15. Clustering II

Motivation

- When we talked about clustering, we discussed two main approaches: k-means and Expectation-Maximization
- Both algorithms required the number K of clusters
- To find a good K, one could try different values for K and decide which is the best on some criterion

Questions:

- is there a more sound (i.e. statistically principled) way to find the number of clusters?
- can we do clustering and estimating of K online?

Motivation

- When we talked about clustering, we discussed two main approaches: k-means and Expectation-Maximization
- Both algorithms required the number K of clusters
- To find a good K, one could try different values for K and decide which is the best on some criterion

Questions:

- is there a more sound (i.e. statistically principled) way to find the number of clusters?
- can we do clustering and estimating of K online?

First step: derive a new algorithm for given (fixed) K

Gibbs Sampling (Rep.)

- Initialize $\{z_i : i = 1, \ldots, M\}$
- For $\tau = 1, \ldots, T$
  - Sample $z_1^{(\tau+1)} \sim p(z_1 \mid z_2^{(\tau)}, \ldots, z_M^{(\tau)})$
  - Sample $z_2^{(\tau+1)} \sim p(z_2 \mid z_1^{(\tau+1)}, \ldots, z_M^{(\tau)})$
  - ...
  - Sample $z_M^{(\tau+1)} \sim p(z_M \mid z_1^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$

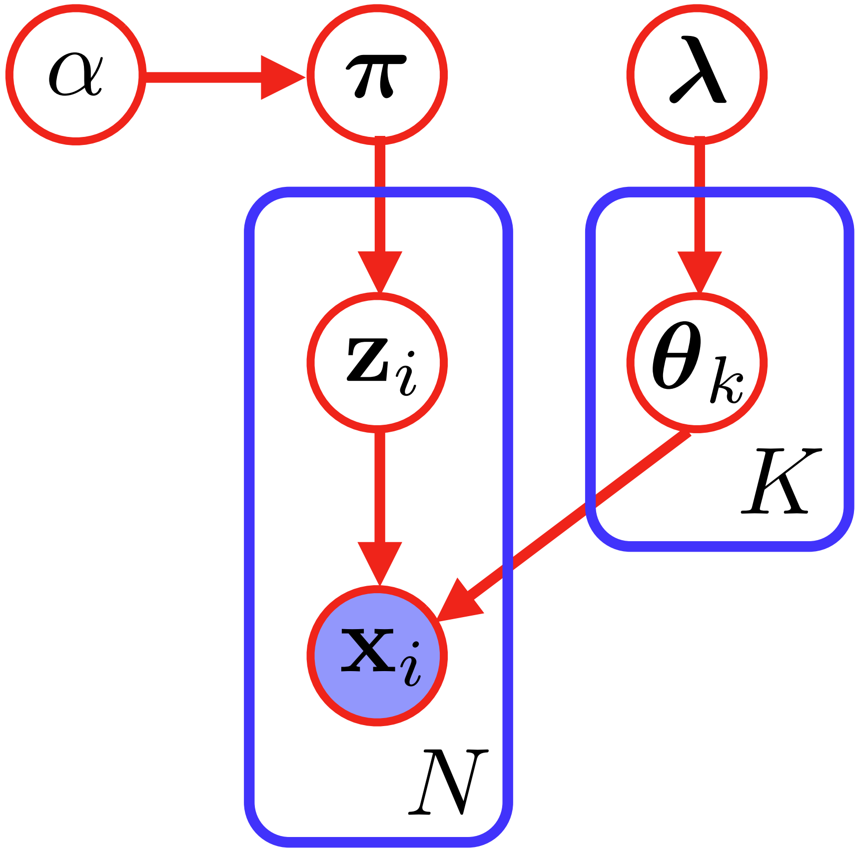
Idea: sample from the full conditional

This can be obtained, e.g. from the Markov blanket in graphical models.

Gibbs Sampling for GMMs

- The full posterior of the Gaussian Mixture Model is

$$p(X, Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \underbrace{p(X \mid Z, \boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\text{data likelihood (Gaussian)}} \underbrace{p(Z \mid \boldsymbol{\pi})}_{\text{correspondence prob. (Multinomial)}} \underbrace{p(\boldsymbol{\pi} \mid \alpha)}_{\text{mixture prior (Dirichlet)}} \underbrace{p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \lambda)}_{\text{parameter prior (Gauss-IW)}}$$



In this model, we use:

- $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$
- $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$
- $(\boldsymbol{\mu}_k, \Sigma_k) = \boldsymbol{\theta}_k$

Gibbs Sampling for GMMs

- The full posterior of the Gaussian Mixture Model is

$$p(X, Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \lambda)$$

- To apply Gibbs sampling we need to first find closed-form expressions for all **full conditionals** (prob. distr. of one variable given all others)

Gibbs Sampling for GMMs

- The full posterior of the Gaussian Mixture Model is

$$p(X, Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \lambda)$$

- To apply Gibbs sampling we need to first find closed-form expressions for all **full conditionals**
- These are:

$$p(z_i = k \mid \mathbf{x}_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) \propto \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k)$$

$$p(\boldsymbol{\pi} \mid \mathbf{z}) = \mathrm{Dir}(\{\alpha_k + \sum_{i=1}^{N} z_{ik}\}_{k=1}^{K})$$

$$p(\boldsymbol{\mu}_k \mid \Sigma_k, Z, X) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, V_k)$$

$$p(\Sigma_k \mid \boldsymbol{\mu}_k, Z, X) = \mathcal{IW}(\Sigma_k \mid S_k, \nu_k)$$

Gibbs Sampling for GMMs

- The full posterior of the Gaussian Mixture Model is

$$p(X, Z, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\mu}, \Sigma) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha) p(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{\lambda})$$

- To apply Gibbs sampling we need to first find closed-form expressions for all **full conditionals**
- These are:

$$p(z_i = k \mid \mathbf{x}_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) \propto \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k)$$

$$p(\boldsymbol{\pi} \mid \mathbf{z}) = \mathrm{Dir}(\{\alpha_k + \sum_{i=1}^{N} z_{ik}\}_{k=1}^{K})$$

$$p(\boldsymbol{\mu}_k \mid \Sigma_k, Z, X) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, V_k)$$

$$p(\Sigma_k \mid \boldsymbol{\mu}_k, Z, X) = \mathcal{IW}(\Sigma_k \mid S_k, \nu_k)$$

A More Efficient Variant

Remember: we have chosen **conjugate** priors

Likelihood	Conjugate Prior
Multinomial $p(Z \mid \pi_1, \ldots, \pi_k) = \prod_{k=1}^{K} \pi_k^{z_k}$	Dirichlet $\text{Dir}(\pi_1, \ldots, \pi_k \mid \alpha_1, \ldots, \alpha_K)$
Multivariate Normal $p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal-Inverse-Wishart $\text{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{m}_0, \kappa_0, \nu_0, S_0)$

This means, we can compute posteriors in closed form and **marginalize out** the model parameters!

Rao-Blackwellization

Instead of computing

$$p(X, Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}, \alpha, \lambda)$$

we compute ("**marginalization**"):

$$\int\int\int p(X, Z, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}, \alpha, \lambda) d\boldsymbol{\mu} d\boldsymbol{\Sigma} d\boldsymbol{\pi}$$

and sample from the resulting full conditionals. This is called **Rao-Blackwellization.** The resulting sampling method is called **collapsed** Gibbs sampling.

Dirichlet Distribution

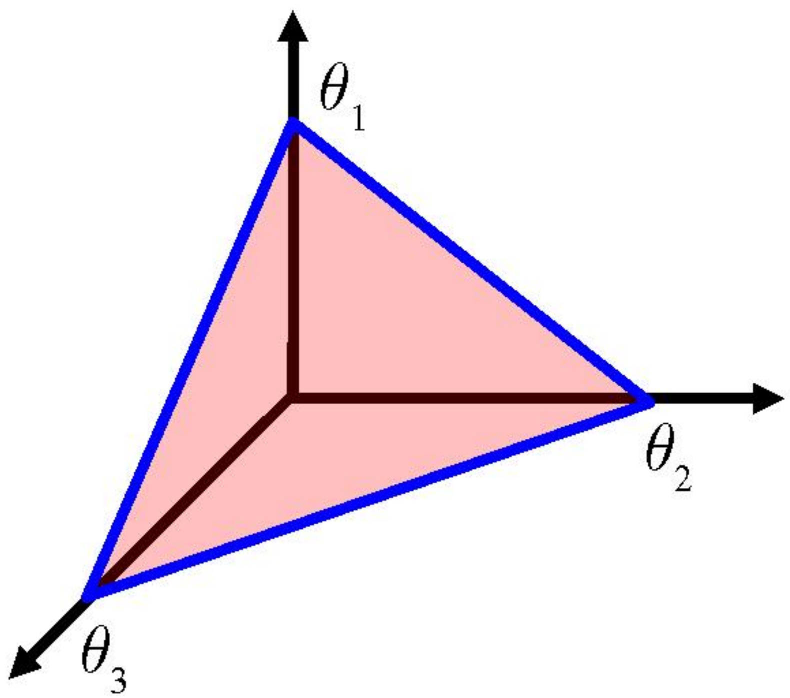
- The Dirichlet distribution is defined as:

$$\mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

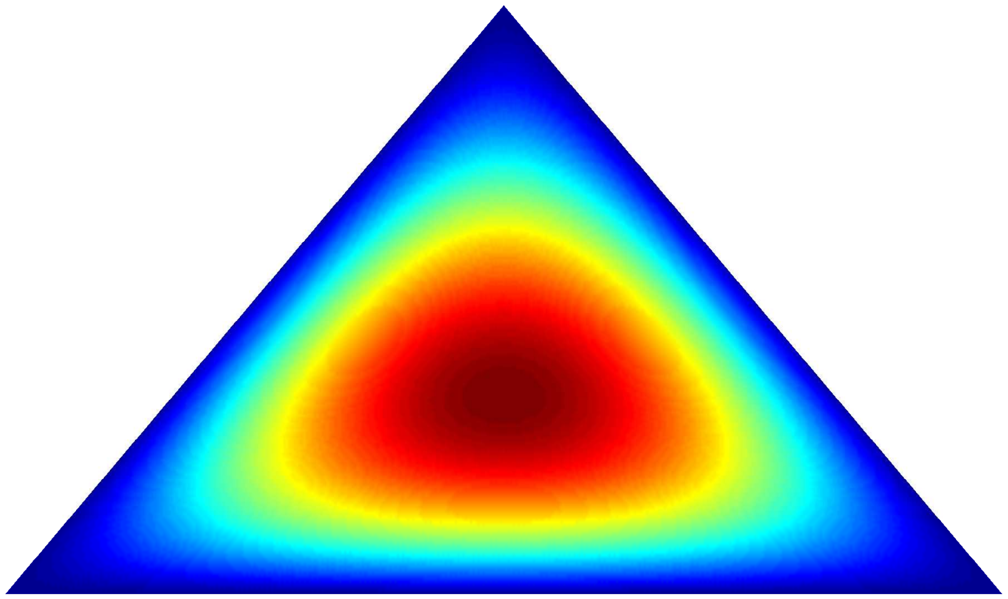
$$0 \leq \pi_k \leq 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

- It is the **conjugate prior** for the multinomial distribution
- The parameter α can be interpreted as the **effective number** of observations for every state

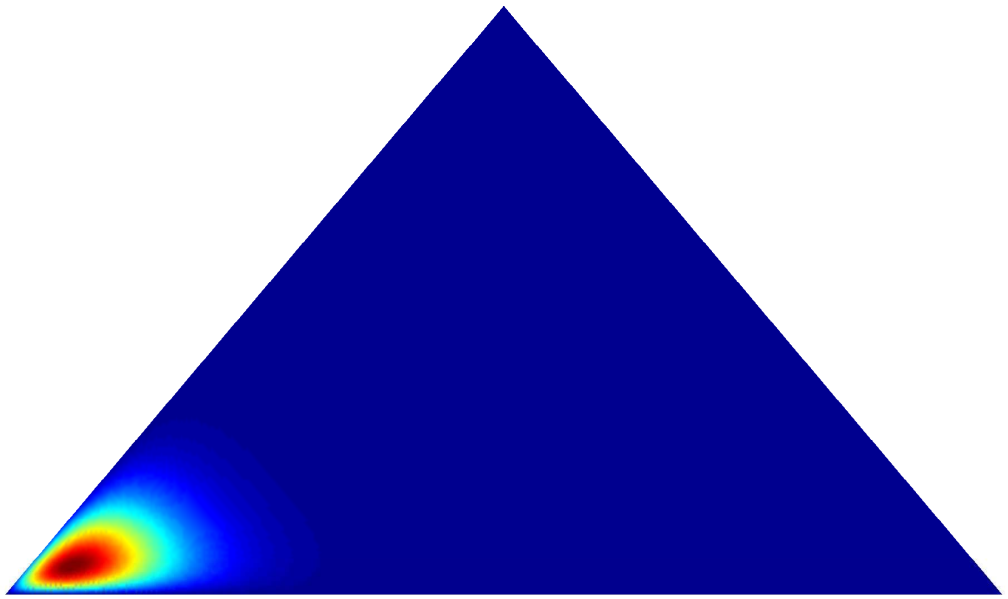


The simplex for K=3

Some Examples

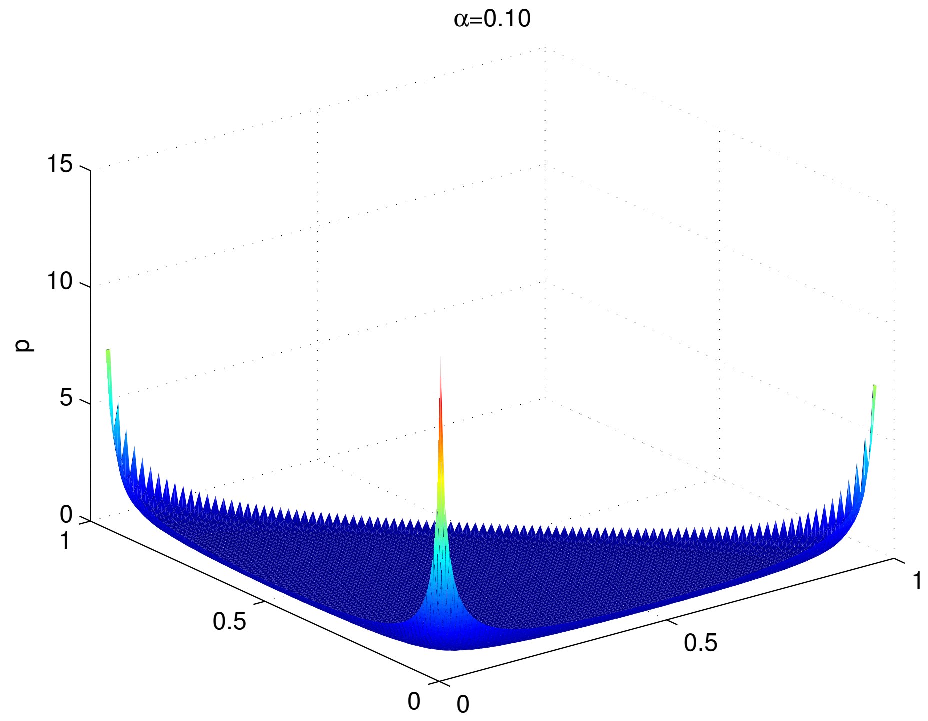


$\alpha = (2, 2, 2)$



$\alpha = (20, 2, 2)$

- α_0 controls the strength of the distribution ("peakedness")
- α_k control the location of the peak



$\alpha = (0.1, 0.1, 0.1)$

Conjugacy

- The Multinomial distribution is defined as:

$$p(\mathbf{z} \mid \pi_1, \ldots, \pi_K) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad \mathbf{z} \in \{0, 1\}^K$$

- Conjugacy means:

$$p(\pi_1, \ldots, \pi_K \mid \mathbf{z}) \propto \underbrace{p(\mathbf{z} \mid \pi_1, \ldots, \pi_K)}_{\text{Multinomial}} \underbrace{p(\pi_1, \ldots, \pi_K \mid \alpha_1, \ldots, \alpha_K)}_{\text{Dirichlet}}$$

Conjugacy

- The Multinomial distribution is defined as:

$$p(\mathbf{z} \mid \pi_1, \ldots, \pi_K) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad \mathbf{z} \in \{0, 1\}^K$$

- Conjugacy means:

$$p(\pi_1, \ldots, \pi_K \mid \mathbf{z}) = \eta^{-1} p(\mathbf{z} \mid \pi_1, \ldots, \pi_K) p(\pi_1, \ldots, \pi_K \mid \alpha_1, \ldots, \alpha_K)$$

Normalizer

$$= \mathrm{Dir}(\pi_1, \ldots, \pi_k \mid \alpha'_1, \ldots, \alpha'_K)$$

where $\alpha'_k = \alpha_k + z_k$

Marginalization

- The normalizer η can be computed as

$$p(Z \mid \alpha_1, \ldots, \alpha_K) = \int \underbrace{p(Z \mid \pi_1, \ldots, \pi_K)}_{\text{Multinomial}} \underbrace{p(\pi_1, \ldots, \pi_K \mid \alpha_1, \ldots, \alpha_K)}_{\text{Dirichlet}} d\boldsymbol{\pi}$$

note: $Z = \mathbf{z}_1, \ldots \mathbf{z}_N$

- This can also be computed in closed form:

$$p(Z \mid \pi_1, \ldots, \pi_K) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} = \prod_{k=1}^{K} \pi_k^{N_k}$$

Marginalization

- The normalizer η can be computed as

$$p(Z \mid \alpha_1, \ldots, \alpha_K) = \int \underbrace{p(Z \mid \pi_1, \ldots, \pi_K)}_{\text{Multinomial}} \underbrace{p(\pi_1, \ldots, \pi_K \mid \alpha_1, \ldots, \alpha_K)}_{\text{Dirichlet}} d\boldsymbol{\pi}$$

note: $Z = \mathbf{z}_1, \ldots \mathbf{z}_\mathrm{N}$

- This can also be computed in closed form:

$$p(Z \mid \pi_1, \ldots, \pi_K) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} = \prod_{k=1}^{K} \pi_k^{N_k}$$

$$\Rightarrow p(Z \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

The Other Pair

- The same operations can be done for the other **likelihood-prior** pair:
- Conjugacy: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid X) = \eta'^{-1} \underbrace{p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\text{Gaussian}} \underbrace{p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\lambda})}_{\text{NIW}}$

The Other Pair

- The same operations can be done for the other **likelihood-prior** pair:
- Conjugacy: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid X) = \eta'^{-1} p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\lambda})$

$$= \mathrm{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\lambda}_N)$$

(we omit details of how to compute $\boldsymbol{\lambda}_N$)

The Other Pair

- The same operations can be done for the other **likelihood-prior** pair:
- Conjugacy: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid X) = \eta'^{-1} p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\lambda})$
- Marginalization:

$$p(X) = \eta' = \int \int p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\lambda}) d\boldsymbol{\mu} d\boldsymbol{\Sigma}$$

The Other Pair

- The same operations can be done for the other **likelihood-prior** pair:
- Conjugacy: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid X) = \eta'^{-1} p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \lambda)$
- Marginalization:

$$p(X) = \eta' = \int \int p(X \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \lambda) d\boldsymbol{\mu} d\boldsymbol{\Sigma}$$

$$= \pi^{-ND/2} \frac{\kappa_0^{D/2} |S_0|^{\nu_0/2}}{\kappa_N^{D/2} |S_N|^{\nu_N/2}} \prod_{i=1}^{D} \frac{\Gamma(\frac{\nu_N+1-i}{2})}{\Gamma(\frac{\nu_0+1-i}{2})}$$

(again, we omit details)

How Can we Use That?

- Our goal is to find the full conditionals:

$$p(\mathbf{z}_i = k \mid Z_{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \propto p(\mathbf{z}_i = k \mid Z_{-i}, \boldsymbol{\alpha}) p(X \mid \mathbf{z}_i = k, Z_{-i}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$$

How Can we Use That?

- Our goal is to find the full conditionals:

$$p(\mathbf{z}_i = k \mid Z_{-i}, X, \alpha, \lambda) \propto p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(X \mid \mathbf{z}_i = k, Z_{-i}, \alpha, \lambda)$$

$$= p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(\mathbf{x}_i \mid X_{-i}, \mathbf{z}_i = k, Z_{-i}, \lambda) p(X_{-i} \mid \cancel{\mathbf{z}_i =} k, Z_{-i}, \lambda)$$

$$\propto p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(\mathbf{x}_i \mid X_{-i}, \mathbf{z}_i = k, Z_{-i}, \lambda)$$

How Can we Use That?

- Our goal is to find the full conditionals:

$$p(\mathbf{z}_i = k \mid Z_{-i}, X, \alpha, \lambda) \propto p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(X \mid \mathbf{z}_i = k, Z_{-i}, \alpha, \lambda)$$

$$= p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(\mathbf{x}_i \mid X_{-i}, \mathbf{z}_i = k, Z_{-i}, \lambda) p(X_{-i} \mid \cancel{\mathbf{z}_i =} k, Z_{-i}, \lambda)$$

$$\propto p(\mathbf{z}_i = k \mid Z_{-i}, \alpha) p(\mathbf{x}_i \mid X_{-i}, \mathbf{z}_i = k, Z_{-i}, \lambda)$$

- We are left with **two** full conditionals that we can compute in closed form and then sample from the product

The First Term

$$p(\mathbf{z}_i = k \mid Z_{-i}, \boldsymbol{\alpha}) = \frac{p(Z \mid \boldsymbol{\alpha})}{p(Z_{-i} \mid \boldsymbol{\alpha})}$$

$\mathbf{z}_i = k$

The First Term

$$p(\mathbf{z}_i = k \mid Z_{-i}, \boldsymbol{\alpha}) = \frac{p(Z \mid \boldsymbol{\alpha})}{p(Z_{-i} \mid \boldsymbol{\alpha})}$$

$\mathbf{z}_i = k$

- We already computed the numerator (see above):

$$p(Z \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

- The denominator is very similar:

$$p(Z_{-i} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N - 1)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + N_{-i,k})}{\Gamma(\alpha_k)}$$

The First Term

$$p(\mathbf{z}_i = k \mid Z_{-i}, \boldsymbol{\alpha}) = \frac{p(Z \mid \boldsymbol{\alpha})}{p(Z_{-i} \mid \boldsymbol{\alpha})}$$

$\mathbf{z}_i = k$

- We already computed the numerator (see above):

$$p(Z \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

- The denominator is very similar:

$$p(Z_{-i} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N - 1)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + N_{-i,k})}{\Gamma(\alpha_k)}$$

- Result:

$$p(\mathbf{z}_i = k \mid Z_{-i}, \boldsymbol{\alpha}) = \frac{N_{-i,k} + \alpha_k}{N + \alpha_0 - 1}$$

The Second Term

$$p(\mathbf{x}_i \mid X_{-i}, \mathbf{z}_i = k, Z_{-i}, \boldsymbol{\lambda}) = p(\mathbf{x}_i \mid X_{-i,k}, \boldsymbol{\lambda})$$

All data samples that belong to cluster k, except the i-th one

- We use the same idea here:

$$p(\mathbf{x}_i \mid X_{-i,k}, \boldsymbol{\lambda}) = \frac{p(X_k \mid \boldsymbol{\lambda})}{p(X_{-i,k} \mid \boldsymbol{\lambda})}$$

- This can be computed again from marginalization (see above). Again, we omit details.

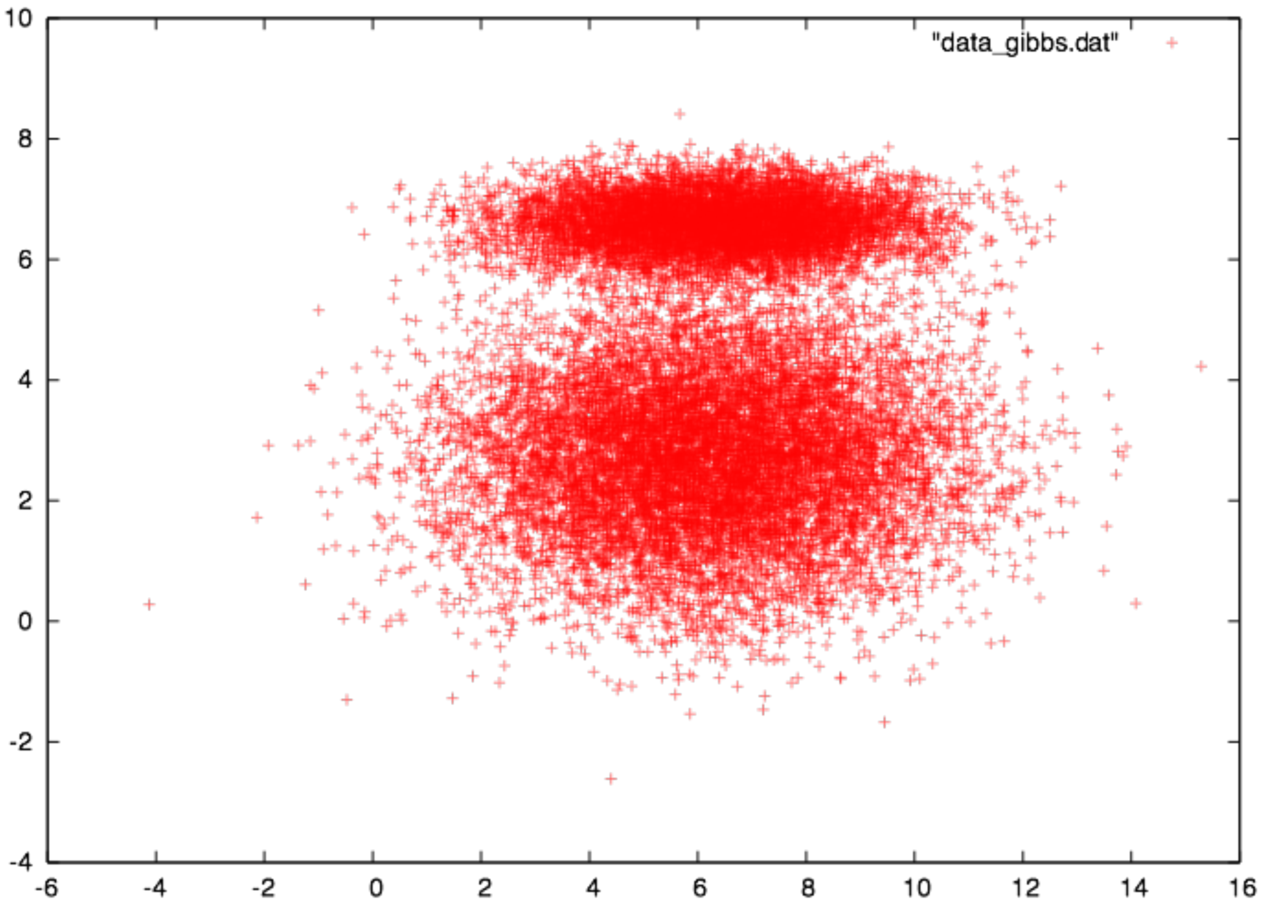
GMM with Collapsed Gibbs Samlping

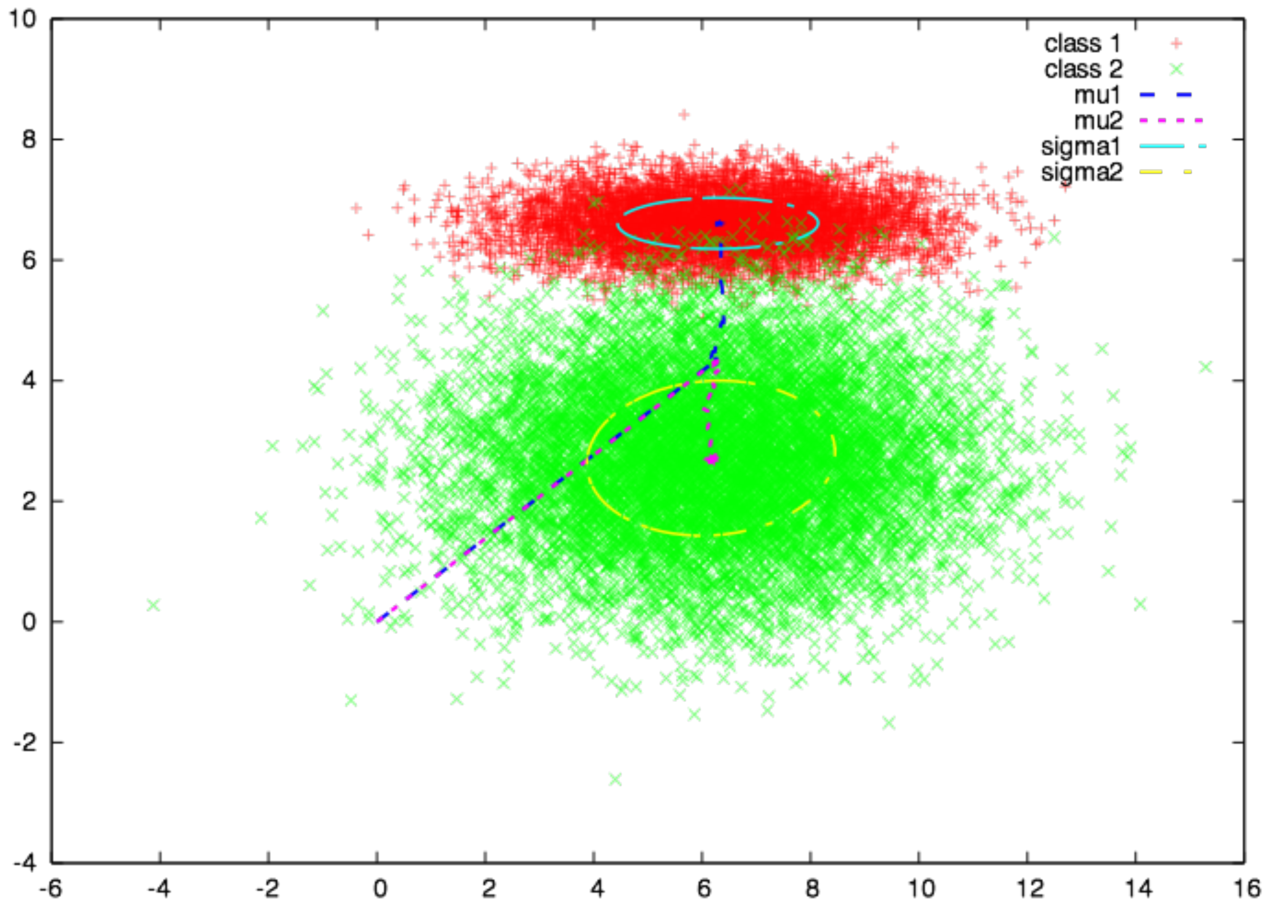
Algorithm 1 Collapsed Gibbs sampler for a finite Gaussian mixture model.

Choose an initial $\mathbf{z}$.
for T iterations **do** ▷ Gibbs sampling iterations
    for $i = 1$ to N **do**
        Remove $\mathbf{x}_i$'s statistics from component z_i. ▷ Old assignment for $\mathbf{x}_i$
        for $k = 1$ to K **do** ▷ Every possible component
            Calculate $P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha})$ using (25).
            Calculate $p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$ in (27) using (14) or (15).
            Calculate $P(z_i = k|\mathbf{z}_{\setminus i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha})\, p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$.
        end for
        Sample k_{new} from $P(z_i|\mathbf{z}_{\setminus i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ after normalizing.
        Add $\mathbf{x}_i$'s statistics to the component $z_i = k_{\text{new}}$. ▷ New assignment for $\mathbf{x}_i$
    end for
end for

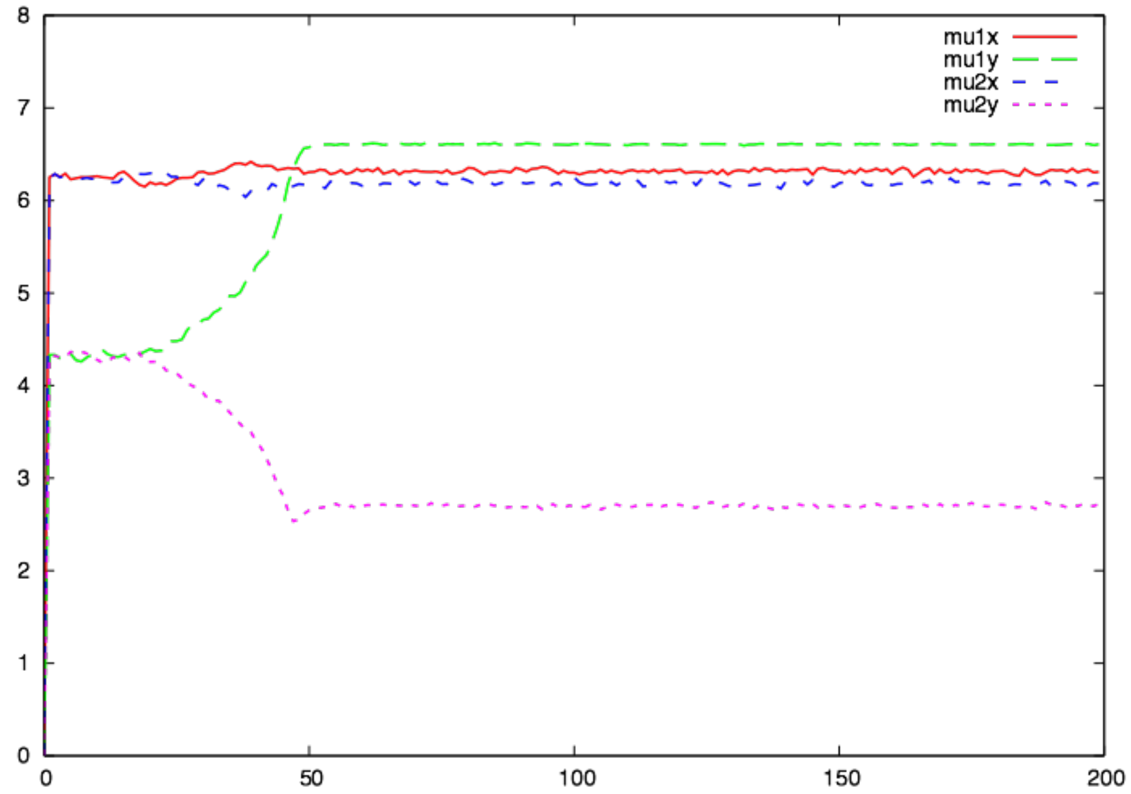
Gibbs Sampling for GMMs

- First, we initialize all variables
- Then we iterate over sampling from each conditional in turn
- In the end, we look at $\boldsymbol{\mu}_k$ and Σ_k





How Often Do We Have To Sample?



- Here: after 50 sample rounds the values don't change any more
- In general, the **mixing time** τ_ϵ is related to the **eigen gap** $\gamma = \lambda_1 - \lambda_2$ of the transition matrix:

$$\tau_\epsilon \leq O(\frac{1}{\gamma} \log \frac{n}{\epsilon})$$

How Can We Get Rid of K?

- We still have the problem that we need the number K of clusters given
- Idea: use the same methodology, but let K go to infinity
- Instead of a Dirichlet distribution, we will then be using a **Dirichlet process**

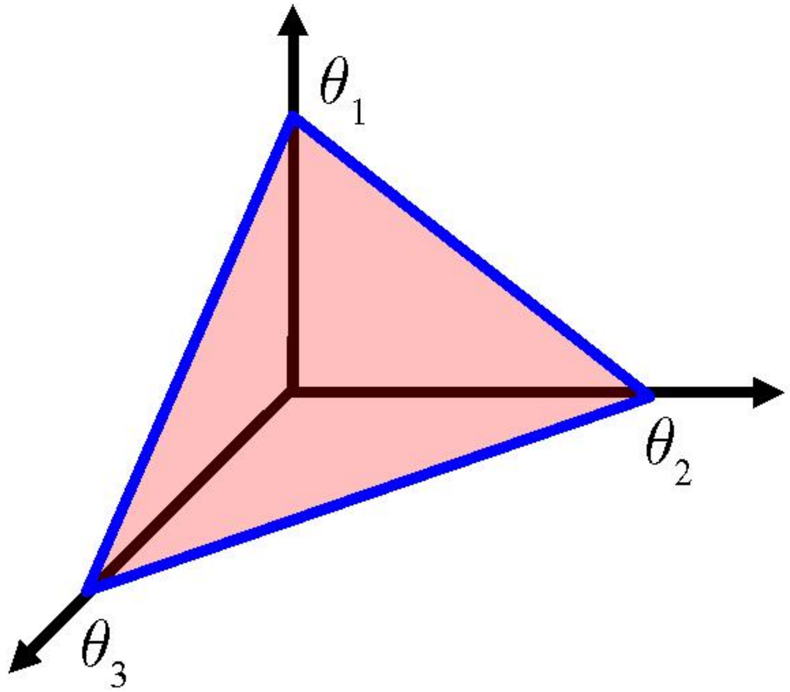
Dirichlet Distribution

- The Dirichlet distribution is defined as:

$$\mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$0 \leq \pi_k \leq 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

- It is the **conjugate prior** for the multinomial distribution
- The parameter α can be interpreted as the **effective number** of observations for every state



The simplex for K=3

Other Properties of the Dirichlet Dist.

- "Agglomerative":

$$p(\mu_1, \ldots, \mu_K) = \mathrm{Dir}(\mu_1, \ldots, \mu_K \mid \alpha_1, \ldots, \alpha_K)$$

$$\Rightarrow p(\mu_1 + \mu_2, \ldots, \mu_K) = \mathrm{Dir}(\mu_1 + \mu_2, \ldots, \mu_K \mid \alpha_1 + \alpha_2, \ldots, \alpha_K)$$

this also holds for general partitions of $1, \ldots, K$

- "Decimative":

$$p(\mu_1, \ldots, \mu_K) = \mathrm{Dir}(\mu_1, \ldots, \mu_K \mid \alpha_1, \ldots, \alpha_K)$$

$$\wedge p(\nu_1, \nu_2) = \mathrm{Dir}(\nu_1, \nu_2 \mid \alpha_1\beta_1, \alpha_1\beta_2) \qquad \beta_1 + \beta_2 = 1$$

$$\Rightarrow p(\mu_1\nu_1, \mu_1\nu_2, \mu_2 \ldots, \mu_K) = \mathrm{Dir}(\mu_1\nu_1, \mu_1\nu_2, \mu_2, \ldots, \mu_K \mid \alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, \ldots, \alpha_K)$$

From Finite to Infinite Dimensions

- Observation: every sample from a Dirichlet distribution represents a distribution over K finite states

- We can generalize this to **infinitely** many states

$$1 \sim \mathrm{Dir}(\mu \mid \alpha)$$

$$(\mu_1, \mu_2) \sim \mathrm{Dir}(\mu_1, \mu_2 \mid \alpha/2, \alpha/2)$$

$$(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}) \sim \mathrm{Dir}(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22} \mid \alpha/4, \alpha/4, \alpha/4, \alpha/4)$$

$$\vdots$$

- The result is a discrete, but infinite distribution

The Dirichlet Process

Definition: A Dirichlet process (DP) is a distribution over probability measures G, i.e. $G(\boldsymbol{\theta}) \geq 0$ and $\int G(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. If for any partition $(T_1, \ldots, T_K)$ it holds:

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K))$$

then G is sampled from a Dirichlet process.

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

where α is the **concentration parameter**
and H is the **base measure**

The Dirichlet Process

Definition: A Dirichlet process (DP) is a distribution over probability measures G, i.e. $G(\boldsymbol{\theta}) \geq 0$ and $\int G(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. If for any partition $(T_1, \dots, T_K)$ it holds:

$$(G(T_1), \dots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \dots, \alpha H(T_K))$$

then G is sampled from a Dirichlet process.

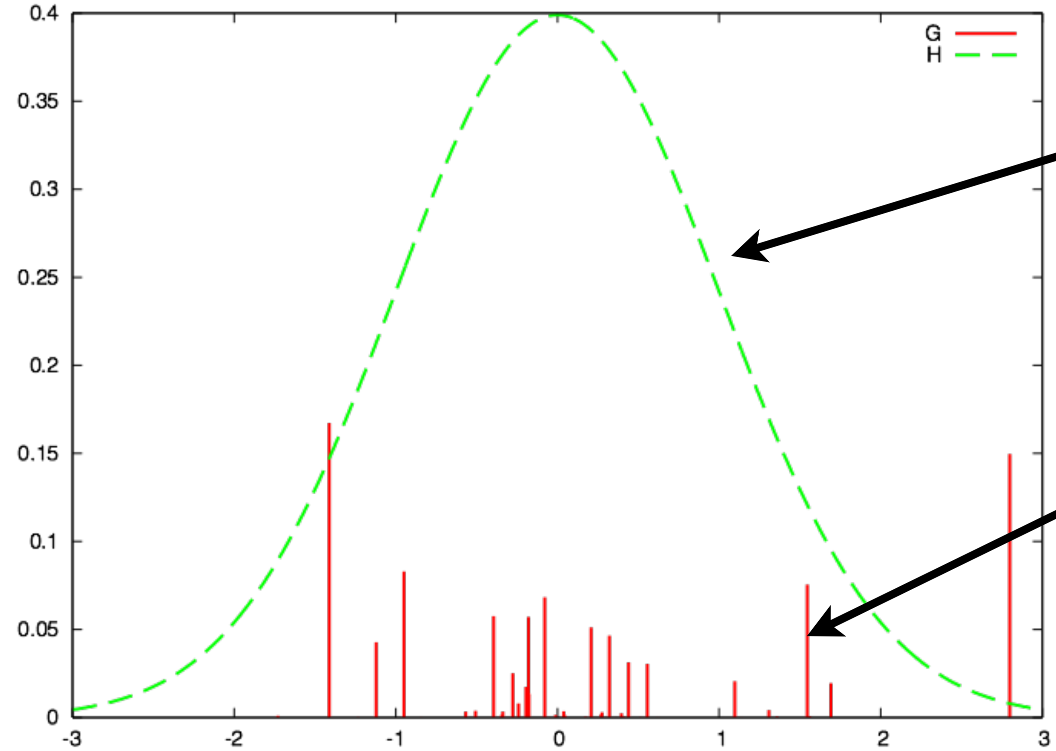
Notation: $G \sim \mathrm{DP}(\alpha, H)$

where α is the **concentration parameter** and H is the **base measure**

Note: This is not a constructive definition!

Intuitive Interpretation

- Every sample from a Dirichlet distribution is a vector of K positive values that sum up to 1, i.e. the sample itself is a finite distribution
- Accordingly, a sample from a Dirichlet process is an infinite (but still discrete!) distribution



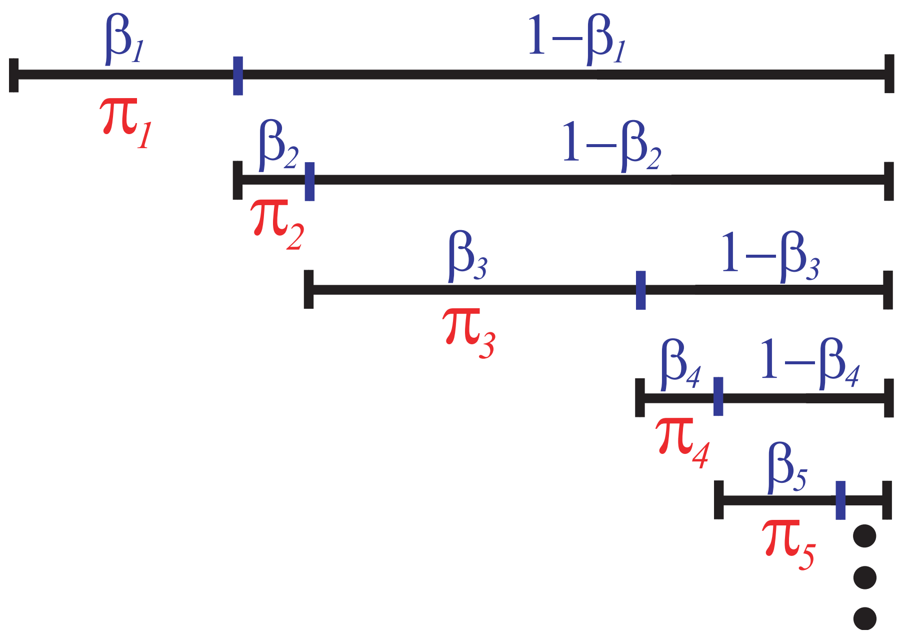
Base distribution (here Gaussian)

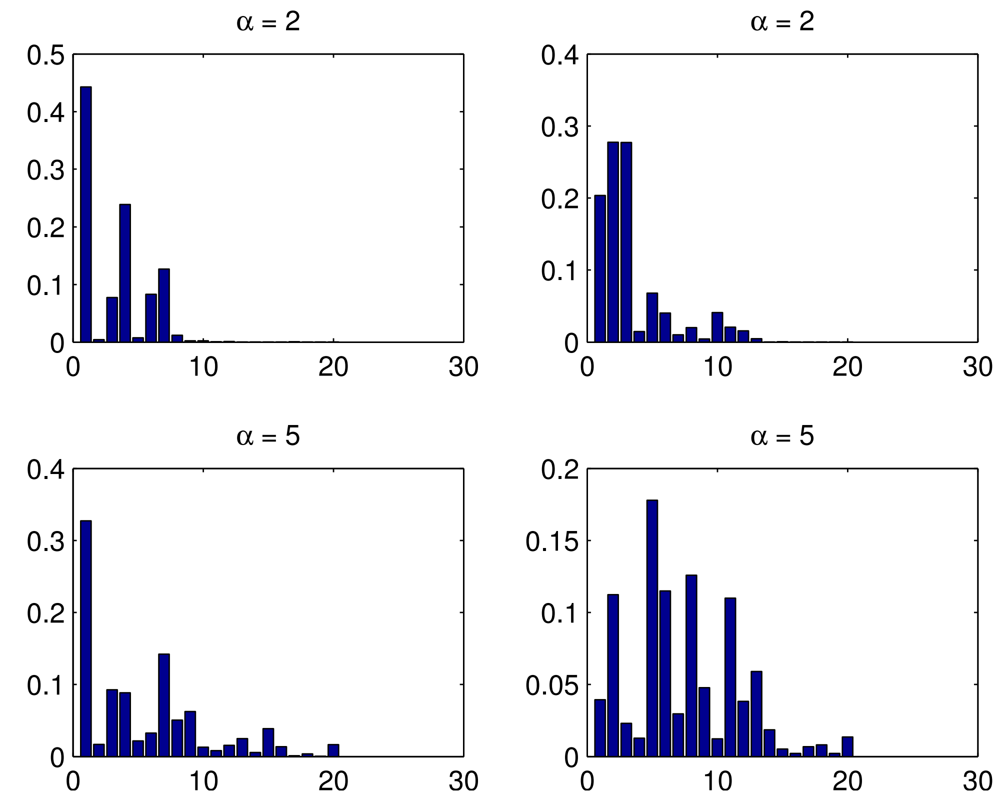
Infinitely many samples (sum up to 1)

Construction of a Dirichlet Process

- The Dirichlet process is only defined **implicitly**, i.e. we can test whether a given probability measure is sampled from a DP, but we can not yet construct one.
- A DP can be constructed using the “stick-breaking” analogy:
  - imagine a stick of length 1
  - we select a random number β between 0 and 1 from a Beta-distribution
  - we break the stick at $\pi = \beta$ * length-of-stick
  - we repeat this infinitely often

The Stick-Breaking Construction





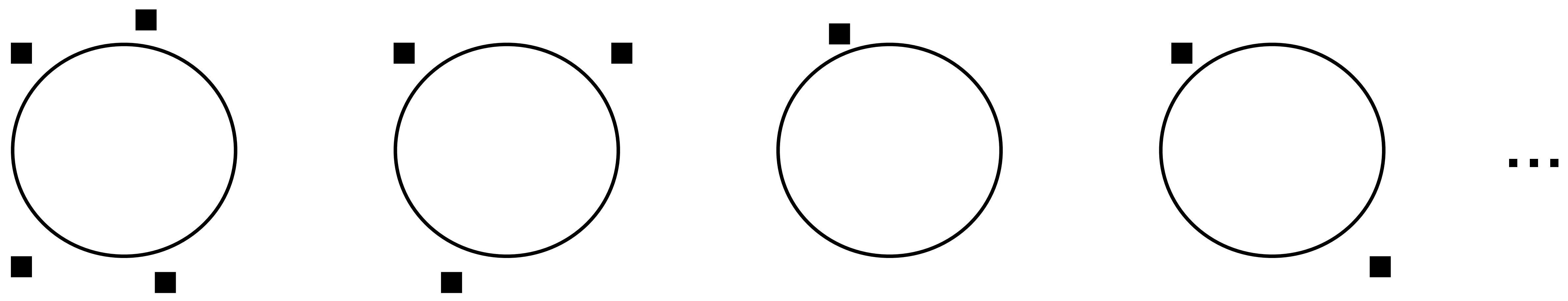
- formally, we have

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k (1 - \sum_{l=1}^{k-1} \pi_l)$$

- now we define

$$G(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k \delta(\boldsymbol{\theta}_k, \boldsymbol{\theta}) \qquad \boldsymbol{\theta}_k \sim H \qquad \text{then: } \quad G \sim \mathrm{DP}(\alpha, H)$$

The Chinese Restaurant Process



- Consider a restaurant with infinitely many tables
- Everytime a new customer comes in, he sits at an **occupied table** with probability **proportional to the number of people** sitting at that table, but he may choose to sit on a **new** table with **decreasing** probability as more customers enter the room.

The Chinese Restaurant Process

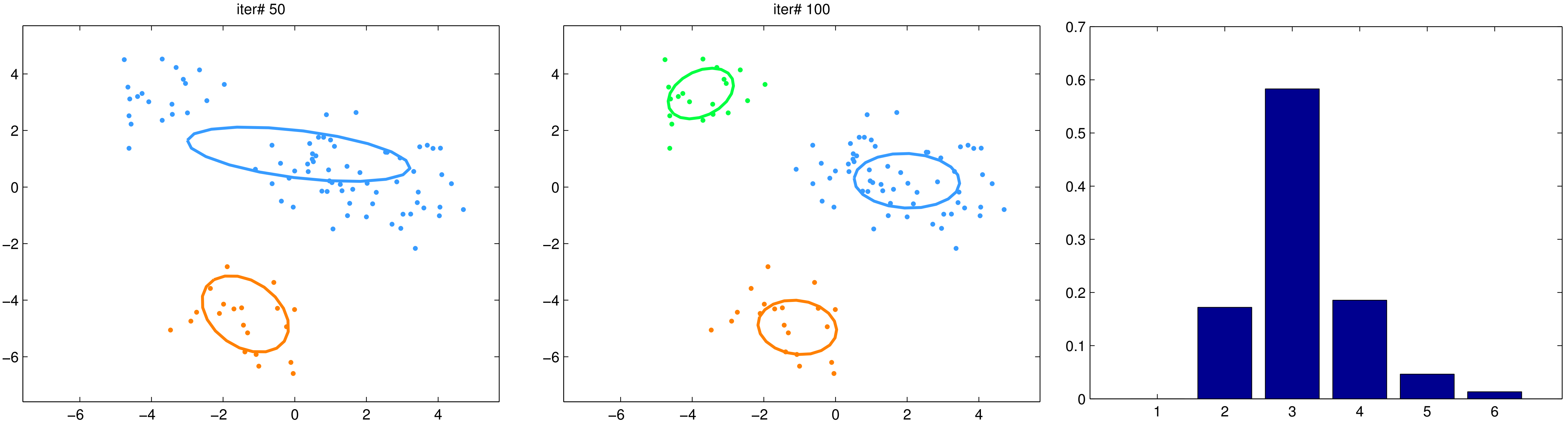
- It can be shown that the probability for a new customer is

$$p(\bar{\boldsymbol{\theta}}_{N+1} = \boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}_{1:N}, \alpha, H) = \frac{1}{\alpha + N} \left(\alpha H(\boldsymbol{\theta}) + \sum_{k=1}^{K} N_k \delta(\bar{\boldsymbol{\theta}}_k, \boldsymbol{\theta}) \right)$$

- This means that currently occupied tables are more likely to get new customers (**rich get richer**)
- The number of occupied tables grows logarithmically with the number of customers

The DP for Mixture Modeling

- Using the stick-breaking construction, we see that we can extend the mixture model clustering to the situation where K goes to infinity
- The algorithm can be implemented using Gibbs sampling



DPMM with Collapsed Gibbs Sampling

Algorithm 2 Collapsed Gibbs sampler for an infinite Gaussian mixture model.

Choose an initial $\mathbf{z}$.
for T iterations **do** ▷ Gibbs sampling iterations
    for $i = 1$ to N **do**
        Remove $\mathbf{x}_i$'s statistics from component z_i. ▷ Old assignment for $\mathbf{x}_i$
        for $k = 1$ to K **do** ▷ Every possible existing component
            Calculate $P(z_i = k | \mathbf{z}_{\setminus i}, \alpha) = \frac{N_{k\setminus i}}{N+\alpha-1}$ as in (34).
            Calculate $p(\mathbf{x}_i | \mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$ in (35) using (14) or (15).
            Calculate $P(z_i = k | \mathbf{z}_{\setminus i}, \mathcal{X}, \alpha, \boldsymbol{\beta}) \propto P(z_i = k | \mathbf{z}_{\setminus i}, \alpha)\, p(\mathbf{x}_i | \mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$.
        end for
        Calculate $P(z_i = k^* | \mathbf{z}_{\setminus i}, \alpha) = \frac{\alpha}{N+\alpha-1}$ as in (34). ▷ Consider a new component
        Calculate $p(\mathbf{x}_i | \boldsymbol{\beta})$ in (36) using (14) or (15).
        Calculate $P(z_i = k^* | \mathbf{z}_{\setminus i}, \mathcal{X}, \alpha, \boldsymbol{\beta}) \propto P(z_i = k^* | \mathbf{z}_{\setminus i}, \alpha)\, p(\mathbf{x}_i | \boldsymbol{\beta})$.
        Sample k_{new} from $P(z_i | \mathbf{z}_{\setminus i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ after normalizing.
        Add $\mathbf{x}_i$'s statistics to the component $z_i = k_{\text{new}}$. ▷ New assignment for $\mathbf{x}_i$
        If any component is empty, remove it and decrease K.
    end for
end for

Summary

- We can use Gibbs sampling to estimate a Gaussian Mixture model for a given data set
- As we are using conjugate priors, we can compute posters in closed form ("**Bayesian approach**")
- To be more efficient, we use collapsed Gibbs sampling, where model parameters are marginalized out ("**Rao-Blackwellization**")
- The same idea can be used to extend the GMM for infinite mixtures (K goes to infinity)
- This results in the Dirichlet Process Mixture Model (DPMM)