Module 3: Bayesian Nonparametrics

Finite Mixture Models

for density estimation

STAT/BIOSTAT 527, University of Washington

Emily Fox

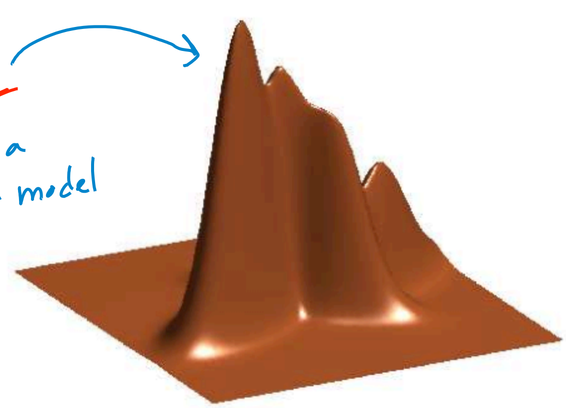
April 29th, 2014

Density Estimation

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

$x_1, \ldots, x_n \sim p$

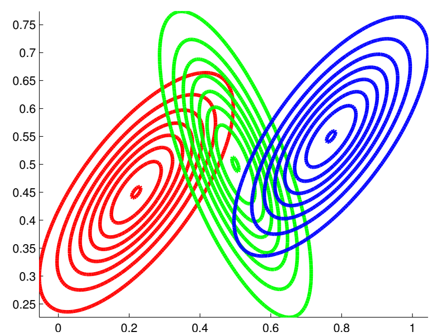
Let's consider a parametric model



Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



K=3

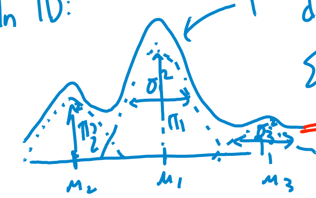
Gaussian kernels, but not centered at obs. like in KDE

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

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

# of mix comp.   mix. weights   shape params

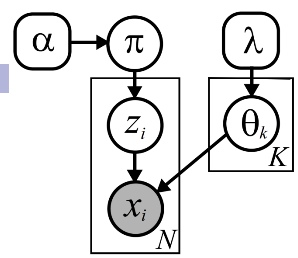
In 1D:



p = target density

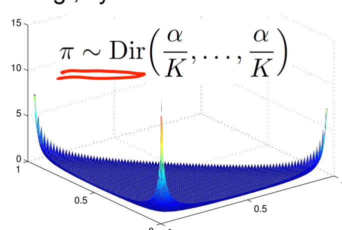
$\sum \pi_k = 1$

Model Summary

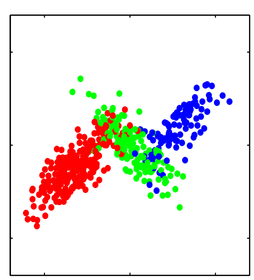


- Prior on model parameters
  - E.g., symmetric Dirichlet for π

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



  - Normal inverse Wishart prior for $\theta_k = \{\mu_k, \Sigma_k\}$



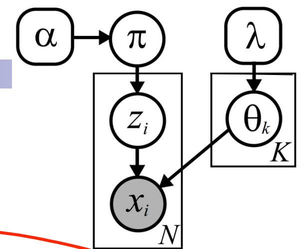
$\{1, \ldots, K\}$

- Sample observations as

$$z_i \sim \pi \quad \text{choose a cluster}$$

$$x_i \mid z_i \sim N(\mu_{z_i}, \Sigma_{z_i}) \quad \text{sim obs. from selected Gauss.}$$

Model In Pictures



- Mixture weights

π

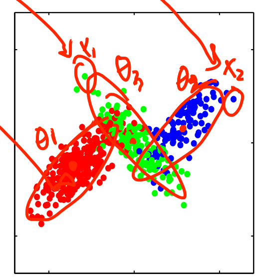
$z_2 = 2$

$z_1 = 3$

- For each observation,

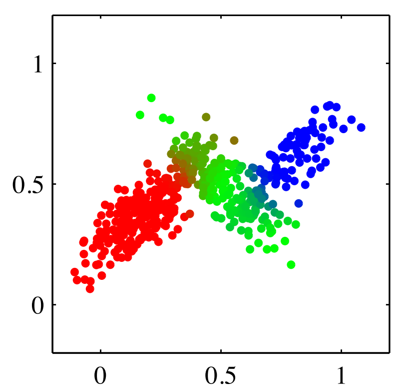
$$z_i \sim \pi$$

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



Clustering our Observations

- We must infer the cluster assignments from the observations



Soft assignments to clusters

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

$$r_{ik} = p(z_i = k \mid x_i, \pi, \theta) =$$

$$= \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

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

$$\hat{\theta}^{MAP} = \arg\max_{\theta} p(\theta \mid x)$$ 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

posterior

$$p(x_{N+1} \mid x_1, \ldots, x_N) = \int p(x_{N+1} \mid \theta)\, p(\theta \mid x_1, \ldots, x_N)\, d\theta$$

- Often no closed-form characterization (e.g., mixture models)
- Alternatives:
  - ☆ Markov chain Monte Carlo (MCMC) providing samples from posterior
  - Variational approximations to posterior

Gibb Sampling

- Let z $= (z_1, \ldots, z_n)$ indicate the set of ***all variables in the model***: e.g., cluster indicators and parameters
- Want draws:

$$(z_1, \ldots, z_n) \sim \pi(z)$$

can think of $\pi(z) = p(\theta \mid x)$ (desired posterior)

issue: can't directly sample

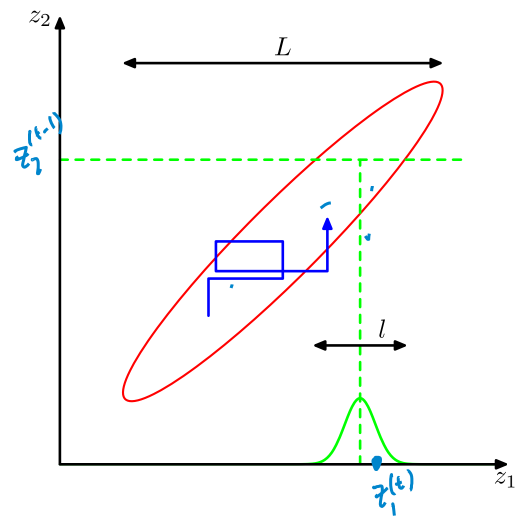
- Construct Markov chain whose steady state distribution is $\pi(z)$
- Simplest case:

for $t = 1, \ldots, N_{iter}$

for $i = 1, \ldots, n$

$$z_i^{(t)} \sim \underbrace{p(z_i \mid z_1^{(t)}, \ldots, z_{i-1}^{(t)}, z_{i+1}^{(t-1)}, \ldots, z_n^{(t-1)})}_{\text{Gibb's sampling assumes that this has closed form + can sample}}$$

Gibbs Sampler for a 2D Gaussian



General Gibbs Sampler

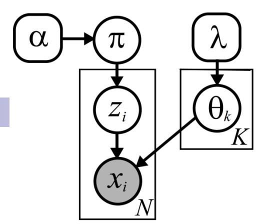
$$z_i^{(t)} \sim p(z_i \mid z_{\backslash i}^{(t-1)}) \qquad i = i(t)$$

$$z_j^{(t)} = z_j^{(t-1)} \qquad j \neq i(t)$$

Under mild conditions, converges assuming all variables are resampled infinitely often (order can be fixed or random)

C. Bishop, Pattern Recognition & Machine Learning

Example – GMM



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

  want these

  - 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

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

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

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

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

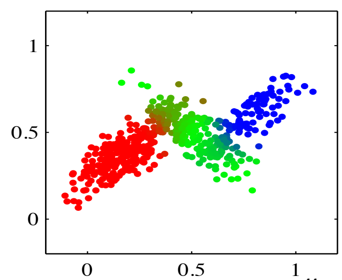
$$r_{ik} = \underbrace{p(z_i = k \mid x_i, \pi, \theta)}_{\text{desired complete cond.}} = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{\ell=1}^{K} \pi_\ell p(x_i \mid \theta_\ell)}$$

$\leftarrow N(x_i \mid \mu_k, \Sigma_k)$

- Sample each cluster indicator as

$z_i \sim r_i \quad i = 1, \ldots, N$

$r_i = [r_{i1}, \ldots, r_{iK}]$



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

- Recall conjugate Dirichlet prior

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

- Dirichlet posterior
  - Assume we condition on cluster indicators $z_i \sim \pi$
  - Count occurrences of $z_i = k$ $\quad N_k = |\{z_i : z_i = k\}|$
  - Then,

$$p(\pi \mid \alpha, z_1, \ldots, z_N) \propto \prod_i p(z_i \mid \pi) p(\pi \mid \alpha)$$
$$\propto \prod_k \prod_{i: z_i = k} \pi_k \cdot \pi_k^{\alpha_k - 1} \propto \prod_k \pi_k^{N_k + \alpha_k - 1}$$
$$= \mathrm{Dir}(N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

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

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

- 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
  - 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,

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

only summing over obs. generated from k^{th} cluster

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

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

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:

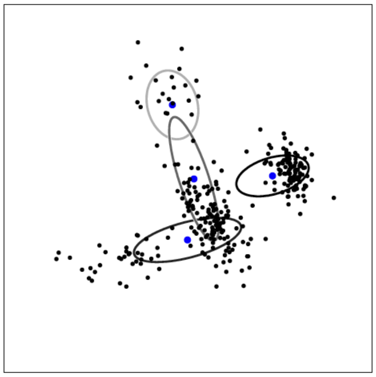
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^N \delta(z_i^{(t)}, k)$$

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

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

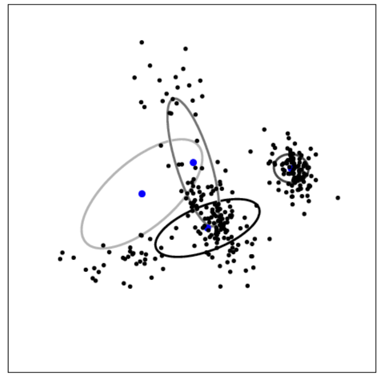
Standard Sampler: 2 Iterations

random init #1



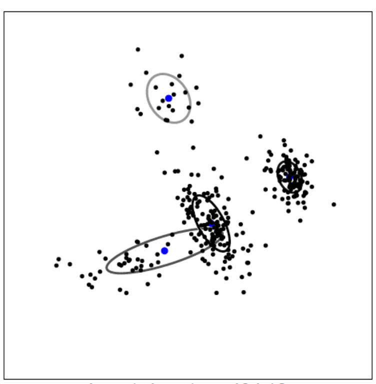
log p(x | π, θ) = –539.17

random init #2



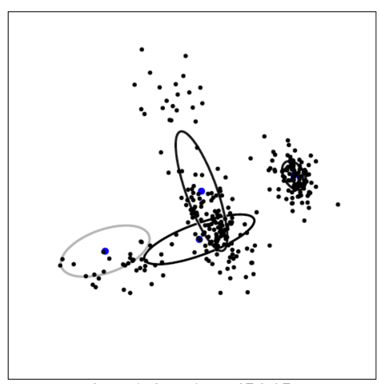
log p(x | π, θ) = –497.77

Standard Sampler: 10 Iterations



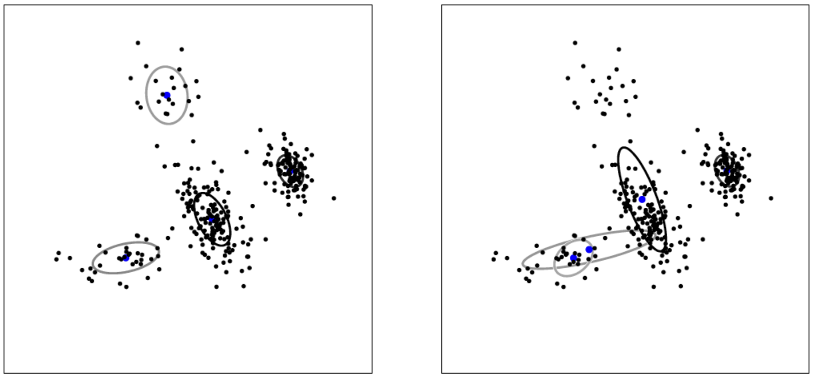
log p(x | π, θ) = –404.18

better



log p(x | π, θ) = –454.15

Standard Sampler: 50 Iterations



log p(x | π, θ) = –397.40   log p(x | π, θ) = –442.89

can get stuck for long time...

Mixtures Induce Partitions

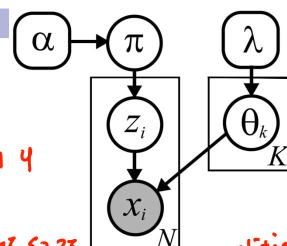
- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

$$z_i \sim \pi$$

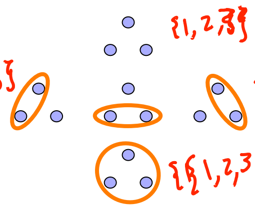
$z_1, \ldots, z_N$

1 5 5 3 2 1 1 4

{{1,6,7},{2,3},{4,5,8}} partition



- 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:

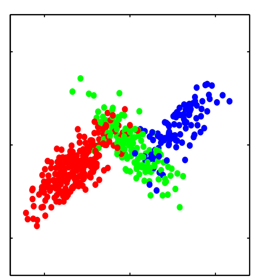


{1,2,3}

{{1,2},{3}}

{1,{2,3}}

{{1,2,3}}



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

Mixtures Induce Partitions

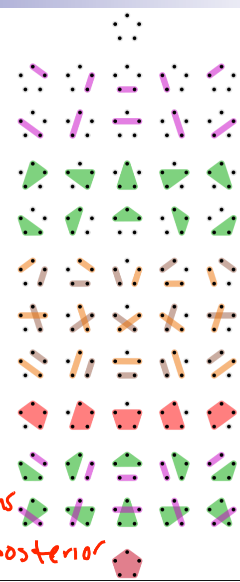
- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

$$z_i \sim \pi$$

- 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.

Note: sampler can switch between eq. partitions.



Courtesy Wikipedia

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

If prior sym., then eq. under posterior

Module 3: Bayesian Nonparametrics

Infinite Mixture Models

going infinite...

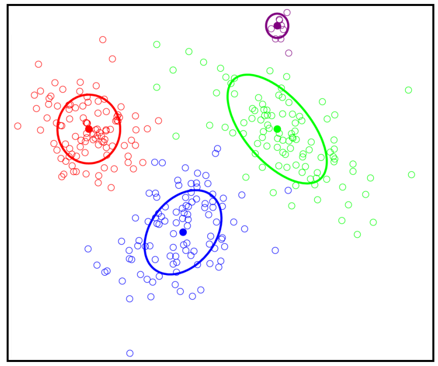
STAT/BIOSTAT 527, University of Washington

Emily Fox

April 29th, 2014

Motivating Nonparametric GMM

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

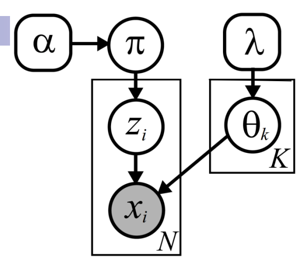


Mixture of Gaussians

allows us to add in new model comp.

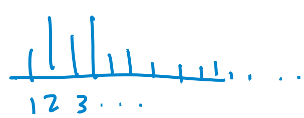
$\theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4 \quad \theta_5 \quad \theta_6 \quad \theta_7 \quad \cdots$

Nonparam. Model In Pictures



- Mixture weights

π

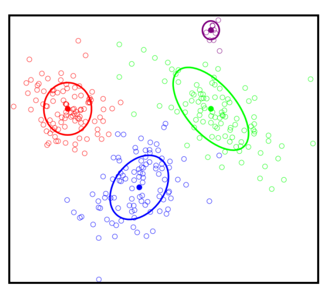


$\{\theta_k\}$ $k=1,2,\ldots$

- For each observation, draw

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

How to define π?



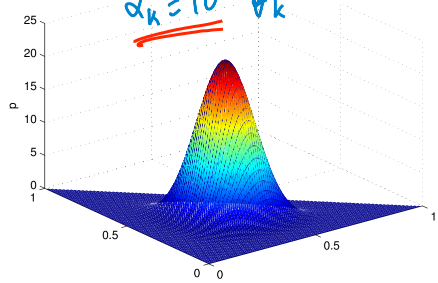
Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

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

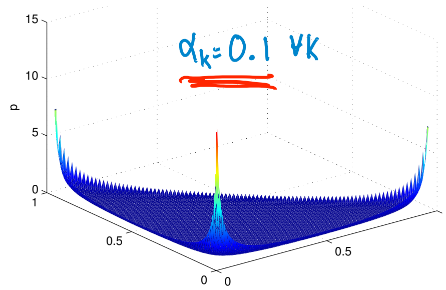
$\Rightarrow \sum \pi_k = 1$

$\alpha_k = 10 \;\; \forall k$



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

$\alpha_k = 0.1 \;\; \forall k$



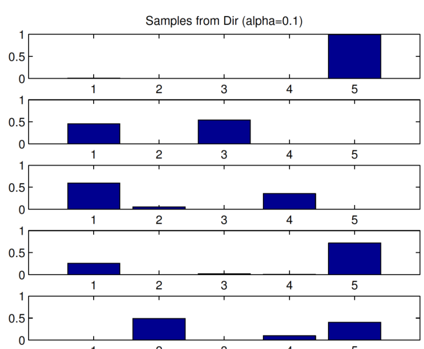
Moments: $\mathbb{E}_\alpha[\pi_k] = \dfrac{\alpha_k}{\alpha_0}$

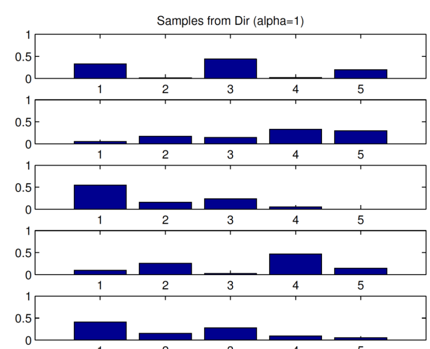
$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)}$$

Dirichlet Samples

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

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

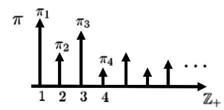


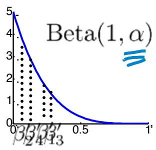


$\mathrm{Dir}(\pi \mid 0.1, 0.1, 0.1, 0.1, 0.1)$ puts mass @ corners

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

Stick-Breaking Process





Stick of unit probability mass

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

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

$\pi_1 = \beta_1$

$\pi_2 = \beta_2(1-\beta_1)$

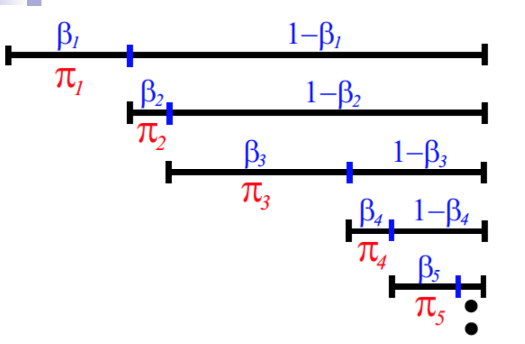
$\vdots$

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

$\vdots$

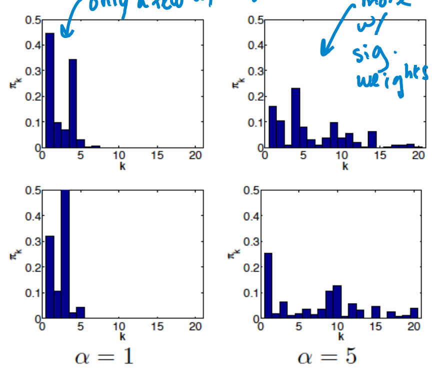
$\boxed{\pi \sim \text{Stick}(\alpha)}$

Stick-Breaking Process Summary



only a few w/ large weights

more w/ sig. weights



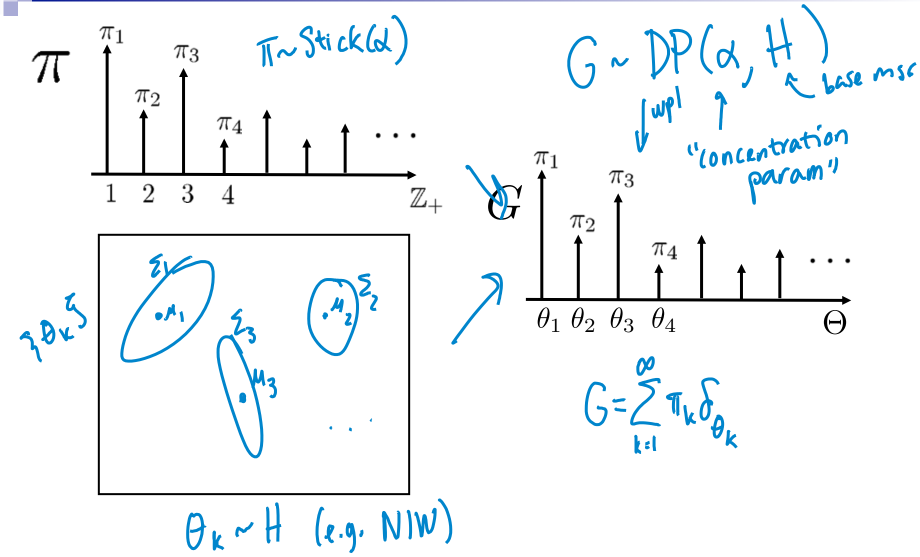
$$\pi_k = \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) = \beta_k\left(1 - \sum_{\ell=1}^{k-1}\pi_\ell\right)$$

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

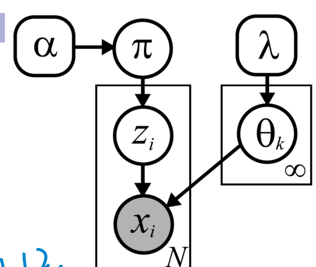
$$\mathbb{E}[\beta_k] = \frac{1}{1+\alpha}$$

For large α, breaking small portions

Stick Breaks + Dirichlet Process



Dirichlet Process Mixture Model



- Place Dirichlet process prior on weights and mixture parameters:

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

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

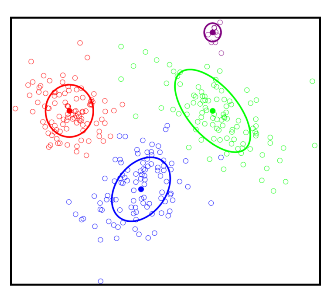
ind. $\pi \sim \text{stick}(\alpha)$

$\theta_k \overset{iid}{\sim} H$ (e.g. NIW) $k=1,2,\ldots$

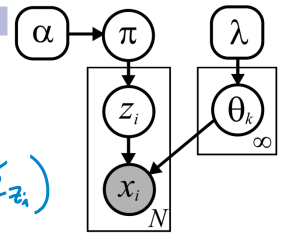
- For each observation, draw

$$z_i \sim \pi$$

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



Finite versus DP Mixtures



Finite Mixture

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

sym.

DP Mixture

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

$$z_i \sim \pi$$

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

e.g. $N(\mu_{z_i}, \Sigma_{z_i})$

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

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

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

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

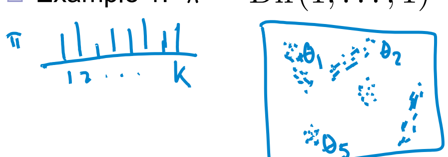
$$G \sim \mathrm{DP}(\alpha, H)$$

Induced Partitions

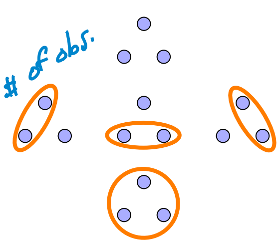
- Recall that mixture models induce partitions of the data

$$z_i \sim \pi$$

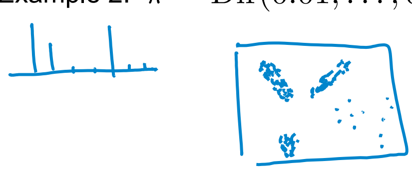
- For a given prior on mixture weights, some partitions are more likely than others apriori
  - Example 1: $\pi \sim \mathrm{Dir}(1, \ldots, 1)$ uniform



expect roughly the same # of obs. in each cluster



  - Example 2: $\pi \sim \mathrm{Dir}(0.01, \ldots, 0.01)$

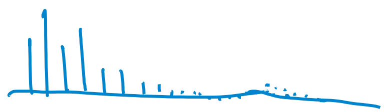


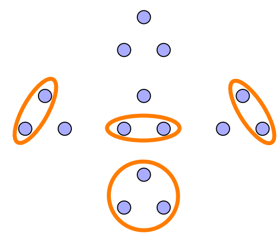
Induced Partitions

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





only a few w/ sig. mass ∞-ly many tiny weights

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

Answer: Chinese restaurant process

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:

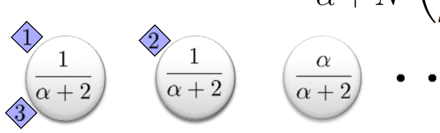
customers	⟷	*observed data to be clustered* x_i
tables	⟷	*distinct clusters* each serving unique dish θ_k

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

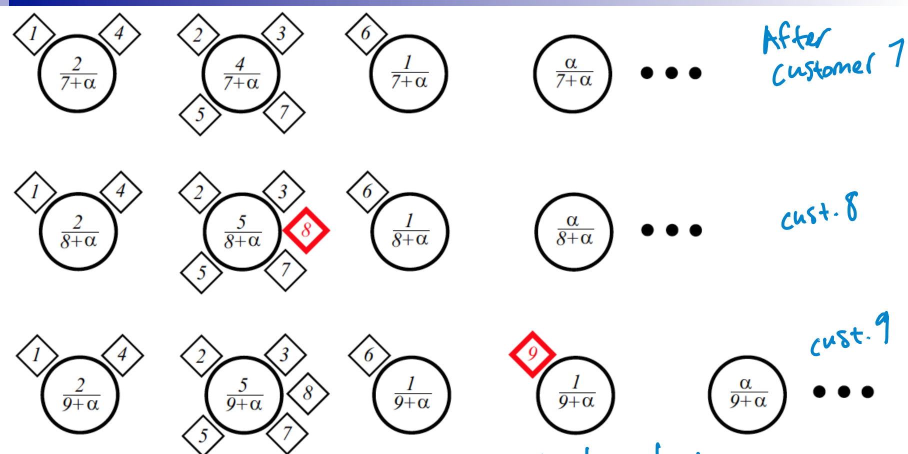
norm.   # of unique clusters in $z_{1:N}$

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

z_i already at table k   weight on new cluster   new cluster index

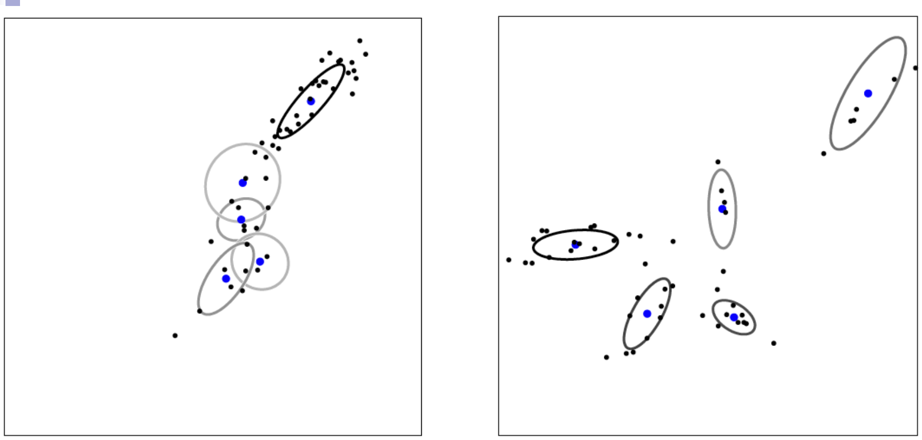


Chinese Restaurant Process (CRP)



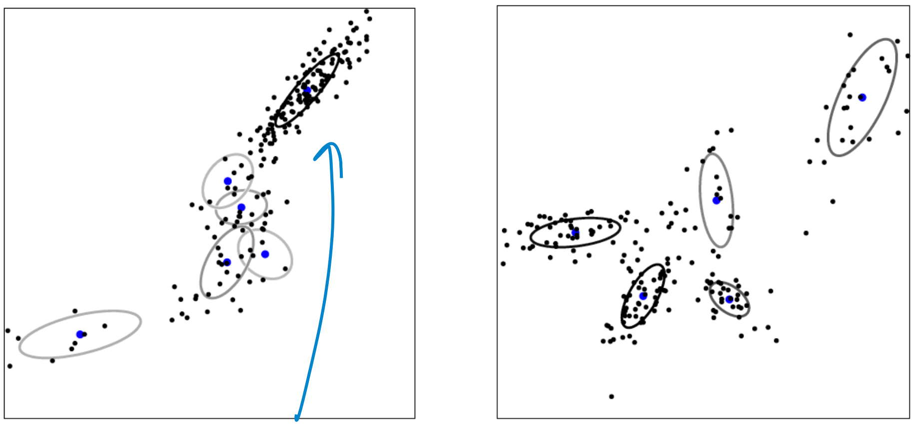
clustering/reinforcement induced by DP
"rich get richer"

Samples from DP Mixture Priors



N=50

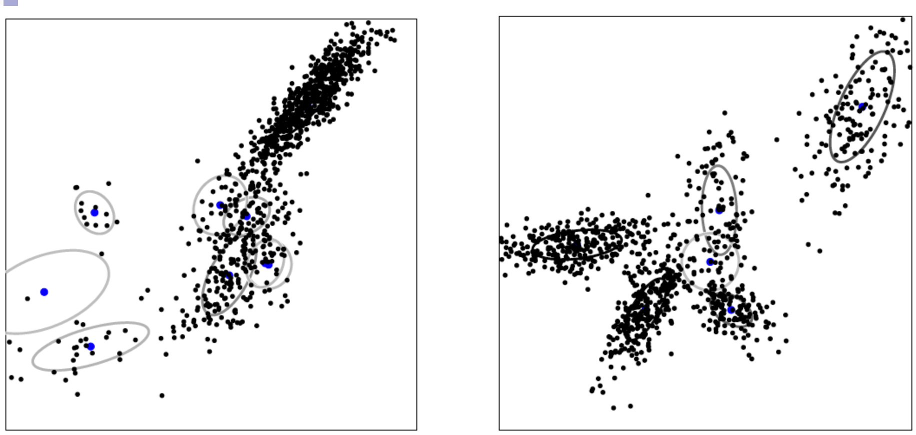
Samples from DP Mixture Priors



"rich get richer"

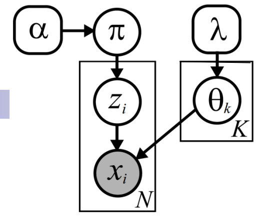
N=200

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:

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

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

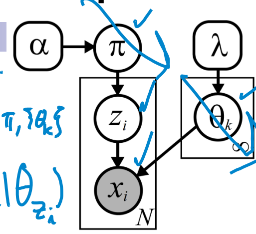
- Iteratively sample

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

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

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

Collapsed DP Mixture Sampler



- Can't sample π directly
- Integrate out all infinite-dimensional params $\pi, \{\theta_k\}$

$$p(z_{1:N}, x_{1:N}) = \int_{\pi} \int_{\theta_1, \theta_2, \dots} \underline{p(\pi \mid \alpha)} \prod_{k=1}^{\infty} p(\theta_k \mid \lambda) \prod_{i=1}^{N} \underline{p(z_i \mid \pi)}\, p(x_i \mid \theta_{z_i})$$

$$= \underbrace{p(z_1, \dots, z_N \mid \alpha)}_{\text{CRP}} \prod_{k=1}^{\infty} \underbrace{p(\{x_i : z_i = k\} \mid \lambda)}_{\text{"likelihood"}}$$

"likelihood"

- Iteratively sample the cluster indicators

treat as last obs. in CRP

$$z_i^{(t)} \sim \underbrace{p(z_i = k \mid z_{\setminus i}^{(t-1)}, \alpha)}_{\text{"prior"}}\, \overbrace{p(x_i \mid \{x_j : z_j = k, i \neq j\})}$$

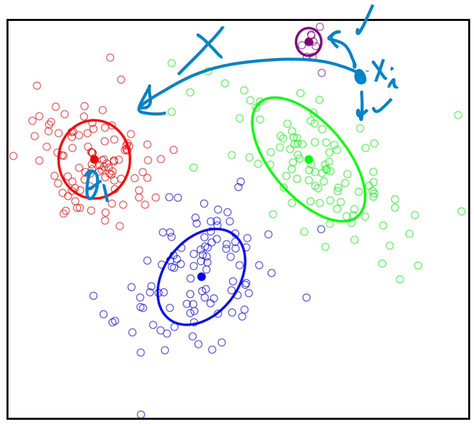
"prior" all other indicators all other obs. assigned to k^{th} cluster

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

Approx π by CRP
→ "prior" is based on cluster occupancy

Approx θ_k
→ "likelihood" based on obs. already assigned to cluster



Predictive Likelihood Term

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

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

- Normal inverse gamma posterior
  → Student t predictive likelihood

$$p(x_i \mid \{x_j : z_j = k, j \neq i\}) = \int \underbrace{p(x_i \mid \theta_k)}_{\text{normal likelihood}} \underbrace{p(\theta_k \mid \{x_j : z_j = k, j \neq i\})}_{\text{NIW posterior}} \, d\theta_k$$

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

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

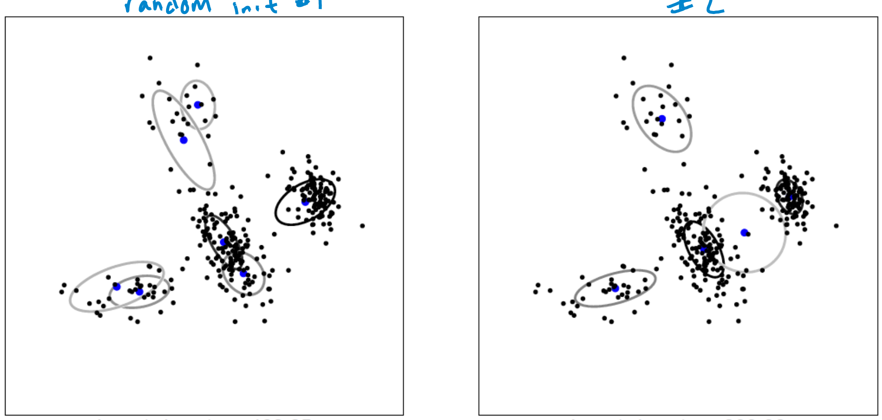
- Conjugacy: This integral is **tractable**

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.

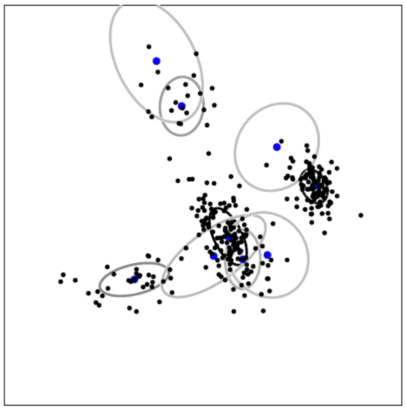
Collapsed DP Sampler: 2 Iterations

random init #1 #2

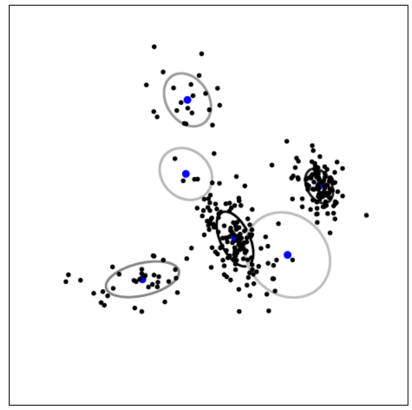


log p(x | π, θ) = –462.25 log p(x | π, θ) = –399.82

Collapsed DP Sampler: 10 Iterations

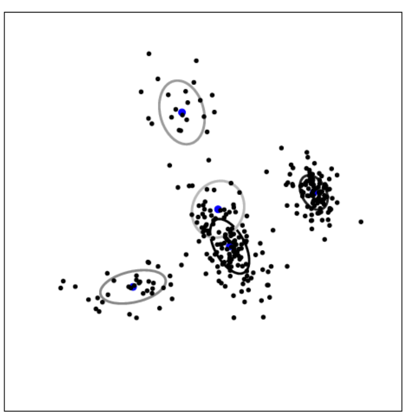


log p(x | π, θ) = –398.32

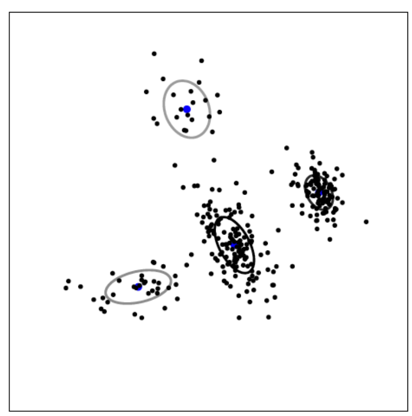


log p(x | π, θ) = –399.08

Collapsed DP Sampler: 50 Iterations

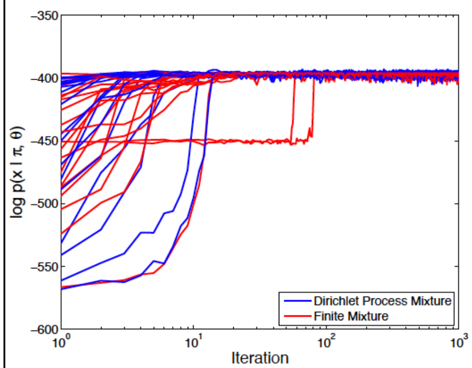


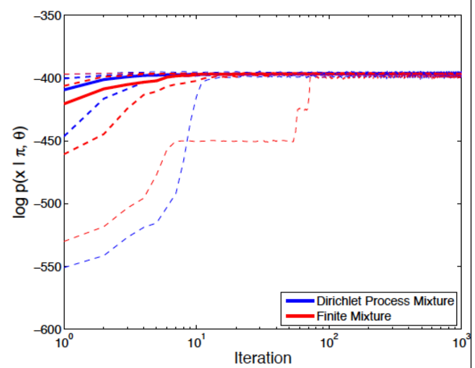
log p(x | π, θ) = –397.67



log p(x | π, θ) = –396.71

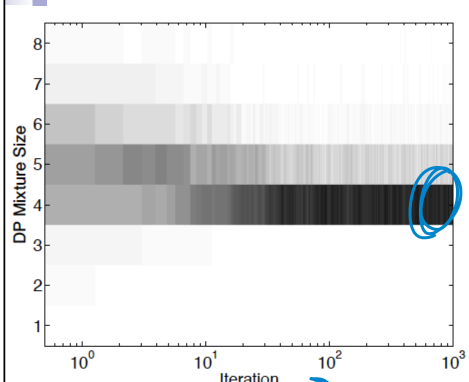
DP vs. Finite Mixture Samplers

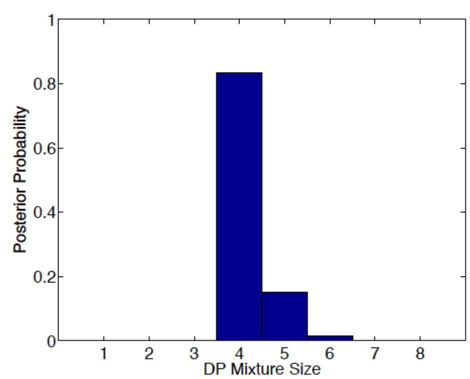




pretty competitive

DP Posterior Number of Clusters

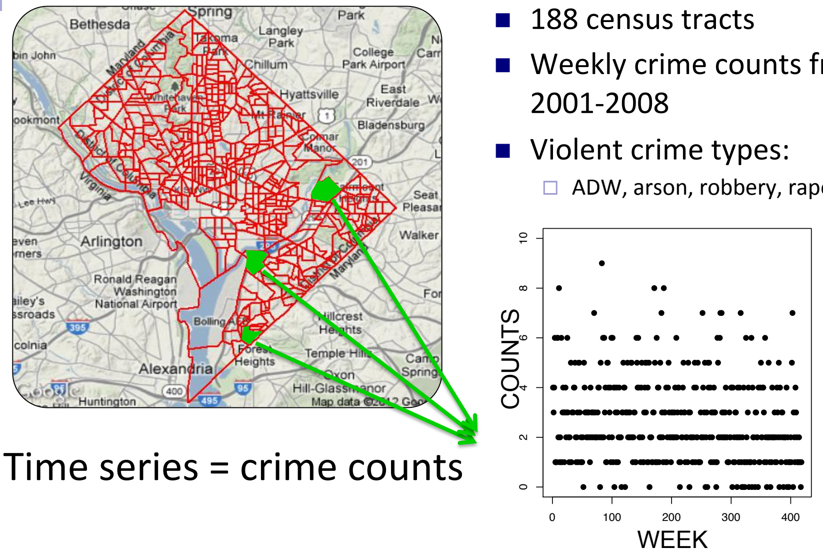




Asy. consistency results for density est. (assuming light tails on target density)

Not asy. consistent on # of comp.

DC Violent Crime Data

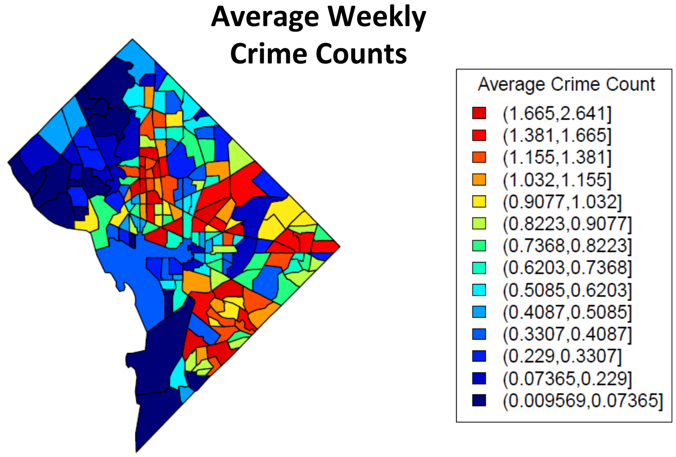


- 188 census tracts
- Weekly crime counts from 2001-2008
- Violent crime types:
  - ADW, arson, robbery, rape

Time series = crime counts

DC Violent Crime Data

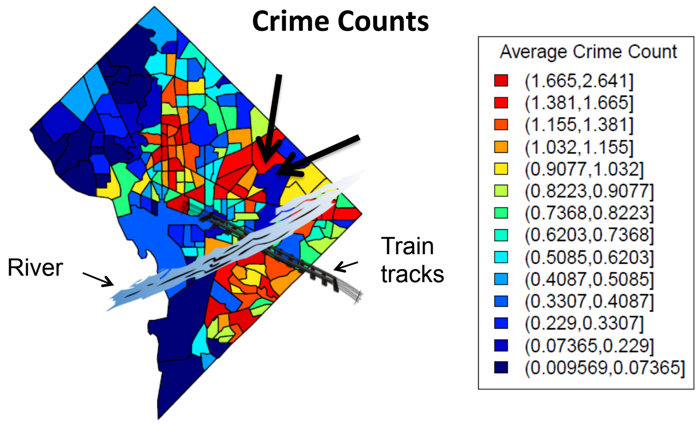
Average Weekly Crime Counts



Goal: Forecast next week's map

DC Violent Crime Data

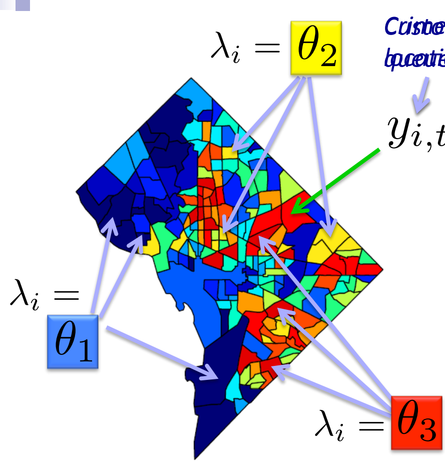
Average Weekly Crime Counts



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

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

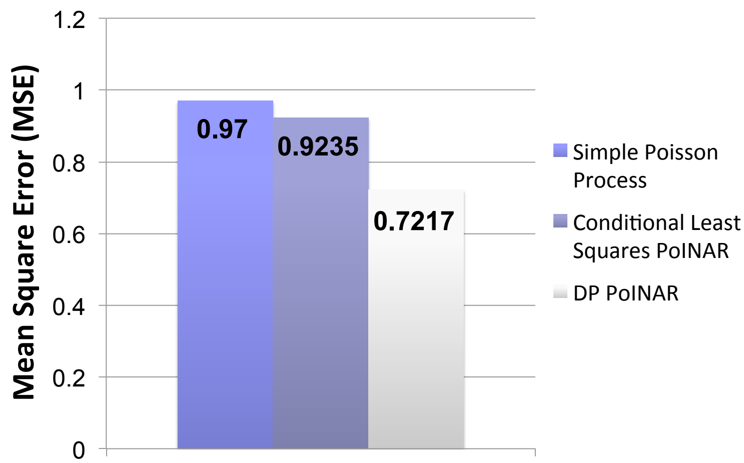
Thinning of customers previously in queue

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

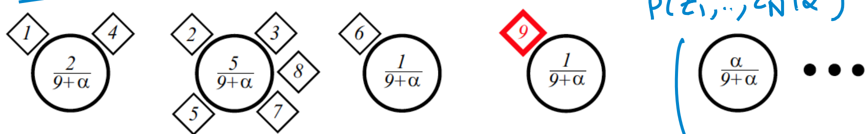
Slides based on parts of the lecture notes of Erik Sudderth for "Applied Bayesian Nonparametrics" at Brown University

CRPs & Exchangeable Partitions

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

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

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



$p(z_1)p(z_2 \mid z_1) \cdots p(z_N \mid z_{1:N-1})$

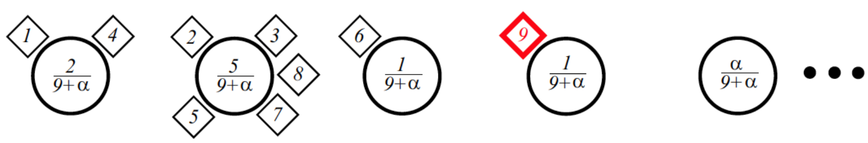
- Denominator terms:

$$1 \cdot \underset{\substack{\uparrow \\ 2^{nd}\text{ cust.}}}{\frac{1}{1+\alpha}} \cdot \underset{\substack{\uparrow \\ 3^{rd}\text{ cust.}}}{\frac{1}{2+\alpha}} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

CRPs & Exchangeable Partitions

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

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

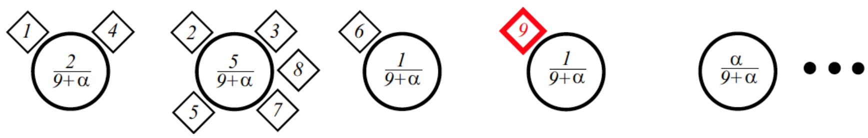


- Denominator terms: $\frac{1}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$
- Number of new tables: K
  Numerator term for each new table: α
  Combined: α^K

CRPs & Exchangeable Partitions

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

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



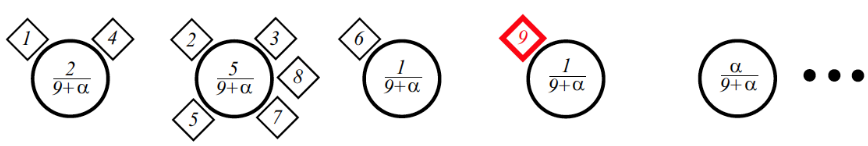
  - Denominator terms: $\frac{1}{1+\alpha}\cdot\frac{1}{2+\alpha}\cdots\frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$
  - New table numerator terms: α^K
  - Customers joining k^{th} occupied table: $1\cdot 2\cdots(N_k - 1) = (N_k-1)! = \Gamma(N_k)$

↑ already 1 person sitting @ k^{th} table

CRPs & Exchangeable Partitions

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

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



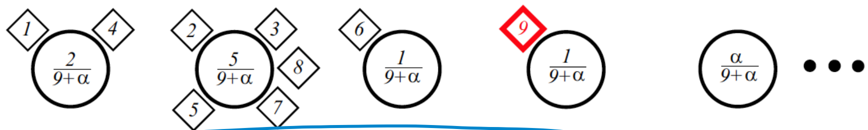
  - Denominator terms: $\frac{1}{1+\alpha}\cdot\frac{1}{2+\alpha}\cdots\frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$
  - New table numerator terms: α^K
  - Customers joining k^{th} occupied table:

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

CRPs & Exchangeable Partitions

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

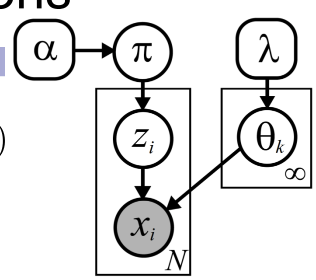
- The probability of a seating arrangement of *N* customers is *independent* of the order they enter the restaurant:



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

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

Finite versus CRP Partitions



Finite Mixture	*DP Mixture*
$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$	$\pi \sim \mathrm{Stick}(\alpha)$

$$z_i \sim \mathrm{Cat}(\pi)$$

$K_+ \longrightarrow$ *number of blocks in cluster*

Chinese Restaurant Process:

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

Finite Dirichlet:

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

- Probability of Dirichlet *indicators* approach zero as $K \to \infty$
- Probability of Dirichlet *partition* approaches CRP as