Module 3: Bayesian Nonparametrics

Infinite Mixture Models

STAT/BIOSTAT 527, University of Washington

Emily Fox

May 7th, 2013

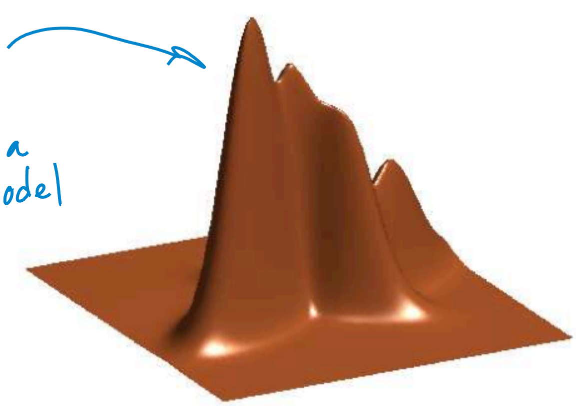
©Emily Fox 2013

Density Estimation

- Estimate a density based on $x_1, \ldots, x_N$

$x_1, \ldots, x_N \sim P$

Let's consider a parametric model

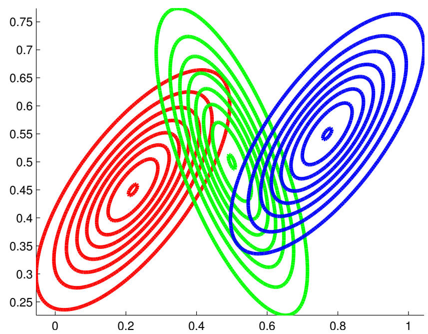


©Emily Fox 2013

Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



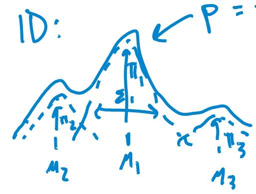
$$P = p(x_i \mid \pi, \mu, \Sigma) =$$

$[\pi_1, \ldots, \pi_K]$ $\{\mu_k, \Sigma_k\}$

$$\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k)$$

Gauss. kernel, just like in KDE, but not centered at obs.

In 1D:

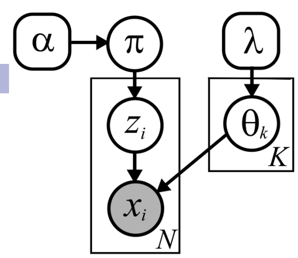


P = target density

$\sum \pi_k = 1$

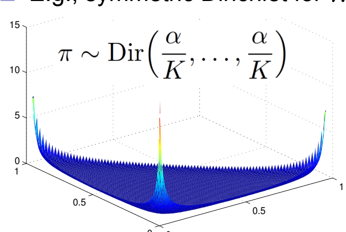
©Emily Fox 2013

Model Summary



- Prior on model parameters
  - E.g., symmetric Dirichlet for π

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

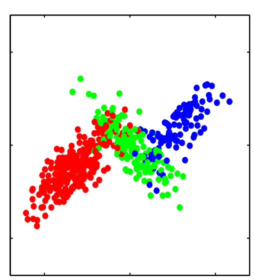


  - Normal inverse Wishart prior for θ_k

- Sample observations as

$$z_i \sim \pi$$

$$x_i \mid z_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$



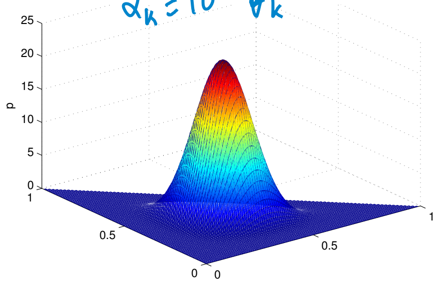
©Emily Fox 2013

Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

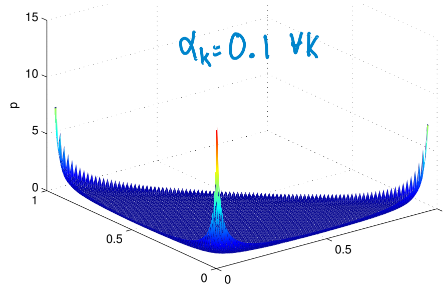
$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$
$\Rightarrow \sum \pi_k = 1$

$\alpha_k = 10 \ \forall k$



$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$\alpha_k = 0.1 \ \forall k$



Moments: $\mathbb{E}_\alpha[\pi_k] = \dfrac{\alpha_k}{\alpha_0}$

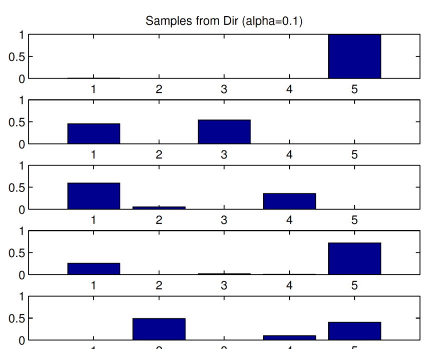
$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)}$$

©Emily Fox 2013

Dirichlet Samples

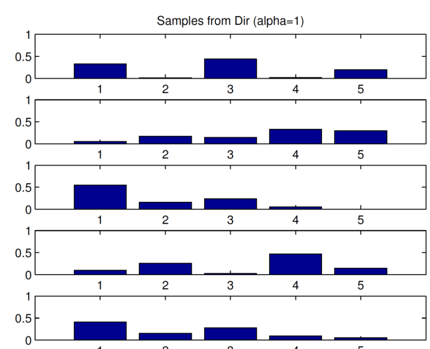
$$\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$$

- Samples are ***sparse*** for small values of α_i



$\mathrm{Dir}(\pi \mid 0.1, 0.1, 0.1, 0.1, 0.1)$

puts mass @ corners

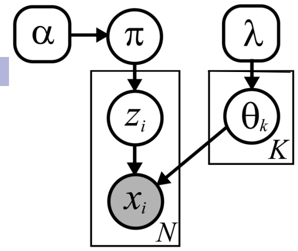


$\mathrm{Dir}(\pi \mid 1.0, 1.0, 1.0, 1.0, 1.0)$

uniform

©Emily Fox 2013

Model In Pictures

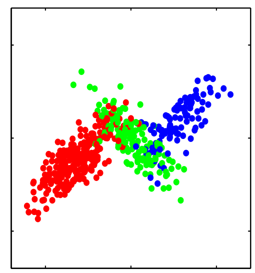


- Mixture weights

$$\pi$$

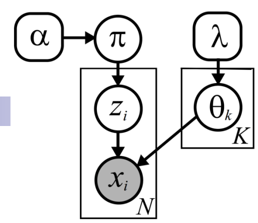
- For each observation,

$$z_i \sim \pi$$
$$x_i \mid z_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$



©Emily Fox 2013

GMM Sampler



- Recall model
  - Observations: $x_1, \dots, x_N$
  - Cluster indicators: $z_1, \dots, z_N$ — want these
  - Parameters: π, θ_k — want these
    - $\pi = [\pi_1, \dots, \pi_K]$
    - $\theta_k = \{\mu_k, \Sigma_k\}$
  - Generative model:

$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z_i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(\lambda) \qquad x_i \mid z_i, \{\theta_k\} \sim N(\mu_{z_i}, \Sigma_{z_i})$$

- Iteratively sample

$$z_i \mid \pi, \{\theta_k\}, \{x_i\} \qquad i=1,\dots,N$$
$$\pi \mid \{z_i\}, \cancel{\{\theta_k\}}, \cancel{\{x_i\}}$$
$$\theta_k \mid \cancel{\pi}, \{z_i\}, \{x_i\} \qquad k=1,\dots,K$$

©Emily Fox 2013

Standard Finite Mixture Sampler

Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample a new set of mixture parameters as follows:

1. Independently assign each of the N data points x_i to one of the K clusters by sampling the indicator variables $z = \{z_i\}_{i=1}^N$ from the following multinomial distributions:

$$z_i^{(t)} \sim \frac{1}{Z_i} \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})\, \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})$$

2. Sample new mixture weights according to the following Dirichlet distribution:

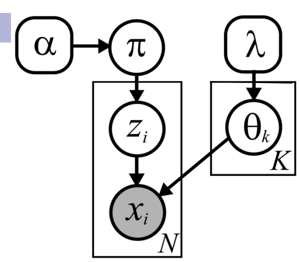
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^{N} \delta(z_i^{(t)}, k)$$

3. For each of the K clusters, independently sample new parameters from the conditional distribution implied by those observations currently assigned to that cluster:

$$\theta_k^{(t)} \sim p(\theta_k \mid \{x_i \mid z_i^{(t)} = k\}, \lambda)$$

©Emily Fox 2013

Mixtures Induce Partitions

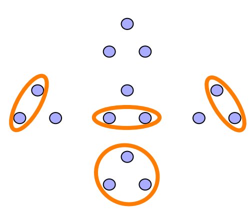


- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

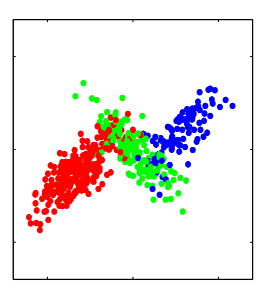
$$z_i \sim \pi$$

- The number of ways of assigning *N* data points to *K* mixture components is K^N

- If $K \geq N$ this is much larger than the number of ways of partitioning that data:



N=3: 5 partitions versus $3^3 = 27$



Mixtures Induce Partitions

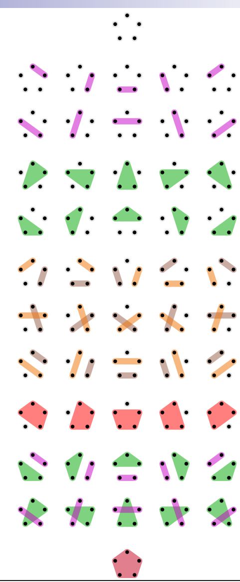
- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

$$z_i \sim \pi$$

- The number of ways of assigning *N* data points to *K* mixture components is K^N
- If $K \geq N$ this is much larger than the number of ways of partitioning that data:

For any clustering, there is a unique partition, but many ways to label that partition's blocks.

N=5: 52 partitions versus $5^5 = 3125$



Courtesy Wikipedia

Module 3: Bayesian Nonparametrics

Infinite Mixture Models

STAT/BIOSTAT 527, University of Washington

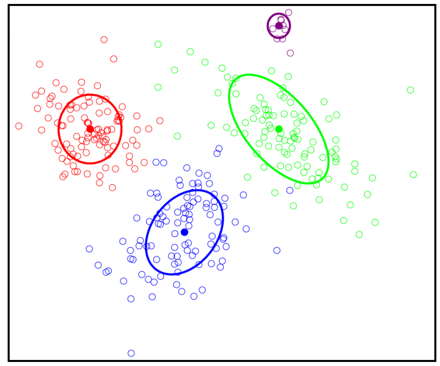
Emily Fox

May 7th, 2013

©Emily Fox 2013

Motivating Nonparametric GMM

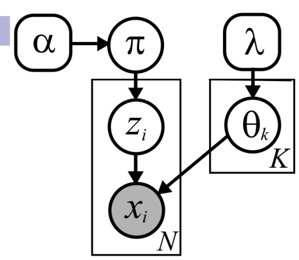
- What if current model doesn't fit new data?
- Bayesian nonparametric approach:
  - Allows infinite # clusters
  - Uses sparse subset
  - Model ***complexity adapts*** to observations



Mixture of Gaussians

$\theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4 \quad \theta_5 \quad \theta_6 \quad \theta_7 \quad \cdots$

Nonparam. Model In Pictures



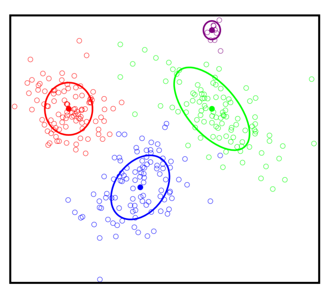
- Mixture weights

π

- For each observation, draw

$$z_i \sim \pi$$

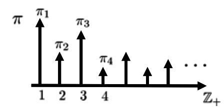
$$x_i \mid z_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$

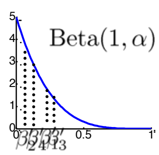


©Emily Fox 2013

Stick-Breaking Process

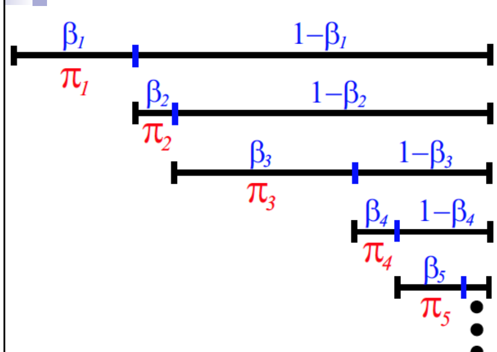
- Start with stick of unit probability mass
- Repeatedly break portions of the remaining stick

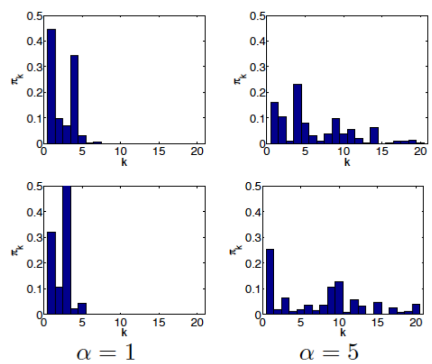




Stick of unit probability mass

Stick-Breaking Process Summary



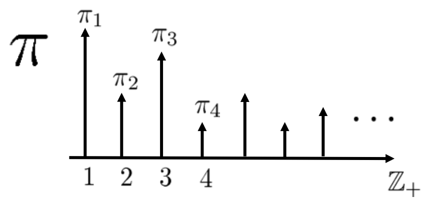


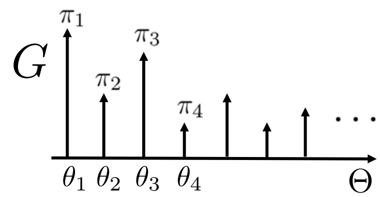
$$\pi_k = \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right)$$

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

$$\mathbb{E}[\beta_k] = \frac{1}{1+\alpha}$$

Stick Breaks + Dirichlet Process





Dirichlet Process Mixture Model

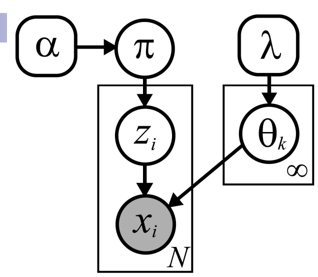
- Place Dirichlet process prior on weights and mixture parameters:

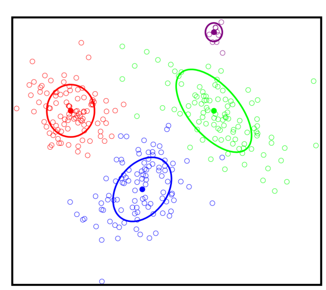
$$G \sim \mathrm{DP}(\alpha, H)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \qquad \begin{array}{l} \pi \\ \theta_k \end{array}$$

- For each observation, draw

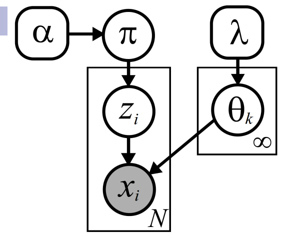
$$z_i \sim \pi$$
$$x_i \mid z_i \sim N(\mu_{z_i}, \Sigma_{z_i})$$





©Emily Fox 2013

Finite versus DP Mixtures



Finite Mixture

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

DP Mixture

$$\pi \sim \mathrm{Stick}(\alpha)$$

$$z_i \sim \pi$$
$$x_i \sim F(\theta_{z_i})$$

THEOREM: For any measureable function *f*, as $K \to \infty$

$$\int_\Theta f(\theta)\, dG^K(\theta) \xrightarrow{\mathcal{D}} \int_\Theta f(\theta)\, dG(\theta)$$

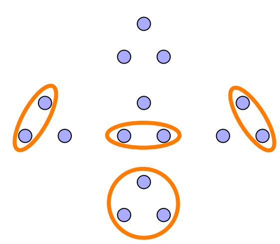
$$G^K(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta) \qquad\qquad G \sim \mathrm{DP}(\alpha, H)$$

Induced Partitions

- Recall that mixture models induce partitions of the data

$$z_i \sim \pi$$

- For a given prior on mixture weights, some partitions are more likely than others apriori
  - Example 1: $\pi \sim \mathrm{Dir}(1, \ldots, 1)$
  - Example 2: $\pi \sim \mathrm{Dir}(0.01, \ldots, 0.01)$

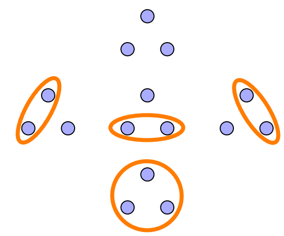


Induced Partitions

- Recall that mixture models induce partitions of the data

$$z_i \sim \pi$$

- For a given prior on mixture weights, some partitions are more likely than others apriori
  - Example 3 (DP mix): $\pi \sim \text{Stick}(\alpha)$



- What is the induced distribution on $z_1, \ldots, z_N$?
  - Do we expect many unique clusters?

Chinese Restaurant Process (CRP)

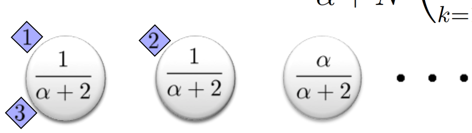
- Distribution on induced partitions described via the CRP
- Visualize clustering as a sequential process of customers sitting at tables in an (infinitely large) restaurant:

customers ⟷ *observed data to be clustered*

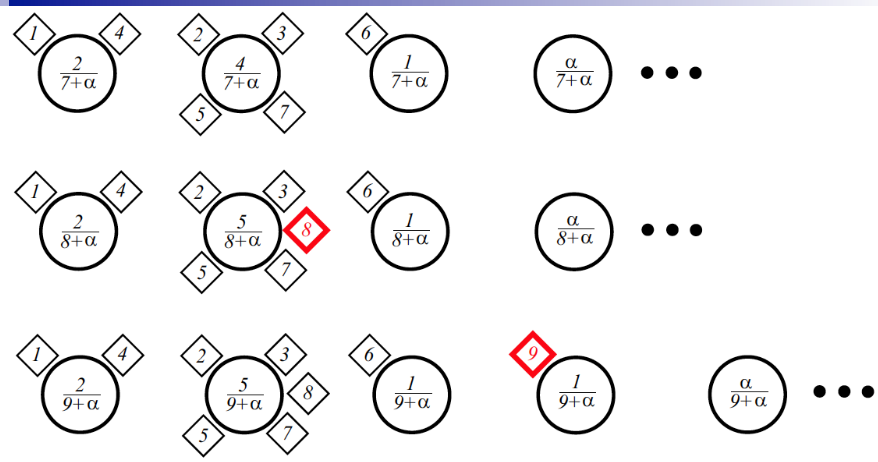
tables ⟷ *distinct clusters*

- The first customer sits at a table. Subsequent customers randomly select a table according to:

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N}\left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k})\right)$$



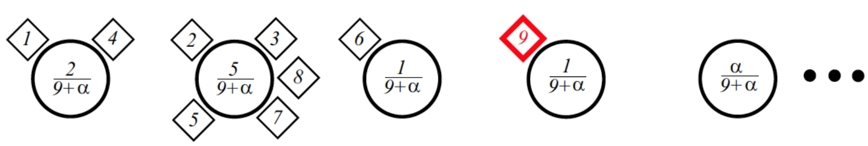
Chinese Restaurant Process (CRP)



CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z,k) + \alpha \delta(z, \bar{k}) \right)$$

- The probability of a seating arrangement of N customers is *independent* of the order they enter the restaurant:

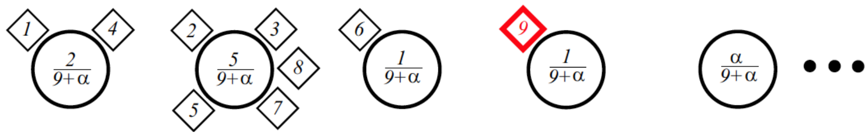


  - Denominator terms:

CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N}\left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k})\right)$$

- The probability of a seating arrangement of *N* customers is *independent* of the order they enter the restaurant:

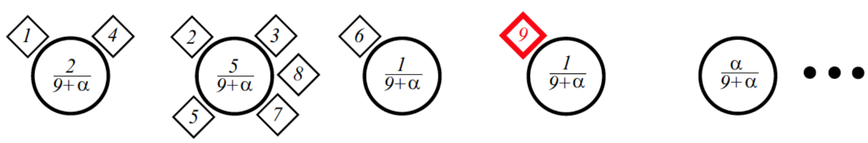


  - Denominator terms: $\frac{1}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$
  - Number of new tables:
    Numerator term for each new table:
    Combined:

CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N}\left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k})\right)$$

- The probability of a seating arrangement of *N* customers is *independent* of the order they enter the restaurant:



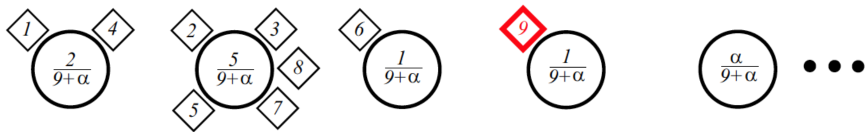
  - Denominator terms: $\frac{1}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$

  - New table numerator terms: α^K
  - Customers joining k^{th} occupied table:

CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z,k) + \alpha \delta(z, \bar{k}) \right)$$

- The probability of a seating arrangement of *N* customers is *independent* of the order they enter the restaurant:



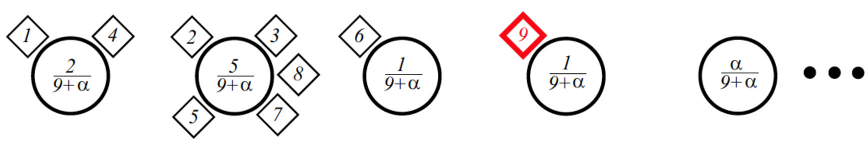
  - Denominator terms: $\frac{1}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$
  - New table numerator terms: α^K
  - Customers joining k^{th} occupied table:

$$1 \cdot 2 \cdots (N_k - 1) = (N_k - 1)! = \Gamma(N_k)$$

CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z,k) + \alpha \delta(z, \bar{k}) \right)$$

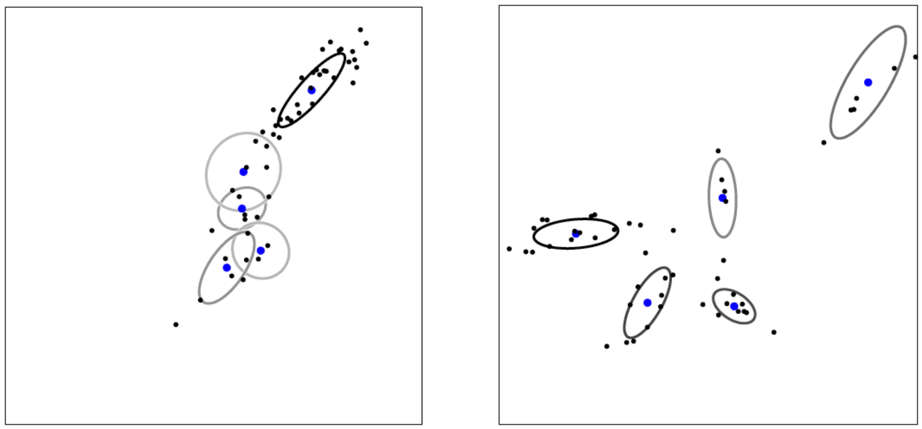
- The probability of a seating arrangement of *N* customers is *independent* of the order they enter the restaurant:



$$p(z_1, \ldots, z_N \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k)$$

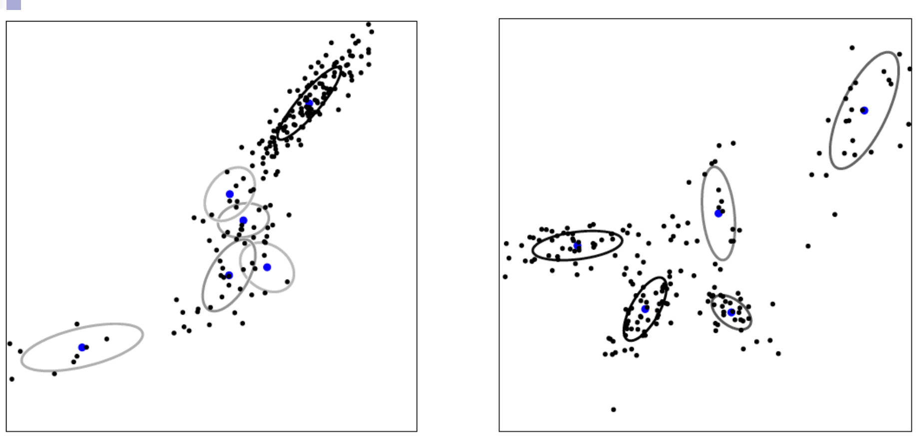
- Thus, the CRP is a prior on an *infinitely exchangeable* sequence

Samples from DP Mixture Priors



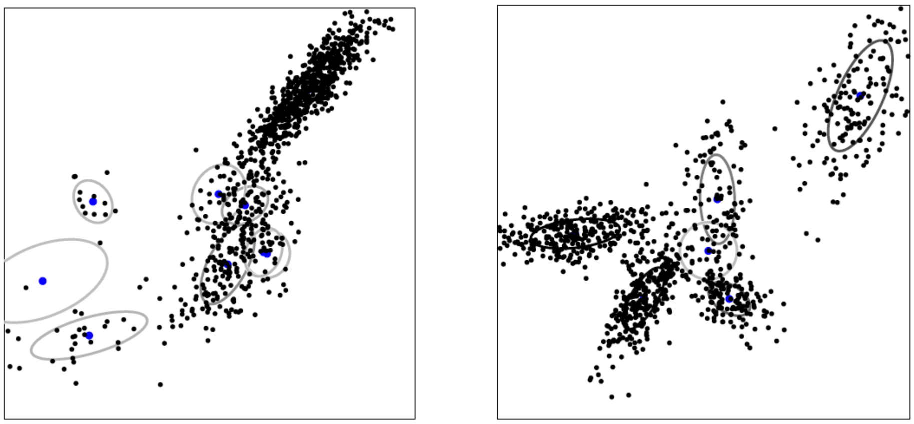
N=50

Samples from DP Mixture Priors



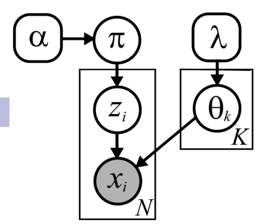
N=200

Samples from DP Mixture Priors



N=1000

Finite GMM Sampler



- Recall model
  - Observations: $x_1, \dots, x_N$
  - Cluster indicators: $z_1, \dots, z_N$
  - Parameters: π, θ_k

want these

$$\pi = [\pi_1, \dots, \pi_K]$$

$$\theta_k = \{\mu_k, \Sigma_k\}$$

  - Generative model:

$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z_i \sim \pi$$

$$\{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(\lambda) \qquad x_i \mid z_i, \{\theta_k\} \sim N(\mu_{z_i}, \Sigma_{z_i})$$

- Iteratively sample

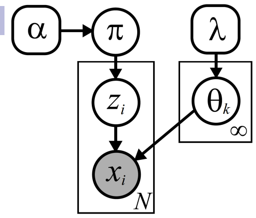
$z_i \mid \pi, \{\theta_k\}, \{x_i\}$ $\quad i=1,\dots,N$

$\pi \mid \{z_i\}, \cancel{\{\theta_k\}}, \cancel{\{x_i\}}$

$\theta_k \mid \cancel{\pi}, \{z_i\}, \{x_i\}$ $\quad k=1,\dots,K$

©Emily Fox 2013

Collapsed DP Mixture Sampler



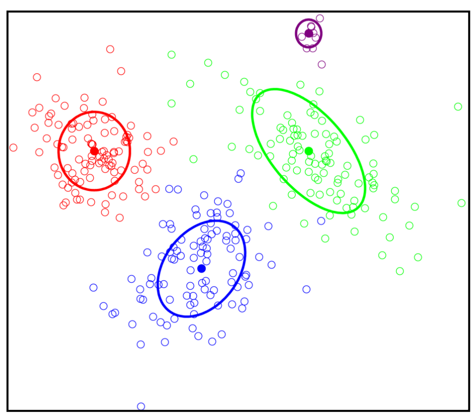
- Can't sample π directly
- Integrate out all infinite-dimensional params

- Iteratively sample the cluster indicators

Collapsed Sampler Intuition

- Previously, $p(z_i = k \mid x_i, \pi, \theta) \propto \pi_k p(x_i \mid \theta_k)$

- If you're not told π, θ_k



©Emily Fox 2013

Predictive Likelihood Term

- Recall NIW prior…Let's consider 1D example → N-IG

$$\mu_k \mid \sigma_k^2 \sim N(0, \gamma\sigma_k^2) \quad \sigma_k^2 \sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 S_0}{2}\right)$$

- Normal inverse gamma posterior
  → Student t predictive likelihood

$$p(x \mid \{x_j | z_j = k, j \neq i\}) = t_{\nu_0 + N_k^{-i}}\left(\frac{1}{\gamma + N_k^{-i}} \sum_{j: z_j = k, j \neq i} x_j,\right.$$

$$\left.\frac{N_k^{-i} + \gamma^{-1} + 1}{(N_k^{-i} + \gamma^{-1})(\nu_0 + N_k^{-i})}\left(\nu_0 S_0 + \sum_{j: z_j = k, j \neq i} x_j^2 - (N_k + \gamma^{-1})^{-1}\Big(\sum_{j: z_j = k, j \neq i} x_j\Big)^2\right)\right)$$

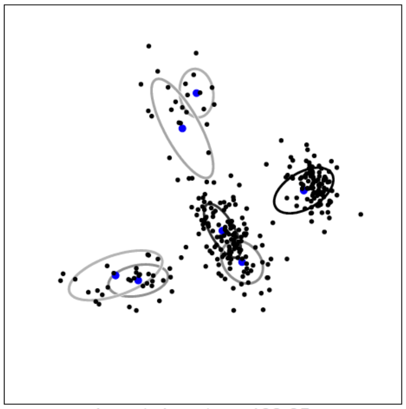
  - Conjugacy: This integral is **tractable**

©Emily Fox 2013

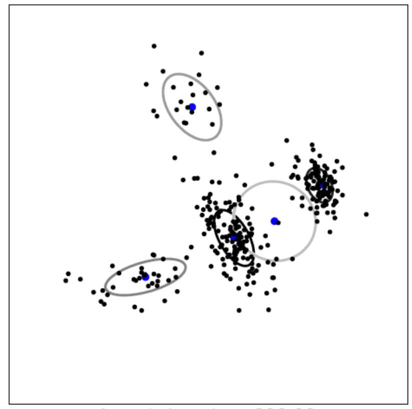
Collapsed DP Mixture Sampler

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $\alpha = \alpha^{(t-1)}$ and $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, resample z_i as follows:
   (a) For each of the K existing clusters, determine the predictive likelihood
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   Also determine the likelihood $f_{\bar{k}}(x_i)$ of a potential new cluster $\bar{k}$
   $$p(x_i \mid \lambda) = \int_\Theta f(x_i \mid \theta)\, h(\theta \mid \lambda)\, d\theta$$
   (b) Sample a new cluster assignment z_i from the following $(K+1)$–dim. multinomial:
   $$z_i \sim \frac{1}{Z_i}\left(\alpha f_{\bar{k}}(x_i)\delta(z_i, \bar{k}) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)\delta(z_i, k)\right) \qquad Z_i = \alpha f_{\bar{k}}(x_i) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)$$
   N_k^{-i} is the number of other observations currently assigned to cluster k.
   (c) Update cached sufficient statistics to reflect the assignment of x_i to cluster z_i. If $z_i = \bar{k}$, create a new cluster and increment K.
3. Set $z^{(t)} = z$.
4. If any current clusters are empty ($N_k = 0$), remove them and decrement K accordingly.

Collapsed DP Sampler: 2 Iterations

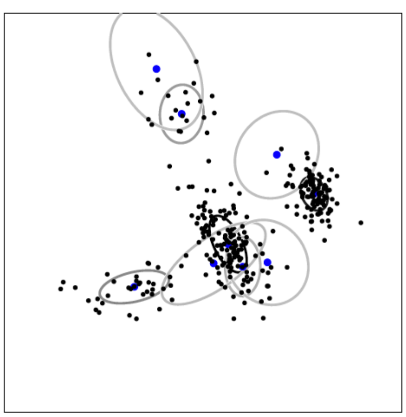


log p(x | π, θ) = –462.25

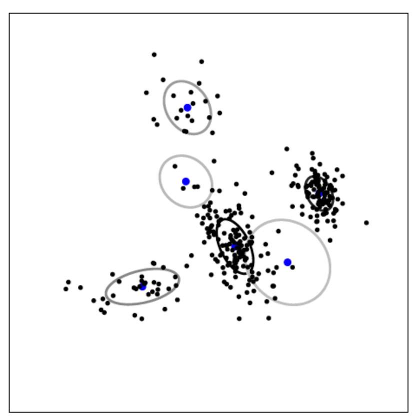


log p(x | π, θ) = –399.82

Collapsed DP Sampler: 10 Iterations

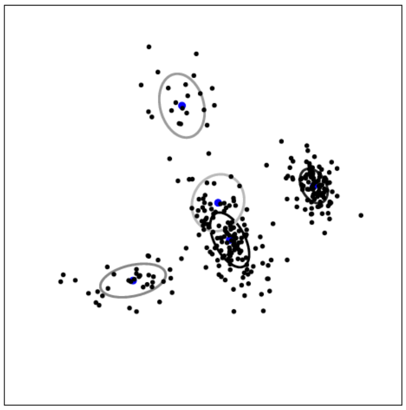


log p(x | π, θ) = –398.32

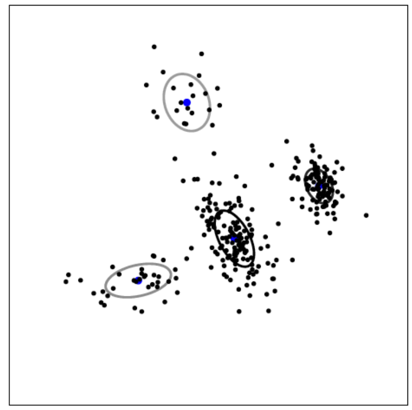


log p(x | π, θ) = –399.08

Collapsed DP Sampler: 50 Iterations

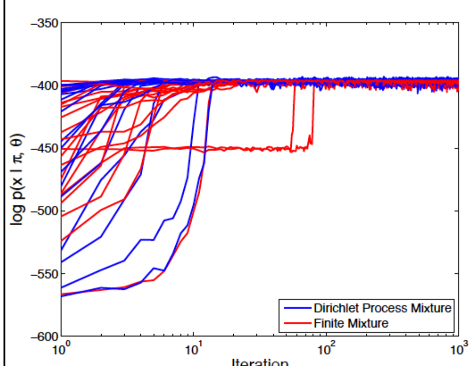


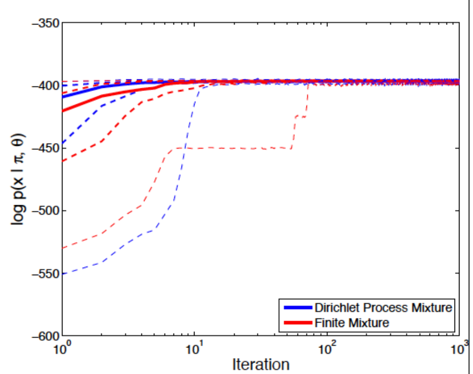
log p(x | π, θ) = –397.67



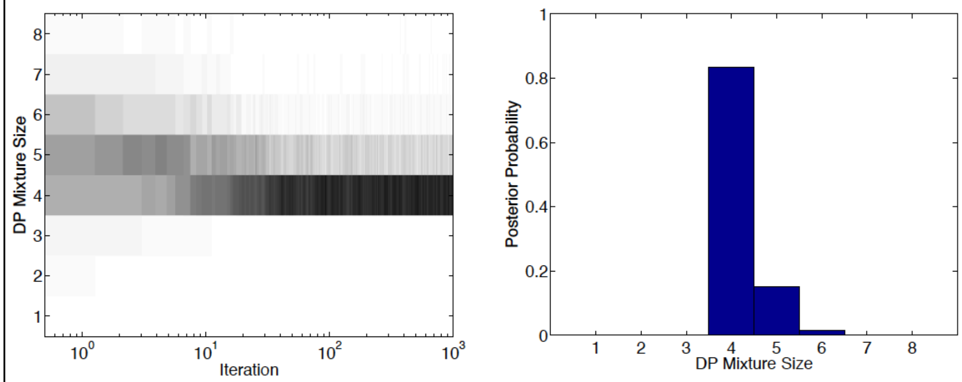
log p(x | π, θ) = –396.71

DP vs. Finite Mixture Samplers

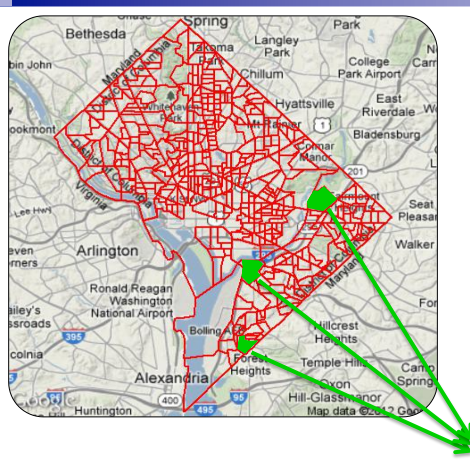




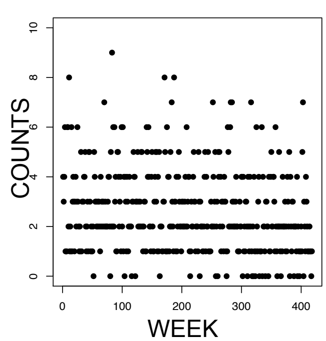
DP Posterior Number of Clusters



DC Violent Crime Data



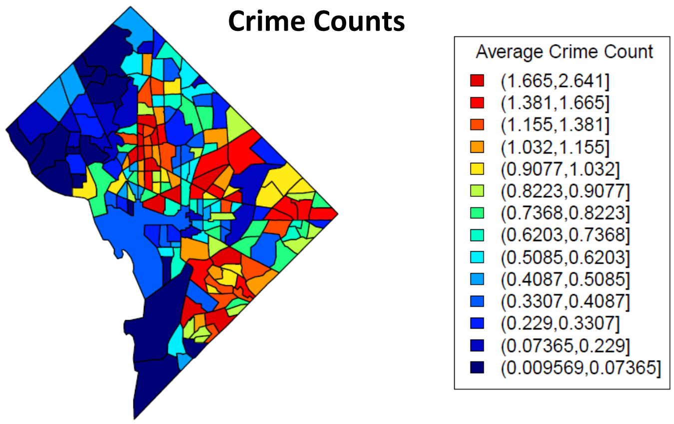
- 188 census tracts
- Weekly crime counts from 2001-2008
- Violent crime types:
  - ADW, arson, robbery, rape



Time series = crime counts

DC Violent Crime Data

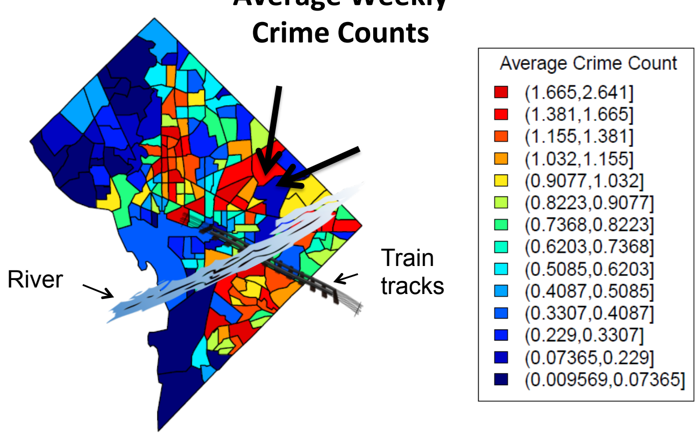
Average Weekly Crime Counts



Goal: Forecast next week's map

DC Violent Crime Data

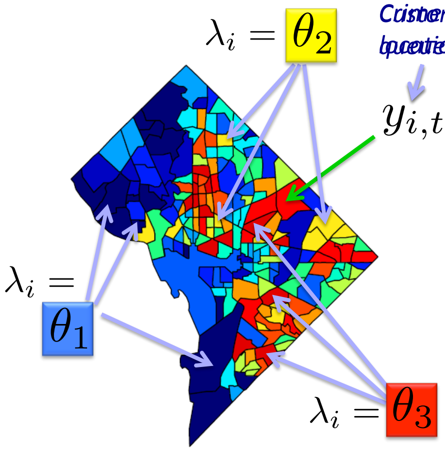
Average Weekly Crime Counts



Similar behavior in spatially disjoint tracts

→ Cluster census tracts

Poisson Integer-Valued Autoregressions



Crime counts in location i at time t

Customers in queue at time t

Poisson new customers

$$y_{i,t} = \alpha \odot y_{i,t-1} + \epsilon_{i,t}(\lambda_i)$$

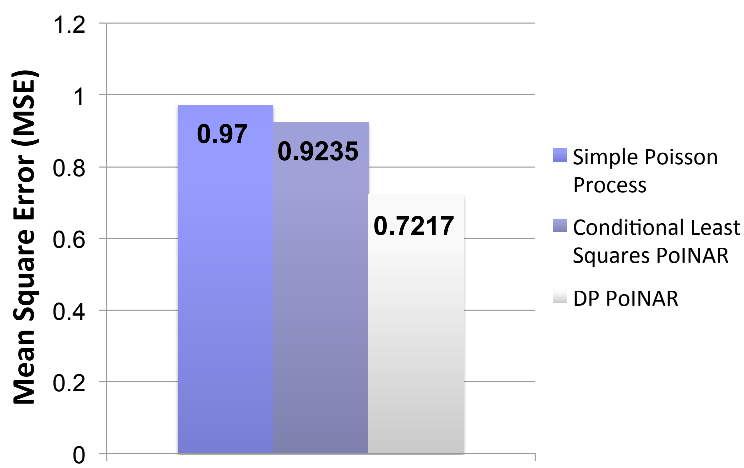
Thinning of customers previously in queue

Rate of arrivals in location i

- Cluster regions based on *rate of arriving crimes*
- Pool info within clusters
- Unknown number of clusters → *Dirichlet process*

Aldor-Noiman, Brown, Fox, and Stine, *arXiv:1304.5642*, April 2013

Prediction Results



Aldor-Noiman, Brown, Fox, and Stine, *arXiv:1304.5642*, April 2013

Acknowledgements

Slides based on parts of the lecture notes of Erik Sudderth for "Applied Bayesian Nonparametrics" at Brown University

©Emily Fox 2013