Mickey Warner
AMS 241

# Problem 1

## The DP mixture model

A simulated dataset consisting of $n = 250$ random draws from the mixture of normals $0.2N(-5,1) + 0.5N(0,1) + 0.3N(3.5,1)$ will be analyzed in this problem. We consider the location normal Dirichlet process mixture model

$$f(\cdot|G,\phi) = \int k_N(\cdot|\theta,\phi)dG(\theta), \quad G|\alpha,\mu,\tau^2 \sim DP(\alpha, G_0 = N(\mu,\tau^2))$$

where $k_N(\cdot|\theta,\phi)$ is the density function of a normal distribution with mean $\theta$ and variance $\phi$. Hence, we are mixing over the location of the normal distribution. The hierarchical version of the model is given by

$$\begin{aligned} y_i|\theta_i,\phi &\overset{ind}{\sim} k_N(y_i|\theta_i,\phi), \quad i = 1,\ldots,n \\ \theta_i|G &\overset{iid}{\sim} G, \quad i = 1,\ldots,n \\ G|\alpha,\mu,\tau^2 &\sim DP(\alpha, G_0 = N(\mu,\tau^2)) \\ \alpha,\mu,\tau^2,\phi &\sim p(\alpha)p(\mu)p(\tau^2)p(\phi) \end{aligned}$$

The priors on $\alpha,\mu,\tau^2,\phi$ are chosen for convenience in the sampling. We will discuss the actual choices in the next section.

Posterior inference is made by sampling from the marginal posterior $p(\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y})$, where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ are the latent mixing parameters and $\mathbf{y} = (y_1,\ldots,y_n)$ are the data. This marginal posterior is found by integrating out the infinite-dimensional parameter $G$ from the full posterior distribution

$$p(\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y}) = \int p(G,\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y})dG$$

or, rather, by noting that the full posterior may be factored, using Bayes' formula, into a product of the full conditional of $G$ and the marginal posterior

$$p(G,\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y}) = p(G|\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi,\mathbf{y})p(\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y}).$$

## Full conditionals of the marginal posterior

To simulate from $p(\boldsymbol{\theta},\alpha,\mu,\tau^2,\phi|\mathbf{y})$ we iteratively draw from the full conditionals of each parameter $\theta_1,\ldots,\theta_n,\alpha,\mu,\tau^2,\phi$. The expressions for the condtional distributions are based on the Pólya urn representation. Before we present the distributions, we introduce some notation.

Since $G$ is almost surely discrete there will be a clustering among the $\theta_i$s and the Gibbs sampler we employ takes advantage of this fact. The following list describes the notation used throughout this section:

- $n^*$ denotes the number of distinct $\theta_i$s
- $\theta_j^*$, $j = 1, \ldots, n^*$ are the distinct $\theta_i$s
- $\mathbf{w} = (w_1, \ldots, w_n)$ is the vector that matches each $\theta_i$ to its corresponding $\theta_j^*$, i.e., $w_i = j$ if and only if $\theta_i = \theta_j^*$
- $n_j$ is the size of the $j$th cluster, $|\{i : w_i = j\}|$, $j = 1, \ldots, n^*$

The vectors $(n^*, \mathbf{w}, \theta_1^*, \ldots, \theta_{n^*}^*)$ and $(\theta_1, \ldots, \theta_n)$ are equivalent. The former will simplify the calculations to follow.

For each $\theta_i, i = 1, \ldots, n$, the full conditional $p(\theta_i|\{\theta_k : k \neq i\}, \alpha, \mu, \tau^2, \phi, \mathbf{y})$ is given by

$$\begin{aligned}
&\frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} h(\theta_i|\mu, \tau^2, \phi, y_i) + \sum_{j=1}^{n^{*-}} \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} \delta_{\theta_j^{*-}}(\theta_i) \\
&= A h(\theta_i|\mu, \tau^2, \phi, y_i) + \sum_{j=1}^{n^{*-}} B_j \delta_{\theta_j^{*-}}(\theta_i)
\end{aligned}$$

where

- $q_j = k_N(y_i|\theta_j^*, \phi)$,
- $q_0 = \int k_N(y_i|\theta, \phi) g_0(\theta|\mu, \tau^2) d\theta$,
- $h(\theta_i|\mu, \tau^2, \phi, y_i) \propto k_N(y_i|\theta_i, \phi) g_0(\theta_i|\mu, \tau^2)$,
- $g_0$ is the density of $G_0 = N(\cdot|\mu, \tau^2)$, and
- The superscript "$-$" denotes the appropriate change to $n^{*-}$, $n_j^-$, and $\theta_j^{*-}$ when omitting $\theta_i$ from their calculations.

We update $p(\theta_i|\{\theta_k : k \neq i\}, \alpha, \mu, \tau^2, \phi, \mathbf{y})$, for $i = 1, \ldots, n$, sequentially by drawing either (1) a new value from $h$ with probability $A$, or (2) $\theta_j^*$ with probability $B_j$ ($A + B_1 + \cdots + B_{n^{*-}} = 1$). With each update of $\theta_i$ we also update the clustering "parameters" $n^*$, $\theta_j^*$, and $n_j$, for $j = 1, \ldots, n^*$ ($\mathbf{w}$ is more or less for bookkeeping and isn't explicitly used in the sampling algorithm).

Note that to use the Gibbs sampler we require conjugacy with $k_N$ and $G_0$. Without conjugacy we would have to resort to other methods for updating $\theta_i$, say an algorithm from Neal (2000).

The functional form of $q_j$ is simply a normal density

$$q_j = (2\pi\phi)^{-1/2} \exp\left\{-\frac{1}{2\phi}(y_i - \theta_j^*)^2\right\}$$

We solve for $q_0$ by integrating out $\theta$ (which has a normal kernel) and re-arranging terms to simplify to a nice normal density

$$\begin{aligned}
q_0 &= \int (2\pi\phi)^{-1/2} \exp\left\{-\frac{1}{2\phi}(y_i - \theta)^2\right\} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\} d\theta \\
&= (4\pi^2\phi\tau^2)^{-1/2} \int \exp\left\{-\frac{1}{2\phi}(y_i - \theta)^2 - \frac{1}{2\tau^2}(\theta - \mu)^2\right\} d\theta \\
&= (4\pi^2\phi\tau^2)^{-1/2} \int \exp\left\{-\frac{1}{2\phi\tau^2}\left[\tau^2 y_i^2 - 2y_i\tau^2\theta + \tau^2\theta^2 + \phi\mu^2 - 2\mu\phi\theta + \phi\theta^2\right]\right\} d\theta \\
&= (4\pi^2\phi\tau^2)^{-1/2} \int \exp\left\{-\frac{1}{2\phi\tau^2}\left[\theta^2(\phi + \tau^2) - 2\theta(\mu\phi + y_i\tau^2)\right] - \frac{\phi\mu^2 + \tau^2 y_i^2}{2\phi\tau^2}\right\} d\theta \\
&= (4\pi^2\phi\tau^2)^{-1/2} \exp\left\{-\frac{\phi\mu^2 + \tau^2 y_i^2}{2\phi\tau^2}\right\} \int \exp\left\{-\frac{\phi + \tau^2}{2\phi\tau^2}\left[\theta^2 - 2\theta\frac{\mu\phi + y_i\tau^2}{\phi + \tau^2}\right]\right\} d\theta \\
&= (4\pi^2\phi\tau^2)^{-1/2} \exp\left\{-\frac{\phi\mu^2 + \tau^2 y_i^2}{2\phi\tau^2}\right\} \int \exp\left\{-\frac{1}{2\sigma^*}\left[\theta^2 - 2\theta\mu^* + \mu^{*2} - \mu^{*2}\right]\right\} d\theta \\
&\qquad\qquad (\text{where } \mu^* = \frac{\mu\phi + \mathrm{y_i}\tau^2}{\phi + \tau^2}, \text{ and } \sigma^* = \frac{\phi\tau^2}{\phi + \tau^2}) \\
&= (4\pi^2\phi\tau^2)^{-1/2} \exp\left\{-\frac{\phi\mu^2 + \tau^2 y_i^2}{2\phi\tau^2}\right\} (2\pi\sigma^*)^{1/2} \exp\left\{\frac{\mu^{*2}}{2\sigma^*}\right\} \\
&= (2\pi\phi\tau^2)^{-1/2} \left(\frac{\phi\tau^2}{\phi + \tau^2}\right)^{1/2} \exp\left\{-\frac{\phi\mu^2 + \tau^2 y_i^2}{2\phi\tau^2} + \frac{(\mu\phi + y_i\tau^2)^2}{2\phi\tau^2(\phi + \tau)}\right\} \\
&= (2\pi(\phi + \tau^2))^{-1/2} \exp\left\{\frac{-(\phi\mu^2 + \tau^2 y_i^2)(\phi + \tau^2) + (\mu\phi + y_i\tau^2)^2}{2\phi\tau^2(\phi + \tau^2)}\right\} \\
&= (2\pi(\phi + \tau^2))^{-1/2} \exp\left\{\frac{-\phi^2\mu^2 - \phi\mu^2\tau^2 - \phi y_i^2\tau^2 - y_i^2\tau^2 + \mu^2\phi^2 + 2\mu\phi y_i\tau^2 + y_i^2\tau^2}{2\phi\tau^2(\phi + \tau^2)}\right\} \\
&= (2\pi(\phi + \tau^2))^{-1/2} \exp\left\{\frac{-\phi\mu^2\tau^2 - \phi y_i^2\tau^2 + 2\mu\phi y_i\tau^2}{2\phi\tau^2(\phi + \tau^2)}\right\} \\
&= (2\pi(\phi + \tau^2))^{-1/2} \exp\left\{-\frac{y_i^2 - 2y_i\mu + \mu^2}{2(\phi + \tau^2)}\right\} \\
&= N(y_i|\mu, \phi + \tau^2)
\end{aligned}$$

And thus the bookkeeping was worth it. More importantly, note the conjugacy requirement for the integral to be tractable.

The density function $h(\theta_i|\cdot)$ has essentially already been derived when finding $q_0$. After dropping all the non $\theta_i$ terms, we are left with the part from $q_0$ that was inside the integral. That is, $h$ is a normal distribution with mean $\mu^*$ and variance $\sigma^*$ given above. This completes the marginal posterior for $\theta_i$.

# References

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9, 249–265.