Markov Chain Sampling Methods for Dirichlet Process Mixture Models

Radford M. Neal, University of Toronto, Ontario, Canada
Presented by Colin DeLong
Outline

- Introduction
- Dirichlet process mixture models
- Gibbs sampling w/ conjugate priors
  - Algorithms 1, 2, and 3
- Methods for handling non-conjugate priors
  - Algorithm 4
- Metropolis-Hastings and partial Gibbs
  - Algorithms 5, 6, and 7
- Gibbs sampling w/ auxiliary parameters
  - Algorithm 8
- Experiments (well, one)
Introduction

- Some problems are more accurately represented with non-conjugate priors
  - Audio interpolation (Godsill & Rayner, 1995)
  - Climatology opinion quantification (Al-Awadhi & Garthwaite, 2001)
  - Financial risk assessment (Siu & Yang, 1999)
- Non-conjugate priors + Gibbs = headache.
  - Update integrals are nasty to compute
Dirichlet process mixture models

- Basic idea
  - Given data $y_1, \ldots, y_n$ ind. drawn from an unknown distribution ($y_i$ may be multivariate)
  - Model the unknown distribution as being drawn from a mixture of distributions $F(\theta)$, w/ mixing distribution over $\theta$ being $G$.
  - Let prior for $G$ be a Dirichlet process w/ concentration parameter $\alpha$ and base distribution $G_0$.
  - Then you have:

$$
\begin{align*}
  y_i \mid \theta_i & \sim F(\theta_i) \\
  \theta_i \mid G & \sim G \\
  G & \sim D(G_0, \alpha)
\end{align*}
$$
Dirichlet process mixture models

- Integrate over $G$ in previous model, giving a representation of the prior distribution of $\theta_i$ in terms of previous $\theta$'s:

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} G_0$$

- $\delta(\theta)$ is distribution concentrated at point $\theta$.

- You might notice the “Chinese Restaurant Process” at work here.
Dirichlet process mixture models

- You can also get here by letting $K$ (# of components) go to $\infty$…

$$y_i \mid c, \phi \sim F(\phi_{c_i})$$

$$c_i \mid p \sim \text{Discrete} \left( p_1, \ldots, p_K \right)$$

$$\phi_c \sim G_0$$

$$p_1, \ldots, p_K \sim \text{Dirichlet} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right)$$

- $c_i$ is the latent class associated with $y_i$
- The parameters $\phi_c$ determine the distribution of observations from $c$
Dirichlet process mixture models

• Integrate over mixing proportions $p_c$ to write prior of $c_i$ as follows:

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) = \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha}$$

• Where $n_{i,c}$ is the number of $c_j$ for $j < i$ equal to $c$. Letting $K$ go to $\infty$, we get $c_i$’s prior as:

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) \rightarrow \frac{n_{i,c}}{i - 1 + \alpha}$$

$$P(c_i \neq c_j \text{ for all } j < i \mid c_1, \ldots, c_{i-1}) \rightarrow \frac{\alpha}{i - 1 + \alpha}$$
Gibbs sampling w/ conjugate priors

- Exact computation of posterior for DP mixture models not feasible, so use Monte Carlo approaches
- Sample from posterior of $\theta_1, \ldots, \theta_n$ by simulating a Markov chain with this posterior as its equilibrium distribution
- Gibbs sampling is the natural approach here for conjugate priors
- 3 main ways of doing this
Algorithm 1: Let the state of the Markov chain consist of $\theta_1, \ldots, \theta_n$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Draw a new value from $\theta_i \mid \theta_{-i}, y_i$ as defined by equation (7).

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

- Where $H_i$ is the posterior for $\theta$ based on the prior $G_0$ and $y_i$, having likelihood $F(y_i, \theta)$ and:

$$
q_{i,j} = b F(y_i, \theta_j)
$$

$$
r_i = b \alpha \int F(y_i, \theta) dG_0(\theta)
$$

- Convergence may be slow due to groups of observations that are highly probably to be associated with the same $\theta$
Algorithm 2 (West, Muller, & Escobar, 1994)

**Algorithm 2:** Let the state of the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: If the present value of $c_i$ is associated with no other observation (i.e., $n_{-i,c_i} = 0$), remove $\phi_{c_i}$ from the state. Draw a new value for $c_i$ from $c_i \mid c_{-i}, y_i, \phi$ as defined by equation (11). If the new $c_i$ is not associated with any other observation, draw a value for $\phi_{c_i}$ from $H_i$ and add it to the state.

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$.

If $c = c_j$ for some $j \neq i$: 

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

$$P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i, \phi) = b \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi) dG_0(\phi)$$  \hspace{1cm} (11)
Algorithm 3 (Neal, 1992)

**Algorithm 3:** Let the state of the Markov chain consist of \( c_1, \ldots, c_n \). Repeatedly sample as follows:

- For \( i = 1, \ldots, n \): Draw a new value from \( c_i \mid c_{-i}, y_i \) as defined by equation (12).

If \( c = c_j \) for some \( j \neq i \):

\[
P(c_i = c \mid c_{-i}, y_i) = b \frac{n-i,c}{n-1+\alpha} \int F(y_i, \phi) dH_{-i,c}(\phi)
\]

\[
P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i) = b \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi) dG_0(\phi)
\]
Methods for handling non-conjugate priors

- If $G_0$ is not the conjugate prior for $F$, the integrals for sampling from the posterior might not be feasible to compute.

- West, Muller, and Escobar suggested a Monte Carlo approximation to compute the integral (1994).
  - Slower convergence
  - New values of $c_i$ are likely to be discarded during following Gibbs iteration, leading to wrong distribution.
Algorithm 4: Let the state of the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Let $k^-$ be the number of distinct $c_j$ for $j \neq i$, and let these $c_j$ have values in $\{1, \ldots, k^-\}$. If $c_i \neq c_j$ for all $j \neq i$, then with probability $k^-/(k^- + 1)$ do nothing, leaving $c_i$ unchanged. Otherwise, label $c_i$ as $k^- + 1$ if $c_i \neq c_j$ for all $j \neq i$, or draw a value for $\phi_{k^-+1}$ from $G_0$ if $c_i = c_j$ for some $j \neq i$. Then draw a new value for $c_i$ from $\{1, \ldots, k^- + 1\}$ using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi_1, \ldots, \phi_{k^-+1}) = \begin{cases} b n_{-i,c} F(y_i, \phi_c) & \text{if } 1 \leq c \leq k^- \\ b [\alpha/(k^- + 1)] F(y_i, \phi_c) & \text{if } c = k^- + 1 \end{cases}$$

where $b$ is the appropriate normalizing constant. Change the state to contain only those $\phi_c$ that are now associated with an observation.

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to $\phi_c$ that leaves this distribution invariant.

Algorithm 4 (MacEachern & Muller, 1998)
Problem with Algorithm 4

- Algorithm 4 has a problem in that assigning $c_i$ to a new component is reduced by a factor of $k^- + 1$.

- However, something similar without this problem is possible.
Metropolis-Hastings and partial Gibbs

- Use Metropolis-Hastings approach to update the $c_i$ using the conditional prior as the proposal distribution.
- Draw a candidate state, compute its acceptance probability. If it’s accepted, use the candidate state, else leave as is.
- We can apply this to the finite model from slide 6, again integrating out $\rho_c$.
Algorithm 5 (Neal, 1998)

**Algorithm 5:** Let the state of the Markov chain consist of \( c_1, \ldots, c_n \) and \( \phi = (\phi_c : c \in \{c_1, \ldots, c_n\}) \). Repeatedly sample as follows:

- For \( i = 1, \ldots, n \), repeat the following update of \( c_i \) \( R \) times: Draw a candidate, \( c_i^* \), from the conditional prior for \( c_i \) given by equation (16). If a \( c_i^* \) not in \( \{c_1, \ldots, c_n\} \) is proposed, choose a value for \( \phi_{c_i^*} \) from \( G_0 \). Compute the acceptance probability, \( a(c_i^*, c_i) \), as in equation (15), and set the new value of \( c_i \) to \( c_i^* \) with this probability. Otherwise let the new value of \( c_i \) be the same as the old value.

- For all \( c \in \{c_1, \ldots, c_n\} \): Draw a new value from \( \phi_c \mid y_i \) s.t. \( c_i = c \), or perform some other update to \( \phi_c \) that leaves this distribution invariant.

\[
a(c_i^*, c_i) = \min \left[ 1, \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right]
\]

If \( c = c_j \) for some \( j \neq i \):
\[
P(c_i = c \mid c_{-i}) = \frac{n_{-i,c}}{n - 1 + \alpha}
\]
\[
P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}) = \frac{\alpha}{n - 1 + \alpha}
\]
Algorithm 6 (Neal, 1998)

Algorithm 6: Let the state of the Markov chain consist of $\theta_1, \ldots, \theta_n$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$, repeat the following update of $\theta_i$ $R$ times: Draw a candidate, $\theta_i^*$, from the following distribution:

  $$
  \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} G_0
  $$

  Compute the acceptance probability

  $$
  a(\theta_i^*, \theta_i) = \min[1, F(y_i, \theta_i^*) / F(y_i, \theta_i)]
  $$

  Set the new value of $\theta_i$ to $\theta_i^*$ with this probability; otherwise let the new value of $\theta_i$ be the same as the old value.
Algorithm 7 (Neal, 1998)

Algorithm 7: Let the state of the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$, update $c_i$ as follows: If $c_i$ is not a singleton (i.e., $c_i = c_j$ for some $j \neq i$), let $c_i^*$ be a newly-created component, with $\phi_{c_i^*}$ drawn from $G_0$. Set the new $c_i$ to this $c_i^*$ with probability

$$ a(c_i^*, c_i) = \min \left[ 1, \frac{\alpha}{n-1} \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right] $$

Otherwise, when $c_i$ is a singleton, draw $c_i^*$ from $c_{-i}$, choosing $c_i^* = c$ with probability $n_i/c / (n-1)$. Set the new $c_i$ to this $c_i^*$ with probability

$$ a(c_i^*, c_i) = \min \left[ 1, \frac{n_i-1}{\alpha} \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right] $$

If the new $c_i$ is not set to $c_i^*$, it is the same as the old $c_i$.

- For $i = 1, \ldots, n$: If $c_i$ is a singleton (i.e., $c_i \neq c_j$ for all $j \neq i$), do nothing. Otherwise, choose a new value for $c_i$ from $\{c_1, \ldots, c_n\}$ using the following probabilities:

$$ P(c_i = c | c_{-i}, y_i, \phi, c_i \in \{c_1, \ldots, c_n\}) = \frac{b_{n-i,c}}{n-1} F(y_i, \phi_c) $$

where $b$ is the appropriate normalizing constant.

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c | y_i$ s.t. $c_i = c$, or perform some other update to $\phi_c$ that leaves this distribution invariant.
Gibbs sampling w/ auxiliary parameters

- More flexible.
  - Basic idea is that we sample from a distribution $\pi_x$ for $x$ by sampling from distribution $\pi_{xy}$ for $(x, y)$.
  - Idea extendable to accommodate auxiliary variables which can be created/discarded during Markov chain simulation.
  - A variable $y$ can be introduced temporarily:
    - Draw a value for $y$ from its conditional given $x$
    - Perform an update of $(x, y)$ leaving $\pi_{xy}$ invariant
    - Discard $y$, leaving $x$.
  - This technique can be used to update $c_i$ for the DPM without having to integrate w.r.t. $G_0$. 
Algorithm 8 (Neal, 1998)

**Algorithm 8:** Let the state of the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Let $k^{-}$ be the number of distinct $c_j$ for $j \neq i$, and let $h = k^{-} + m$. Label these $c_j$ with values in $\{1, \ldots, k^{-}\}$. If $c_i = c_j$ for some $j \neq i$, draw values independently from $G_0$ for those $\phi_c$ for which $k^{-} < c \leq h$. If $c_i \neq c_j$ for all $j \neq i$, let $c_i$ have the label $k^{-} + 1$, and draw values independently from $G_0$ for those $\phi_c$ for which $k^{-} + 1 < c \leq h$. Draw a new value for $c_i$ from $\{1, \ldots, h\}$ using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi_1, \ldots, \phi_h) = \begin{cases} 
    b \frac{n-i,c}{n-1+\alpha} F(y_i, \phi_c) & \text{for } 1 \leq c \leq k^{-} \\
    b \frac{\alpha/m}{n-1+\alpha} F(y_i, \phi_c) & \text{for } k^{-} < c \leq h
\end{cases}$$

where $n_{-i,c}$ is the number of $c_j$ for $j \neq i$ that are equal to $c$, and $b$ is the appropriate normalizing constant. Change the state to contain only those $\phi_c$ that are now associated with one or more observations.

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to $\phi_c$ that leaves this distribution invariant.
The Experiment

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time per iteration in microseconds</th>
<th>Autocorrelation time for $k$</th>
<th>Autocorrelation time for $\theta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 4 (“no gaps”)</td>
<td>7.6</td>
<td>13.7</td>
<td>8.5</td>
</tr>
<tr>
<td>Alg. 5 (Metropolis-Hastings, $R = 4$)</td>
<td>8.6</td>
<td>8.1</td>
<td>10.2</td>
</tr>
<tr>
<td>Alg. 6 (M-H, $R = 4$, no $\phi$ update)</td>
<td>8.3</td>
<td>19.4</td>
<td>64.1</td>
</tr>
<tr>
<td>Alg. 7 (mod M-H &amp; partial Gibbs)</td>
<td>8.0</td>
<td>6.9</td>
<td>5.3</td>
</tr>
<tr>
<td>Alg. 8 (auxiliary Gibbs, $m = 1$)</td>
<td>7.9</td>
<td>5.2</td>
<td>5.6</td>
</tr>
<tr>
<td>Alg. 8 (auxiliary Gibbs, $m = 2$)</td>
<td>8.8</td>
<td>3.7</td>
<td>4.7</td>
</tr>
<tr>
<td>Alg. 8 ($m = 30$, approximates Alg. 2)</td>
<td>38.0</td>
<td>2.0</td>
<td>2.8</td>
</tr>
</tbody>
</table>

- $k$ is the number of distinct $c_i$, $\theta_1$ is the parameter associated with $y_1$
- Algorithm 8 with $m=1$ superior to algorithm 4 (“no gaps”)
- Performance much worse for algorithm 6, where no updates for $\varphi_c$ are included
- With $m=30$, algorithm 8 takes longer, but performance is great.