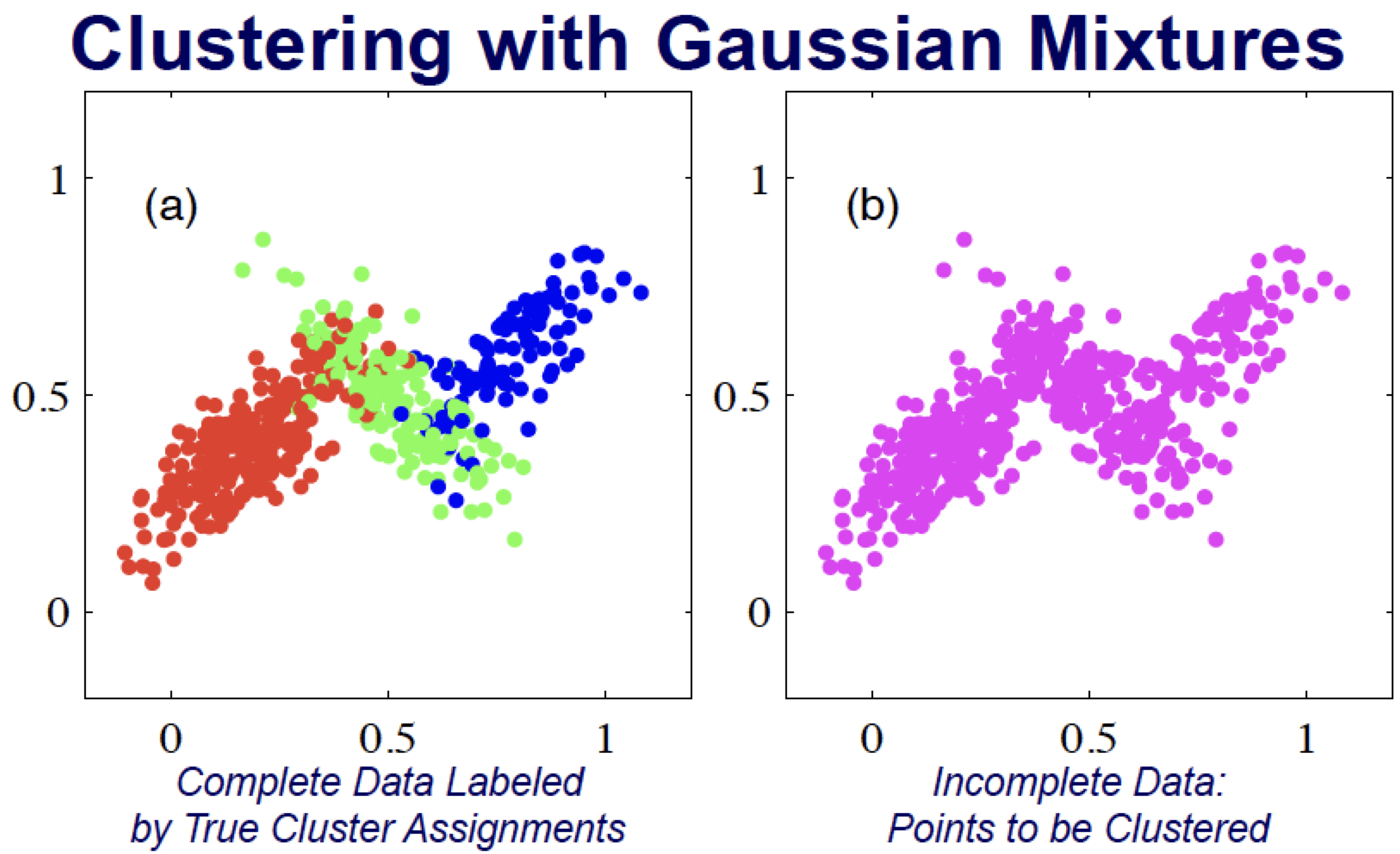
Dirichlet process

Dirichlet process: What's this good for?

- Principled, Bayesian method for fitting a mixture model with an unknown number of clusters

- Because it's Bayesian, can build hierarchies (e.g. HDPs) and integrate with other random variables in a principled way

Dirichlet process: What's this good for? Recall Clustering with GM



Dirichlet process: What's this good for? Recall Clustering with GM

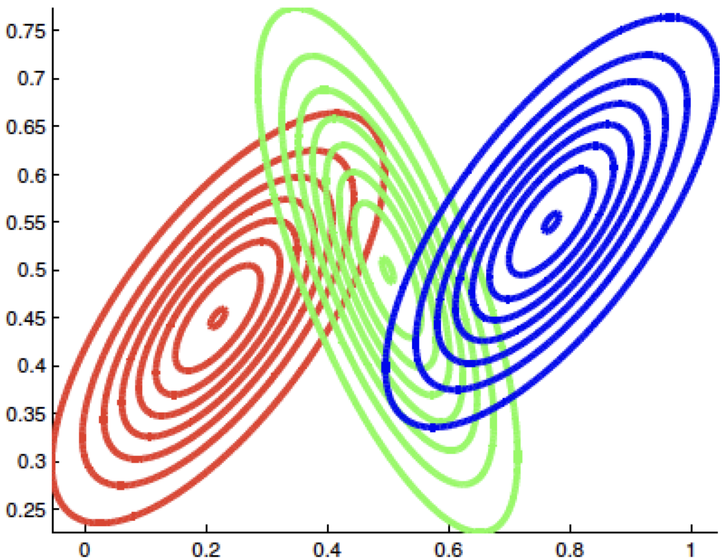
- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances: $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities: $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$
- Gaussian mixture marginal likelihood:

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

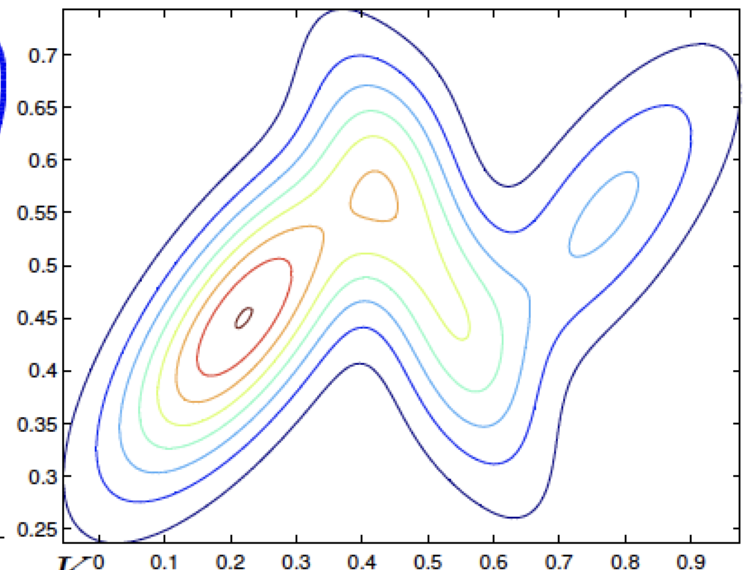
$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Dirichlet process: What's this good for? Recall Clustering with GM

Mixture of 3 Gaussian Distributions in 2D



Contour Plot of Joint Density, Marginalizing Cluster Assignments



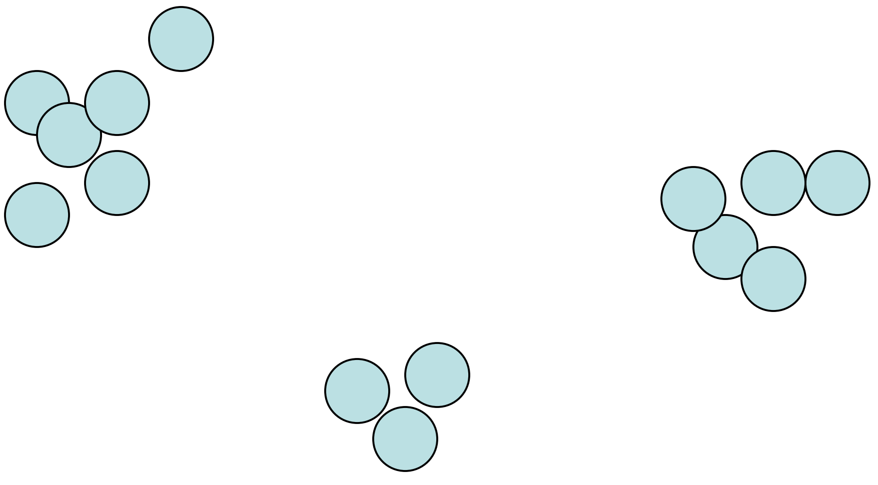
$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

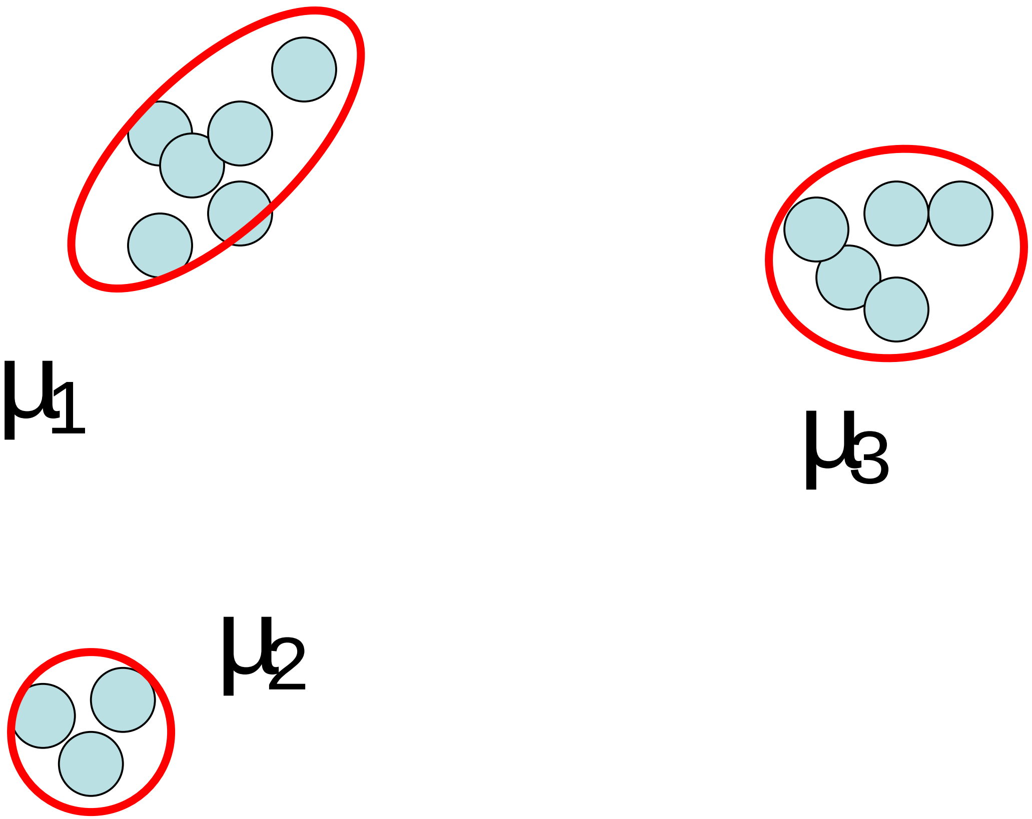
Aren't there other ways to count the number of clusters?

- Adhoc approaches (e.g., hierarchical clustering)
  - they do often yield a data-driven choice of K
  - but there is little understanding of how good these choices are
- Methods based on objective functions (M-estimators)
  - e.g., K-means, spectral clustering
  - do come with some frequentist guarantees
  - but it's hard to turn these into data-driven choices of K
- Parametric likelihood-based approaches
  - finite mixture models, Bayesian variants thereof
  - various model choice methods: hypothesis testing, cross-validation, bootstrap
    etc
  - but do the assumptions underlying the method really apply to this setting? (not often)
- Let's try something different...

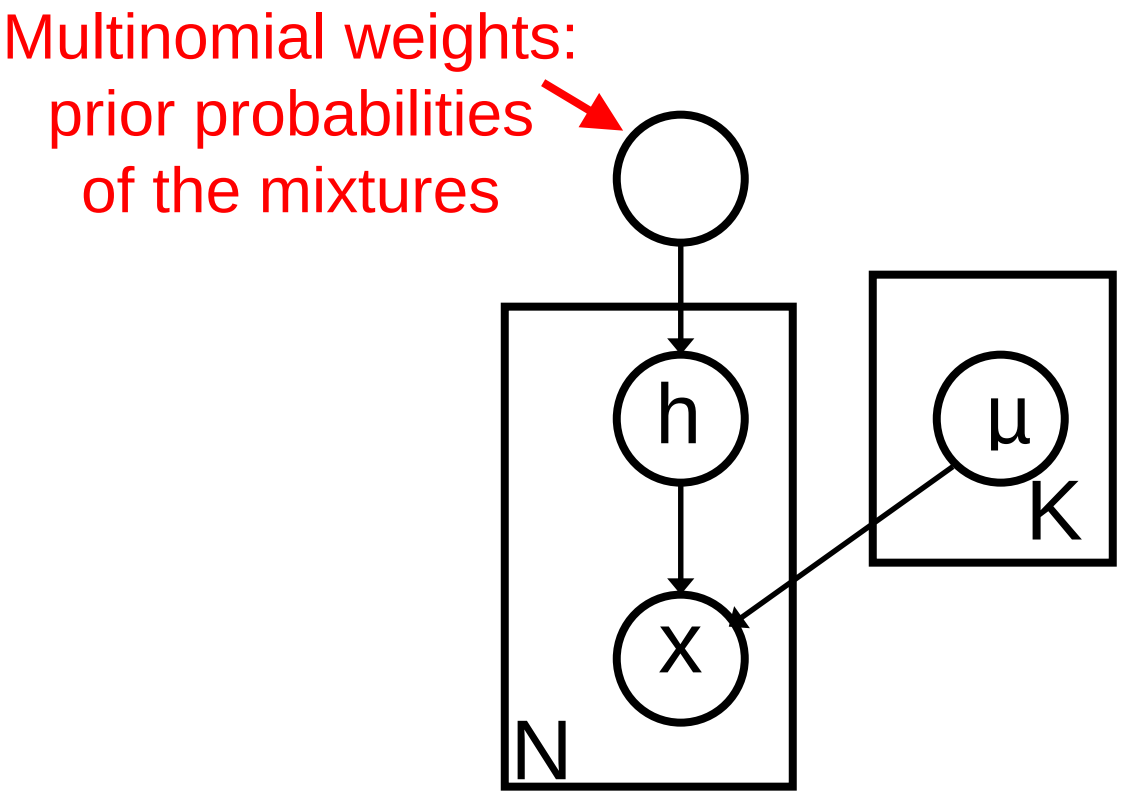
Gaussian mixture model, revisited



Gaussian mixture model, revisited

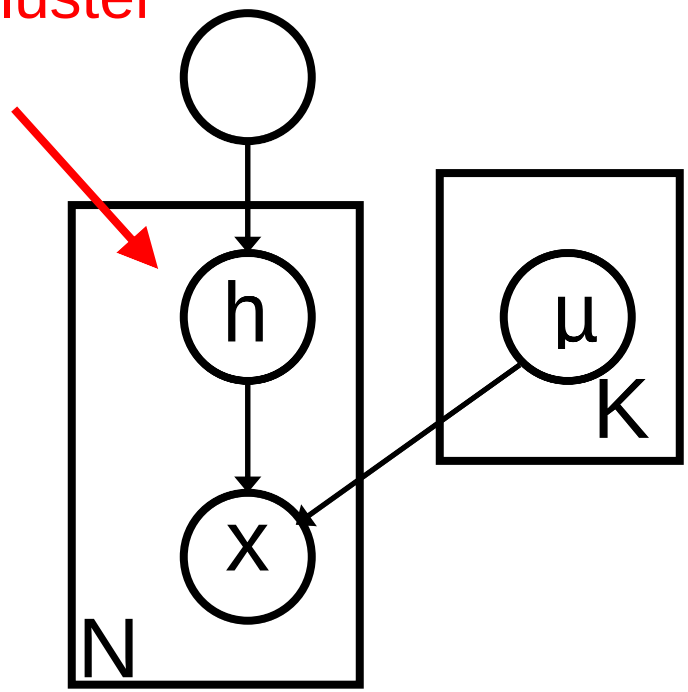


Gaussian mixture model, revisited

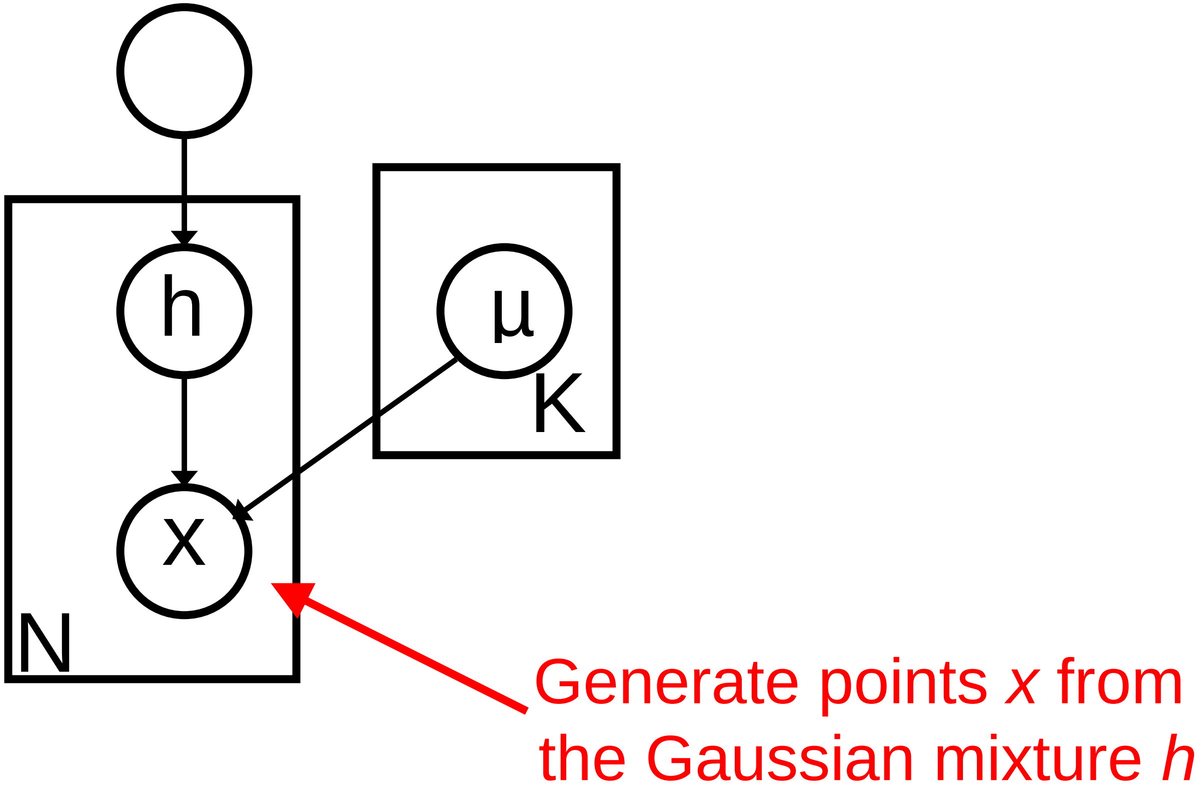


Gaussian mixture model, revisited

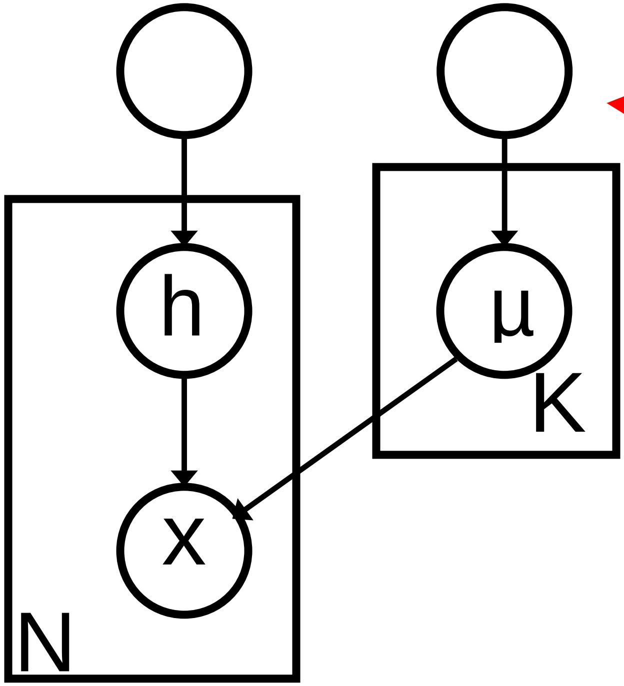
For each data point, choose cluster center *h*



Gaussian mixture model, revisited



Let us be more Bayesian…

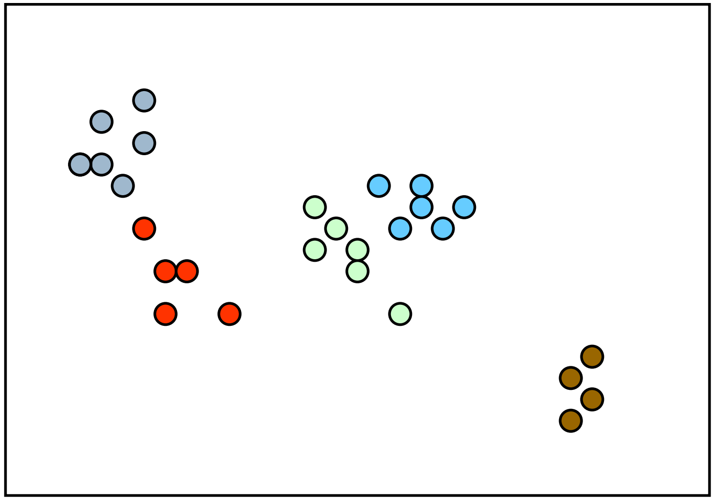


Put a prior over mixture parameters

For Gaussian mixtures, this is a normal inverse-Wishart density

Motivation

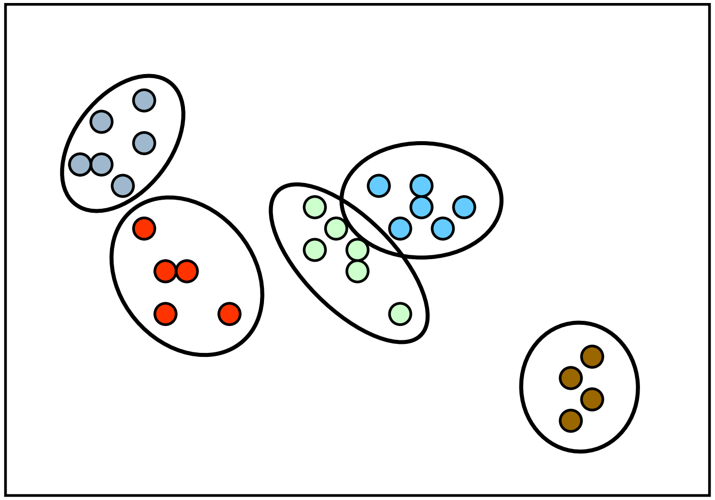
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

Motivation

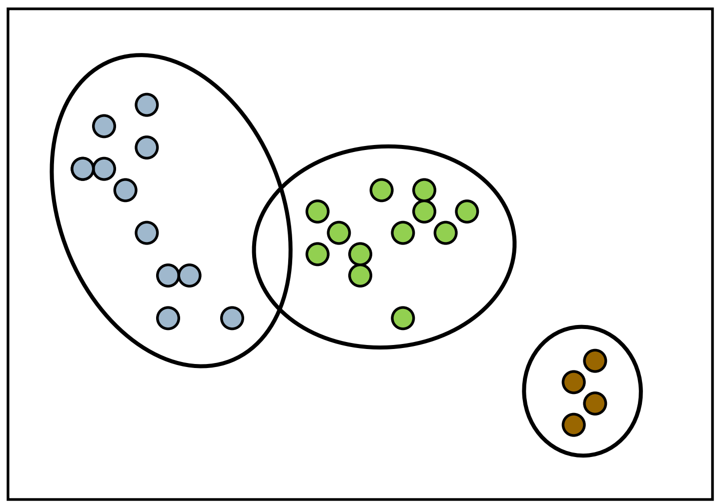
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

The Dirichlet Distribution

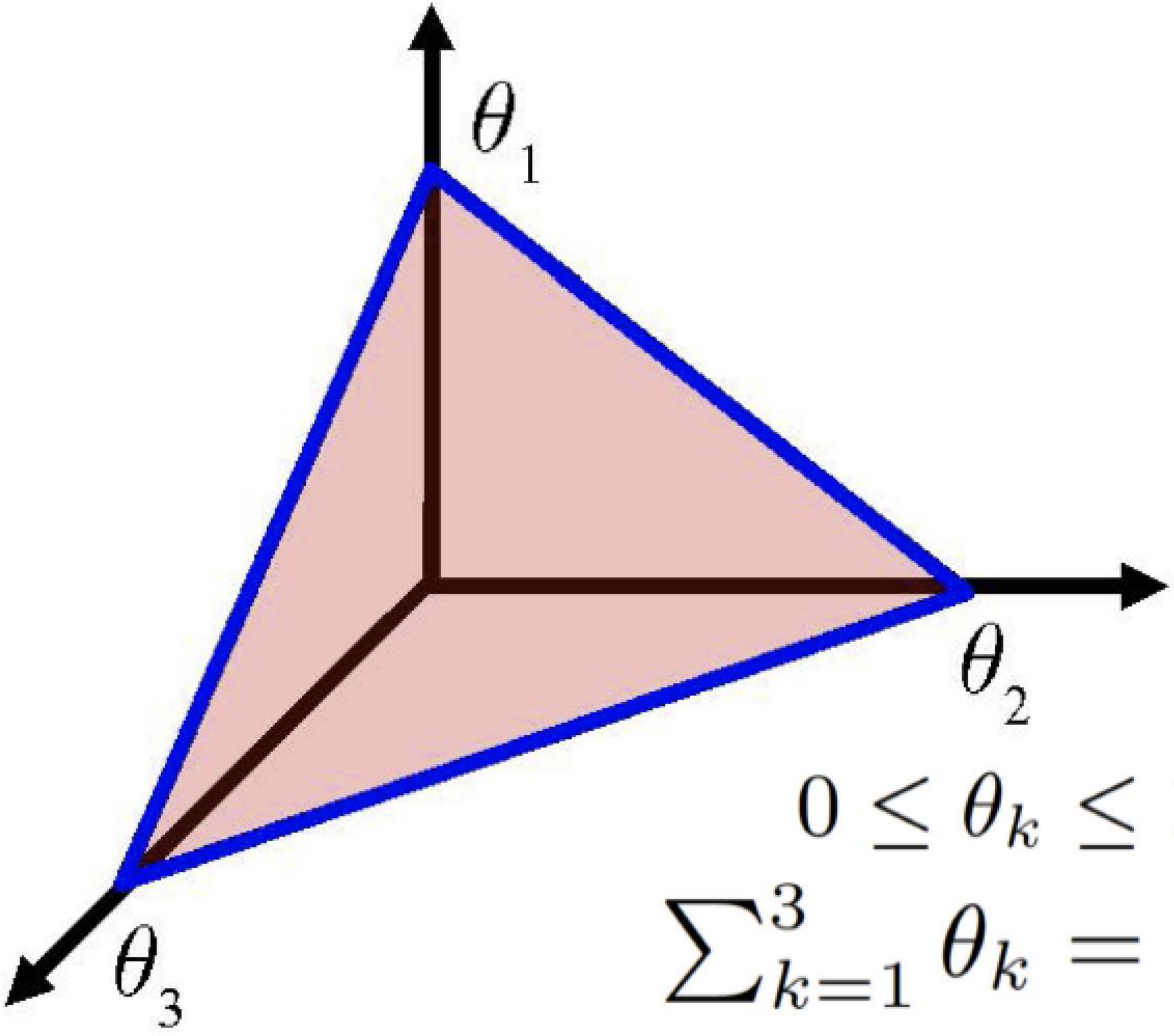
- Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\Theta \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

$$P(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Samples from the distribution lie in the *m-1* dimensional probability simplex

Multinomial Simplex



$$0 \le \theta_k \le 1$$

$$\sum_{k=1}^{3} \theta_k = 1$$

The Dirichlet Distribution

- Let $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\boldsymbol{\Theta} \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.
- Beta distribution is the special case of a Dirichlet for 2 dimensions.
- Thus, it is in fact a "distribution over distributions."

Dirichlet Process

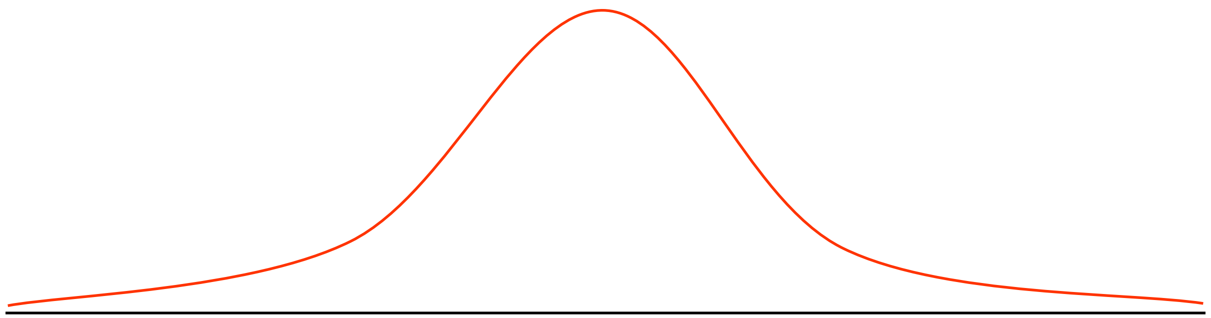
- A *Dirichlet Process* is also a distribution over distributions.
- Let G be Dirichlet Process distributed:

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

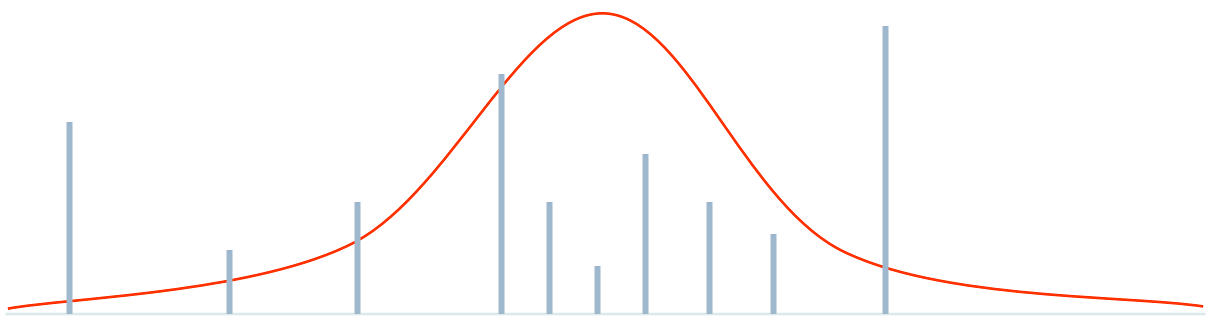
  - G_0 is a base distribution
  - α is a positive scaling parameter
- G is a random probability measure that has the same support as G_0

Dirichlet Process

- Consider Gaussian G_0

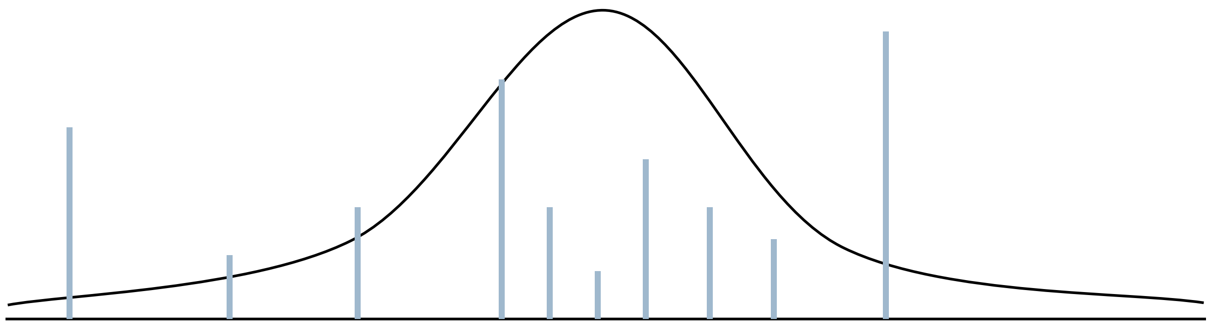


- $G \sim DP(\alpha, G_0)$



Dirichlet Process

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



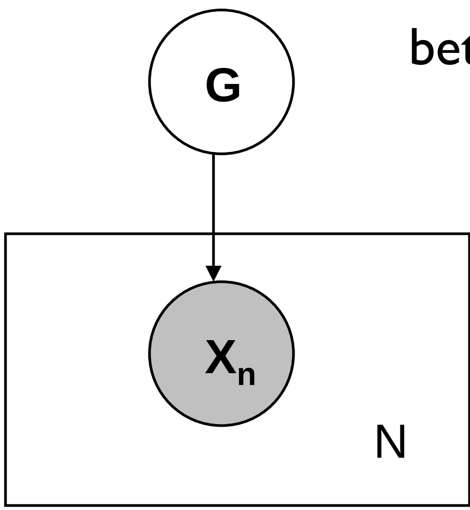
- G_0 is continuous, so the probability that any two samples are equal is precisely zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses [Blackwell]
  - Therefore, there is always a non-zero probability of two samples colliding

Samples from a Dirichlet Process

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

$X_n \mid G \sim G$ for $n = \{1, \ldots, N\}$ (iid given G)

Marginalizing out G introduces dependencies between the X_n variables



$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G)dG$$

Samples from a Dirichlet Process

$$P(X_1, \dots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n | G) dG$$

Assume we view these variables in a specific order, and are interested in the behavior of X_n given the previous n - 1 observations.

$$X_n | X_1, \dots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be K unique values for the variables:

$$X_k^* \text{ for } k \in \{1, \dots, K\}$$

Samples from a Dirichlet Process

$$X_n | X_1, \dots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$P(X_1, \dots, X_N) = P(X_1)P(X_2|X_1)\dots P(X_N|X_1, \dots, X_{N-1})$$

Chain rule

$$= \frac{\alpha^K \prod_{k=1}^{K} (\text{num}(X_k^*) - 1)!}{\alpha(1+\alpha)\dots(N-1+\alpha)} \prod_{k=1}^{K} G_0(X_k^*)$$

P(partition) **P(draws)**

Notice that the above formulation of the joint distribution does not depend on the order we consider the variables.

Samples from a Dirichlet Process

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be *K* unique values for the variables:

$$X_k^* \text{ for } k \in \{1, \ldots, K\}$$

Can rewrite as:

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_k^* & \text{with probability } \frac{\mathrm{num}_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Blackwell-MacQueen Urn Scheme

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

$$X_n \mid G \sim G$$

- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step *n*:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.

[Blackwell and Macqueen, 1973]

References

David Blackwell and James B. MacQueen. "Ferguson Distributions via Polya Urn Schemes." *Annals of Statistics* 1(2), 1973, 353-355.

David M. Blei and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." *Bayesian Analysis* 1(1), 2006.

Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems" *Annals of Statistics* 1(2), 1973, 209-230.

Zoubin Gharamani. "Non-parametric Bayesian Methods." UAI Tutorial July 2005.

Teg Grenager. "Chinese Restaurants and Stick Breaking: An Introduction to the Dirichlet Process"

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249-265, 2000.

C.E. Rasmussen. The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems* 12, 554-560. (Eds.) Solla, S. A., T. K. Leen and K. R. Müller, MIT Press (2000).

Y.W. Teh. "Dirichlet Processes." Machine Learning Summer School 2007 Tutorial and Practical Course.

Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. "Hierarchical Dirichlet Processes." *J. American Statistical Association* 101(476):1566-1581, 2006.