

# STATS 225: Bayesian Analysis

## Dirichlet Process Mixture Models

Babak Shahbaba

Department of Statistics, UCI

- In this lecture, we are going to introduce Dirichlet process mixture models, which is usually used for nonparametric density estimation and clustering.
- A Dirichlet process,  $\mathcal{D}(G_0, \gamma)$ , with *base distribution*  $G_0$  and *scale or concentration* parameter  $\gamma$  is a distribution over distributions.
- More formally, we say a random probability measure  $G$  is distributed according to  $\mathcal{D}(G_0, \gamma)$  if for any finite partition of  $(A_1, \dots, A_n)$  of space  $\Omega$ , the random vector  $(G(A_1), \dots, G(A_n))$  has a finite dimensional Dirichlet distribution as follows:

$$(G(A_1), \dots, G(A_n)) \sim \text{Dir}(\gamma G_0(A_1), \dots, \gamma G_0(A_n))$$

- Using the Polya urn scheme, Blackwell and MacQueen (1973) showed that the distributions sampled from a Dirichlet process are discrete almost surely.

- Ferguson (1973) introduced the Dirichlet process as a class of prior distributions for which the support is large, and the posterior distribution is manageable analytically,

$$G|\theta_1, \dots, \theta_n \sim \mathcal{D}\left(\frac{\gamma}{\gamma + n} G_0 + \frac{n}{\gamma + n} P_n, \gamma + n\right)$$

where  $P_n$  is the empirical distribution.

- Antoniak (1974) proposed the idea of using a Dirichlet process as the prior for the mixing proportions of a simple distribution (e.g., Gaussian).
- First, let's review simple mixture distributions from the Bayesian point of view.

# Mixture distributions

- Consider a random sample,  $x_1, \dots, x_n$ , drawn independently from some unknown distribution.
- We can use a mixture of simple distributions to model the density of  $X$ :

$$P(x_i|\pi, \phi) = \sum_{k=1}^K \pi_k P_k(x_i|\phi_k)$$

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

Here,  $\pi_k$  are the mixing proportion, and  $P_k$  is a simple distribution such as normal with  $\phi = (\mu, \sigma)$ .

- A common prior for  $\pi_k$  is a symmetric Dirichlet distribution

$$P(\pi_1, \dots, \pi_K) = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \prod_{k=1}^K \pi_k^{(\gamma/K)-1}$$

- $\phi_c$  are assumed to be independent under the prior with distribution  $G_0$ .

- For a finite  $K$ , we can represent the mixture model as follows:

$$\begin{aligned}\phi_k &\sim G_0 \\ \pi_1, \dots, \pi_K &\sim \text{Dirichlet}(\gamma/K, \dots, \gamma/K) \\ z_i | \pi_1, \dots, \pi_K &\sim \text{Discrete}(\pi_1, \dots, \pi_K) \\ x_i | z_i, \phi &\sim P_k(x_i | \phi_{k_i})\end{aligned}$$

- By integrating over the Dirichlet prior, we obtain the following conditional distribution for  $z_i$ :

$$P(z_i = k | z_1, \dots, z_{i-1}) = \frac{n_{ik} + \gamma/K}{i - 1 + \gamma}$$

where,  $n_{ik}$  represents the number of data points previously (i.e., before the  $i^{th}$ ) assigned to component  $k$ .

- We might not know the number of components,  $K$ , *a priori*. We can assume  $K \rightarrow \infty$ ,

$$\begin{aligned}P(z_i = k | z_1, \dots, z_{i-1}) &\rightarrow \frac{n_k}{i-1+\gamma} \\P(z_i \neq z_j, \forall j < i | z_1, \dots, z_{i-1}) &\rightarrow \frac{\gamma}{i-1+\gamma}\end{aligned}$$

- Now consider the specific parameters associated with observation  $i$ , denoted as  $\theta_i$ . Note that  $\theta$ 's are not unique: multiple subjects can have the same parameters. We refer to unique values of  $\theta$  as  $\phi$ )
- The successive conditional distribution of  $\theta_i$  is

$$\theta_i | \theta_1, \dots, \theta_{i-1} \sim \frac{1}{i-1+\gamma} \sum_{j < i} \delta(\theta_j) + \frac{\gamma}{i-1+\gamma} G_0$$

# Dirichlet process mixture

- Because of exchangeability, we can treat  $i$  as the last observation so we can use this to write the conditional distribution of  $\theta_i$  given all other  $\theta$ 's,

$$\theta_i | \theta_{-i} \sim \frac{1}{n-1+\gamma} \sum_{j \neq i} \delta(\theta_j) + \frac{\gamma}{n-1+\gamma} G_0$$

- The resulting model is equivalent to the Dirichlet process mixture model.

# Dirichlet process mixture model

- The idea of using a Dirichlet process as the prior for the mixing proportions of a simple distribution (e.g., Gaussian) was first introduced by Antoniak (1974).
- Suppose we have  $x_1, \dots, x_n$  observations from some unknown distribution.
- We can model the unknown distribution of  $x$  as a mixture of simple distributions of the form  $F(\theta)$
- We denote the mixing distribution over  $\theta$  as  $G$  and let the prior over  $G$  be a Dirichlet process:

$$x_i | \theta_i \sim F(\theta_i)$$

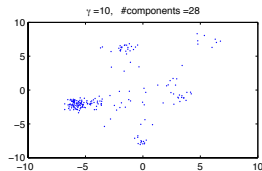
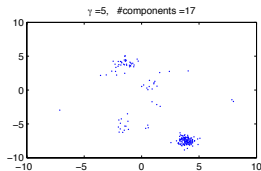
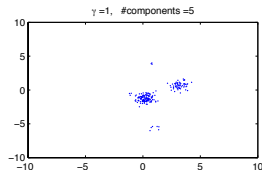
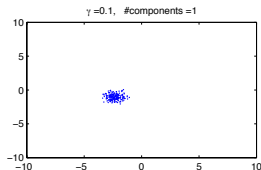
$$\theta_i | G \sim G$$

$$G \sim \mathcal{D}(G_0, \gamma)$$



# Samples from Dirichlet process mixture prior

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



# Chinese restaurant process

- Dirichlet process is sometimes explained as a “Chinese restaurant” process.
- Consider a restaurant with infinite possibilities, i.e., infinite menu items and infinite number of tables.
- The rule is that people at the same table must order the same food.
- First customer comes and sits anywhere she wants and order whatever she wants. The second customer can sit with the first one (therefore, order the same food), or he can sit by himself and order something different, and so on.
- Note that tables with more customers are more attractive.

# Posterior sampling

- Given a finite set of observations, we want to find the posterior distribution of model parameters.
- Using the Chinese restaurant process analogy, we assume that our objective is to cluster customers according to their taste of food.
- We start by randomly allocating customers to a set of tables.
- Then, one-by-one we let them move around and change their table if they want until they find a table that matches their taste.
- When this procedure converges, customers form few clusters.
- Neal (2000) discussed several MCMC algorithms for Dirichlet process mixtures.

- When conjugate priors are used, the conditional distribution of  $\theta_i$  becomes

$$\theta_i | \theta_{-i}, x_i \sim \sum_{j \neq i} q_{ij} \delta(\theta_j) + r_i H_i$$

- Here,  $H_i$  is the posterior distribution of  $\theta$  given the conjugate prior  $G_0$  and the single observation  $x_i$ .
- Using the likelihood  $F(x_i, \theta)$ , we obtain  $q_{ij}$  and  $r_i$  as follows:

$$\begin{aligned} q_{ij} &= b F(x_i, \theta_j) \\ r_i &= b \gamma \int F(x_i, \theta) dG_0(\theta) \end{aligned}$$

such that  $\sum_{j \neq i} q_{ij} + r_i = 1$  (we use this to find  $b$ ).

# Gibbs sampling for conjugate priors

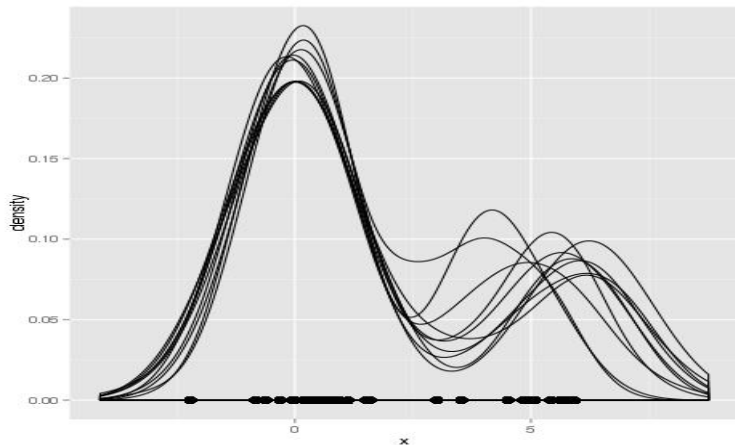
- Given the current state of the Markov chain,  $c = \{c_1, \dots, c_n\}$  and  $\phi = (\phi_c)$ , a simple Gibbs sampling algorithm is as follows (Algorithm 2 in Neal, 2000):
  - For  $i = 1, \dots, n$ , find  $n_{-i, c_i}$ , i.e., the number of observations  $j \neq i$  that are assigned to  $c_i$
  - If  $n_{-i, c_i} = 0$ , remove  $\phi_{c_i}$  from the state
  - Then sample a new value for  $c_i$  according to the following conditional probabilities:

$$P(c_i = c | c_{-i}, x_i, \phi) = b \frac{n_{-i, c}}{n - 1 + \gamma} F(y_i, \phi_c)$$

$$P(c_i \neq c_j, \forall j \neq i | c_{-i}, x_i, \phi) = b \frac{\gamma}{n - 1 + \gamma} \int F(x_i, \theta) dG_0(\theta)$$

- For all  $c = \{c_1, \dots, c_n\}$ , sample a new value from  $\phi_c | x_i$  such that  $c_i = c$
- The last step is called “remixing”: after allocating the subjects into clusters with unique  $\phi$ 's, we perform draw a new value from  $\phi_c | x_i$  given observations with  $c_i = c$  only.

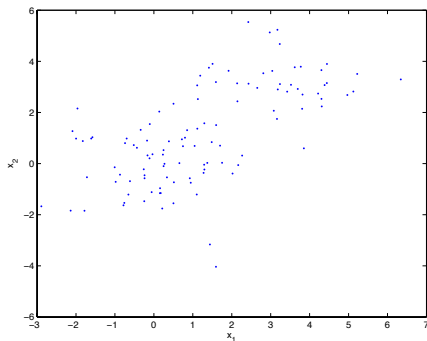
# DPM of univariate Gaussians



10 Posterior samples using the Gibbs algorithm of Escobar and West (1995), which is Algorithm 1 in Neal (2000).

# Dirichlet process mixture models for clustering

- In the following movie (click on the figure), data are sampled from two different bivariate normals (i.e., two clusters). A Dirichlet process mixture model (MCMC samples are shown) correctly identifies these two clusters: blue and red.



- Neal (2000) discusses several algorithms for non-conjugate priors. Here is Algorithm 8:
  - ▶ For  $i = 1, \dots, n$ , find  $n_{-i, c_i}$
  - ▶ If  $n_{-i, c_i} \neq 0$  (i.e.,  $i$  is not the only observation in its cluster), let  $k$  be the number of unique remaining  $\phi$ 's; sample  $m$  (e.g.,  $m = 5$ ) new  $\phi$ 's from  $G_0$  as *auxiliary* components.
  - ▶ If  $n_{-i, c_i} = 0$  (i.e.,  $i$  is the only observation in its cluster), remove  $\phi_{c_i}$  from the state; let  $k$  be the number of unique remaining  $\phi$ 's; sample  $m - 1$  auxiliary components  $\phi$ 's from  $G_0$ , and set the  $m^{\text{th}}$  one to  $\phi_{c_i}$ .
  - ▶ Then sample a new value for  $c_i$  according to the following conditional probabilities:

$$P(c_i = c | c_{-i}, x_i, \phi) = \begin{cases} b \frac{n_{-i, c}}{n-1+\gamma} F(x_i, \phi_c) & \text{for existing components} \\ b \frac{\gamma/m}{n-1+\gamma} F(x_i, \phi_c) & \text{for auxiliary components} \end{cases}$$

- ▶ Only keep the components that are associated with at least one observation.
- ▶ For all  $c = \{c_1, \dots, c_n\}$ , sample a new value from  $\phi_c | x_i$  such that  $c_i = c$



# Sampling the scale parameter

- For the scale parameter  $\gamma$ , we have (Antoniak, 1974)

$$P(k|\gamma) \propto \gamma^k \frac{\Gamma(\gamma)}{\Gamma(\gamma + n)}$$

where  $k$  is the number of unique  $c_i$  (or  $\phi$ 's).

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