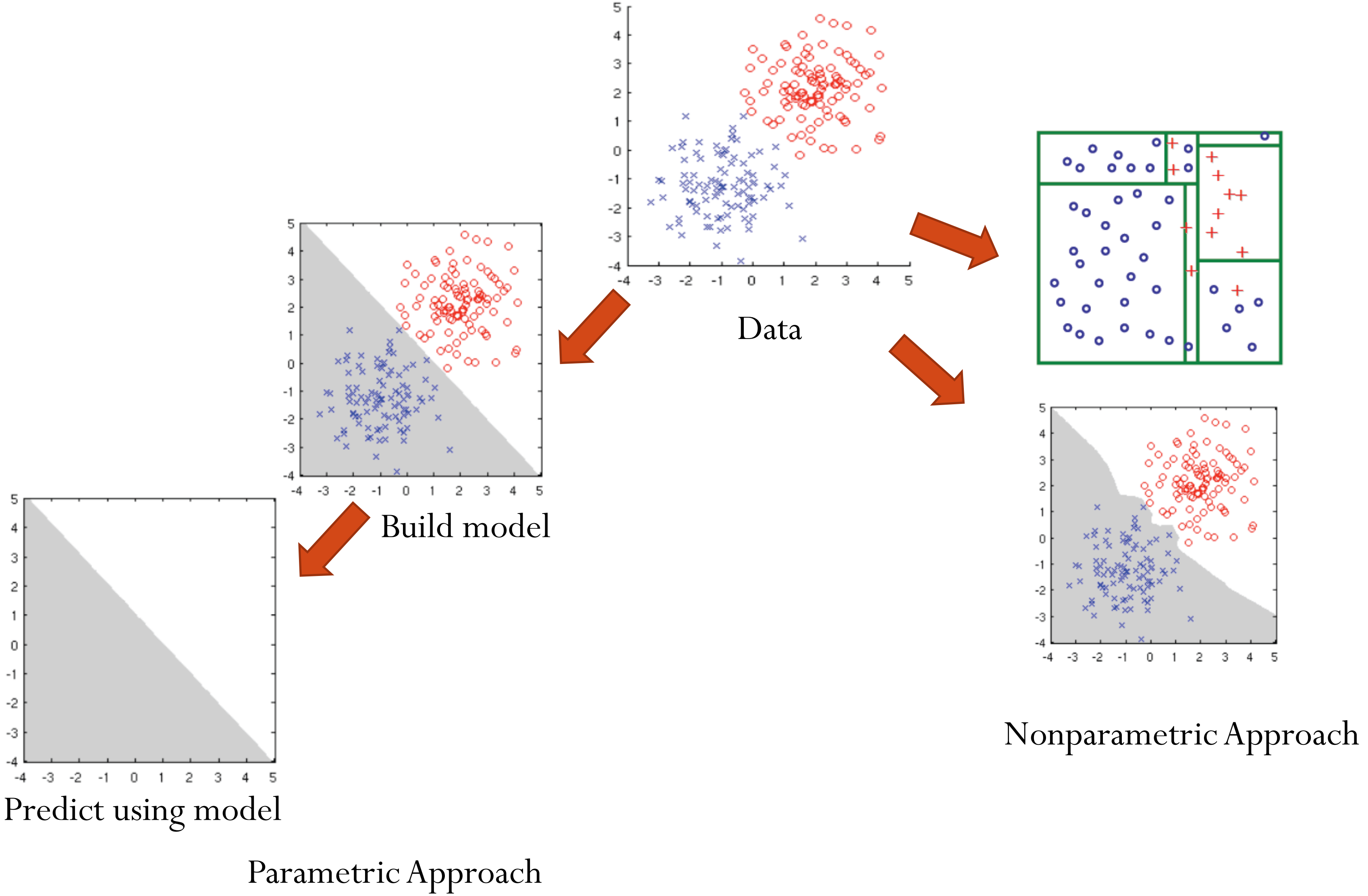
- “Nonparametric Bayesian methods”?
- What does that mean?

- So now we know what Bayesian means, but what does nonparametric mean?

Nonparametric

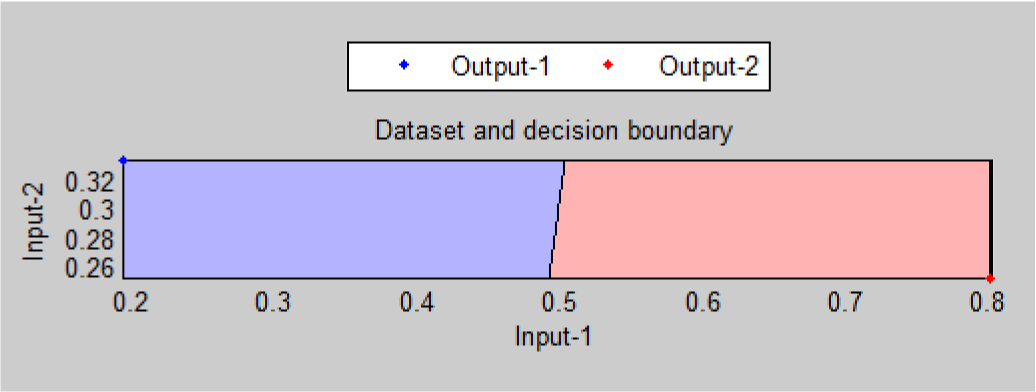
- Nonparametric:
  - Does NOT mean there are no parameters

Example: Classification



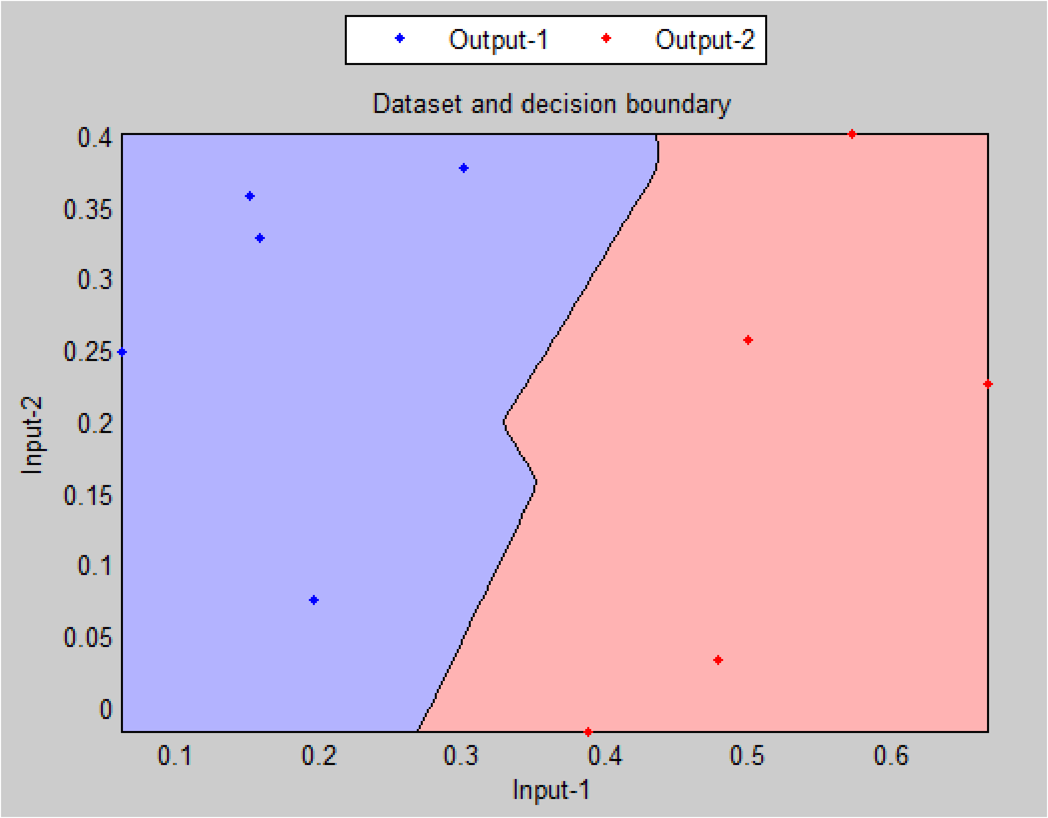
Complexity of 1-NN

- 2 samples



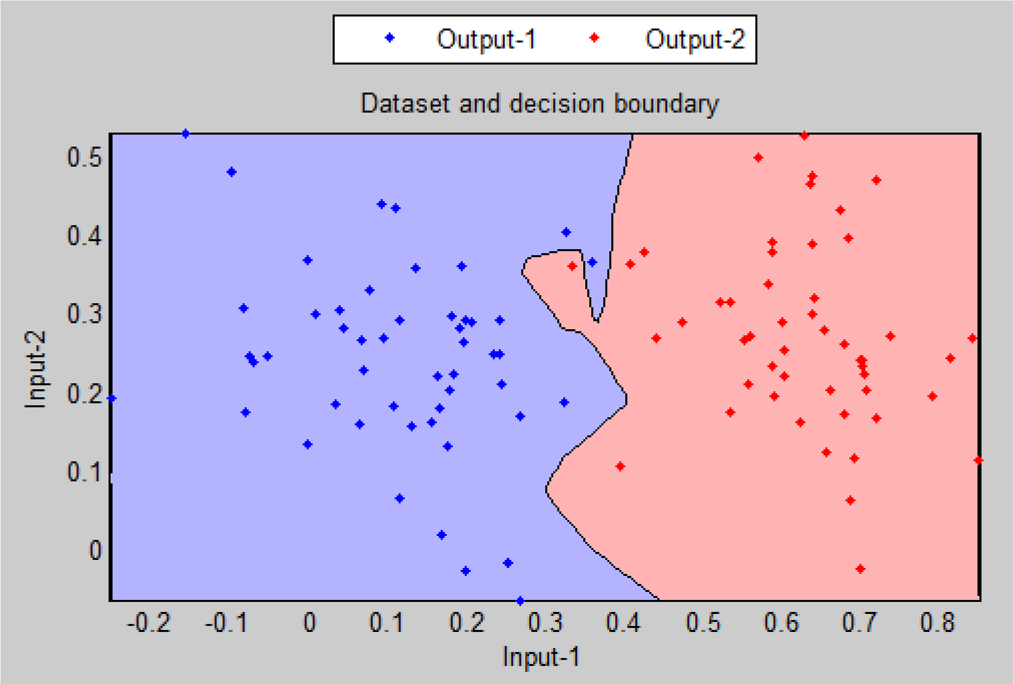
Complexity of 1-NN

- 10 samples



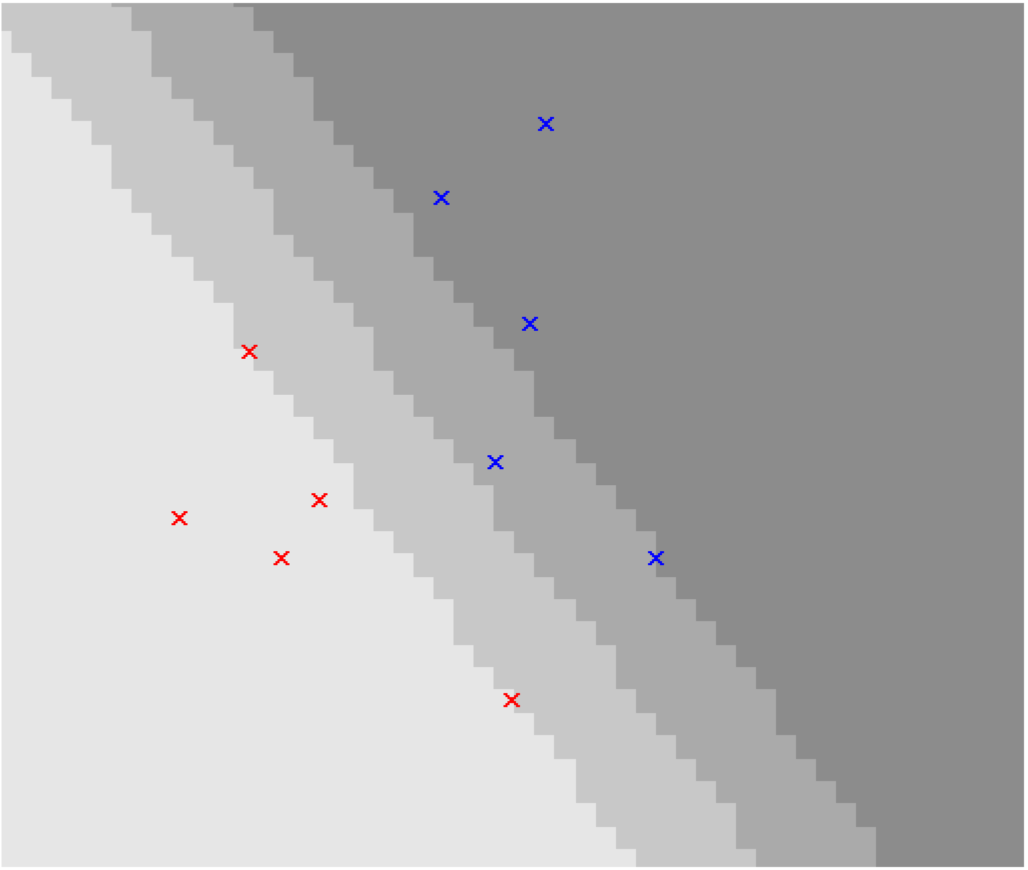
Complexity of 1-NN

- 100 samples



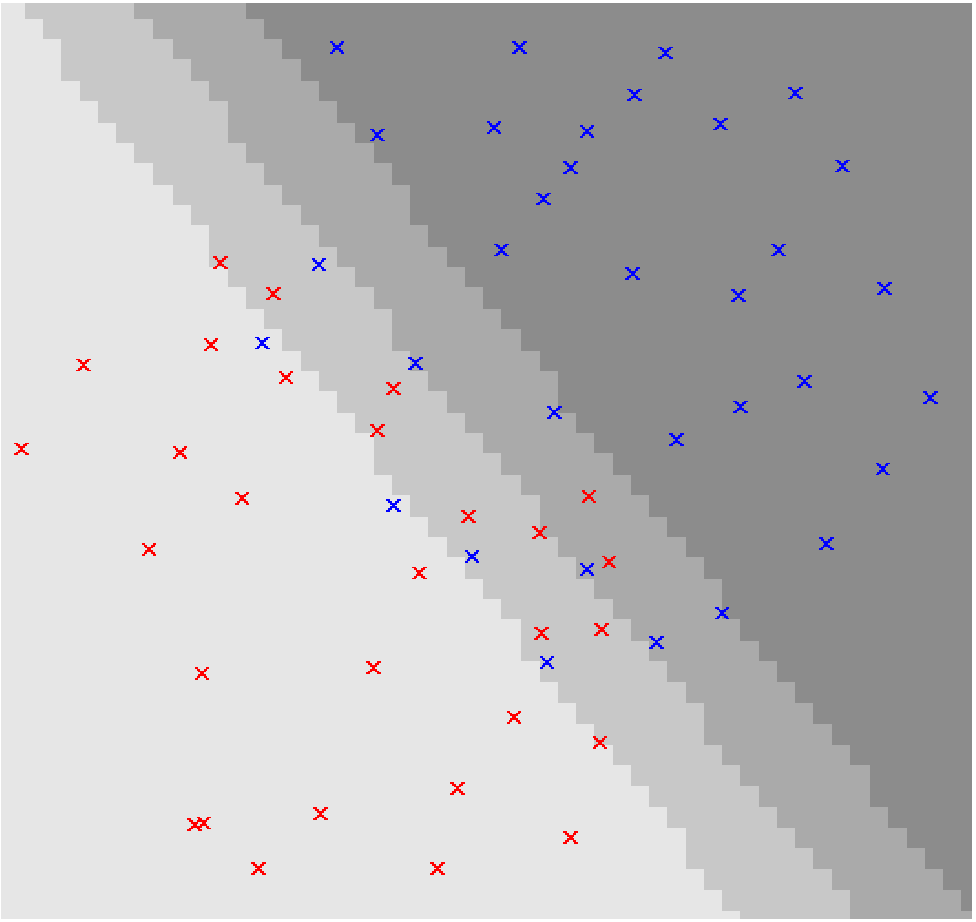
How about linear SVM?

- 10 samples

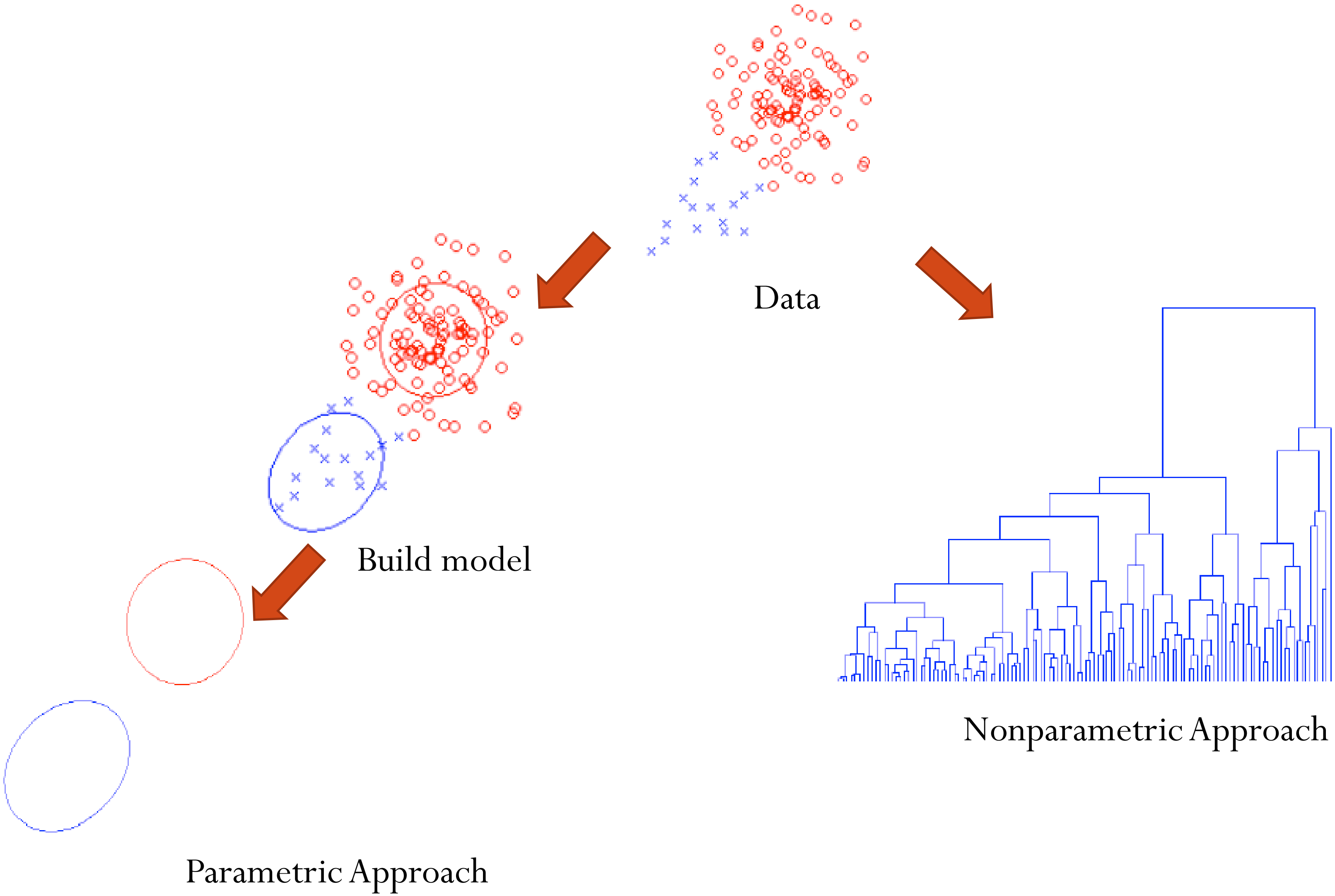


How about linear SVM?

- A lot of samples (inseparable)

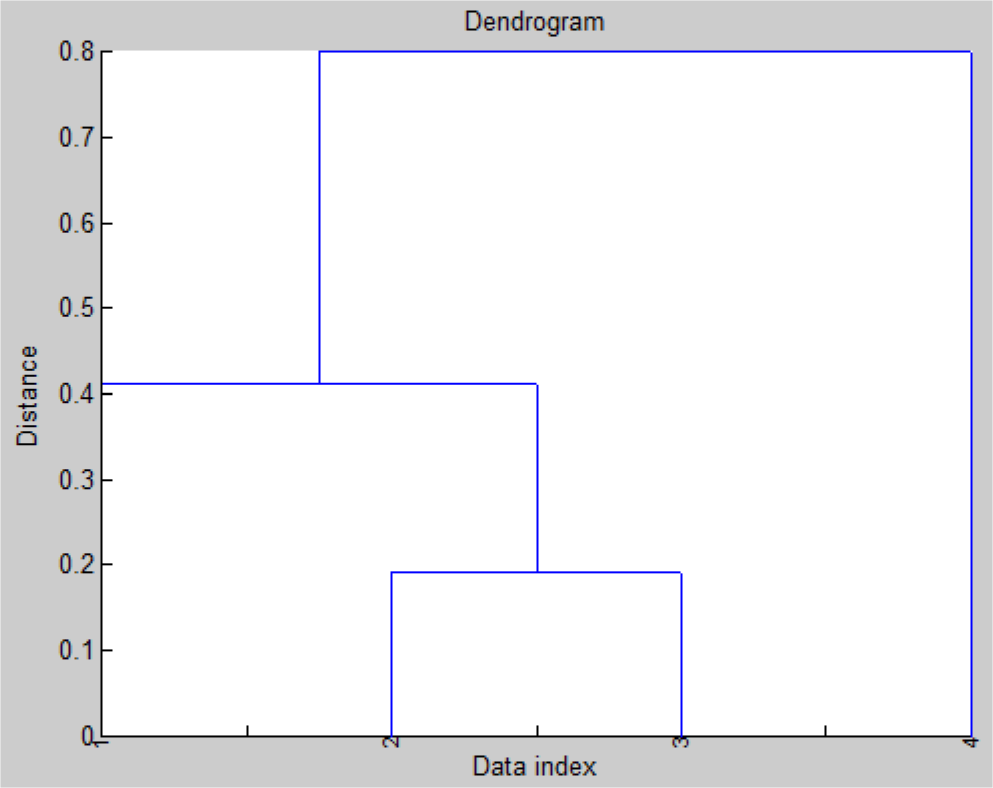


Example: Clustering



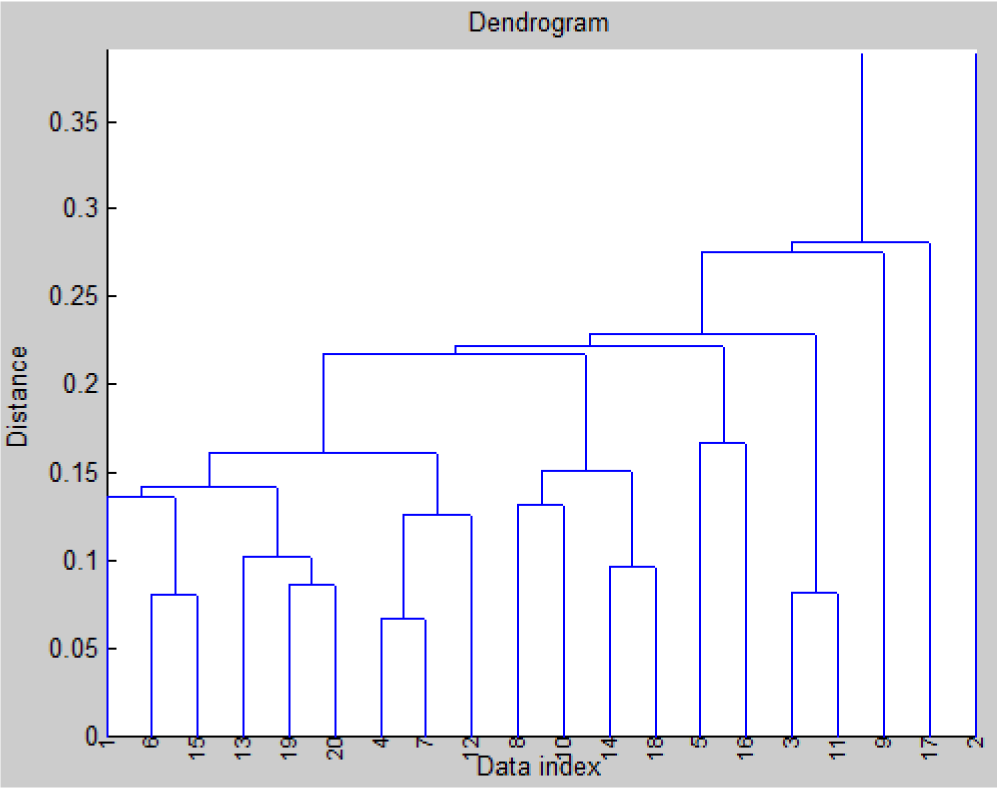
Complexity of Hierarchical Clustering

- 4 samples

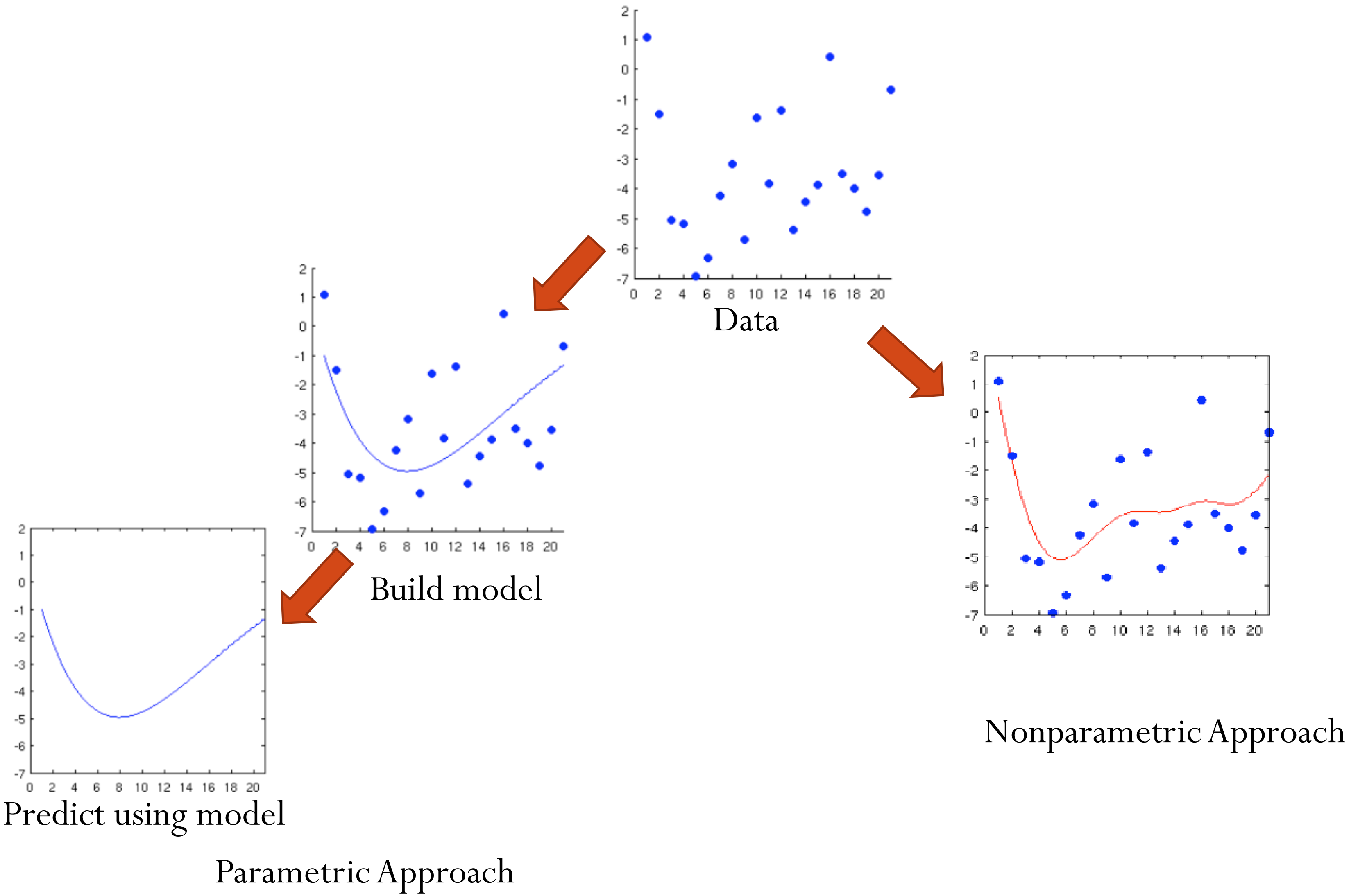


Complexity of Hierarchical Clustering

- 20 samples

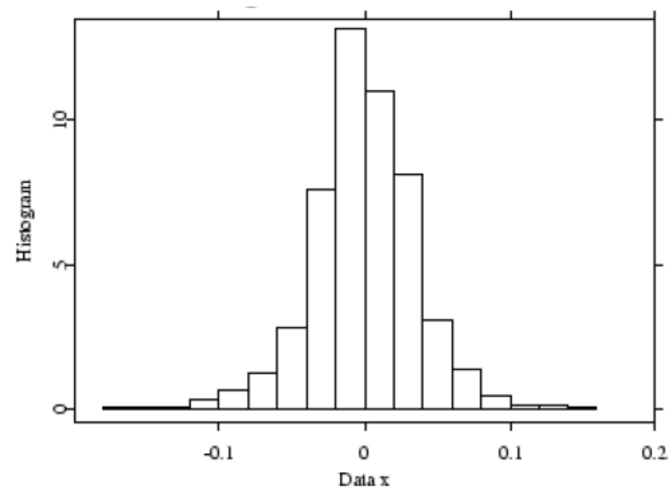


Example: Regression



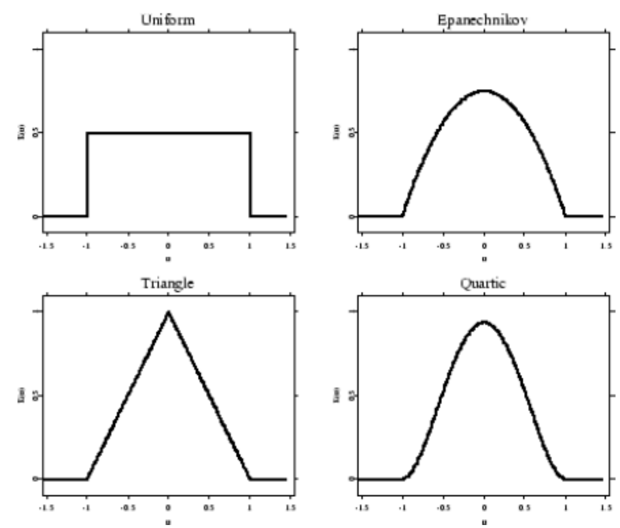
Other Examples: Density Estimation

- Histogram
  - Issue with binwidth
  - Issue with origins of bins
  - Issue with discreteness



- Smoothing techniques to improve
  - Averaged shifted histogram
  - Kernel density estimation

$$\hat{f}(x) = \sum_{i=1}^{N} K_h(x - x_i)$$

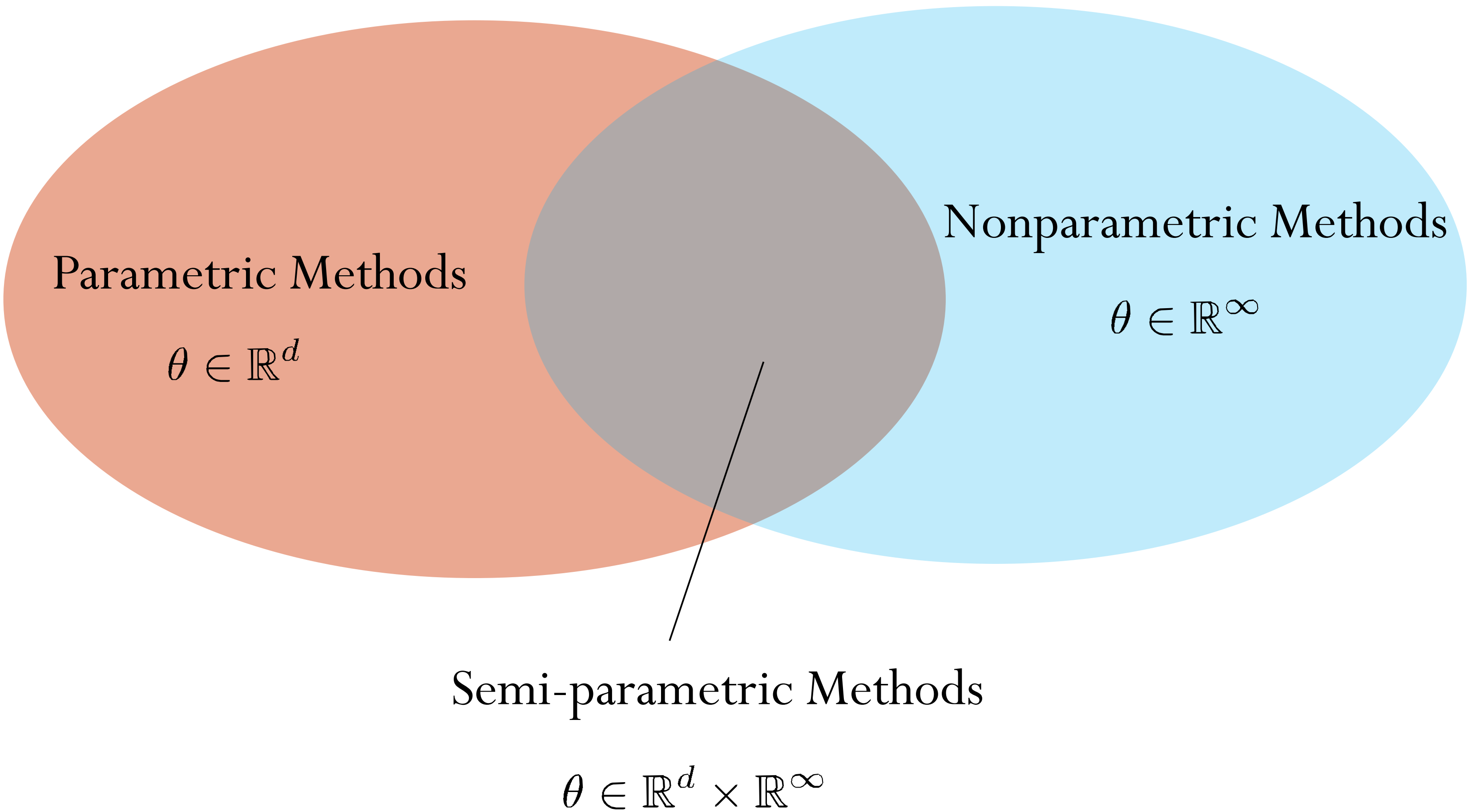


[Chap 3. Nonparametric and Semi-parametric Models, W. Hardel et al., 2004]

Various Paradigms

- Parametric Models
  - the parameters are belonging to a fixed finite dimensional space, e.g., a subset of $\mathbb{R}^d$
- Nonparametric Models
  - the parameters belong to some space, not necessarily finite dimensional
  - Principe of "let the data speak for themselves"
- Semi-parametric Models
  - the parameters have both finite dimensional component and infinite dimensional component
  - E.g., (sparse) additive models for regression

Various Paradigms



Pros & Cons

- Parametric Models
  - If underlying assumptions are correct, the models are simple and easy to interpret
  - If not, estimates may be inconsistent and give misleading results

- Nonparametric Models:
  - Avoid restrictive assumptions
  - Usually hard to interpret and yield inaccurate estimates

- Semi-parametric Models:
  - Keep the easy interpretability the former and retain some of the flexibility of the latter.

Nonparametric Bayesian Methods

- Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?
  - Complexity of our model should be allowed to grow as we get more data
  - Place a prior on an unbounded number of parameters

Nonparametric Bayesian Methods overview

- Dirichlet Process/Chinese Restaurant Process
  - Latent class models – often used in the clustering context
- Beta Process/Indian Buffet Process
  - Latent feature models
- Gaussian Process (optional)
  - Regression and Classification

Dirichlet Process

- A nonparametric approach to clustering.
- It can be used in any probabilistic model for clustering.

Outline

- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

A Bayesian Approach to Clustering

- We must specify two things:

  - the likelihood model (how data is affected by the parameters)

$$p(\mathcal{D}|\theta)$$

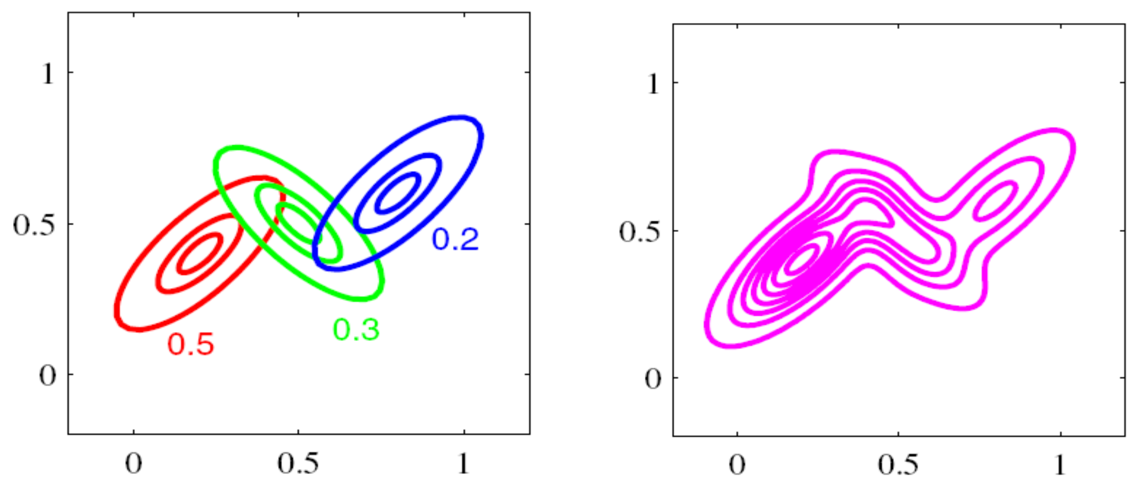
  - The prior distribution (the prior belief on the parameters)

$$p(\theta)$$

Clustering – A Parametric Approach

- Guassian Mixture Models with K components
  - a distribution over classes/clusters: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$
  - each cluster has a mean and covariance $\phi_k = (\mu_k, \Sigma_k)$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$



  - Using EM to maximize the likelihood of the data to estimate $(\boldsymbol{\pi}, \phi)$

[Figure credit: Bishop, 2006]

Clustering – A Parametric Approach

- Guassian Mixture Models with K components
- An alternative definition

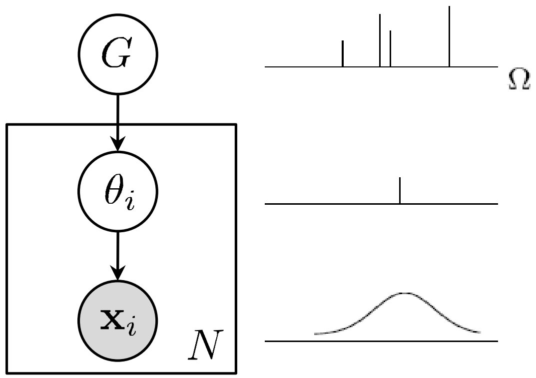
$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

where is δ_{ϕ_k} an *atom* at ϕ_k

- Then,

$$\theta_i \sim G$$

$$\mathbf{x}_i \sim p(\mathbf{x}|\theta_i)$$



Clustering – A Parametric Approach

- Bayesian Approach: Bayesian Gaussian Mixture Models with K mixtures
  - a distribution over classes/clusters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$

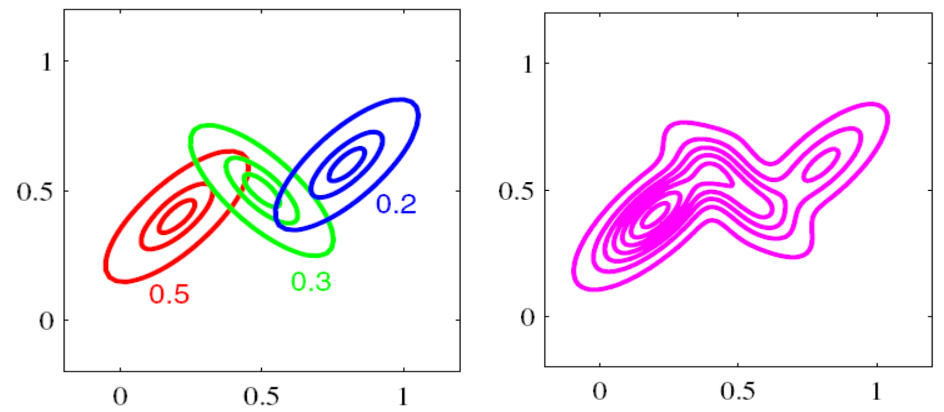
$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

  - each cluster has a mean and covariance $\phi_k = (\mu_k, \Sigma_k)$

$$(\mu_k, \Sigma_k) \sim \text{Normal-Inverse-Wishart}(\nu)$$

- We still have

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$



Clustering – A Parametric Approach

- Bayesian Approach: Bayesian Gaussian Mixture Models with K mixtures
- The Alternative Definition
  - G is now a random measure

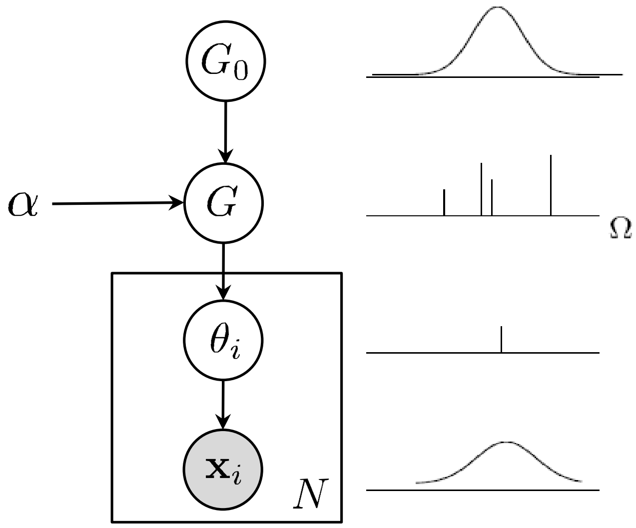
$$\phi_k \sim G_0$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

$$\theta_i \sim G$$

$$\mathbf{x}_i \sim p(\mathbf{x}|\theta_i)$$



The Dirichlet Distribution

- We have $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- A Dirichlet distribution has the form

$$p(\pi|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \cdots \pi_K^{\alpha_K - 1}$$

where $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$

- The expectation is

$$\mathbb{E}[\pi_i] = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}$$

- *Beta distribution is a special case with $K = 2$.*

Key Property of Dirichlet Distribution

- Aggregation Property
  - If

$$(\pi_1, \ldots, \pi_i, \pi_{i+1}, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_K)$$

  - Then

$$(\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K)$$

  - This is valid for any aggregation

$$(\pi_1 + \pi_2, \sum_{i=3}^{K} \pi_i) \sim \text{Beta}(\alpha_1 + \alpha_2, \sum_{i=3}^{K} \alpha_i)$$

Multinomial-Dirichlet Conjugacy

- Let

$$X \sim \text{Multinomial}(\pi), \text{ and } \pi \sim \text{Dirichlet}(\alpha)$$

- The posterior

$$\begin{aligned} p(\pi|X) &\propto p(X|\pi)p(\pi) \\ &\propto (\pi_1^{x_1} \cdots \pi_K^{x_K})(\pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1}) \end{aligned}$$

which is $\text{Dirichlet}(\alpha + \mathbf{x})$

Clustering – A Parametric Approach

- Bayesian Approach: Bayesian Gaussian Mixture Models with K mixtures
- The Alternative Definition
  - G is now a random measure

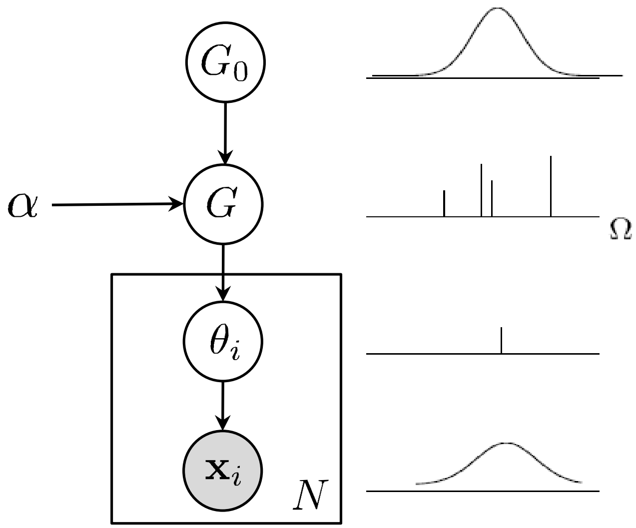
$$\phi_k \sim G_0$$

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

$$\theta_i \sim G$$

$$\mathbf{x}_i \sim p(\mathbf{x}|\theta_i)$$



Bayesian Mixture Models

- We no longer want just the maximum likelihood parameters, we want the full posterior:

$$p(\pi, \phi | \mathcal{D}) \propto p(\mathcal{D} | \pi, \phi) p(\pi, \phi)$$

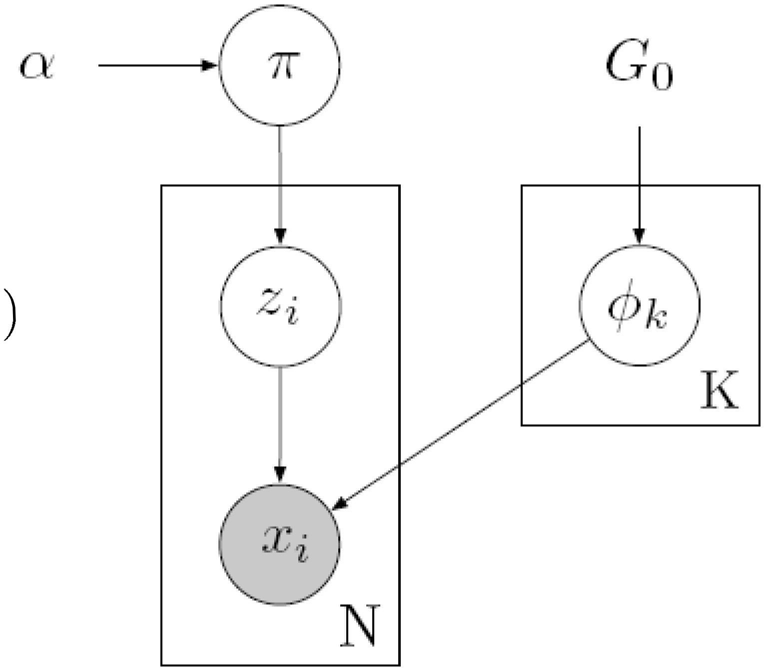
  - Unfortunately, this is not analytically tractable

- Two main approaches to approximate inference
  - Markov Chain Monte Carlo (MCMC) methods
  - Variational approximations

Bayesian Mixture Models – MCMC inference

- Introduce "membership" indicators z_i ,where $z_i \sim \text{Multinomial}(\pi)$ indicates which cluster data point *i* belongs to
- The model is equivalently represented as

$$p(\pi, Z, \phi|\mathcal{D}) \propto p(\mathcal{D}|Z, \phi)p(Z|\pi)p(\pi, \phi)$$



Gibbs Sampling for the Bayesian Mixture Models

- Randomly initialize Z, π, ϕ. Repeat until we have enough samples
  - Sample z_i from

$$p(z_i | Z_{-i}, \pi, \phi, \mathcal{D}) \propto \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i | \phi_k) \delta_{z_i, k}$$

  - Sample π from

$$p(\pi | Z, \phi, \mathcal{D}) = \text{Dirichlet}(n_1 + \alpha/K, \ldots, n_K + \alpha/K)$$

    where n_i is the number of points assigned to cluster i.
  - Sample each ϕ_k from the NIW posterior based on $(Z, \mathcal{D})$

Derivations

- For z_i , it's easy to derive

$$p(z_i|Z_{-i}, \pi, \phi, \mathcal{D}) \propto \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i|\phi_k)\delta_{z_i,k}$$

- For π , it's also easy due to conjugacy

$$p(\pi|Z, \phi, \mathcal{D}) = \text{Dirichlet}(n_1 + \alpha/K, \ldots, n_K + \alpha/K)$$

- For ϕ , it's also easy due to conjugacy
  - The Normal-Inverse-Wishart (NIW) distribution

$$\begin{aligned} \Sigma_k|\kappa, W &\sim \mathcal{IW}(\Sigma; \kappa, W^{-1}), \\ \mu_k|\Sigma_k, \mu_0, \rho &\sim \mathcal{N}(\mu; \mu_0, \Sigma_k/\rho) \end{aligned}$$

$$\mathcal{IW}(\Sigma; \kappa, W^{-1}) = \frac{|W|^{\kappa/2}}{2^{\frac{\kappa M}{2}} \Gamma_M(\frac{\kappa}{2}) |\Sigma|^{\frac{\kappa+M+1}{2}}} \exp(-\frac{1}{2}\text{Tr}(W\Sigma^{-1}))$$

Conjugacy of NIW and Gaussians

- Details

$$
\begin{aligned}
p(\mu_k, \Sigma_k | \mathbf{Z}, \pi, \mathcal{D}) &\propto p_0(\mu_k, \Sigma_k) \prod_i p(\mathbf{x}_i | z_i, \phi)^{\delta_{z_i,k}} \\
&= \mathcal{NIW}(\mu_0, \rho, \kappa, W) \prod_i p(\mathbf{x}_i | z_i, \phi)^{\delta_{z_i,k}} \\
&= \mathcal{NIW}(\mu_0^k, \rho_k, \kappa_k, W_k),
\end{aligned}
$$

$$
\begin{aligned}
\mu_0^k &= \frac{\rho}{\rho + n_k}\mu_0 + \frac{n_k}{\rho + n_k}\bar{\mathbf{x}}_k \\
\rho_k &= \rho + n_k \\
\kappa_k &= \kappa + n_k \\
W_k &= W + Q_k + \frac{\rho n_k}{\rho + n_k}(\bar{\mathbf{x}}_k - \mu_0)(\bar{\mathbf{x}}_k - \mu_0)^\top
\end{aligned}
$$

$$
n_k = \sum_i \delta_{z_i,k}
$$

$$
\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_i \delta_{z_i,k} \mathbf{x}_i
$$

$$
Q_k = \sum_i \delta_{z_i,k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top
$$

$$n_k = \sum_i \delta_{z_i,k} \qquad \bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_i \delta_{z_i,k}\mathbf{x}_i$$

More details …

$$Q_k = \sum_i \delta_{z_i,k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top$$

$$p(\mu_k, \Sigma_k|\mathbf{Z}, \pi, \mathcal{D}) \propto |\Sigma_k|^{-\frac{1}{2}} \exp(-\frac{1}{2}\rho(\mu_k - \mu_0)^\top \Sigma_k^{-1}(\mu_k - \mu_0))|\Sigma_k|^{-\frac{\kappa+M+1}{2}} \exp(-\frac{1}{2}\mathrm{Tr}(W\Sigma_k^{-1})$$

$$|\Sigma_k|^{n_k} \exp(-\frac{1}{2}\sum_i \delta_{z_i,k}(\mathbf{x}_i - \mu_k)\Sigma_k^{-1}(\mathbf{x}_i - \mu_k))$$

$$-\frac{1}{2}\rho(\mu_k - \mu_0)^\top \Sigma_k^{-1}(\mu_k - \mu_0) - \frac{1}{2}\sum_i \delta_{z_i,k}(\mathbf{x}_i - \mu_k)\Sigma_k^{-1}(\mathbf{x}_i - \mu_k)$$

$$= -\frac{1}{2}\rho(\mu_k - \mu_0)^\top \Sigma_k^{-1}(\mu_k - \mu_0) - \frac{1}{2}n_k(\mu_k - \bar{\mathbf{x}}_k)^\top \Sigma_k^{-1}(\mu_k - \bar{\mathbf{x}}_k) - \frac{1}{2}\mathrm{Tr}(Q_k\Sigma_k^{-1})$$

$$= -\frac{1}{2}(\rho + n_k)(\mu_k - \mu_0^k)^\top \Sigma_k^{-1}(\mu_k - \mu_0^k) - \frac{1}{2}\frac{\rho n_k}{\rho + n_k}(\bar{\mathbf{x}}_k - \mu_0)^\top \Sigma_k^{-1}(\bar{\mathbf{x}}_k - \mu_0) - \frac{1}{2}\mathrm{Tr}(Q_k\Sigma_k^{-1})$$

$$p(\mu_k, \Sigma_k|\mathbf{Z}, \pi, \mathcal{D}) \propto |\Sigma_k|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\rho + n_k)(\mu_k - \mu_0^k)^\top \Sigma_k^{-1}(\mu_k - \mu_0^k))$$

$$\times|\Sigma_k|^{-\frac{(\kappa+n_k)+M+1}{2}} \exp(-\frac{1}{2}\mathrm{Tr}(W_k\Sigma_k^{-1})$$

$$\Sigma_k|\kappa_k, W_k \sim \mathcal{IW}(\Sigma; \kappa_k, W_k^{-1}),$$

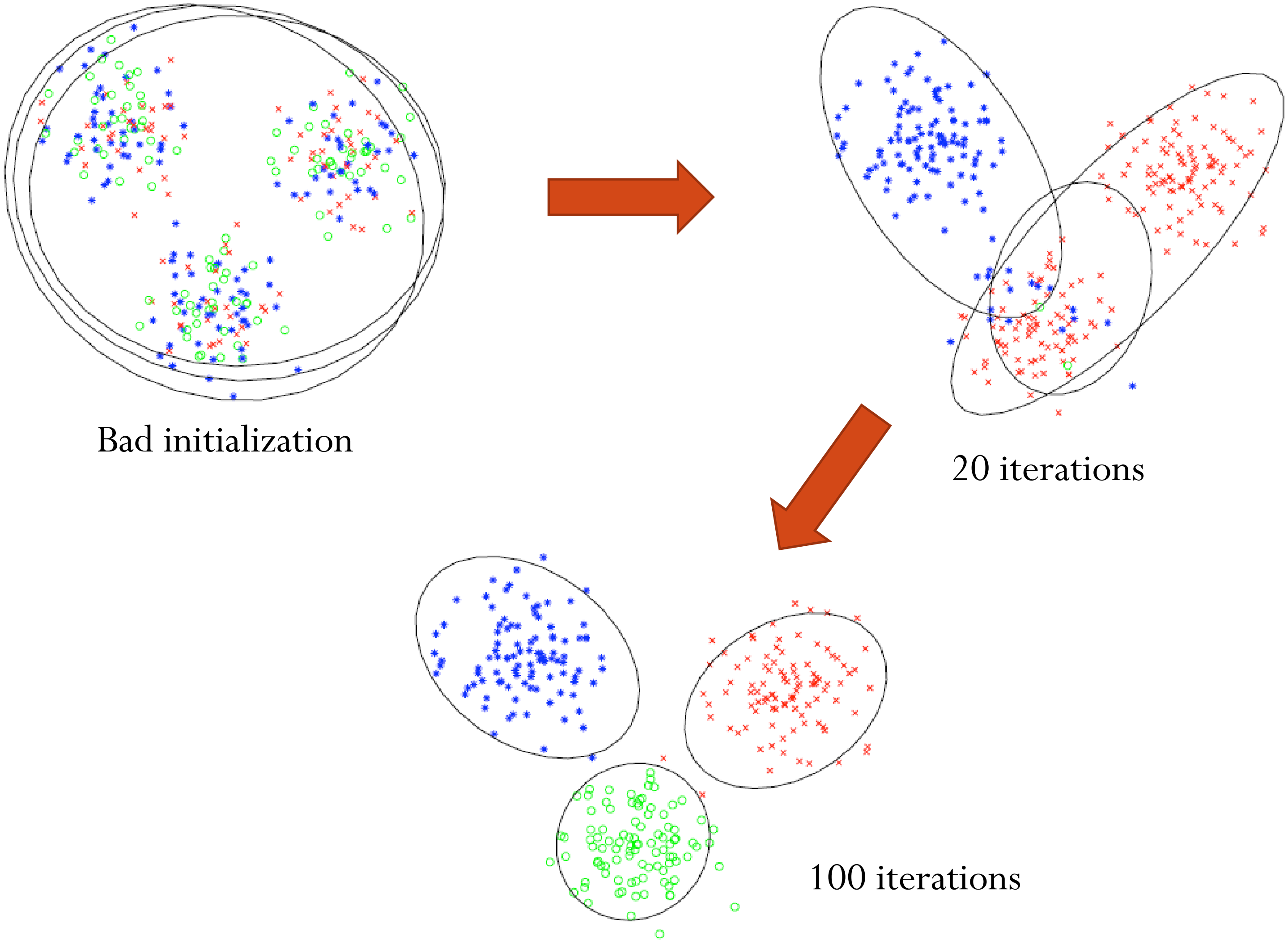
$$\mu_k|\Sigma_k, \mu_0^k, \rho_k \sim \mathcal{N}(\mu; \mu_0^k, \Sigma_k/\rho_k)$$

$$\mu_0^k = \frac{\rho}{\rho + n_k}\mu_0 + \frac{n_k}{\rho + n_k}\bar{\mathbf{x}}_k$$

$$\rho_k = \rho + n_k, \quad \kappa_k = \kappa + n_k$$

$$W_k = W + Q_k + \frac{\rho n_k}{\rho + n_k}(\bar{\mathbf{x}}_k - \mu_0)(\bar{\mathbf{x}}_k - \mu_0)^\top$$

Example



Collapsed Gibbs Sampler

- Idea for an improvement:
  - we can marginalize out some variables due to conjugacy, so do not need to sample it. This is called a collapsed sampler. Here marginalize out π

- Randomly initialize Z, ϕ . Repeat:
  - Sample each z_i from

$$p(z_i|Z_{-i}, \phi, \mathcal{D}) \propto \sum_{k=1}^{K} (n_{-i}^k + \alpha/K) p(\mathbf{x}_i|\phi_k) \delta_{z_i,k}$$

    - n_{-i}^k : # of data points assigned to component k, except i
  - Sample each ϕ_k from the NIW posterior based on $(Z, \mathcal{D})$

Details

- For ϕ, the conditional doesn't change.
- For Z, we have

$$p(\phi, \mathbf{Z}, \mathcal{D}) = \int_\pi p(\pi, \phi, \mathbf{Z}, \mathcal{D}) = p(\phi) \prod_i p(\mathbf{x}_i | z_i, \phi) \int_\pi p(\pi) \prod_i p(z_i | \pi)$$

$$\int_\pi p(\pi) \prod_i p(z_i|\pi) \propto \int_\pi \prod_k \pi_k^{\alpha_k/K + n_k} = \frac{\prod_k \Gamma(\alpha_k/K + n_k)}{\Gamma(\sum_k \alpha_k/K + N)}$$

$$\Rightarrow \quad \int_\pi p(\pi) \prod_i p(z_i|\pi) \propto \prod_k \Gamma(\frac{\alpha_k}{K} + n_k)$$

$$\Rightarrow \quad p(\phi, z_i = k, \mathbf{Z}_{-i}, \mathcal{D}) = p(\phi) \prod_i p(\mathbf{x}_i|z_i, \phi) \Gamma(\frac{\alpha_k}{K} + n_{-i}^k + 1) \prod_{j \neq k} \Gamma(\frac{\alpha_j}{K} + n_{-i}^j)$$

$$= p(\phi) p(\mathbf{x}_i|z_i, \phi)(\frac{\alpha_k}{K} + n_{-i}^k) \prod_j \Gamma(\frac{\alpha_j}{K} + n_{-i}^j) \prod_{j \neq i} p(\mathbf{x}_j|z_j, \phi)$$

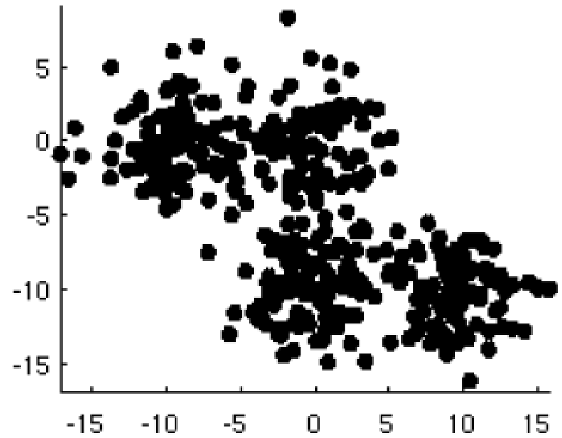
$$p(\phi, z_i = k, \mathbf{Z}_{-i}, \mathcal{D}) \propto p(\mathbf{x}_i|z_i, \phi)(\frac{\alpha_k}{K} + n_{-i}^k)$$

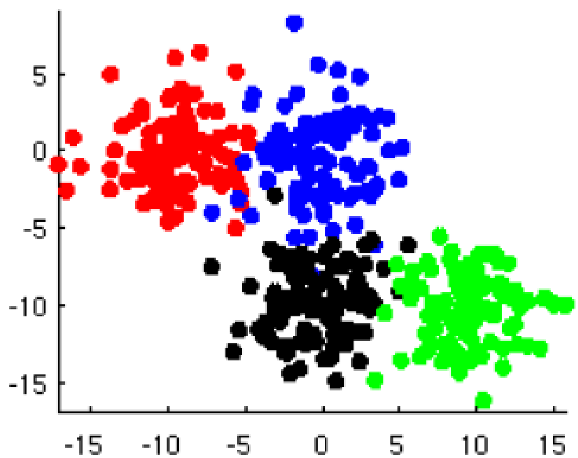
Summary: parametric Bayesian clustering

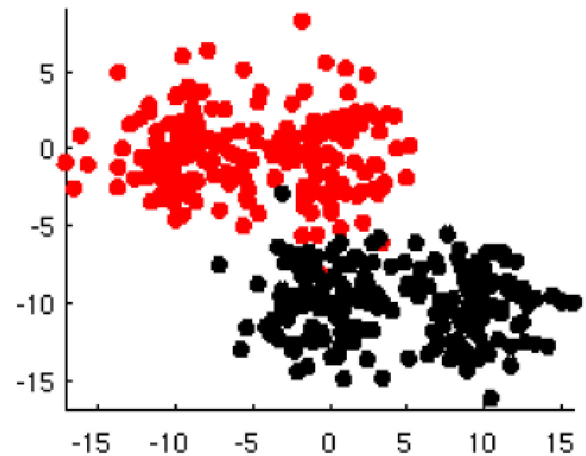
- First specify the likelihood - application specific.
- Next specify a prior on all parameters.
- Exact posterior inference is intractable. Can use a Gibbs sampler for approximate inference.

How to choose K?

- How many clusters?







How to choose *K*?

- Generic model selection:
  - cross-validation, AIC, BIC, MDL, etc.
- Can place of parametric prior on *K*.
- What if we just let $K \to \infty$ in our parametric model?

Outline

- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

A Nonparametric Bayesian Approach to Clustering

- We must again specify two things:
  - The likelihood function (how data is affected by the parameters):

$$p(\mathcal{D}|\theta)$$

    Identical to the parametric case.
  - The prior (the prior distribution on the parameters):

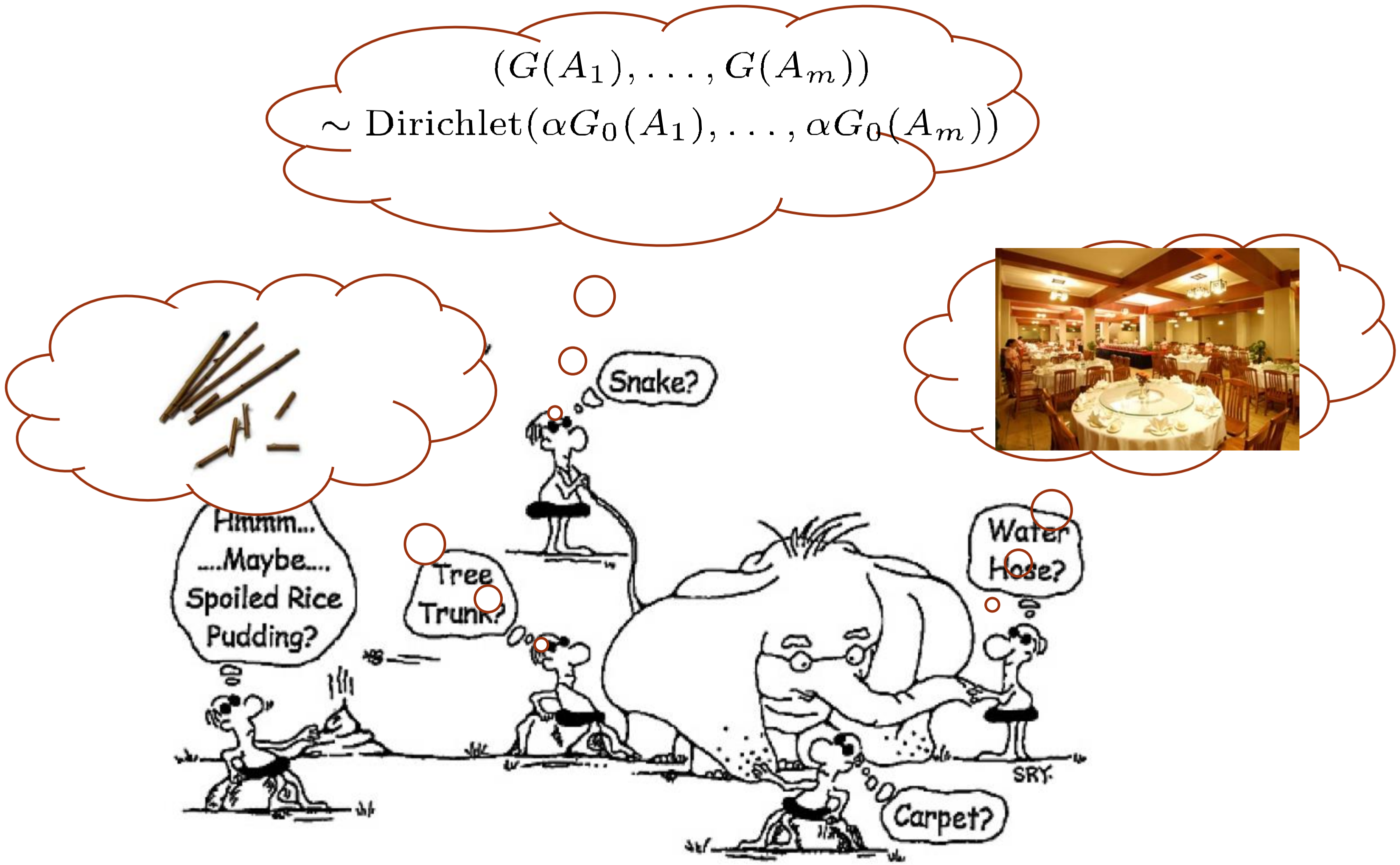
$$p(\theta)$$

    The Dirichlet Process!
- Exact posterior inference is still intractable. But we have can derive the Gibbs update equations!

What is Dirichlet Process?

[http://www.nature.com/nsmb/journal/v7/n6/fig_tab/nsb0600_443_F1.html]

What is Dirichlet Process?



[http://www.nature.com/nsmb/journal/v7/n6/fig_tab/nsb0600_443_F1.html]

Dirichlet Process

- A flexible, nonparametric prior over an infinite number of clusters/classes as well as the parameters for those classes.

- The Dirichlet Process (DP) is a distribution over distributions. We write

$$G \sim DP(\alpha, G_0)$$

to indicate G is a random distribution drawn from the DP

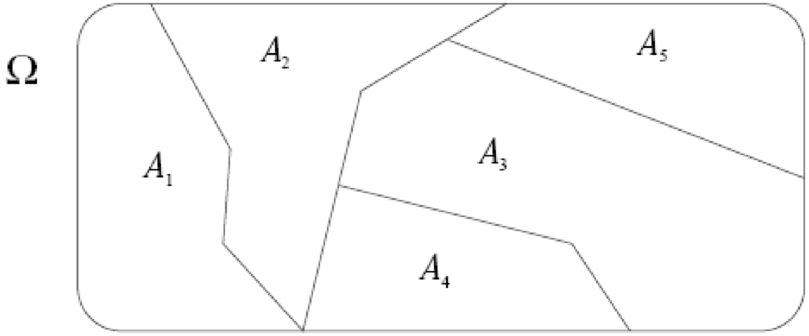
- Parameters:
  - α - the concentration parameter
  - G_0 - the base distribution. A prior for the cluster-specific parameters

Dirichlet Process

- Definition: Let G be a probability measure on the measurable space (Ω, B) and $\alpha \in \mathbb{R}_+$.

- The Dirichlet Process $DP(\alpha, G_0)$ is the distribution on probability measure G such that for any finite partition $(A_1, \ldots, A_m)$ of Ω

$$(G(A_1), \ldots, G(A_m)) \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G(A_m))$$



[Ferguson, Annals of Stats., 1973]

Mathematical Property of DP

- Suppose we sample

$$G \sim DP(\alpha, G_0)$$
$$\theta_1 \sim G$$

- What is the posterior distribution of G given θ_1?

$$G|\theta_1 \sim DP\Big(\alpha + 1, \frac{\alpha}{\alpha+1}G_0 + \frac{1}{\alpha+1}\delta_{\theta_1}\Big)$$

- More generally

$$G|\theta_1, \ldots, \theta_n \sim DP\Big(\alpha + n, \frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i}\Big)$$

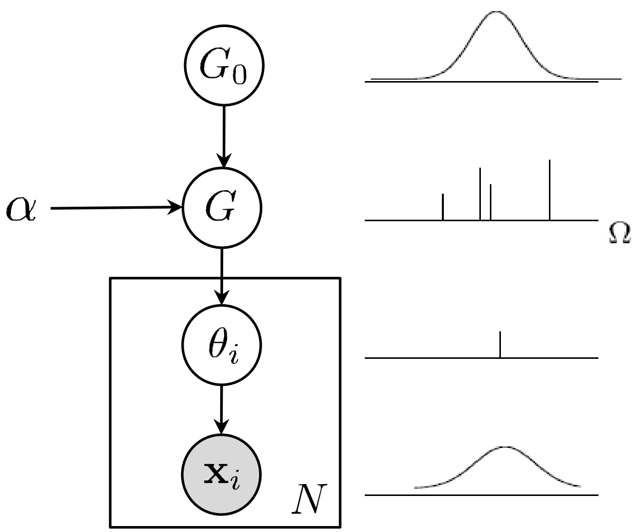
[Ferguson, Annals of Stats., 1973]

Mathematical Property of DP

- With probability 1, a sample $G \sim DP(\alpha, G_0)$ is of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

- This is why DP can used for clustering!



[Sethuraman, Statistica Sinica, 1994]

The Stick-Breaking Process

- Define an infinite sequence of Beta random variables:

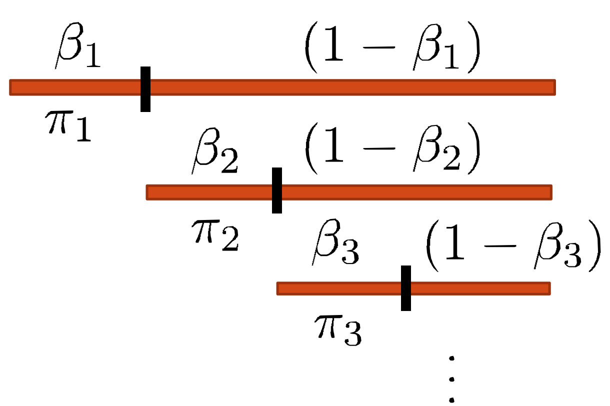
$$\beta_k \sim \mathrm{Beta}(1, \alpha), \;\; k = 1, 2, \ldots$$

- And then define an infinite sequence of mixing proportions as:

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i), \;\; k = 2, 3, \ldots$$

- This can be viewed as breaking off portions of a stick:



The Stick-Breaking Process

- We now have an explicit form of π

$$\pi_1 = \beta_1$$

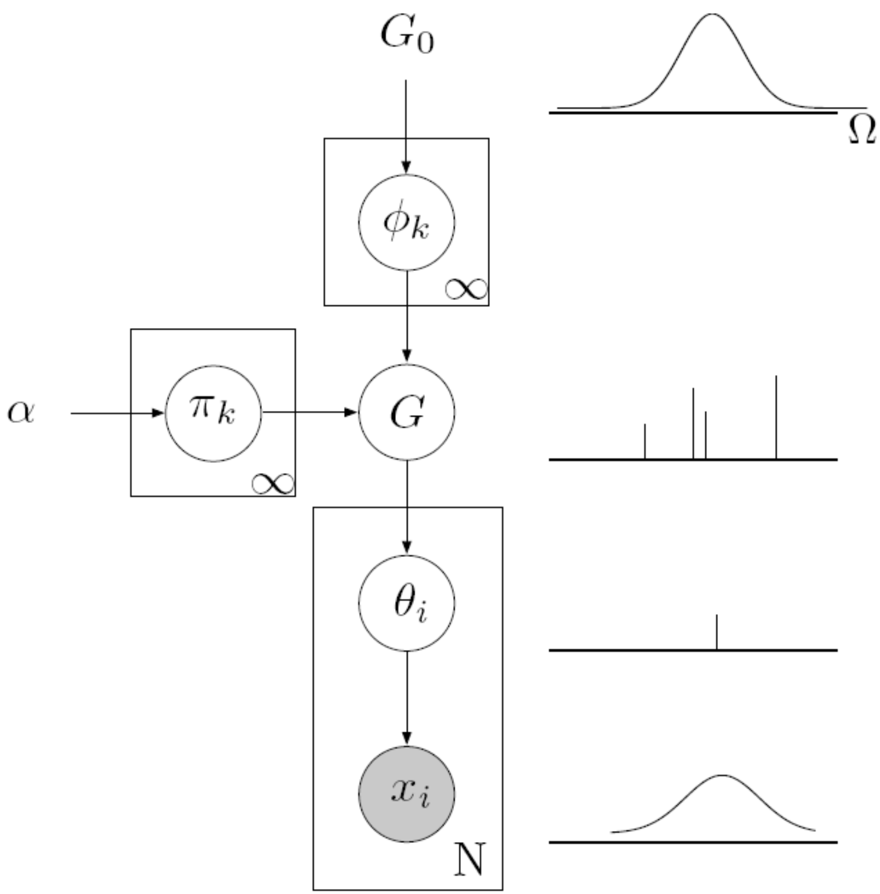
$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i), \ k = 2, 3, \ldots$$

- We can also easily see that $\sum_{k=1}^{\infty} \pi_k = 1$ with probability 1
  - *How to prove?*

- So, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ is a random measure

The Stick-Breaking Process

- Equivalent representation of DP mixtures



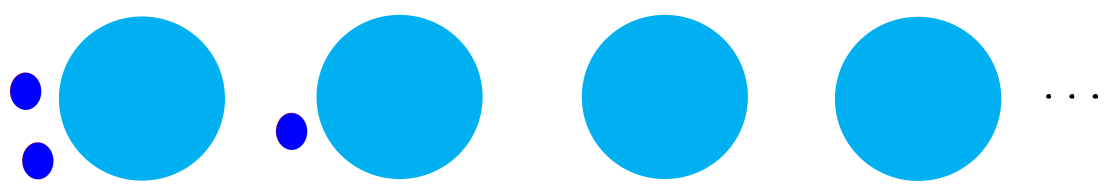
The Chinese Restaurant Process (CRP)

- A random process in which *n* customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - the *n*th customer chooses a table with probability

$$p(z_i = k) = \frac{n_k}{n - 1 + \alpha}, \text{ for a pre-occupied table } k$$

$$p(z_i = k) = \frac{\alpha}{n - 1 + \alpha}, \text{ for an empty table } k$$

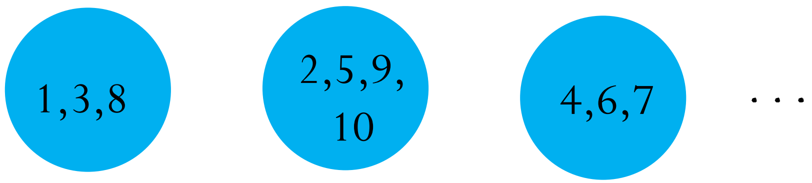
  - where n_k is the number of people sitting at table k.



$\frac{2}{\alpha+3}$ $\frac{1}{\alpha+3}$ $\frac{\alpha}{\alpha+3}$

CRP defines a Partition

- With 10 customers, after sampling, we have



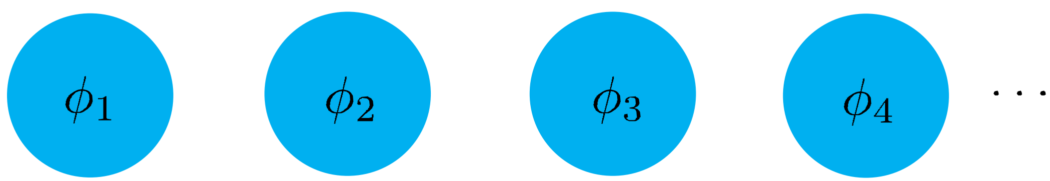
$$p(z_1, z_2, \ldots, z_{10}) = p(z_1)p(z_2|z_1)\ldots p(z_{10}|z_1, \ldots, z_9)$$

$$= \frac{\alpha}{\alpha}\frac{\alpha}{1+\alpha}\frac{1}{2+\alpha}\frac{\alpha}{3+\alpha}\frac{1}{4+\alpha}\frac{1}{5+\alpha}\frac{2}{6+\alpha}\frac{2}{7+\alpha}\frac{2}{8+\alpha}\frac{3}{9+\alpha}$$

- Properties:
  - Any seating arrangement creates a partition
  - Permutation invariant: relabeling the customers doesn't change the distribution
  - Expected number of occupied tables: $O(\alpha \log n)$

The CRP and Clustering

- Data points are customers; tables are clusters
  - CRP defines a prior distribution on the partitioning of the data and on the number of tables
- This prior can be completed with:
  - a likelihood – e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters – a customer to sit at table *k* chooses the parameter vector for that table from the prior



- So we now have a distribution for any quantity that we might care about in the clustering setting

Relation between CRP and DP

- Important fact:
  - The CRP is *exchangeable*.
- Infinite Exchangeability:

$$\forall n,\ \forall \sigma,\ p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

- De Finetti's Theorem (1955): if $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

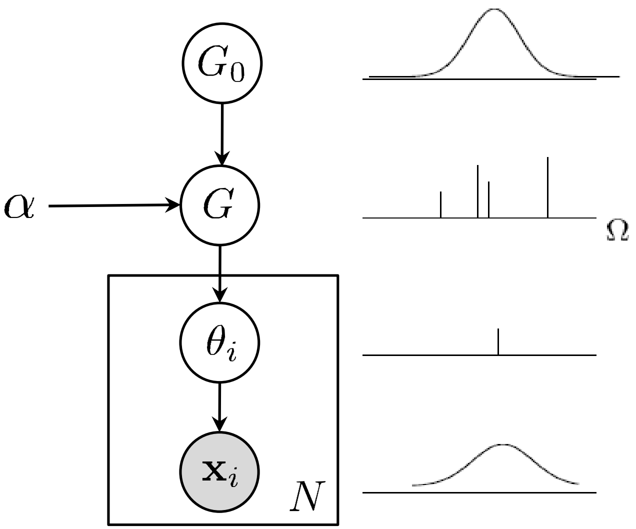
$$p(x_1, \ldots, x_n) = \int \Big(\prod_{i=1}^{n} p(x_i|\theta) \Big) dP(\theta)$$

for some random variable θ

Relation between CRP and DP

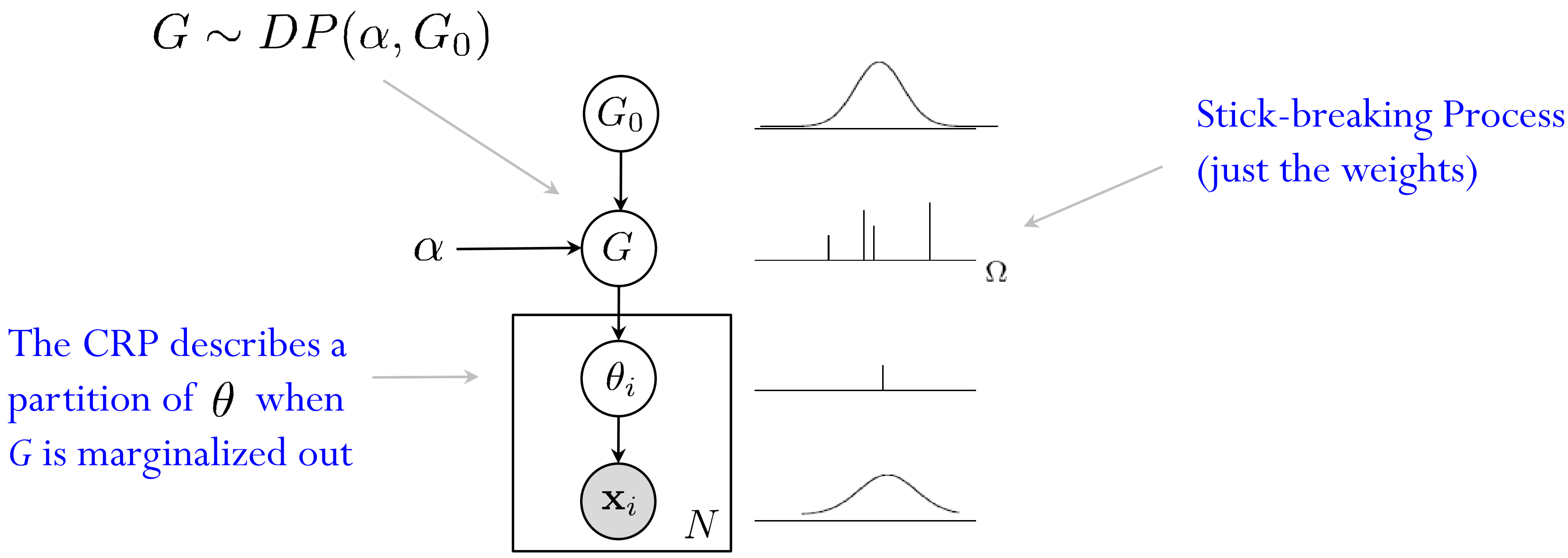
- The Dirichlet Process is the *De Finetti mixing distribution* for the CRP.
- That means, when we integrate out G, we get the CRP

$$p(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} p(\theta_i | G) dP(G)$$



The DP, CRP and Stick-Breaking Process

- Three birds on the same stone



Inference for DP Mixtures – Gibbs sampler

- We introduce the indicators z_i and use the CRP representation.
- Randomly initialize Z, θ . Repeat:
  - sample each z_i from

$$z_i | Z_{-i}, \theta, X \propto \sum_{k=1}^{K} n_{-i}^{k} p(\mathbf{x}_i | \theta_k) \delta_{z_i, k} + \alpha f(\mathbf{x}_i | G_0) \delta_{z_i, K+1}$$

  - Sample each θ_k based on Z and X only for occupied clusters

- This is the sampler we saw earlier, but now with some theoretical basis.

Inference for DP Mixtures – Gibbs sampler

- More Details
  - For the component j with $n_{-i,j} > 0$

$$p(z_i = j|\mathbf{Z}_{-i}, \theta, X) \propto p(z_i = j|\mathbf{Z}_{-i}, \alpha)p(\mathbf{x}_i|\theta_j)$$
$$= \frac{n^j_{-i}}{N-1+\alpha} p(\mathbf{x}_i|\theta_j)$$

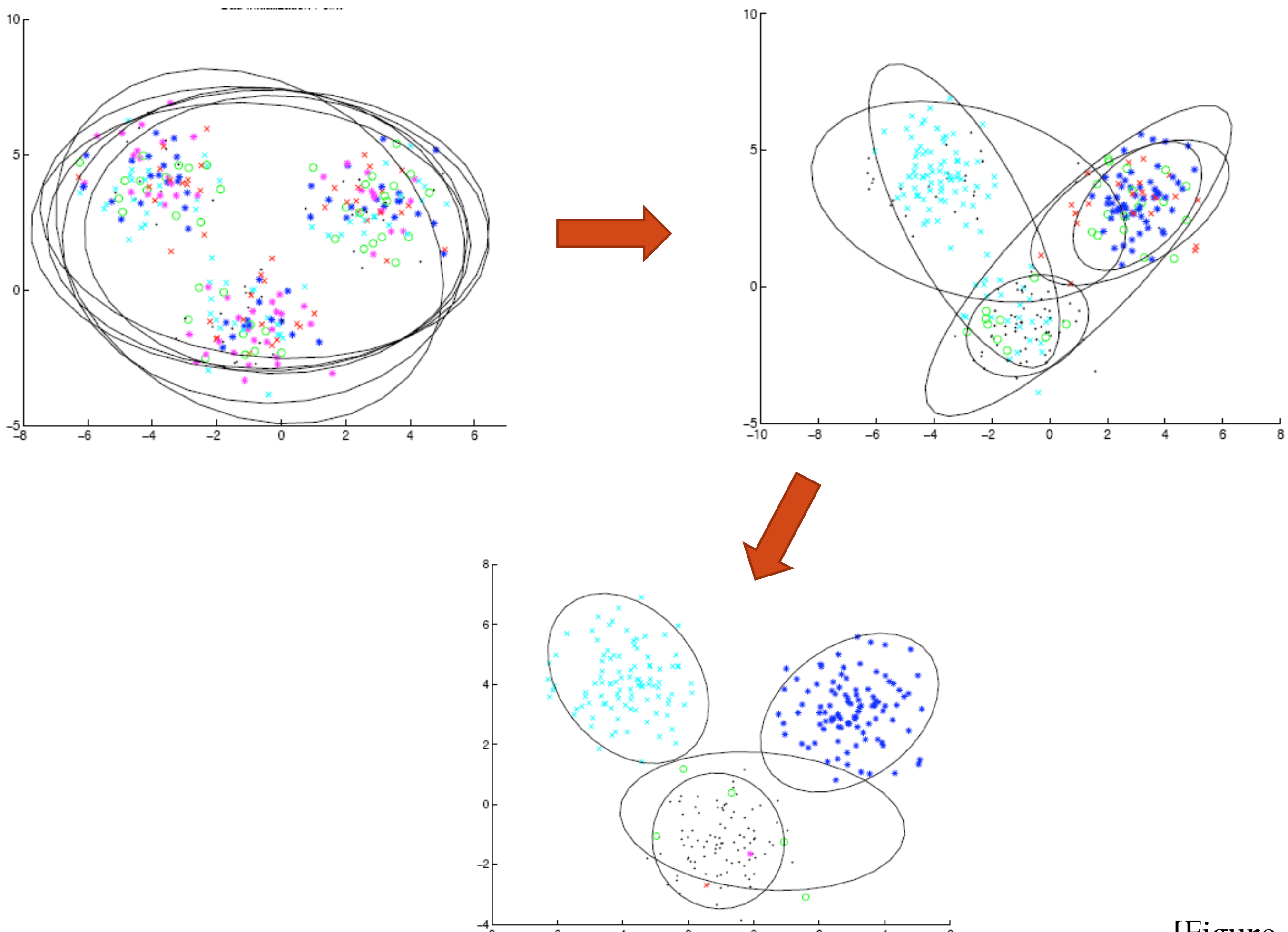
  - For a new component
    - Let $A = \{z_i \neq z_{i'} \text{ for all } i \neq i'\}$

$$p(A|\mathbf{Z}_{-i}, X) = \int p(A, \theta|\mathbf{Z}_{-i}, X)d\theta \propto p(A|\mathbf{Z}_{-i}) \int p_0(\theta)p(\mathbf{x}_i|\theta)d\theta$$
$$\propto \frac{\alpha}{N-1+\alpha} \int p(\mathbf{x}_i|\theta)p_0(\theta)d\theta$$

$$z_i|Z_{-i}, \theta, X \propto \sum_{k=1}^{K} n^k_{-i} p(\mathbf{x}_i|\theta_k)\delta_{z_i,k} + \alpha f(\mathbf{x}_i|G_0)\delta_{z_i,K+1}$$

MCMC in Action for DP

- Matlab demo:



[Figure credit: Miller, 2010]

Improvements to the MCMC Algorithm

- Collapsed Gibbs sampler – collapse out the θ_k if conjugate model
- Split-merge algorithms

Summary: Nonparametric Bayesian Clustering

- First specify the likelihood - application specific.
- Next specify a prior on all parameters - the Dirichlet Process!
- Exact posterior inference is intractable.
  - Can use a Gibbs sampler for approximate inference. This is based on the CRP representation.
  - Can use variational methods for approximate inference. This is based on the Stick-Breaking representation

Outline

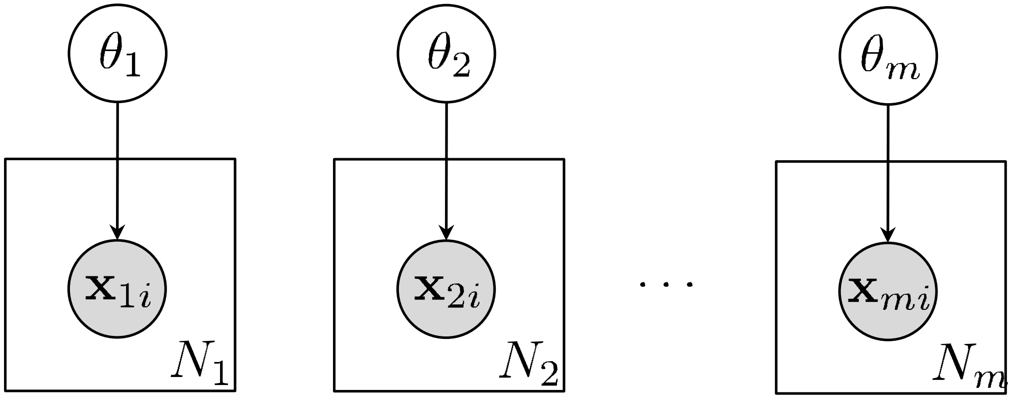
- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

Hierarchical Bayesian Models

- Original Bayesian idea
  - View parameters as random variables - place a prior on them.
- Problem?
  - Often the priors themselves need parameters.
- Solution
  - Place a prior on these parameters!

Multiple Learning Problems

- Example: $\mathbf{x}_i \sim \mathcal{N}(\theta_i, \sigma^2)$ in *m* different groups

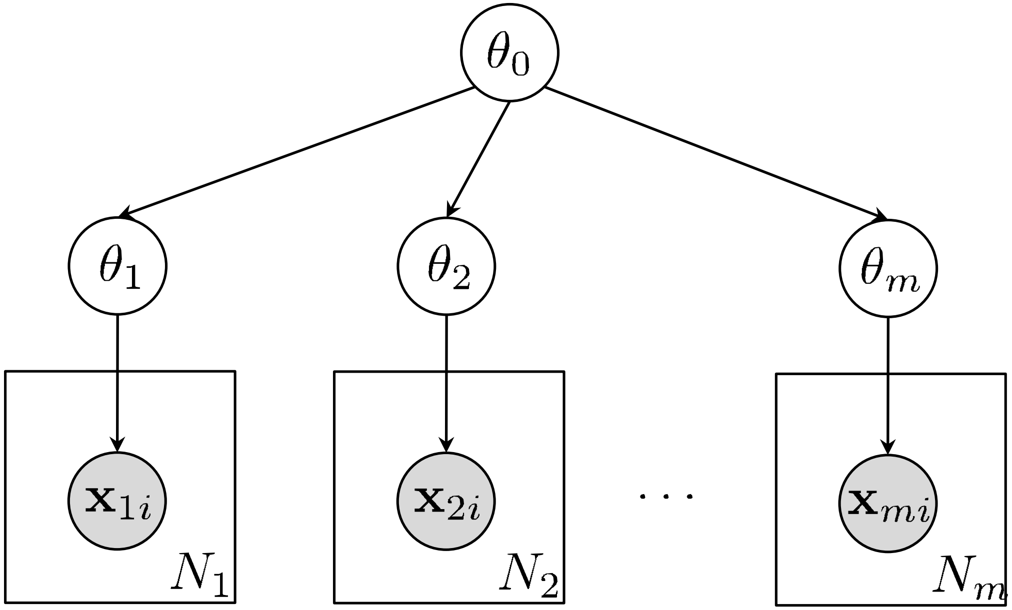


- How to estimate θ_i for each group?

Multiple Learning Problems

- Example: $\mathbf{x}_i \sim \mathcal{N}(\theta_i, \sigma^2)$ in m different groups
- Treat θ_i as random variables sampled from a common prior

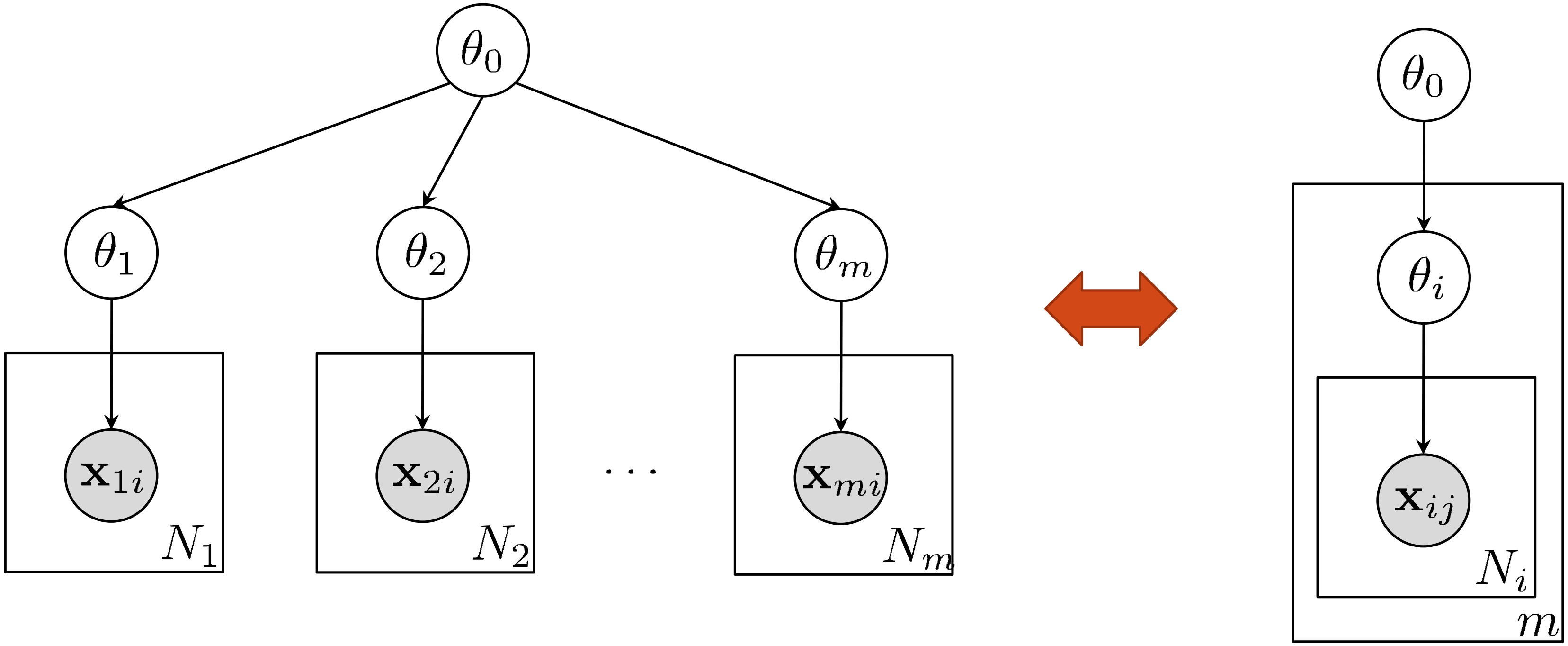
$$\theta_i \sim \mathcal{N}(\theta_0, \sigma_0^2)$$



Multiple Learning Problems

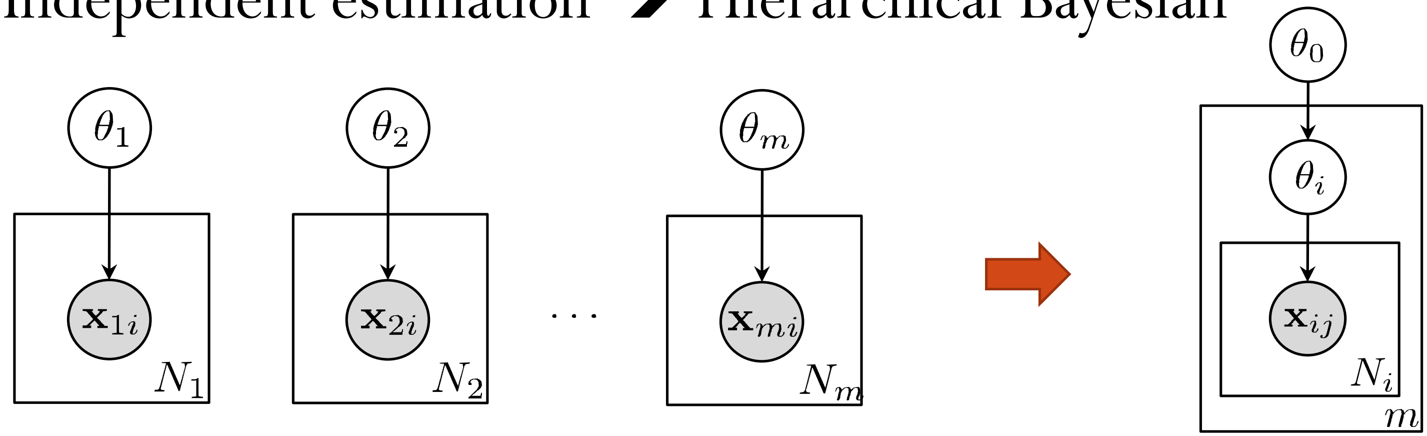
- Example: $\mathbf{x}_i \sim \mathcal{N}(\theta_i, \sigma^2)$ in m different groups
- Treat θ_i as random variables sampled from a common prior

$$\theta_i \sim \mathcal{N}(\theta_0, \sigma_0^2)$$

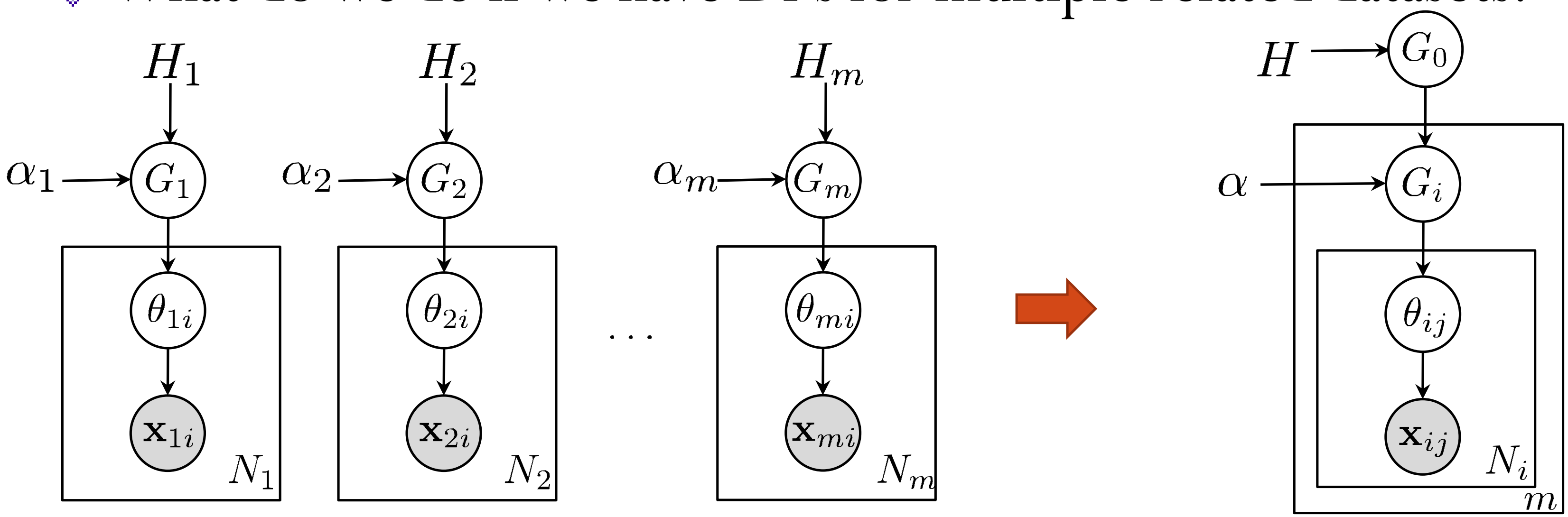


Multiple Learning Problems

- Independent estimation ➔ Hierarchical Bayesian



- What do we do if we have DPs for multiple related datasets?



Hierarchical Dirichlet Process

- What kind of distribution do we use for G_0 ?

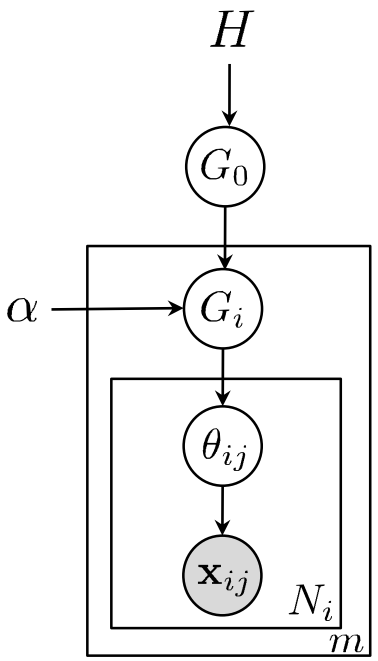
- Attempt 1:
  - Suppose θ_{ij} are mean parameters for a Gaussian where

$$G_i \sim DP(\alpha, G_0)$$

and G_0 is a Gaussian with unknown mean?

$$G_0 = \mathcal{N}(\mu_0, \sigma_0^2)$$

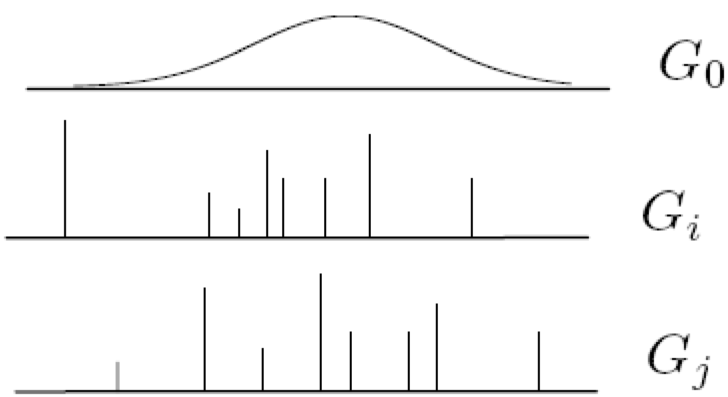
How about this one?



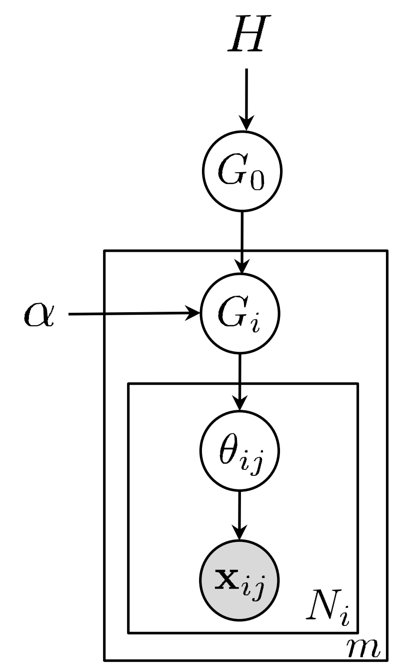
Hierarchical Dirichlet Process

- What kind of distribution do we use for G_0 ?

- Attempt 1:
  - Problem: if G_0 is continuous, then with probability ZERO, G_i and G_j share atoms

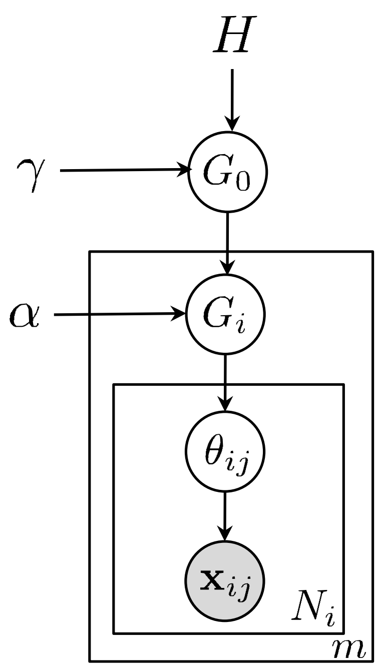


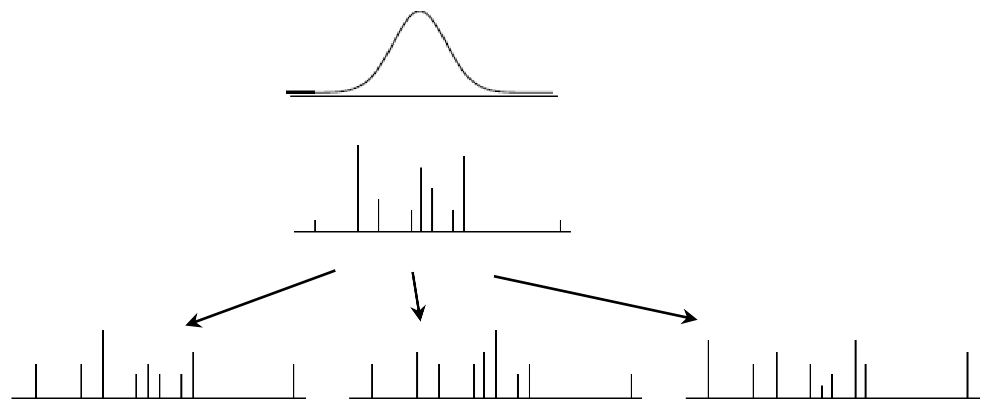
  - There is NO clustering between groups!



Hierarchical Dirichlet Process

- What kind of distribution do we use for G_0 ?
- So, G_0 must be discrete!
- Solution – the *Hierarchical Dirichlet Process*:

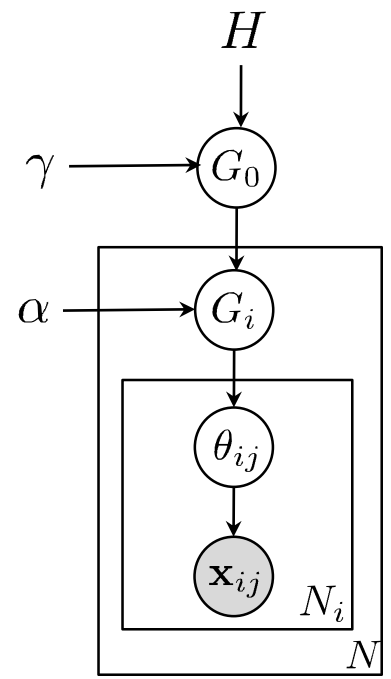




$$G_0 \sim DP(\gamma, H)$$
$$G_i \sim DP(\alpha, G_0)$$
$$\theta_{ij} \sim G_i$$
$$\mathbf{x}_{ij}|\theta_{ij} \sim p(\mathbf{x}_{ij}|\theta_{ij})$$

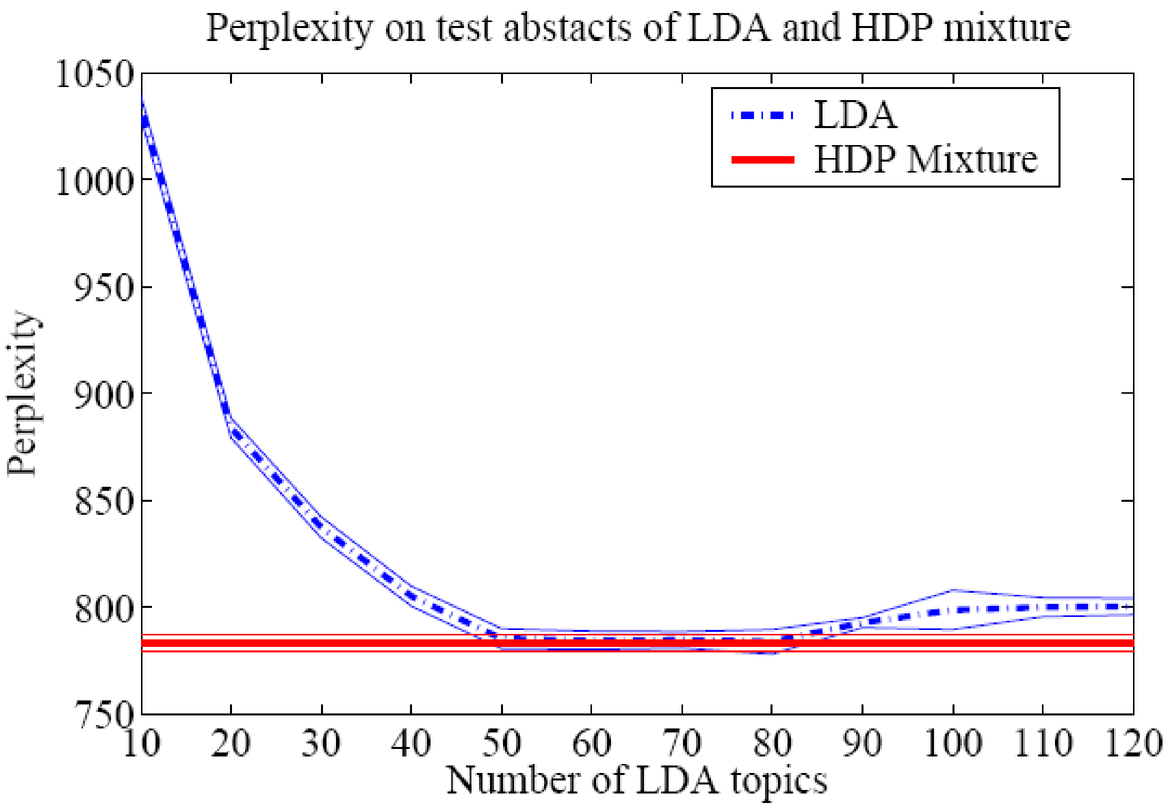
Example 1: HDP topic model

- H – a measure on multinomial probability vectors, e.g., V-dimensional Dirichlet distribution
- G_0 provides a countably infinite collection of multinomial probability vectors (i.e., topics)
- G_i selects a document-specific subset of topics
- θ_{ij} is a particular topic



Example 1: HDP topic model

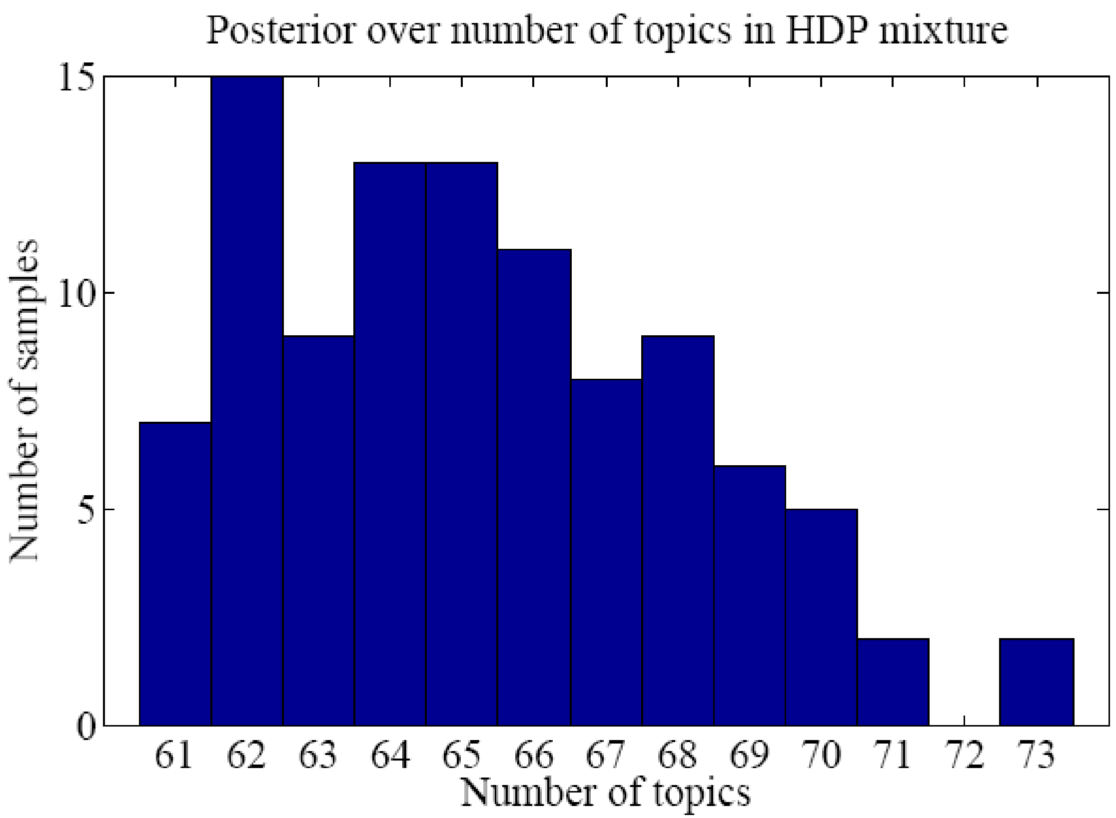
- Results on 5838 biology abstracts



[Teh, Jordan, Beal, & Blei, JASA, 2006]

Example 1: HDP topic model

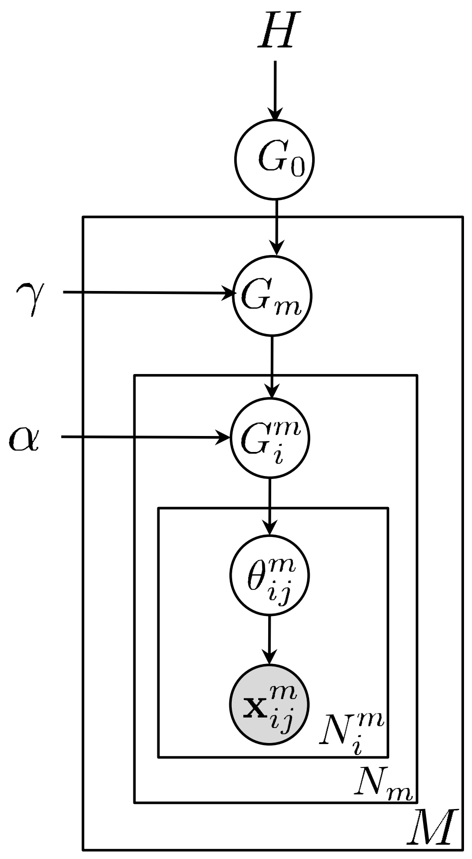
- Results on 5838 biology abstracts



[Teh, Jordan, Beal, & Blei, JASA, 2006]

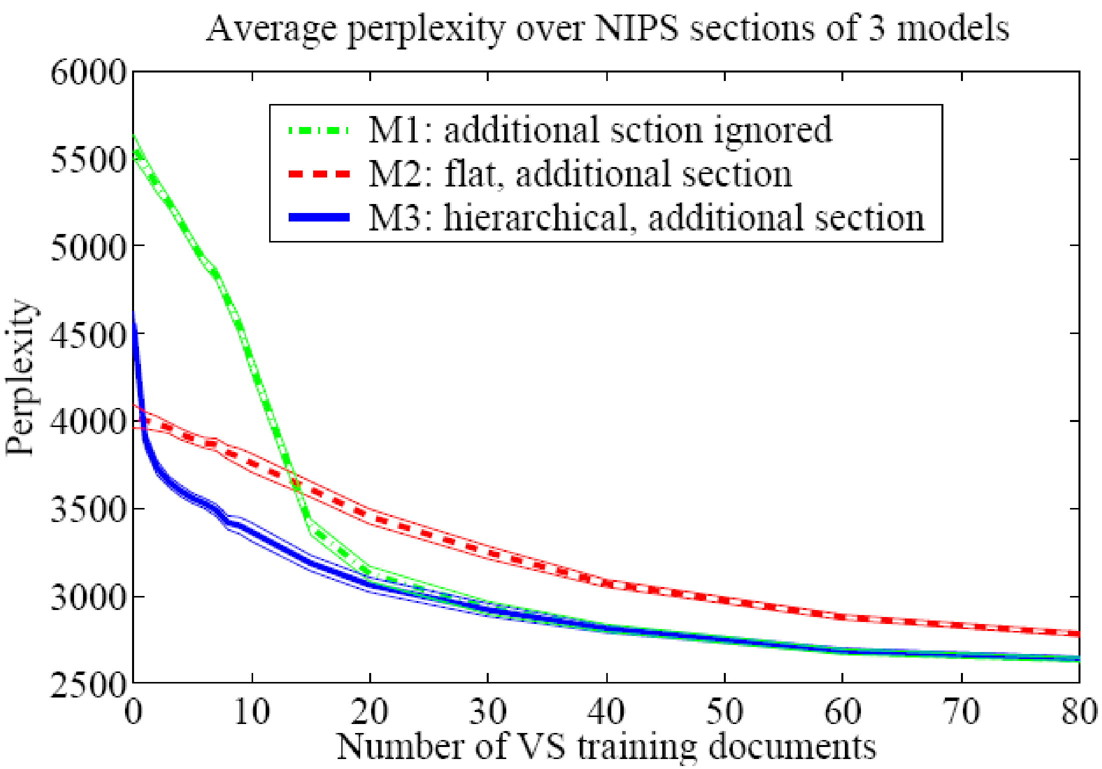
Example 2: HDP topic model for multi-corpora

- H – a measure on multinomial probability vectors, e.g., V-dimensional Dirichlet distribution
- G_0 provides a countably infinite collection of multinomial probability vectors (i.e., topics)
- G_m selects a corpus-specific subset of topics
- G_i^m selects a document-specific subset of topics
- θ_{ij}^m is a particular topic



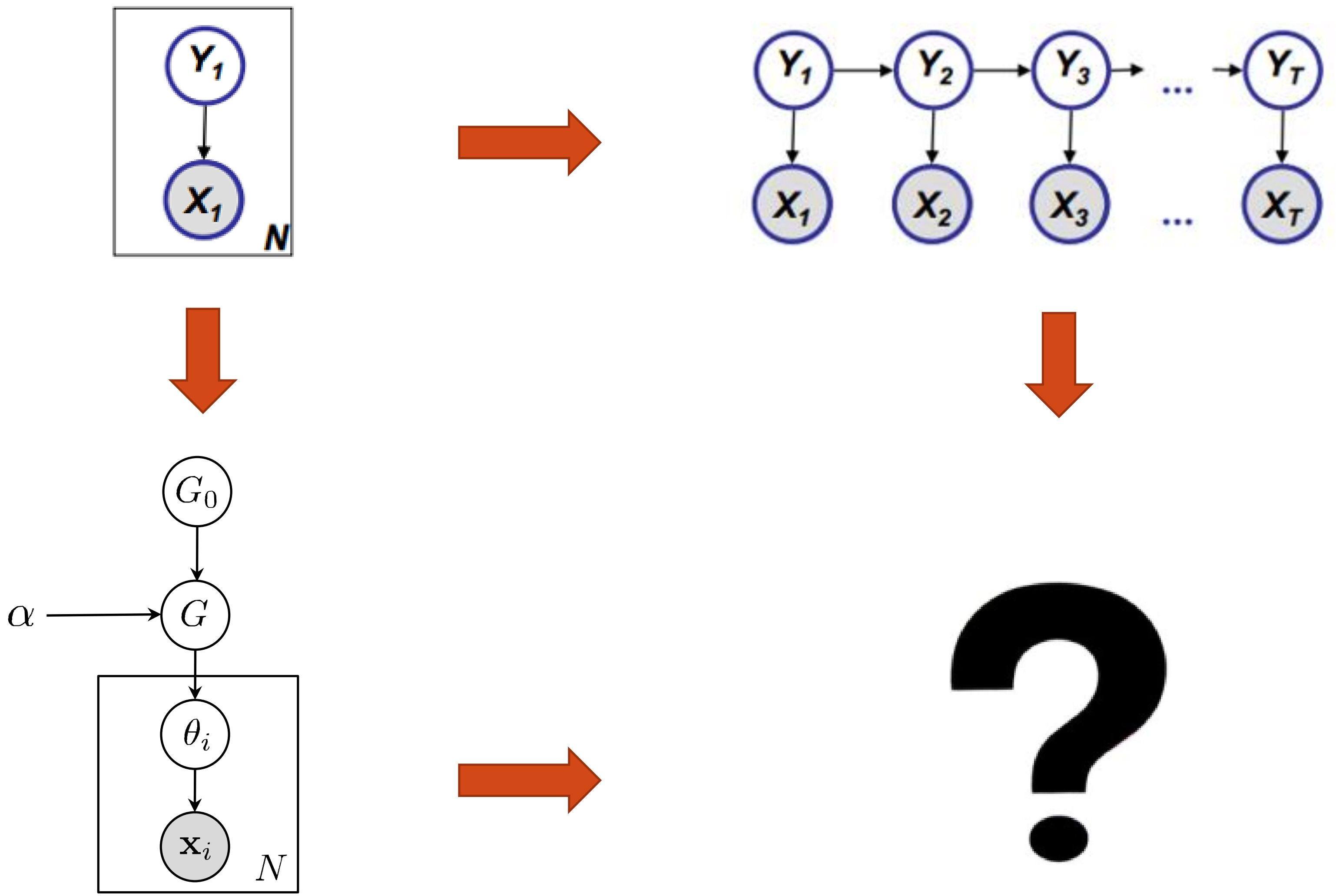
Example 2: HDP topic model for multi-corpora

- Results on NIPS conference proceedings (1988-1999)

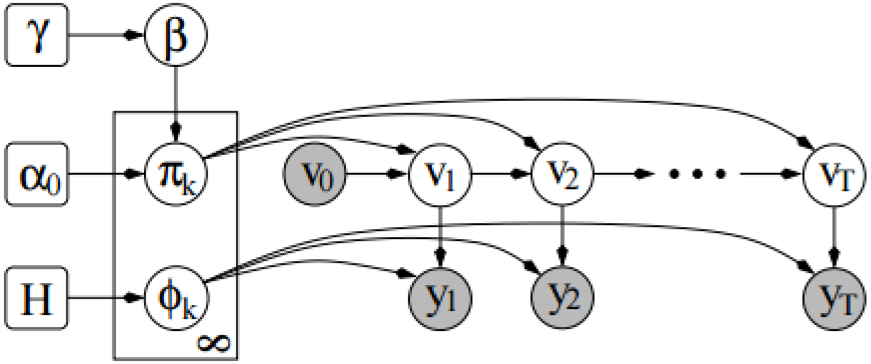


[Teh et al., 2006]

Example 3: Infinite HMMs



Infinite HMMs



$$\boldsymbol{\beta} \mid \gamma \sim \mathrm{GEM}(\gamma)$$

$$\boldsymbol{\pi}_k \mid \alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$$

$$v_t \mid v_{t-1}, (\boldsymbol{\pi}_k)_{k=1}^{\infty} \sim \boldsymbol{\pi}_{v_{t-1}}$$

$$y_t \mid v_t, (\phi_k)_{k=1}^{\infty} \sim F(\phi_{v_t})$$

Questions about HDP?

- Sampling algorithms?
- Variational inference algorithms?
- Stick-breaking construction representation?

References

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics, 1(2):209–230.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics, 2(6):1152–1174.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4:639–650.
- Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In Advances in Neural Information Processing Systems, volume 12.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9:249–265.
- Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1(1):121–144.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.
- http://npbayes.wikidot.com/references
- http://stat.columbia.edu/~porbanz/talks/npb-tutorial.html