

Module 3: Bayesian Nonparametrics

Finite Mixture Models

for density estimation

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 29th, 2014

©Emily Fox 2014

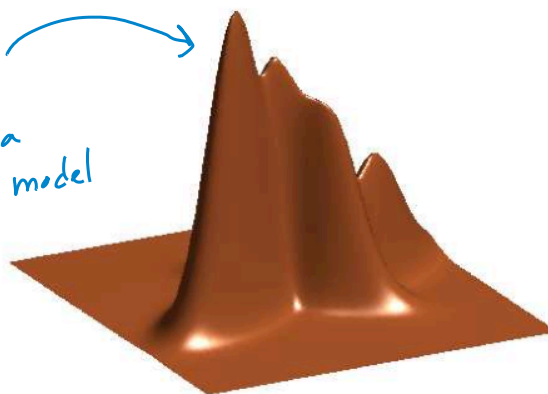
1

Density Estimation

- Estimate a density based on x_1, \dots, x_N

$x_1, \dots, x_n \sim P$

*Let's consider a
parametric model*



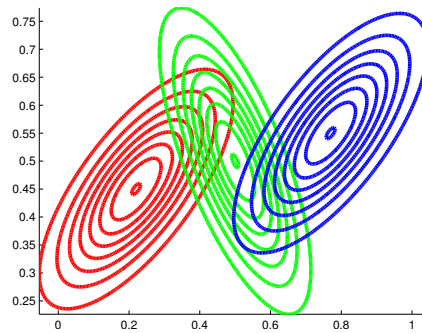
©Emily Fox 2014

2

Density as Mixture of Gaussians

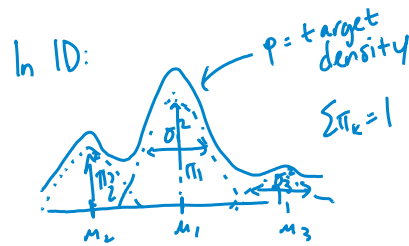
- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



$p =$
 $p(x_i | \pi, \mu, \Sigma) = \sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k)$

Handwritten notes:
 $\{\pi_1, \dots, \pi_K\}$
 $\{\mu_k, \Sigma_k\}$
 Gaussian kernels, but not centered at obs. like in KDE

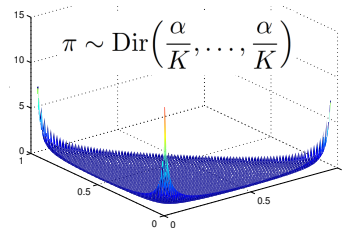


©Emily Fox 2014

3

Model Summary

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

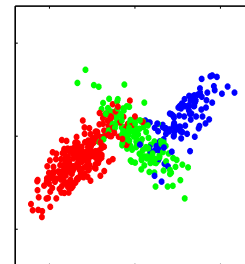
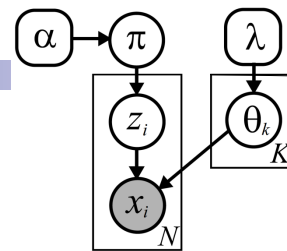


- Normal inverse Wishart prior for θ_k

- Sample observations as

$$z_i \sim \pi$$

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



©Emily Fox 2014

4

Model In Pictures

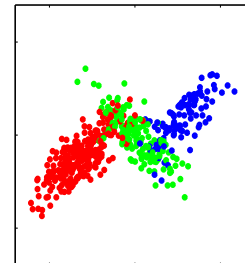
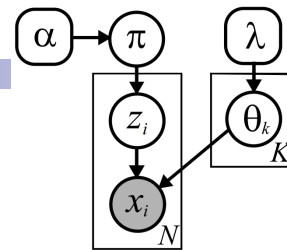
- Mixture weights

π

- For each observation,

$$z_i \sim \pi$$

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

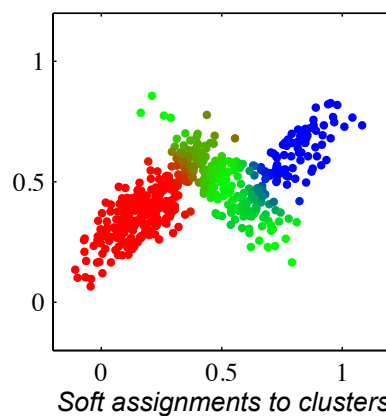


©Emily Fox 2014

5

Clustering our Observations

- We must infer the cluster assignments from the observations



- Posterior probabilities of assignments to each cluster *given* model parameters:

$$r_{ik} = p(z_i = k | x_i, \pi, \theta) = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_i | \mu_j, \Sigma_j)}$$

motivates an iterative alg.

C. Bishop, Pattern Recognition & Machine Learning

Posterior Computations

- From our observations, we want to infer model params
- MAP estimation can be done using expectation maximization (EM) algorithm: *MAP version*

$$\hat{\theta}^{MAP} = \arg \max_{\theta} p(\theta | x) \quad \text{point estimation}$$

- What if we want a full characterization of the posterior?
 - Maintain a measure of uncertainty
 - Estimators other than posterior mode (different loss functions)
 - Predictive distributions for future observations
- Often no closed-form characterization (e.g., mixture models)
- Alternatives:
 - Markov chain Monte Carlo (MCMC) providing samples from posterior
 - Variational approximations to posterior

©Emily Fox 2014

7

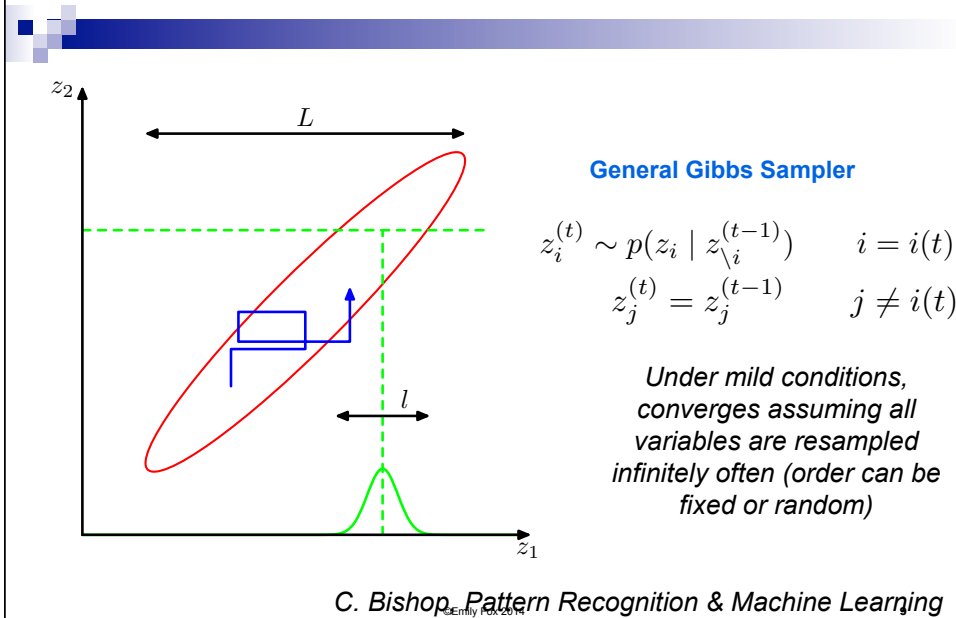
Gibb Sampling

- Let z indicate the set of **all variables in the model**: e.g., cluster indicators and parameters
- Want draws:
 - Construct Markov chain whose steady state distribution is
 - Simplest case:

©Emily Fox 2014

8

Gibbs Sampler for a 2D Gaussian



Example – GMM

■ Recall model

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

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

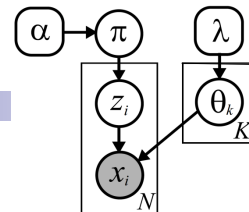
$$\theta_k = \{\mu_k, \Sigma_k\}$$

- Generative model:

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad z_i \sim \pi$$

$$\{\mu_k, \Sigma_k\} \sim \text{NIW}(\lambda) \quad x_i | z_i, \{\theta_k\} \sim N(\mu_{z_i}, \Sigma_{z_i})$$

■ Iteratively sample



Complete Conditional $p(z_i \mid \pi, \{\theta_k\}, \{x_i\})$

- We have

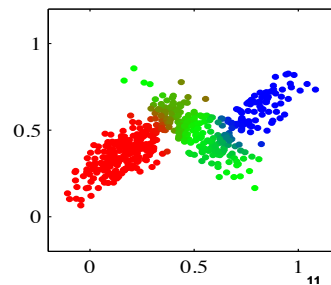
$$z_i \sim \pi$$

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

- As before, we can compute the “responsibility” of each cluster to the observation

$$r_{ik} = p(z_i = k \mid x_i, \pi, \theta) = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{\ell=1}^K \pi_\ell p(x_i \mid \theta_\ell)}$$

- Sample each cluster indicator as



©Emily Fox 2014

Complete Conditional $p(\pi \mid \{z_i\})$

- Recall conjugate Dirichlet prior

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

- Dirichlet posterior

- Assume we condition on cluster indicators $z_i \sim \pi$
- Count occurrences of $z_i = k$
- Then,

$$p(\pi \mid \alpha, z_1, \dots, z_N) \propto$$

- Conjugacy: This **posterior** has same form as **prior**

©Emily Fox 2014

12

Complete Conditional $p(\theta_k \mid \{z_i\}, \{x_i\})$

- Recall NIW prior...Let's consider 1D example \rightarrow N-IG

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

- Normal inverse gamma posterior

- Consider observation indices i such that $z_i = k$
- For these observations, $x_i \mid z_i = k \sim N(\mu_k, \Sigma_k)$
- Then,

$$\mu_k \mid \sigma_k^2, \{z_i\}, \{x_i\} \sim N\left(\frac{1}{N_k + \gamma^{-1}} \sum_{i:z_i=k} x_i, \frac{1}{N_k + \gamma^{-1}} \sigma_k^2\right)$$

$$\sigma_k^2 \mid \{z_i\}, \{x_i\} \sim \text{IG}\left(\frac{\nu_0 + N_k}{2}, \frac{\nu_0 S_0 + \sum_{i:z_i=k} x_i^2 - (N_k + \gamma^{-1})^{-1} (\sum_{i:z_i=k} x_i)^2}{2}\right)$$

- Conjugacy: This **posterior** has same form as **prior**

©Emily Fox 2014

13

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:

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

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

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

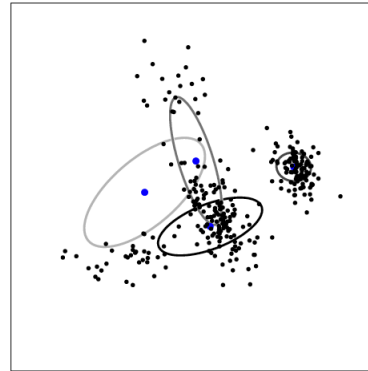
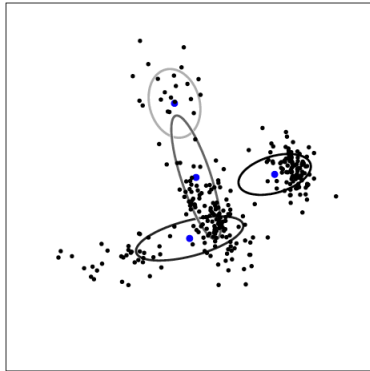
3. For each of the K clusters, independently sample new parameters from the conditional distribution implied by those observations currently assigned to that cluster:

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

©Emily Fox 2014

14

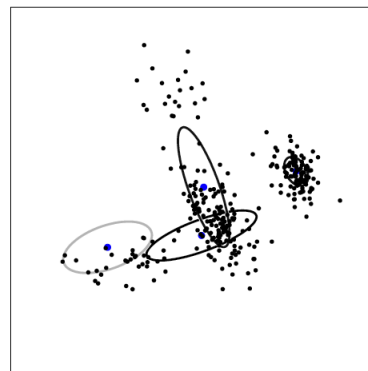
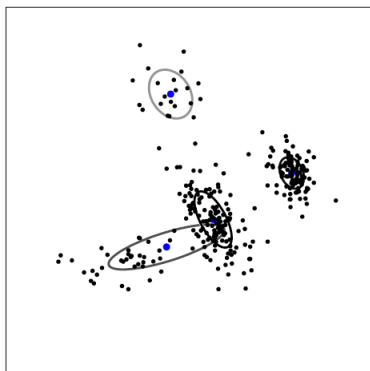
Standard Sampler: 2 Iterations



©Emily Fox 2014

15

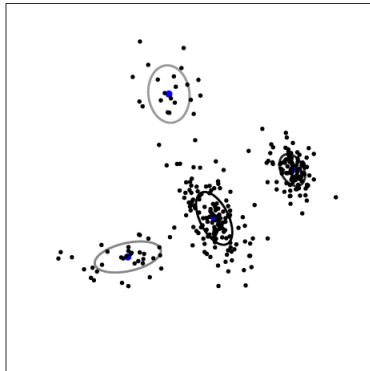
Standard Sampler: 10 Iterations



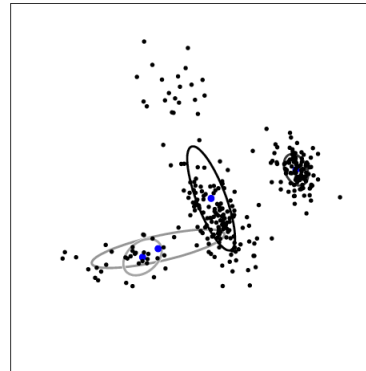
©Emily Fox 2014

16

Standard Sampler: 50 Iterations



$\log p(x | \pi, \theta) = -397.40$



$\log p(x | \pi, \theta) = -442.89$

©Emily Fox 2014

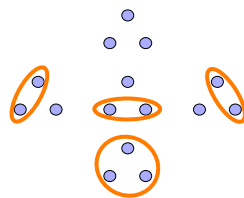
17

Mixtures Induce Partitions

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

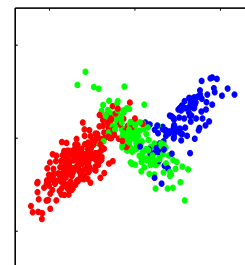
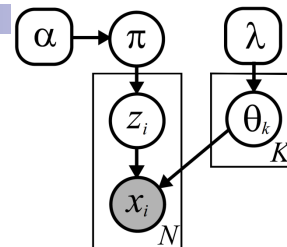
$$z_i \sim \pi$$

- 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 $3^3 = 27$

©Emily Fox 2014



18

Mixtures Induce Partitions

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

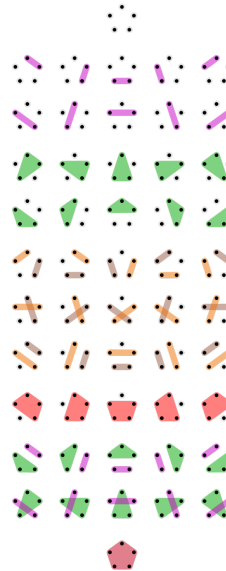
$$z_i \sim \pi$$

- 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 $5^5 = 3125$

©Emily Fox 2014



Courtesy
Wikipedia

Module 3: Bayesian Nonparametrics

Infinite Mixture Models

STAT/BIOSTAT 527, University of Washington

Emily Fox

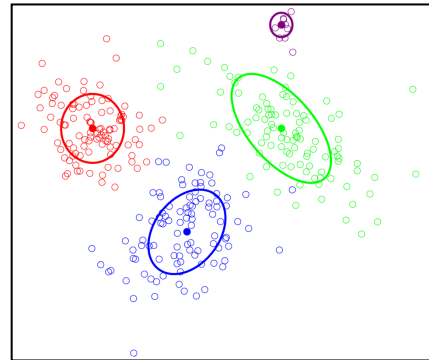
April 29th, 2014

©Emily Fox 2014

20

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



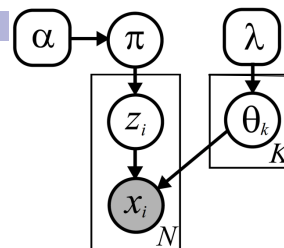
Mixture of Gaussians

θ_1 θ_2 θ_3 θ_4 θ_5 θ_6 θ_7 \dots

Nonparam. Model In Pictures

- Mixture weights

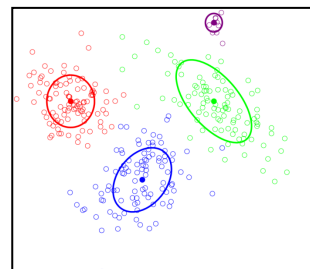
π



- For each observation, draw

$$z_i \sim \pi$$

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



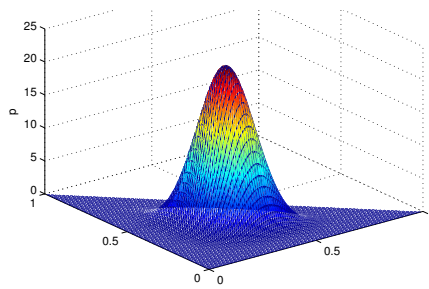
Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

$$\alpha_k = 10 \quad \forall k$$

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$$

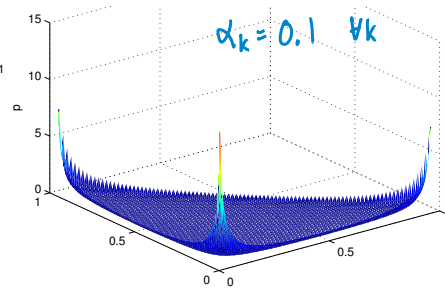
$$\Rightarrow \sum \pi_k = 1$$



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

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

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



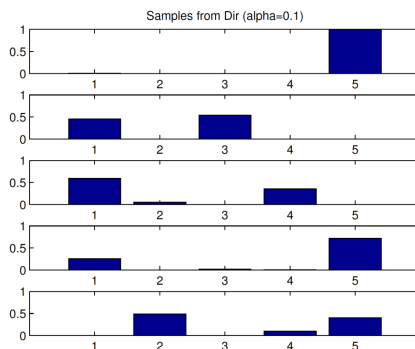
©Emily Fox 2014

23

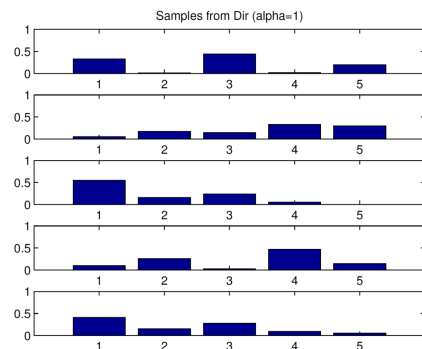
Dirichlet Samples

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

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



$\text{Dir}(\pi | 0.1, 0.1, 0.1, 0.1, 0.1)$
puts mass at corners

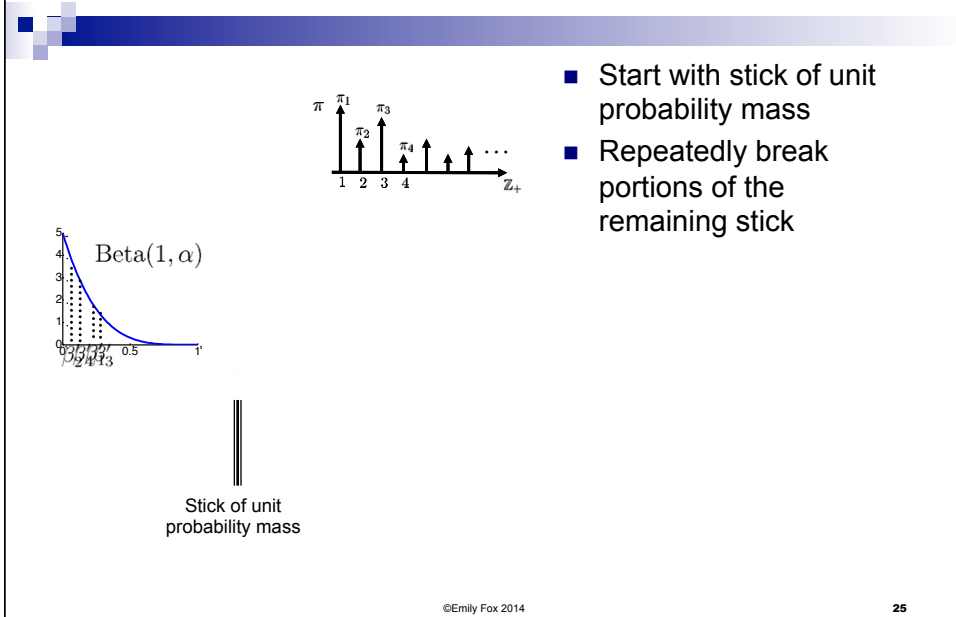


$\text{Dir}(\pi | 1.0, 1.0, 1.0, 1.0, 1.0)$
uniform

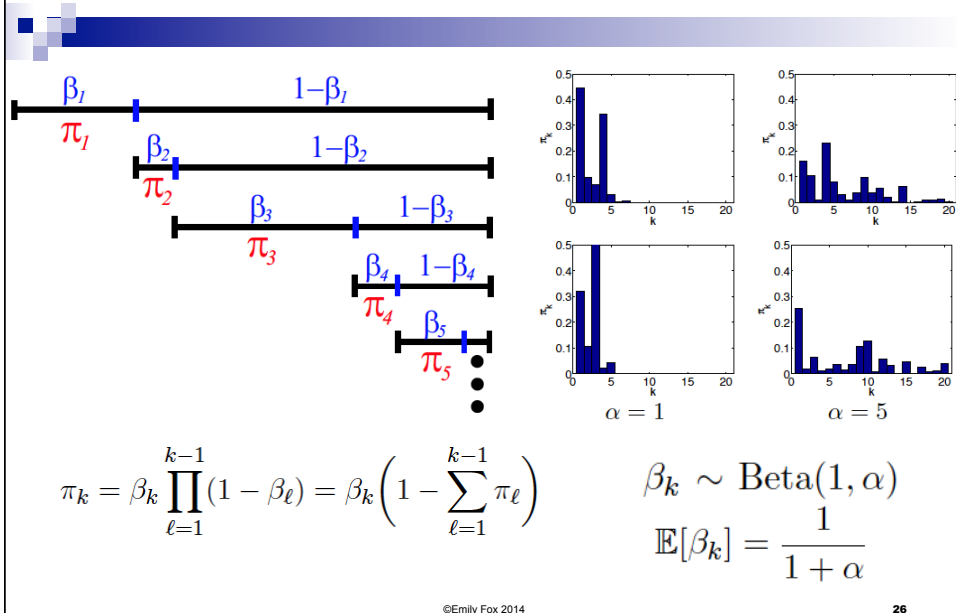
©Emily Fox 2014

24

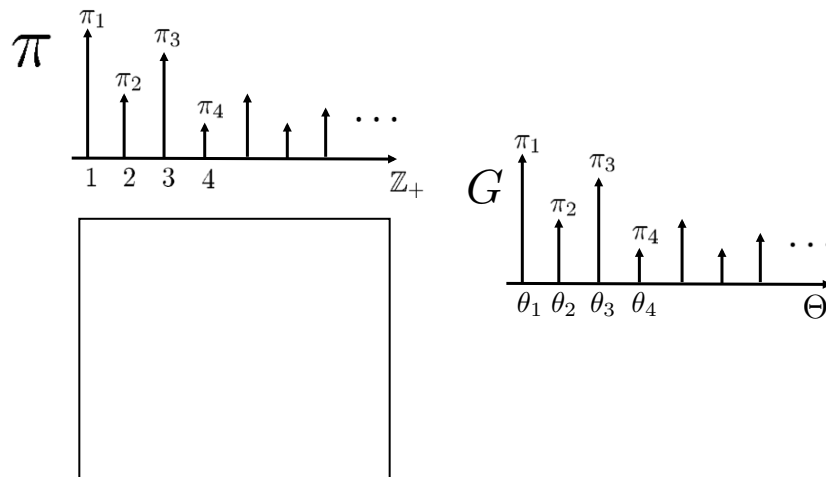
Stick-Breaking Process



Stick-Breaking Process Summary



Stick Breaks + Dirichlet Process



©Emily Fox 2014

27

Dirichlet Process Mixture Model

- Place Dirichlet process prior on weights and mixture parameters:

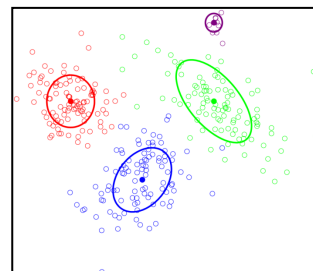
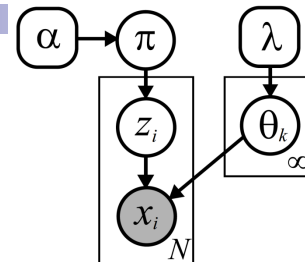
$$G \sim \text{DP}(\alpha, H)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad \begin{matrix} \pi \\ \theta_k \end{matrix}$$

- For each observation, draw

$$z_i \sim \pi$$

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



©Emily Fox 2014

28

Finite versus DP Mixtures

Finite Mixture

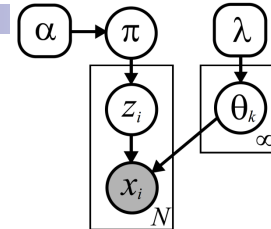
$$\pi \sim \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$z_i \sim \pi$$

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

DP Mixture

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



THEOREM: For any measurable function f , as $K \rightarrow \infty$

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

$$G^K(\theta) = \sum_{k=1}^K \pi_k \delta_{\theta_k}(\theta)$$

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

©Emily Fox 2014

29

Induced Partitions

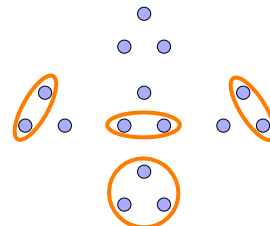
- Recall that mixture models induce partitions of the data

$$z_i \sim \pi$$

- For a given prior on mixture weights, some partitions are more likely than others apriori

- Example 1: $\pi \sim \text{Dir}(1, \dots, 1)$

- Example 2: $\pi \sim \text{Dir}(0.01, \dots, 0.01)$



©Emily Fox 2014

30

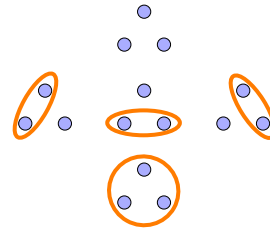
Induced Partitions

- 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, \dots, z_N ?

- Do we expect many unique clusters?

©Emily Fox 2014

31

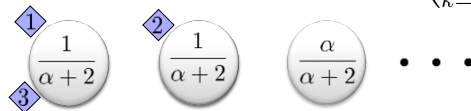
Chinese Restaurant Process (CRP)

- 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 \longleftrightarrow *observed data to be clustered*
tables \longleftrightarrow *distinct clusters*

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

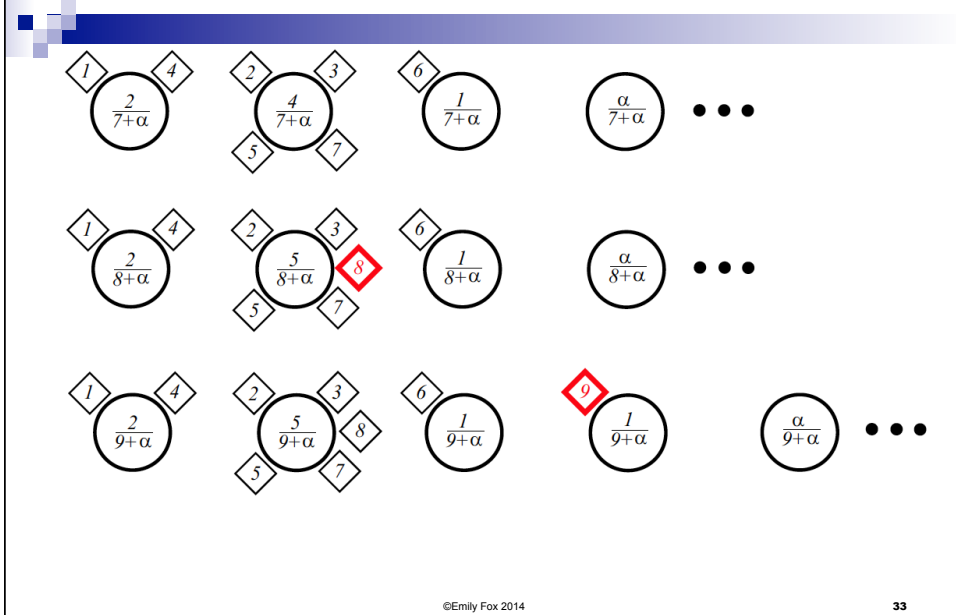
$$p(z_{N+1} = z \mid z_1, \dots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^K N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$



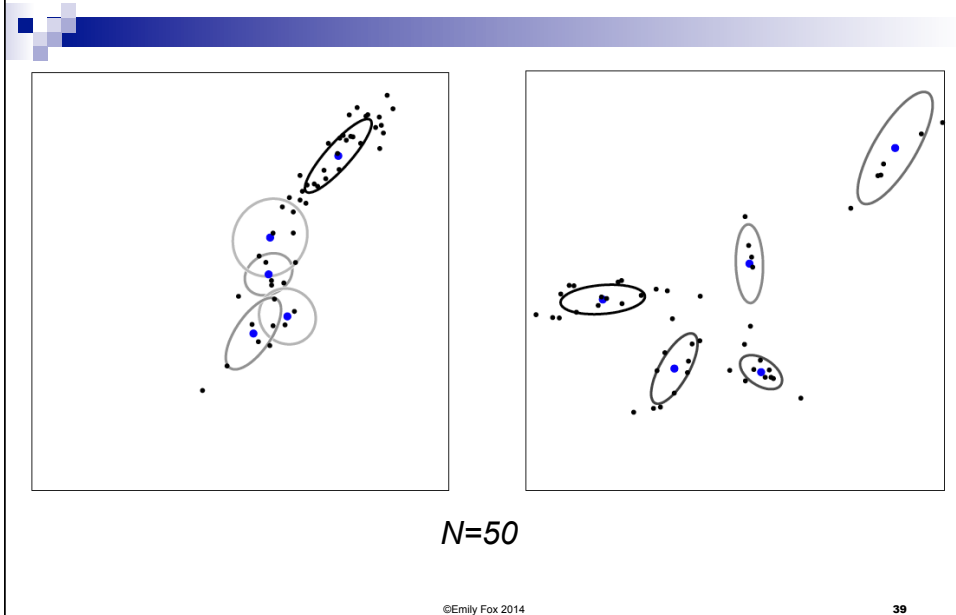
©Emily Fox 2014

32

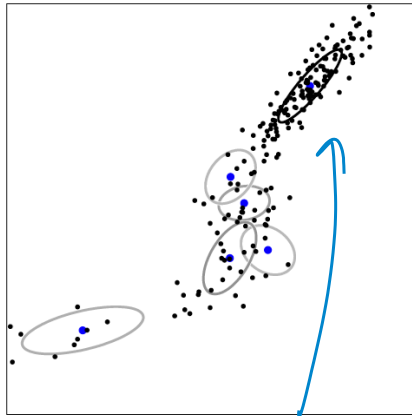
Chinese Restaurant Process (CRP)



Samples from DP Mixture Priors



Samples from DP Mixture Priors

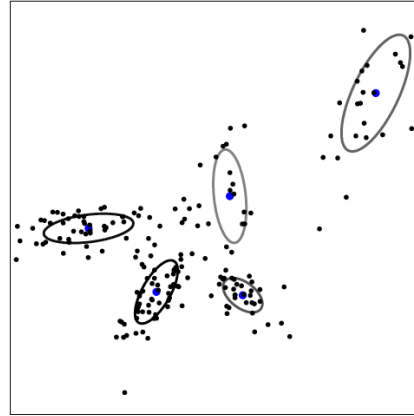


"rich
get
richer"

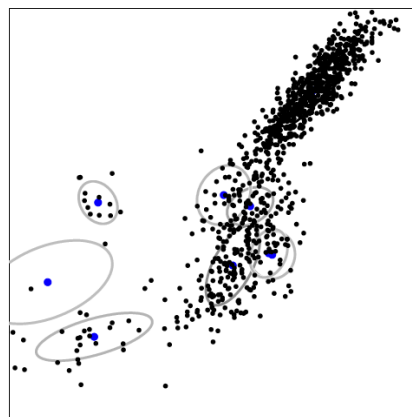
$N=200$

©Emily Fox 2014

40



Samples from DP Mixture Priors

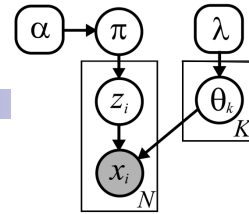


$N=1000$

©Emily Fox 2014

41

Finite GMM Sampler



Recall model

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

Generative model:

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad z_i \sim \pi$$

$$\{\mu_k, \Sigma_k\} \sim \text{NIW}(\lambda) \quad x_i | z_i, \{\theta_k\} \sim N(\mu_{z_i}, \Sigma_{z_i})$$

Iteratively sample

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

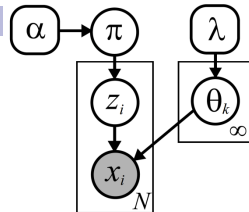
$$\pi | \{z_i\}, \{x_i\}$$

$$\theta_k | \{z_i\}, \{x_i\} \quad k=1, \dots, K$$

©Emily Fox 2014

42

Collapsed DP Mixture Sampler



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

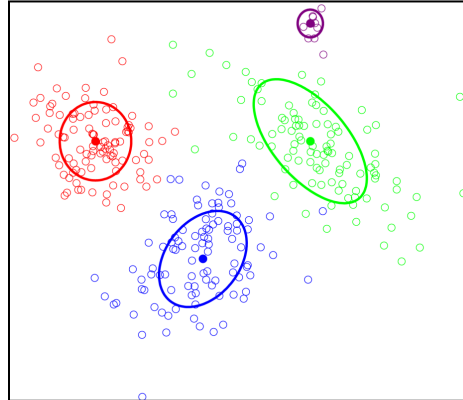
Iteratively sample the cluster indicators

©Emily Fox 2014

43

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 2014

44

Predictive Likelihood Term

- Recall NIW prior...Let's consider 1D example \rightarrow N-IG

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

- Normal inverse gamma posterior
 \rightarrow Student t predictive likelihood

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

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

- Conjugacy: This integral is **tractable**

©Emily Fox 2014

45

Collapsed DP Mixture Sampler

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \dots, N\}$.
2. Set $\alpha = \alpha^{(t-1)}$ and $z = z^{(t-1)}$. For each $i \in \{\tau(1), \dots, \tau(N)\}$, 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}

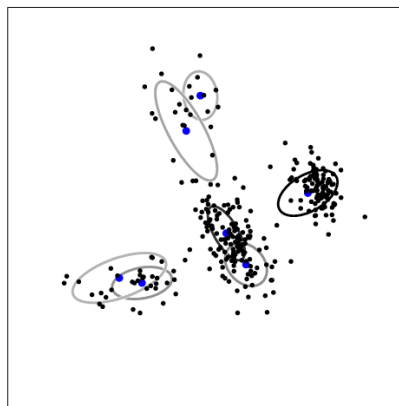
$$p(x_i \mid \lambda) = \int_{\Theta} f(x_i \mid \theta) h(\theta \mid \lambda) d\theta$$
 - (b) Sample a new cluster assignment z_i from the following $(K+1)$ -dim. multinomial:

$$z_i \sim \frac{1}{Z_i} \left(\alpha f_{\bar{k}}(x_i) \delta(z_i, \bar{k}) + \sum_{k=1}^K N_k^{-i} f_k(x_i) \delta(z_i, k) \right) \quad Z_i = \alpha f_{\bar{k}}(x_i) + \sum_{k=1}^K N_k^{-i} f_k(x_i)$$
 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.

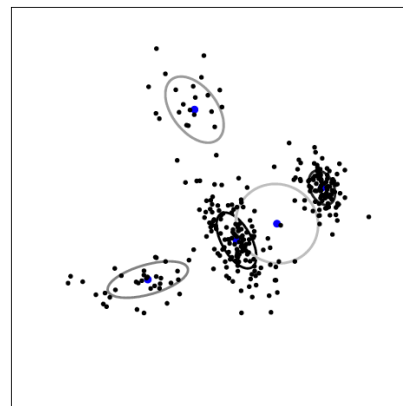
©Emily Fox 2014

46

Collapsed DP Sampler: 2 Iterations



$\log p(x \mid \pi, \theta) = -462.25$

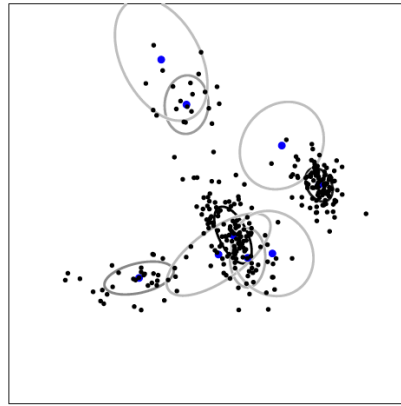


$\log p(x \mid \pi, \theta) = -399.82$

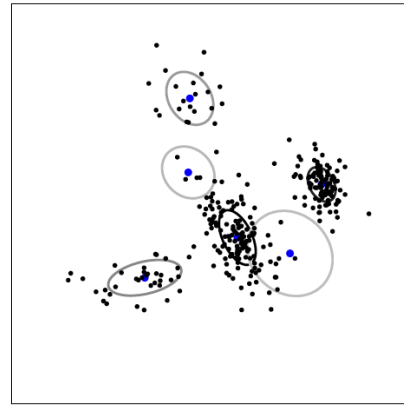
©Emily Fox 2014

47

Collapsed DP Sampler: 10 Iterations



$$\log p(x \mid \pi, \theta) = -398.32$$

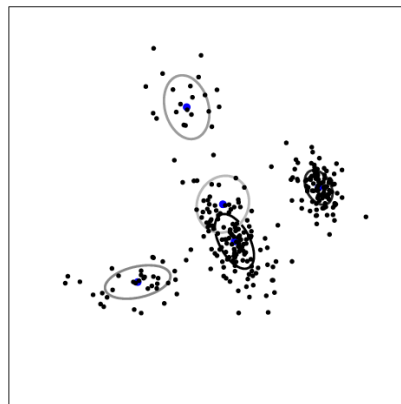


$$\log p(x \mid \pi, \theta) = -399.08$$

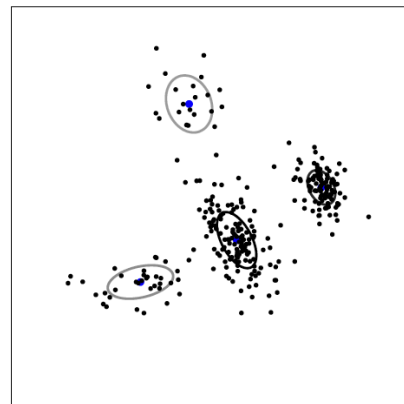
©Emily Fox 2014

48

Collapsed DP Sampler: 50 Iterations



$$\log p(x \mid \pi, \theta) = -397.67$$

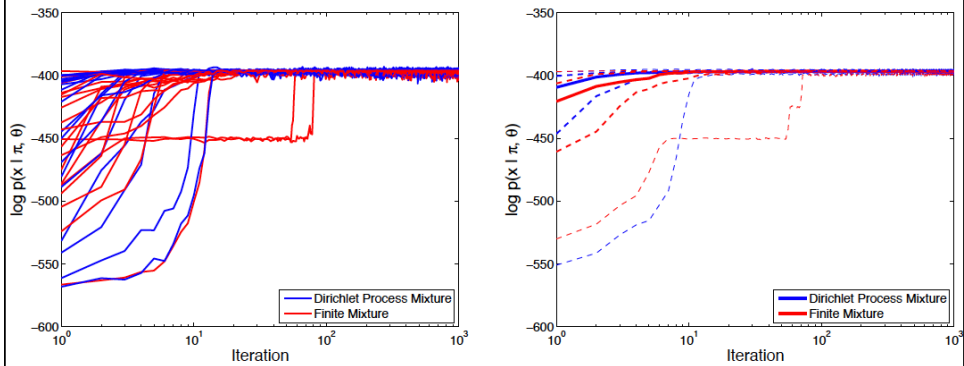


$$\log p(x \mid \pi, \theta) = -396.71$$

©Emily Fox 2014

49

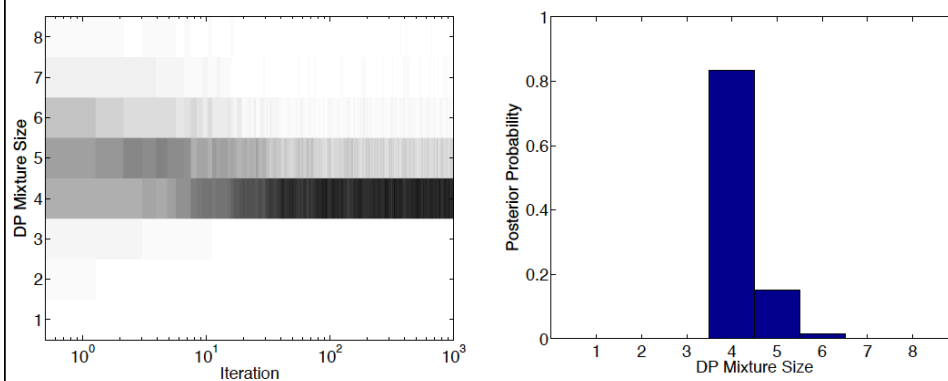
DP vs. Finite Mixture Samplers



©Emily Fox 2014

50

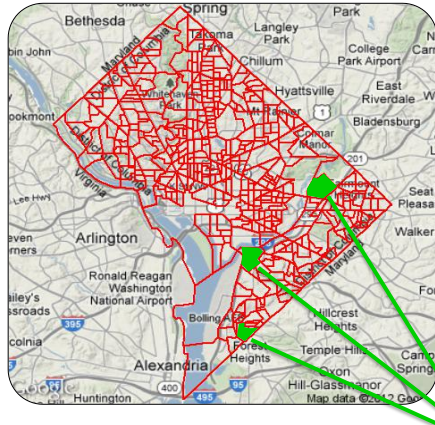
DP Posterior Number of Clusters



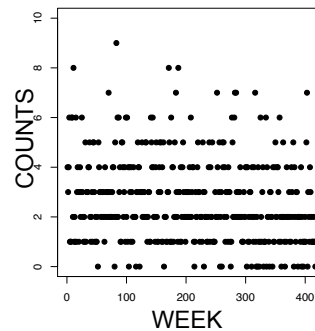
©Emily Fox 2014

51

DC Violent Crime Data



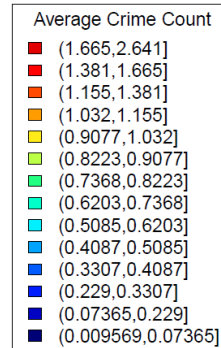
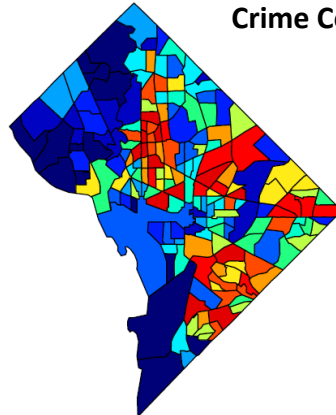
- 188 census tracts
- Weekly crime counts from 2001-2008
- Violent crime types:
 - ADW, arson, robbery, rape



Time series = crime counts

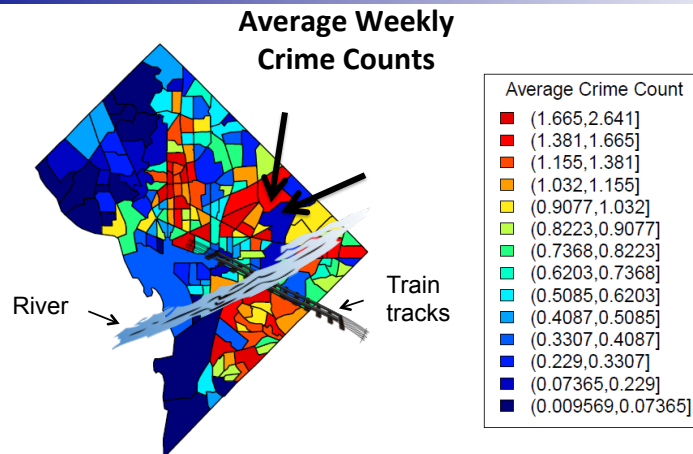
DC Violent Crime Data

Average Weekly
Crime Counts



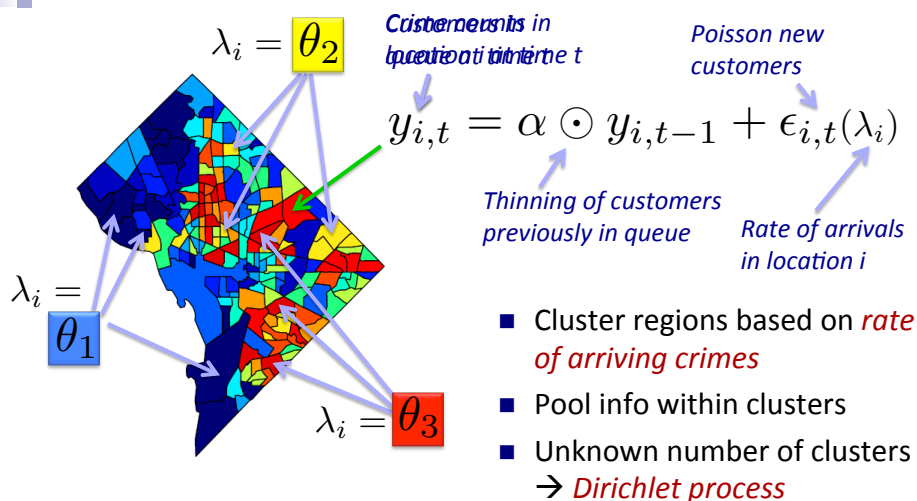
Goal: Forecast next week's map

DC Violent Crime Data



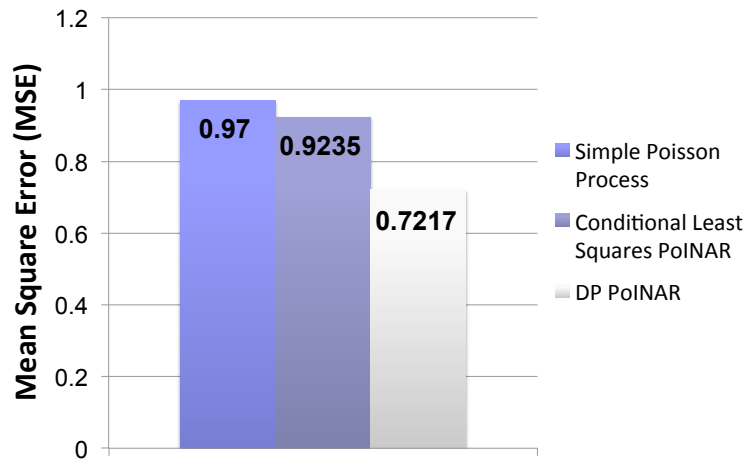
Similar behavior in spatially disjoint tracts
 → *Cluster census tracts*

Poisson Integer-Valued Autoregressions



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