# Bayesian Theory and Computation

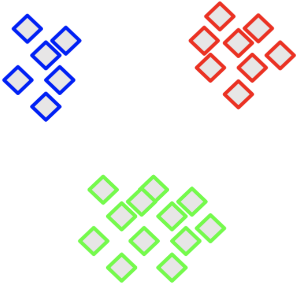
# Lecture 18: Dirichlet Processes

**Cheng Zhang**

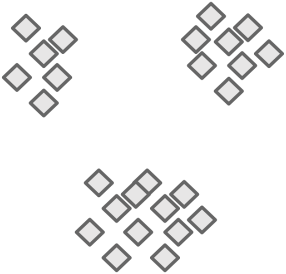
School of Mathematical Sciences, Peking University

June 01, 2021

How to choose the number of clusters?

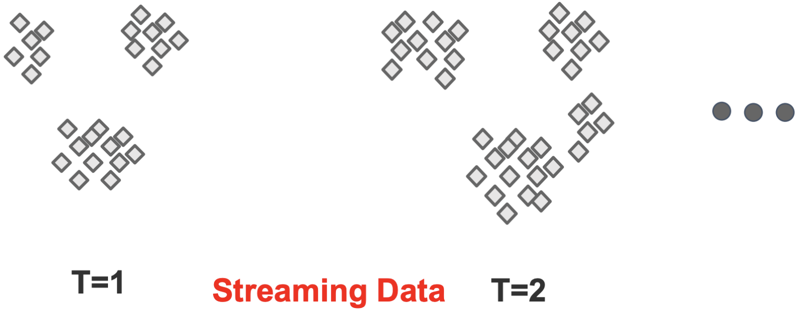


How to choose the number of clusters?



# Motivation via Clustering

How to choose the number of clusters?



# Finite Mixture Models

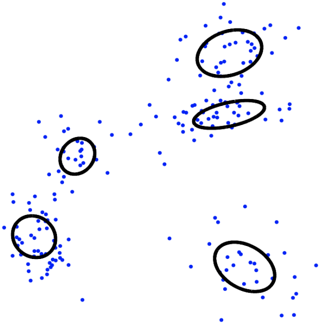


- A generative approach to clustering
  - pick one of $K$ clusters from a distribution $\pi = (\pi_1, \dots, \pi_K)$
  - generate a data point from a cluster-specific probability distribution
- This yields a finite mixture model:

$$p(x|\phi, \pi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$

  where $\pi$ and $\phi = (\phi_1, \dots, \phi_K)$ are the parameters, and here we assume the sae parameterized family for each cluster for simplicity.
- Data $\{x_i\}_{i=1}^N$ are assumed to be generated conditionally iid from this mixture model.

- For Gaussian mixtures, $\phi_k = (\mu_k, \Sigma_K)$ and $p(x|\phi_k)$ is a Gaussian density with mean $\mu_k$ and covariance matrix $\Sigma_k$



- Mixture models make the assumption that each data point arises from a single mixture component, i.e., the $k$th cluster is by definition the set of data points arising from the $k$th mixture component.
- Can capture this explicitly via a latent multinomial variable $Z$:

$$\begin{aligned} p(x|\phi, \pi) &= \sum_{k=1}^{K} p(Z = k|\pi) p(x|Z = k, \phi) \\ &= \sum_{k=1}^{K} \pi_k p(x|\phi_k) \end{aligned}$$

- Another way to express this: define an underlying measure

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

where $\delta_{\phi_k}$ is an *atom* (Dirac delta function) at $\phi_k$.

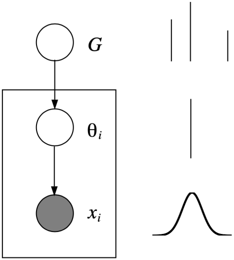
- Now we can redefine the process of obtaining a sampling from a finite mixture model as follows. For $i = 1, \ldots, n$:

$$\theta_i \sim G$$
$$x_i \sim p(\cdot|\theta_i)$$

- Note that each $\theta_i$ is equal to one of the underlying $\phi_k$. Indeed, the subset of $\{\theta_i\}$ that maps to $\phi_k$ is exactly the $k$th cluster

# Finite Mixture Models





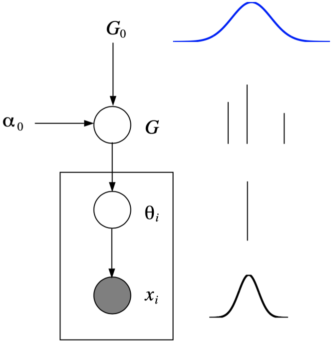

$$
\begin{aligned}
G &= \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot \,|\, \theta_i)
\end{aligned}
$$

Adapted from M. I. Jordan

- Bayesian approaches allow us to integrate out model parameters
- Need to place priors on the parameters $\phi$ and $\pi$
- The choice of prior for $\phi$ is model-specific; e.g., we may use conjugate normal/inverse-gamma priors for a Gaussian mixture model. Let us denote this prior as $G_0$.
- What to choose for the mixture weights $\pi$? A common choice is a symmetric Dirichlet prior, $\mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K)$
  - the symmetry accords with the common assumption of the order-free of the labels of the mixture components
  - the concentration parameter $\alpha_0$ controls concentration level of the labels

# Bayesian Finite Mixture Models

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi_k &\sim \mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \\
G &= \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot \,|\, \theta_i)
\end{aligned}
$$




- Note that $G$ is now a random measure

- Posterior distributions can't be found analytically; nor can predictive distributions (for future observations)
- However, a variety of MCMC sampling algorithms are available
- Use the indicators $Z$ within a Gibbs sampler. Give $Z$, we know which data points belong to which cluster, so:
  - $p(\pi|Z, \phi)$: standard multinomial-Dirichlet conjugacy
  - $p(\phi|Z, \pi)$: separate updates for each cluster; i.e., for each $\phi_k$ (and conjugacy of $G_0$ and $p(\cdot|\phi)$ can make this easy)
  - $p(Z|\pi, \phi)$: multinomial classification
- We can also use variational inference.

- How to choose $K$, the number of mixture components?
- Various generic model selection methods can be considered: e.g., cross-validation, bootstrap, AIC, BIC, DIC, Laplace, bridge sampling, etc
- Or we can place a parametric prior on $K$ (e.g., Poisson) and use Bayesian methods
- The Dirichlet process provides a nonparametric Bayesian alternative.

- Make sure we always have more clusters than we need.
- How about infinite clusters a priori?

$$p(x|\phi, \pi) = \sum_{k=1}^{\infty} \pi_k p(x|\phi_k)$$

- A finite data set will always use a finite, but random, number of clusters.
- How to choose the prior?
- We need something like a Dirichlet prior, but with an infinte number of components.

- Relation to gamma distribution: If $\eta_k \sim \text{Gamma}(\alpha_k, \beta)$ independently, then

$$S = \sum_k \eta_k \sim \text{Gamma}\left(\sum_k \alpha_k, \beta\right)$$

and

$$V = (v_1, \ldots, v_k) = (\eta_1/S, \ldots, \eta_k/S) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$$

- Therefore, if $(\pi_1, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ then

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

This is known as the collapsing property.

- The beta distribution is a Dirichlet distribution on the 1-simplex
- Let $(\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ and $\theta \sim \mathrm{Beta}(\alpha_1 b, \alpha_1(1-b)),\ 0 < b < 1$.
- Then

$$(\pi_1\theta, \pi_1(1-\theta), \pi_2, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 b_1, \alpha_1(1-b_1), \alpha_2, \ldots, \alpha_K)$$

- More generally, if $\theta \sim \mathrm{Dir}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_N),\ \sum_i b_i = 1$, then

$$(\pi_1\theta_1, \ldots, \pi_1\theta_N, \pi_2, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 b_1, \ldots, \alpha_1 b_N, \alpha_2, \ldots, \alpha_K)$$

This is known as the splitting property.

- Renormalization. If $(\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$, and

$$V = (V_2, V_3, \ldots, V_K), \quad V_k = \frac{\pi_k}{\sum_{k \geq 2} \pi_k}$$

- What is the distribution of V?

$$V \sim \mathrm{Dir}(\alpha_2, \ldots, \alpha_K)$$

- All these properties can be easily verified using the aforementioned gamma distribution representation.

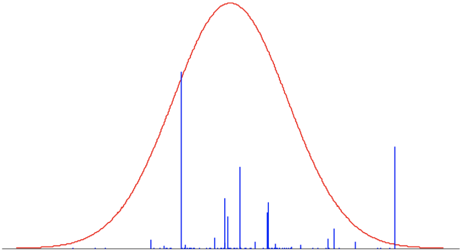
- Let $G_0$ be a distribution on some space $\Omega$, e.g. a Gaussian distribution on the real line.
- Assume that $\pi, \phi$ have the following distributions

$$\phi_k \sim G_0$$
$$\pi \sim \lim_{K\to\infty} \mathrm{Dir}\left(\frac{\alpha_0}{K}, \dots, \frac{\alpha_0}{K}\right)$$

- Then $G := \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ defines an infinite distribution over $G_0$.
- We say (informally) that $G$ follows a Dirichlet Process

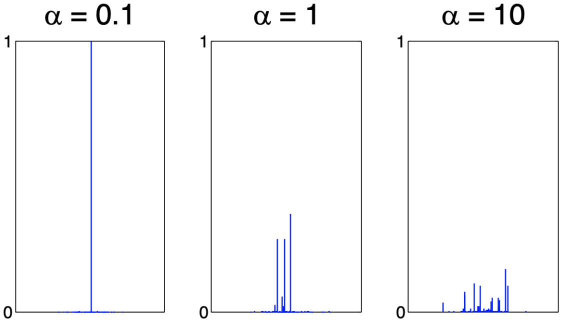
$$G \sim \mathrm{DP}(\alpha_0 G_0)$$

- Samples from the Dirichlet process are *discrete*.
- We call the point masses in the resulting distribution, atoms.



- The *base measure* $G_0$ determines the *locations* of the atoms.

- The *concentration parameter* $\alpha_0$ determines the distribution over atom sizes.
- Small values of $\alpha_0$ gives *sparse* distributions.



# Dirichlet Process: A Formal Definition

- Let $(\Omega, \mathcal{B})$ be a measurable space, with $G_0$ a probability measure on the space, and let $\alpha_0$ be a positive real number.
- A Dirichlet process is the distribution of a random probability measure $G$ over $(\Omega, \mathcal{B})$ such that, for any finite partition $(A_1, \ldots, A_r)$ of $\Omega$, the random vector $(G(A_1), \ldots, G(A_r))$ follows a finite-dimensional Dirichlet distribution:

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G(A_r))$$

- We write $G \sim \mathrm{DP}(\alpha_0 G_0)$, and call $G_0$ the base measure, $\alpha_0$ the concentration parameter.

- Let $A_1, \ldots, A_K$ be a partition of $\Omega$. Let $G(A_k)$ be the mass assigned by $G \sim \mathrm{DP}(\alpha_0 G_0)$ to partition $A_k$. Then

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_K))$$

- If we see an observation in the $j$-th segment, then

$$\begin{aligned}(G(A_1), \ldots, G(A_K)|\theta_1 \in A_j) \\ \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G(A_j) + 1, \ldots, \alpha_0 G_0(A_K)).\end{aligned}$$

- This is true for all possible partitions of $\Omega$.
- Therefore, the posterior distribution of $G$, given an observation $\phi$, is given by

$$G|\theta_1 = \phi \sim \mathrm{DP}(\alpha_0 G_0 + \delta_\phi)$$

# Predictive Distribution

- The Dirichlet process clusters observations.
- A new data point can either join an existing cluster, or start a new cluster.
- Question: What is the predictive distribution for a new data point?
- Assume $G_0$ is a continuous distribution on $\Omega$. This means for every point $\phi$ in $\Omega$, $G_0(\phi) = 0$.
- First data point:
  - Start a new cluster
  - Sample a parameter $\phi_1 \sim G_0$ for that cluster.

- We have now split our parameter space in two: the singleton $\phi_1$, and everything else.
- Let $\pi_1$ be the size of atom at $\phi_1$.
- The combined mass of all the other atoms is $\pi_* = 1 - \pi_1$.
- According to the DP,

$$(\pi_1, \pi_*) \sim \mathrm{Dir}(0, \alpha_0)$$

- Given $\theta_1 = \phi_1$, the posterior is

$$(\pi_1, \pi_*) | \theta_1 = \phi_1 \sim \mathrm{Dir}(1, \alpha_0)$$

# Predictive Distribution

- If we integrate out $\pi_1$, we get

$$
\begin{aligned}
p(\theta_2 = \phi_k | \theta_1 = \phi_1) &= \int p(\theta_2 = \phi_k | (\pi_1, \pi_*)) p((\pi_1, \pi_*) | \theta_1 = \phi_1) d\pi_1 \\
&= \int \pi_k \mathrm{Dir}((\pi_1, 1 - \pi_1) | 1, \alpha_0) d\pi_1 \\
&= \mathbb{E}_{\mathrm{Dir}(1,\alpha_0)} \pi_k \\
&= \begin{cases} \frac{1}{1+\alpha_0} & \text{if } k = 1 \\ \frac{\alpha_0}{1+\alpha_0} & \text{for new } k. \end{cases}
\end{aligned}
$$

# Predictive Distribution



- Lets say we choose to start a new cluster, and sample a new parameter $\phi_2 \sim G_0$. Let $\pi_2$ be the size of the atom at $\phi_2$.
- Similarly, the posterior is

$$(\pi_1, \pi_2, \pi_*)|\theta_1 = \phi_1, \theta_2 = \phi_2 \sim \mathrm{Dir}(1, 1, \alpha_0)$$

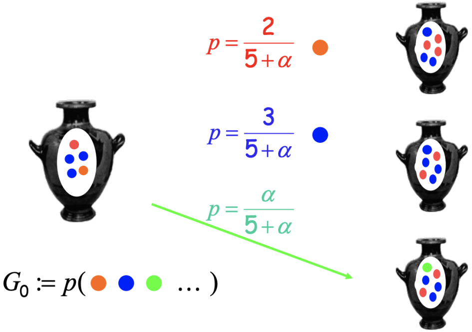
- If we integrate out $\pi = (\pi_1, \pi_2, \pi_*)$, we get

$$\begin{aligned} p(\theta_3 = \phi_k|\theta_1 = \phi_1, \theta_2 = \phi_2) \\ &= \int p(\theta_3 = \phi_k|\pi)p(\pi|\theta_1 = \phi_1, \theta_2 = \phi_2)d\pi \\ &= \mathbb{E}_{\mathrm{Dir}(1,1,\alpha_0)}\pi_k \\ &= \begin{cases} \frac{1}{2+\alpha_0} & \text{if } k = 1 \\ \frac{1}{2+\alpha_0} & \text{if } k = 2 \\ \frac{\alpha_0}{2+\alpha_0} & \text{for new } k. \end{cases} \end{aligned}$$

- In general, if $m_k$ is the number of times we have seen $X_i = k$, and $K$ is the total number of observed values,

$$
\begin{aligned}
p(\theta_{n+1} = \phi_k | \theta_1, \ldots, \theta_n) &= \int p(\theta_{n+1} = \phi_k | \pi) p(\pi | \theta_1, \ldots, \theta_n) d\pi \\
&= \mathbb{E}_{\mathrm{Dir}(m_1, \ldots, m_K, \alpha_0)} \pi_k \\
&= \begin{cases} \frac{m_k}{n+\alpha_0} & \text{if } k \leq K \\ \frac{\alpha_0}{n+\alpha_0} & \text{for new cluster.} \end{cases}
\end{aligned}
$$

- We tend to see observations that we have seen before, i.e., rich-get-richer property
- We can always add new features, a typical nonparametric behavior.



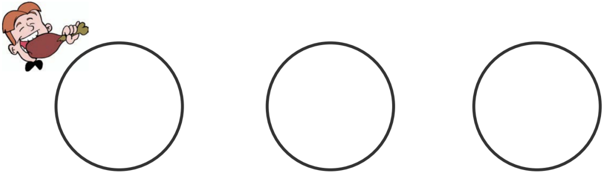

Adapted from Eric Xing

- Joint: $G($  $) \sim \mathrm{DP}(\alpha_0 G_0)$
- Marginal: $\theta_{n+1} | \theta_{\leq n}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{n+\alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{n+\alpha_0} G_0.$

- The resulting distribution over data points can be thought of using the following urn scheme (Blackwell and MacQueen, 1973).
- An urn initially contains a black ball of mass $\alpha_0$.
- For $n = 1, 2, \ldots$, sample a ball from the urn with probability proportional to its mass.
- If the ball is black, choose a previously unseen color, record that color, and return the black ball plus a unit-mass ball of the new color to the urn.
- If the ball is not black, record it's color and return it, plus another unit-mass ball of the same color, to the urn.

- The distribution over partitions can also be described in terms of the following restaurant metaphor:

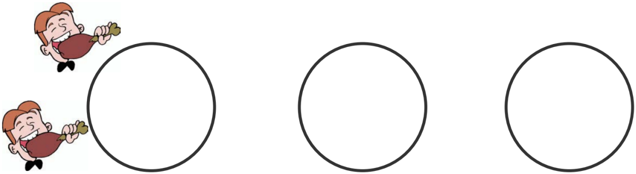
- The distribution over partitions can also be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.



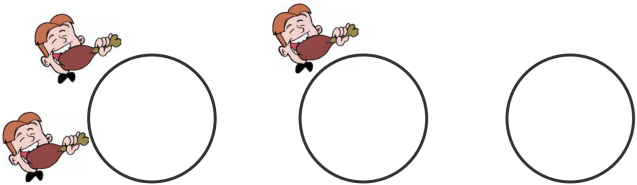
# Chinese Restaurant Process



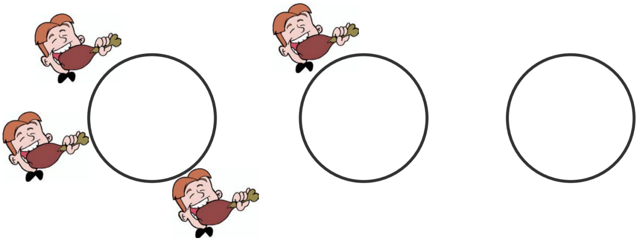
- The distribution over partitions can also be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
- The $n$-th customer enters the restaurant. He sits at an existing table with probability $\frac{m_k}{n-1+\alpha_0}$, where $m_k$ is the number of people sat the table $k$. He starts a new table with probability $\frac{\alpha_0}{n-1+\alpha_0}$.



- The distribution over partitions can also be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
- The $n$-th customer enters the restaurant. He sits at an existing table with probability $\frac{m_k}{n-1+\alpha_0}$, where $m_k$ is the number of people sat the table $k$. He starts a new table with probability $\frac{\alpha_0}{n-1+\alpha_0}$.



- The distribution over partitions can also be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
- The $n$-th customer enters the restaurant. He sits at an existing table with probability $\frac{m_k}{n-1+\alpha_0}$, where $m_k$ is the number of people sat the table $k$. He starts a new table with probability $\frac{\alpha_0}{n-1+\alpha_0}$.



- An interesting fact: the distribution over the clustering of the first $N$ customers does not depend on the order in which they arrived.
- However, the customers are not independent. They tend to sit at popular tables.
- We say that distributions like this are *exchangeable*.

$$p(\theta_1, \ldots, \theta_N) = p(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)})$$

- By **de Finetti**'s theorem, there exists a random distribution $G$ and a prior $P(G)$ such that

$$p(\theta_1, \ldots, \theta_N) = \int \prod_{i=1}^{N} G(\theta_i) dP(G)$$

- In our setting, the prior $p(G)$ is just $\mathrm{DP}(\alpha_0 G_0)$, thus establishing existence.

# A Second Perspective: Stick Breaking



- Define an infinite sequence of Beta random variables:

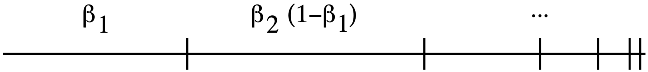
$$\beta_k \sim \text{Beta}(1, \alpha_0), \quad k = 1, 2, \ldots$$

- Now define an infinite sequence of mixing proportions as:

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell), \quad k = 2, 3, \ldots$$

- This can be viewed as breaking off portions of a stick:



- We now have an explicit formula for each $\pi_k$:

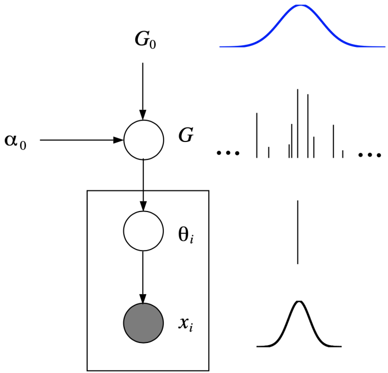
$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

- We can easily see that $\sum_{k=1}^{\infty} \pi_k = 1$:

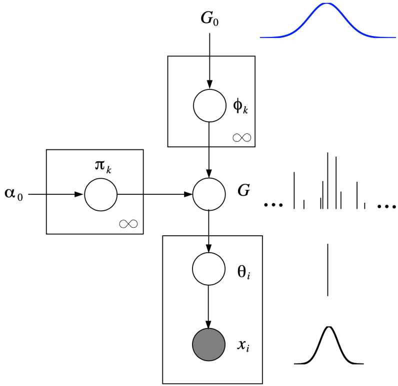
$$\begin{aligned} 1 - \sum_{k=1}^{\infty} \pi_k &= 1 - \beta_1 - \beta_2(1 - \beta_1) - \beta_3(1 - \beta_1)(1 - \beta_2) - \cdots \\ &= (1 - \beta_1)(1 - \beta_2 - \beta_3(1 - \beta_2) - \cdots) \\ &= \prod_{k=1}^{\infty} (1 - \beta_k) = 0 \end{aligned}$$

- Let $\phi_k \sim G_0, \forall k$, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure. In fact,

$$G \sim \mathrm{DP}(\alpha_0 G_0).$$




Polya urn construction




Stick breaking construction

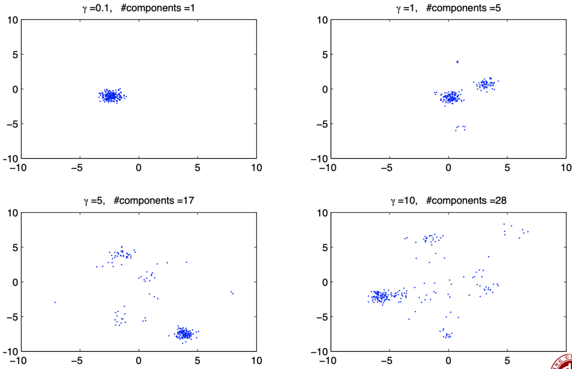
- Now we can use a Dirichlet process as the prior for an unknown mixture distribution (with potentially infinite mixture components).
- Suppose we have $x_1, \ldots, x_n$ observations from some unknown distribution.
- We can model the unknown distribution of $x$ as a mixture of simple distributions of the form $f(\cdot|\theta)$.
- We denote the mixing distribution over $\theta$ as $G$ and let the prior over $G$ be a Dirichlet process

$$x_i|\theta_i \sim f(\theta_i)$$
$$\theta_i|G \sim G$$
$$G \sim \mathrm{DP}(\alpha_0 G_0)$$

# Samples from DP Mixture Prior



- Multiple subjects can be mapped to the same $\phi$. This creates a clustering of subjects.
- The following graphs shows 4 different data sets ($n = 200$) randomly generated from distributions sampled from Dirichlet process mixture priors with $\alpha_0$.



- We can integrate out $G$ to get the CRP. Note that the CRP is exchangeable, which induces the conditional priors

$$p(\theta_i|\theta_{-i},\alpha_0,G_0) = \frac{\alpha_0}{n-1+\alpha_0}G_0(\theta_i) + \sum_{k=1}^{K^{(-i)}} \frac{m_k^{(-i)}}{n-1+\alpha_0}\delta_{\phi_k^{(-i)}}$$

- Let $z_i$ be the cluster allocation of the $i$-th data point. The collapsed Gibbs sampler alternates between
  - update $z_i$

$$p(z_i = k|x_i, z_{-i}, \phi_{1:K}) \propto \begin{cases} m_k^{(-i)} f(x_i|\phi_k^{(-i)}) & k \leq K^{(-i)} \\ \alpha_0 \int f(x_i|\theta)dG_0(\theta) & k = K^{(-i)} + 1 \end{cases}$$

  - update $\phi_k$

$$p(\phi_k|z_{1:n}, x_{1:n}) \propto G_0(\phi_k) \prod_{i:z_i=k} f(x_i|\phi_k)$$

- If $G_0$ is conjugate to $f$, the above steps can be evaluated accurately.

- For the concentration parameter $\alpha_0$, we have

$$p(K|\alpha_0) \propto \alpha_0^K \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)}$$

where $K$ is the number of unique $\phi$'s (e.g., the number of clusters).

- Therefore, given $K$ and the prior distribution $P(\alpha_0)$ we can sample from the posterior distribution of $\alpha_0$ using the MH algorithm or the Gibbs sampling method of Escobar and West (1995).

- We are only updating one data point at a time.
- Imagine two "true" clusters are merged into a single cluster, a single data point is unlikely to "break away".
- Getting to the true distribution involves going through low probability states, i.e., mixing can be slow.
- If the likelihood is not conjugate, integrating out parameter values for new features can be difficult.
- Neal (2000) offers a variety of algorithms.

- The stick-breaking representation orders the mixture components so that the weights are stochastically decreasing. For a sufficiently large $T$, we will have $\sum_{k>T} \pi_k \approx 0$.
- Therefore, we can truncate the stick-breaking construction at a fixed value $T$ and let $\beta_T = 1$.
- This implies $\pi_k = 0, \forall k > T$, and the distribution of

$$G_T = \sum_{k=1}^{T} \pi_k \delta_{\phi_k}$$

is known as a truncated Dirichlet process.

- Variational distance between distributions of marginals from a DP and from its truncation at $T$ is approximately $4n \exp(-(T-1)/\alpha_0)$. $T$ doesn't have to be very large to get a good approximation.

- State of the Markov chain: $(\beta_{1:T-1}, \phi_{1:T}, z_{1:n})$.
- Update $z_i$ by multinomial sampling with

$$p(z_i = k|\beta, \phi, x_i) \propto \pi_k f(x_i|\phi_k)$$

- Update $\beta_k$ by sampling from the conditional posterior

$$\beta_k \sim \mathrm{Beta}(1 + m_k, \alpha_0 + \sum\nolimits_{j>k} m_j)$$

- Update $\phi_k$ by sampling from the conditional posterior

$$p(\phi_k|z_{1:n}, x_{1:n}) \propto G_0(\phi_k) \prod_{i:z_i=k} f(x_i|\phi_k)$$

- One can monitor $\max_i z_i$ to verify that truncation at $T$ is good enough, and increase $T$ if necessary.

- We can also use truncated steak-breaking representation to form a mean field approximation of DP mixtures

$$q(\beta, \phi, z) = \prod_{i=1}^{n} q(z_i|w_i) \prod_{k=1}^{T} q(\phi_k|\tau_k) \prod_{k=1}^{T-1} q(\beta_k|\gamma_k)$$

- For a conjugate DP mixture in the exponential family

$$\gamma_{k,1} = 1 + \sum\nolimits_{i=1}^{n} w_{i,k}, \quad \gamma_{k,2} = \alpha_0 + \sum\nolimits_{i=1}^{n} \sum\nolimits_{j>k} w_{i,j}$$
$$\tau_{k,1} = \lambda_1 + \sum\nolimits_{i=1}^{n} w_{i,k} t(x_i), \quad \tau_{k,2} = \lambda_2 + \sum\nolimits_{i=1}^{n} w_{i,k}$$
$$w_{i,k} \propto \exp(S_k)$$

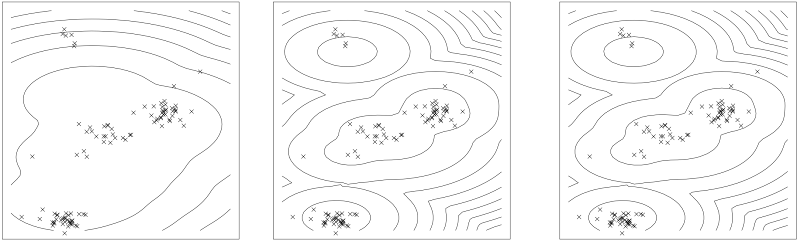
where

$$S_k = \mathbb{E} \log \beta_k + \sum\nolimits_{j<k} \mathbb{E} \log(1 - \beta_j) + \mathbb{E}\phi_k^T t(x_i) - \mathbb{E}A(\phi_k)$$

# Example: DP Gaussian Mixture



- The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.



Initial state | 1st iteration | 5th (and last) iteration

- Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174..
- Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. The Annals of Statistics, 1(2):353–355.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1:209–230.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90:577–588.

- Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265.
- Blei, D. M. and Jordan, M. I. (2005). Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1(1):121-144.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581, 2006.