Gibbs Sampling Methods for Dirichlet Process Mixture Model: Technical Details

Xiaodong Yu
University of Maryland, College Park

September 12, 2009

1 Introduction

The Neal's review paper [4] presents three algorithms of Gibbs sampling for Dirichlet Process Mixture Models (DPMM) when conjugate priors are used. But he did not give technical details for these algorithms and thus it is unclear how to use them in a practical problem for a newbie. Rasmussen's paper [6], Teh's tutorial course [9] and Ranganathan's thesis appendix [5] provide us more technical details of these algorithms. Sudderth has a good review about the two Gibbs sampling methods with Chinese restaurant process in the Chapter 2 of his thesis[8]. The purpose of this report is thus to combine these materials into a self-contained tutorial for the techniques of Gibbs sampling for Dirichlet Process Mixture Models (DPMM), especially when conjugate priors are used.

Later on, I found several papers, which are perhaps the original papers about the algorithms described here:

- The paper by Escobar and West [1] related to Section 4.1
- The paper by West, Muller and Escobar [11] seems to be related to Section 4.2
- The paper by MacEarchern [3] seems to be related to Section 4.3

The papers [11, 3] also gives examples on normal distribution, which would be very helpful to understand the algorithms. But I have no time to read these papers in details. So this report is still based on Neal, Rasmussen, Teh, Ranganathan, and Sudderth's materials.

2 Dirichlet Process and Its Representations

$G \sim DP(\alpha_0, G_0)$ denotes a Dirichlet Process (DP) if G is a DP-distributed random probability measure. This definition has two key points:

- First, G is a probability measure over a subsets of a space $\mathbb{X}$, which can be loosely viewed as a generalized probability distribution;
- Second, any finite set of partitions of $\mathbb{X}$, $A_1 \bigcup ... \bigcup A_k = \mathbb{X}$, we require $(G(A_1), \bigcup ..., G(A_k))$ to be Dirichlet distributed.

Thus a DP can be viewed as a distribution of distribution. A DP has two parameters

- Base distribution G_0, which is like the mean of the DP because $\mathbb{E}[G(A)] = G_0(A)$;
- Strength parameter α_0, which is like an inverse-variance of the DP because $\mathbb{V}[G(A)] = \frac{G_0(A)(1-G_0(A))}{\alpha_0+1}$.

A DP can be represented from various schemes, as summarized in [10, 9]. They are briefly reviewed in the rest of this section.

2.1 Pólya urn scheme

The Pólya urn scheme (a.k.a. Blackwell-MacQueen urn scheme) describes a process that produces a sequence of i.i.d. random variables $\phi_1, \phi_2, ...$ distributing according to G:

- Start with no balls in the urn.
- With probability $\propto \alpha_0$, draw $\phi_n \sim G_0$, and add a ball of that color into the urn.
- With probability $\propto n-1$, pick a ball at random from the urn, record ϕ_n to be its color, return the ball into the urn and place a second ball of the same color into the urn.

The process can be summarized as the following conditional distribution:

$$\phi_n | \phi_{1:n-1} \sim \frac{\alpha_0 G_0}{\alpha_0 + n - 1} + \frac{\sum_{j=1}^{n-1} \delta(\phi_n - \phi_j)}{\alpha_0 + n - 1} \tag{2.1}$$

where $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. This process provides a method to predict the new sample based on the existing samples and leads to a Gibbs sampling method, as shown in next section.

2.2 Chinese restaurant process

The above generating process shows that the random variables $\phi_1, ..., \phi_n$ drawn from a Pólya urn scheme have probability of being equal to one of the previous draws. Suppose n draws of ϕ_i can take on $K < n$ distinct values and denote them as $\theta_1, ..., \theta_K$. This defines a partition of $1, ..., n$ into K clusters. The induced distribution over such partitions is a Chinese restaurant process, which is described as follows:

- Imagine a Chinese restaurant that has unlimited number of tables.
- First customer sits at the first table.
- Customer n sits at:
  - Table k with probability $\frac{n_k}{\alpha_0+n-1}$, where n_k is the number of customers at table k.
  - A new table $K+1$ with probability $\frac{\alpha_0}{\alpha_0+n-1}$.
- In this metaphor, customers are analogies of integers and tables of clusters.

This process can also be summarized as follows:

$$p(\text{customer } n \text{ sat at table } k|\text{past } n-1 \text{ customers }) = \begin{cases} \frac{n_k}{\alpha_0+n-1}, & \text{if occupied table;} \\ \frac{\alpha_0}{\alpha_0+n-1}, & \text{if new table.} \end{cases} \tag{2.2}$$

The Chinese restaurant process illustrate the "cluster" property of the DP, i.e., the more customers sit at a table, the higher chance a new customer will choose to sit at this table and most probably, and thus only a limited number of tables will be occupied although there are unlimited number of tables in the restaurant. This property makes it feasible for us to sample from a DP mixture, as shown in next section.

2.3 Stick-breaking construction

Both of the above representations refer to the draws from G, while the stick-breaking construction shows the property of G explicitly:

$$G(\theta) = \sum_{i=1}^{\infty} \pi_k \delta(\theta - \theta_k), \text{ where } \theta_k \sim G_0. \tag{2.3}$$

The mixture weights $\{\pi_k\}_{k=1}^{\infty}$ can constructed as follows:

- Start with a unit-length stick, break the stick according to the proportion β_1 where $\beta_1 \sim \text{Beta}(1, \alpha_0)$, and assign β_1 to π_1;
- The remaining stick is broken according the proportion $\beta_k \sim \text{Beta}(1, \alpha_0)$, and assign the β_k portion of the remaining stick to π_k.

This procedure can be summarized as follows:

$$\begin{aligned} \beta_k &\sim \text{Beta}(1, \alpha_0) \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \end{aligned} \tag{2.4}$$

The sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one and can be writen as $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$, named after Griffiths, Engen, and McCloskey. The stick-breaking construction reveal the discrete nature of the random measure G.

3 Dirichlet Process Mixture Modeling

Dirichlet process mixture model (DPMM) can be considered as an infinite extension of finite mixture model (FMM). So it is easier to understand a DPMM by when starting from a FMM.

A FMM can be described with the graphical representation in Figure 1, which is equivalent to the following distributions:

$$\boldsymbol{\pi}|\alpha_0 \sim \mathrm{Dir}(\alpha_0/K, ..., \alpha_0/K) \tag{3.1}$$

$$z_i|\boldsymbol{\pi} \sim \pi \tag{3.2}$$

$$\theta_k|\lambda \sim G_0(\lambda) \tag{3.3}$$

$$x_i|z_i, \{\theta_k\}_{k=1}^K \sim F(\theta_{z_i}) \tag{3.4}$$

In this model, each datum x_i is generated by first selecting one of K clusters, say, cluster k, according to the multinomial distribution that is parameterized by π as in (3.2), and then sampling from the distribution of this cluster $F(\theta_{z_i})$ that is parameterized by θ_k, as in (3.4). In this equation, an indicator variable $z_i \in \{1, ..., K\}$ is introduced to specify the cluster associated with x_i. The mixture weight π is given a symmetric Dirichlet prior with a hyperparameter α_0, as in (3.1) and the cluster parameters θ_k are given a common prior distribution $G_0(\lambda)$ with parameter λ, as in (3.3). In practice, $F(\theta)$ is typically some exponential family of densities, and $G_0(\lambda)$ a corresponding conjugate prior.

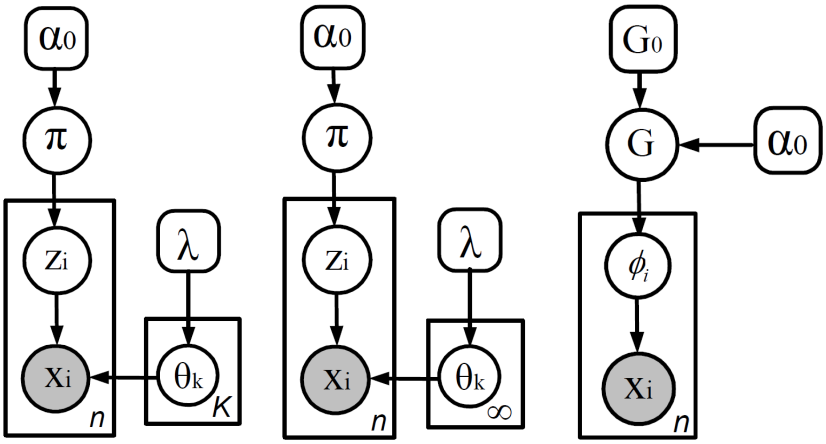

Figure 1: Finite mixture model (left), Dirichlet process mixture model in stick-breaking representation (right) and Dirichlet process mixture model in Pólya urn representation (right)

Let K go to infinity, the FMM becomes a DPMM, whose graphical representation is in Figure 1. The generating process of the DPMM is the same as those of the FMM, except that the number of clusters is not a fixed value K. Thus in the DPMM, the Dirichlet prior for $\boldsymbol{\pi}$ is replaced by a stick-breaking

construction, $\boldsymbol{\pi} \sim \text{GEM}(1, \alpha_0)$ and the conditional distributions of a DPMM are:

$$\begin{aligned} \boldsymbol{\pi}|\alpha_0 &\sim \text{GEM}(1, \alpha_0) \\ z_i|\boldsymbol{\pi} &\sim \pi \\ \theta_k|\lambda &\sim G_0(\lambda) \\ x_i|z_i, \{\theta_k\}_{k=1}^{\infty} &\sim F(\theta_{z_i}). \end{aligned} \tag{3.5}$$

If we do not use indicator variables and explicitly present the generative process of the cluster parameters, we can let $\phi_i = \theta_{z_i}$ and get the DPMM in the Pólya urn representation. The graphical representation is in Figure 1 and its conditional probabilities are:

$$\begin{aligned} G|G_0, \alpha_0 &\sim G_0 \\ \phi_i &\sim G \\ x_i|\phi_i &\sim F(\phi_i). \end{aligned} \tag{3.6}$$

The connection between the DPMM and the Chinese restaurant process can be explicitly illustrated from the conditional distributions of the indicator variables. In the FMM,

$$P(z_i = k|\mathbf{z}_{-i}, \alpha_0) = \frac{n_{k,-i} + \alpha_0/K}{n + \alpha_0 - 1}, \tag{3.7}$$

where $\mathbf{z}_{-i}$ denotes the number of points in the k-th cluster excluding the i-th point. The details of derivation of this result can be found in my technical report "Derivation of Gibbs Sampling for Finite Gaussian Mixture Model". Let K go to infinity, the conditional distributions of the indicator variables reaches the following limits:

$$\begin{aligned} \text{for cluster } k \text{ with } n_{k,-i} > 0: \qquad P(z_i = k|\mathbf{z}_{-i}, \alpha_0) &= \frac{n_{k,-i}}{n + \alpha_0 - 1} \\ \text{for all the other clusters:} P(z_i \neq z_j \text{ for all } j \neq i|\mathbf{z}_{-i}, \alpha_0) &= \frac{\alpha_0}{\alpha_0 + n - 1} \end{aligned} \tag{3.8}$$

Since the order of z_i does not matter in the DPMM, we can always imagine z_i be the last one and then Equation (3.8) and (2.2) are equivalent.

4 Three Gibbs Sampling Methods for Mixture Models

4.1 Gibbs sampling based on Pólya urn representation

In the Pólya urn representation of DPMM (3.6), the only unknown variables are $\{\phi_i\}_{i=1}^n$. This leads to a simple Gibbs sampling method: alternatively draw ϕ_i from its posterior distribution conditioned on the other variables $\boldsymbol{\phi}_{-i}$ and all the

observations. To achieve this goal, we need to combine a prior of ϕ_i conditioned on $\boldsymbol{\phi}_{-i}$ and the likelihood for ϕ_i given the corresponding observation x_i, i.e. $F(x_i|\phi_i)$. Such prior can be derived from (2.1) by imaging that the i is the last one in the n observations without changing the distribution form, since all the ϕ_i are exchangeable. Thus we have:

$$\phi_i = \phi|\boldsymbol{\phi}_{-i} \sim \frac{\alpha_0 G_0(\phi_)}{\alpha_0 + n - 1} + \frac{\sum_{j \neq i} \delta(\phi - \phi_j)}{\alpha_0 + n - 1}. \tag{4.1}$$

Combined with the likelihood, we get the posterior of ϕ_i conditioned on $\boldsymbol{\phi}_{-i}$:

$$\begin{aligned} p(\phi_i|\boldsymbol{\phi}_{-i}, x_i) &= b\alpha_0 q_0 H(\phi_i|x_i) + b\sum_{j \neq i} F(x_i|\phi_j)\delta(\phi_i - \phi_j) \\ H(\phi_i|x_i) &= \frac{G_0(\phi_i)F(x_i|\phi_i)}{\int_\phi G_0(\phi)F(x_i|\phi)} \\ q_0 &= \int_\phi G_0(\phi)F(x_i|\phi) \\ b &= \left(\alpha_0 q_0 + \sum_{j \neq i} F(x_i|\phi_j)\right)^{-1} \end{aligned} \tag{4.2}$$

When G_0 is a conjugate prior for $F(x_i|\phi_i)$, the posterior distribution of ϕ_i, $H(\phi_i|x_i)$ and the marginal distribution of x_i, q_0, have analytical forms and the Gibbs sampling can be easily performed.

In summary, the tasks in each iteration of this sampling method is illustrated in Algorithm 1.

Algorithm 1 Gibbs sampling for DPMM based on the Pólya urn representation

Given $\{\phi_i^{(t-1)}\}_{i=1}^n$ from the previous iteration, sample a new set of $\{\phi_i^{(t)}\}_{i=1}^n$ as follows:

1. For $i = 1, ..., n$
   (a) draw a new sample for $\phi_i^{(t)}$ from the following distribution:
   $\phi_i^{(t)} \sim b\alpha_0 q_0 H(\phi_i|x_i) + b\sum_{j \neq i} F(x_i|\phi_j^{(t-1)})\delta(\phi_i - \phi_j^{(t-1)})$

In the appendix of the thesis [5], Ranganathan illustrates an example for this algorithm. In this example, x_i is 1D real number, F is a univariate normal distribution with unknown mean μ but known variance equal to unity. Thus the ϕ contains only one random variable, μ. The base measure G_0 is taken to be

the standard normal distribution. The model can then be described as follows:

$$\begin{aligned} x_i|\mu_i &\sim \mathcal{N}(\mu_i, 1) && (4.3)\\ \mu_i &\sim G(\mu) && (4.4)\\ G &\sim \mathrm{DP}(\alpha_0 G_0(\mu)) && (4.5)\\ G_0 &= \mathcal{N}(0,1) && (4.6) \end{aligned}$$

Using the equations (4.2), we get

$$\begin{aligned} q_0 &= \int_\mu \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right) \\ &= \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{x_i^2}{4}\right) \times \frac{1}{\sqrt{2\pi\frac{1}{2}}} \int_\mu \exp\left(-\frac{(\mu-\frac{1}{2}x_i)^2}{2\times\frac{1}{2}}\right) \\ &= \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{x_i^2}{4}\right) \end{aligned}$$

and

$$\begin{aligned} H(\mu_i|x_i) &= \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_i)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu_i^2}{2}\right)}{\frac{1}{2\sqrt{\pi}} \exp\left(-\frac{x_i^2}{4}\right)} \\ &= \frac{1}{\sqrt{2\pi\times\frac{1}{4}}} \exp\left(-\frac{(\mu_i-\frac{1}{2}x_i)^2}{2\times\frac{1}{4}}\right) \\ &= \mathcal{N}\left(\frac{1}{2}x_i, \frac{1}{2}\right). \end{aligned}$$

A easier way to compute the above q_0 and $H(\mu_i|x_i)$ is to start with $H(\mu_i|x_i)$ and use the property of conjugate prior to derive $H(\mu_i|x_i)$ directly and then compute q_0 use $H(\mu_i|x_i)$. Use the property of the conjugate prior of Gaussian distribution with known variance and unknown mean

$$p(\mu|\mathbf{x}, \sigma^2, \mu_0, \sigma_0^2) = \mathcal{N}\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right),$$

and substitute $\mu_0 = 0, \sigma_0^2 = 1, \sigma^2 = 1, n = 1$ into the above equation, the posterior $H(\mu_i|x_i)$ is also a Gaussian with updated mean and variance: S $H(\mu_i|x_i) = \mathcal{N}(\frac{1}{2}x_i, \frac{1}{2})$. Then we get $q_0 = G_0(\mu)F(x_i|\mu_i)/H(\mu_i|x_i) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{x_i^2}{4}\right)$. Finally, the Gibbs sampler becomes

$$P(\mu_i^{(t)}|\boldsymbol{\mu}_{-i}^{(t-1)}, x_i) \propto \alpha_0 q_0 H(\mu_i|x_i) + \sum_{j\neq i} F(x_i|\mu_j^{(t-1)})\delta(\mu_i - \mu_j^{(t-1)}).$$

In each iteration, we need to sample μ_i from $P(\mu_i|\boldsymbol{\mu_{-i}}, x_i)$ for $i = 1, ..., n$ in turn.

4.2 Gibbs sampling using latent indicator variables

Though the Gibbs sampling based on Pólya urn representation is very simple to implement, it is very inefficient. In each iteration, we need to sample the cluster parameter for n times and each time we only change the parameter for a single data point. As we know, there are usually lots of data points share the same cluster parameter. A more efficient way is obvious to operate the data points belonging the same cluster simultaneously. To do this, we need to employ the DPMM in stick-breaking representation, where cluster parameters are moved outside of the plate of x_i and the indicator variables are used to identify the cluster x_i associated to.

Before discussing Gibbs sampling for DPMM, we first see the case of FMM. In a FMM, the data points $\mathbf{x} = \{x_i\}_{i=1}^n$ are observed and the cluster indicators $\mathbf{z} = \{z_i\}_{i=1}^n$ are latent. Thus the Gibbs sampling involves iterations that alternately draw samples from one of the following variables while keeping others fixed: the cluster indicators $\mathbf{z} = \{z_i\}_{i=1}^n$, the cluster parameters $\{\theta_k\}_{k=1}^K$ and the mixture weights π. The first step towards Gibbs sampling is to derive the conditional posterior distributions for these variables. The hyperparameter α_0 and λ are assumed known in this process. By exploiting the Markov properties of the FMM and employing the Bayes rule, these distributions will be simplified to great extents.

For each indicator variable z_i, we need to derive its conditional posterior:

$$\begin{aligned}
& p(z_i = k|\mathbf{z}_{-i}, \mathbf{x}, \pi, \{\theta_k\}_{k=1}^K, \alpha_0, \lambda) \\
= \; & p(z_i = k|x_i, \pi, \{\theta_k\}_{k=1}^K) && (4.7) \\
\propto \; & p(z_i = k|\pi, \{\theta_k\}_{k=1}^K) p(x_i|z_i = k, \pi, \{\theta_k\}_{k=1}^K) && (4.8) \\
= \; & p(z_i = k|\pi) p(x_i|\theta_k) && (4.9) \\
= \; & \pi_k F(x_i|\theta_k). && (4.10)
\end{aligned}$$

In the above derivation, (4.7) exploits the Markov property of the FMM, (4.8) uses the Bayes rule that posterior $\propto$ prior $\times$ likelihood, (4.9) uses the Markov property and uses the definition of indicator variables.

For the mixture weight π, we need to derive its conditional posterior:

$$\begin{aligned}
p(\pi|\mathbf{z}, \mathbf{x}, \{\theta_k\}_{k=1}^K, \alpha_0, \lambda) &= p(\pi|\mathbf{z}, \alpha_0) && (4.11) \\
&= \mathrm{Dir}(n_1 + \alpha_0/K, ..., n_K + \alpha_0/K), && (4.12)
\end{aligned}$$

where $n_k = \sum_{i=1}^n \delta(z_i - k)$. Here (4.11) results from Markov property and (4.12) employs the property of the conjugate Dirichlet prior.

For the cluster parameters, we need to derive its conditional posterior. In [2], it is shown that the mixture weights π and parameters $\{\theta_k\}_{k=1}^K$ are mutually independent conditioning on the indicator variables $\mathbf{z}$:

$$p(\pi, \{\theta_k\}_{k=1}^K|\mathbf{z}, \mathbf{x}, \alpha_0, \lambda) = p(\pi|\mathbf{z}, \alpha_0) \prod_{k=1}^K p(\theta_k|\mathbf{x}_k, \lambda). \tag{4.13}$$

This result shows that the conditional posterior of the parameter of the k-th cluster, θ_k, only depends on the observations belonging to this cluster, i.e., $\mathbf{x}_k$. Thus

$$\begin{aligned} p(\theta_k|\boldsymbol{\theta}_{-k}, \pi, \mathbf{z}, \mathbf{x}, \alpha_0, \lambda) &= p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{z}, \mathbf{x}, \lambda) && (4.14) \\ &= p(\theta_k|\mathbf{x}_k, \lambda) && (4.15) \\ &\propto G_0(\theta_k|\lambda)\mathcal{L}(\mathbf{x}_k|\theta_k) && (4.16) \end{aligned}$$

Here, (4.14) uses the Markov property, (4.15) uses the results in (4.13), and (4.16) uses the Bayesian rule. If $G(\lambda)$ is a conjugate prior of θ_k, the posterior of θ_k will be the same form of $G(\lambda)$ with parameters updated by the x_i's in $\mathbf{x}_k$

In summary, the tasks in each iteration of this sampling method is illustrated in Algorithm 2.

Algorithm 2 Gibbs sampling for FMM using latent indicator variables

Given $\pi^{(t-1)}$, $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample a new set of $\pi^{(t)}$ and $\{\theta_k^{(t-1)}\}_{k=1}^K$ as follows:

1. For $i = 1, ..., n$
   (a) Draw a new sample for z_i from the distribution:
   $$p(z_i^{(t)} = k) \propto \pi_k^{(t-1)} F(x_i|\theta_k^{(t-1)})$$
2. Sample new mixture weight $\pi^{(t)}$ from the following distribution:
   $$\pi^{(t)} \sim \text{Dir}(n_1^{(t)} + \alpha_0/K, ..., n_K^{(t)} + \alpha_0/K) \qquad n_k^{(t)} = \sum_{i=1}^{n} \delta(z_i^{(t)} - k)$$
3. For $k = 1, ..., K$
   (a) Sample cluster parameter of each cluster, θ_k, from the following distribution:
   $$\theta_k^{(t)} \propto G_0(\theta_k|\lambda)\mathcal{L}(\mathbf{x}_k^{(t)}|\theta_k^{(t-1)})$$

In Algorithm 2, the mixture weight π is explicitly sampled from a Dirichlet distribution. However, such sampling is difficult when K goes to infinite. One option is to integrate π out. This requires us to derive z_i's conditional posterior

of z_i as follows:

$$\begin{aligned} & p(z_i = k|\mathbf{z}_{-i}, \mathbf{x}, \{\theta_k\}_{k=1}^K, \alpha_0, \lambda) \\ = \; & p(z_i = k|\mathbf{z}_{-i}, x_i, \theta_k, \alpha_0) && (4.17) \\ = \; & p(z_i = k|\mathbf{z}_{-i}, \alpha_0, \theta_k) p(x_i|z_i = k, \mathbf{z}_{-i}, \theta_k, \alpha_0) && (4.18) \\ = \; & p(z_i = k|\mathbf{z}_{-i}, \alpha_0) p(x_i|\theta_k) && (4.19) \\ = \; & \frac{n_{k,-i} + \alpha_0/K}{n + \alpha_0 - 1} F(x_i|\theta_k). && (4.20) \end{aligned}$$

Here, (4.17) uses the Markov property, the property of indicator variable, and results implied from (4.13), (4.18) uses the Bayesian rule, (4.19) uses the Markov property, and (4.20) use the results in (3.7).

In summary, the tasks in each iteration of this sampling method is illustrated in Algorithm 3.

Algorithm 3 Gibbs sampling for FMM with mixture weight integrated out

Given $\{\theta_k^{(t-1)}\}_{k=1}^K$ and $\{z_i^{(t-1)}\}_{i=1}^n$ from the previous iteration, sample a new set of $\{\theta_k^{(t-1)}\}_{k=1}^K$ and $\{z_i^{(t)}\}_{i=1}^n$ as follows:

1. Set $z = z^{(t-1)}$
2. For $i = 1, ..., n$
   (a) Remove data item x_i from the cluster z_i, since we are going to sample a new z_i for x_i.
   (b) Draw a new sample for z_i from the distribution:

$$p(z_i = k) \propto \frac{n_{k,-i} + \alpha_0/K}{n + \alpha_0 - 1} F(x_i|\theta_k^{(t-1)}) \qquad n_{k,-i} = \sum_{j \neq i} \delta(z_j - k)$$

3. For $k = 1, ..., K$
   (a) Sample cluster parameter of each cluster, θ_k, from the following distribution:

$$\theta_k^{(t)} \propto G_0(\theta_k|\lambda) \mathcal{L}(\mathbf{x}_k^{(t)}|\theta_k^{(t-1)})$$

4. Set $z^{(t)} = z$
5. After the burn-in period, optionally, we can sample $\pi^{(t)}$ via Step 2 in Algorithm 2 using $\{z_i^{(t)}\}_{i=1}^n$.

Now we can generalize FMM to DPMM by letting K go to infinity. By doing so, the conditional prior of z_i evolves from (3.7) to (3.8). When z_i is assigned to one of the current K clusters, the conditional posterior of z_i can be obtained by replacing the conditional prior $p(z_i = k|\mathbf{z}_{-i}, \alpha_0)$ in (4.20) by $\frac{n_{k,-i}}{n+\alpha_0-1}$. If

z_i is assigned to a new cluster index, which we denote as $K+1$ without loss generality, we need to derive z_i's conditional posterior in this case:

$$\begin{aligned} & p(z_i = K+1|\mathbf{z}_{-i}, \mathbf{x}, \alpha_0, \lambda) \\ = \; & p(z_i = K+1|\mathbf{z}_{-i}, x_i, \alpha_0, \lambda) && (4.21) \\ = \; & p(z_i = K+1|\mathbf{z}_{-i}, \alpha_0, \lambda) p(x_i|z_i = K+1, \mathbf{z}_{-i}, \alpha_0, \lambda) && (4.22) \\ = \; & p(z_i = K+1|\mathbf{z}_{-i}, \alpha_0) p(x_i|\lambda) && (4.23) \\ = \; & \frac{\alpha_0}{n+\alpha_0-1} \int F(x_i|\theta) G_0(\theta|\lambda) d\theta. && (4.24) \end{aligned}$$

Here, (4.21) uses the property of indicator variable, (4.22) uses the Bayesian rule, (4.23) uses the Markov property and the property of indicator variable, and (4.24) use the results in (3.8) and definition of marginal distribution. When z_i is assigned to a new cluster K, we should draw a new parameter ϕ_i chosen from $H(\phi_i|x_i)$, the posterior distribution based on the prior G_0 and the single observation x_i, as defined in (4.2), assign it to the cluster parameter of this cluster, i.e., $\theta_{K+1} = \phi$, and increase K by 1.

Generally, DPMM is robust the concentration parameter α_0. However, the number of clusters, K, is quite sensitive to α_0 [1]. Thus, in many applications, it is useful to choose a weakly informative prior for α_0, and sample from its posterior while learning cluster parameters. If $\alpha_0 \sim \text{Gamma}(a, b)$ is assigned a Gamma prior, its posterior is simple function of K, and samples are easily drawn via auxiliary variable method [1]. An alternative method using Adaptive Rejection Sampling is described in [6].

In summary, the tasks in each iteration of this sampling method is illustrated in Algorithm 4.

In [6], this Gibbs sampling algorithm is applied to a univariate Gaussian mixture. In this model, the cluster parameters includes cluster mean and precision μ, s, the hyperparameter λ, r for the conjugate prior of μ, the hyperparameter β, w for the conjugate prior of s, and hyperparameter α_0 for the Dirichlet conjugate prior of π. Since we give conjugate priors to all the cluster parameters, the posterior of the parameter of the k-th cluster, θ_k, in (4.16) and the marginal distribution of x_i, q_0, in (4.2) both have analytical form. The detailed derivations can refer to my technical report "Derivation of Gibbs Sampling for Finite Gaussian Mixture Model".

4.3 Collapsed Gibbs sampling

When using conjugate prior, we can often integrate out the cluster parameters θ_k and then we need sample z_i only. This method is called as collapsed Gibbs sampling. The justification of integrating out the cluster parameters is due to the Rao-Blackwell Theorem [7], which states that marginalization of some variables from a joint distribution always reduces the variance of later estimates. The idea of this theorem can be illustrated by the following example [8].

Let $p(x, z)$ be a joint distribution of two random variables, where $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. Given L independent samples $\{x^{(\ell)}, z^{(\ell)}\}_{\ell=1}^{L}$ from this joint distribution,

Algorithm 4 Direct Gibbs sampling for DPMM

Given $\alpha_0^{(t-1)}$, $\{\theta_k^{(t-1)}\}_{k=1}^K$ and $\{z_i^{(t-1)}\}_{i=1}^n$ from the previous iteration, sample a new set of $\{\theta_k^{(t-1)}\}_{k=1}^K$ and $\{z_i^{(t)}\}_{i=1}^n$ as follows:

1. Set $z = z^{(t-1)}$, $\alpha_0 = \alpha_0^{(t-1)}$
2. For $i = 1, ..., n$
   (a) Remove data item x_i from the cluster z_i, since we are going to sample a new z_i for x_i.
   (b) If x_i is the only data in its current cluster, this cluster becomes empty after Step (2.a). This cluster is then removed, together with its parameter, and K is decreased by 1.
   (c) Re-arrange cluster indices so that $1, ..., K$ are active (i.e., non-empty)
   (d) Draw a new sample for z_i from the following probabilities:

$$p(z_i = k, k \leq K) \propto \frac{n_{k,-i}}{n + \alpha_0 - 1} F(x_i|\theta_k^{(t-1)}) \qquad n_{k,-i} = \sum_{j \neq i} \delta(z_j - k)$$

$$p(z_i = K + 1) \propto \frac{\alpha_0}{n + \alpha_0 - 1} \int F(x_i|\theta) G_0(\theta) d\theta$$

   (e) If $z_i = K + 1$, we get a new cluster. Index this cluster as $K + 1$, sample a new cluster parameter ϕ_i from $H(\phi_i|x_i)$ defined in (4.2), assign it to θ_{K+1}, and increase K by 1
3. For $k = 1, ..., K$
   (a) Sample cluster parameter of each cluster, θ_k, from the following distribution:

$$\theta_k^{(t)} \propto G_0(\theta_k|\lambda) \mathcal{L}(\mathbf{x}_k^{(t)}|\theta_k^{(t-1)})$$

4. Set $z^{(t)} = z$
5. If $\alpha_0 \sim \text{Gamma}(a, b)$, sample $\alpha_0^{(t)} \sim p(\alpha_0|K, n, a, b)$ via auxiliary variable method [1].

our goal is to estimate a statistic of $f(x,z)$ that equals

$$\begin{aligned}\mathbb{E}_p[f(x,z)] &= \int_{\mathcal{Z}}\int_{\mathcal{X}} f(x,z)p(x,z)dxdz &(4.25)\\ &\approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)},z^{(\ell)}) = \mathbb{E}_{\tilde{p}}[f(x,z)] &(4.26)\end{aligned}$$

Sometimes, the conditional density $p(x|z)$ has a tractable analytic form. In this case, we can consider to sample L independent samples $\{z^{(\ell)}\}_{\ell=1}^{L}$ from the marginal distribution $p(z)$ to replace the samples from the joint distribution:

$$\begin{aligned}\mathbb{E}_p[f(x,z)] &= \int_{\mathcal{Z}}\int_{\mathcal{X}} f(x,z)p(x|z)p(z)dxdz &(4.27)\\ &= \int_{\mathcal{Z}}\left[\int_{\mathcal{X}} f(x,z)p(x|z)dx\right]p(z)dz &(4.28)\\ &\approx \frac{1}{L}\sum_{\ell=1}^{L} f(x,z^{(\ell)})p(x|z^{(\ell)})dx = \mathbb{E}_{\tilde{p}}[\mathbb{E}_p[f(x,z)|z]] &(4.29)\end{aligned}$$

Both estimators are unbiased and converge to $\mathbb{E}_p[f(x,z)]$ almost surely as $L\to\infty$. However, with the Rao-Blackwell Theorem, we know that the latter one has lower variance. Also intuitively, the underlying sample space of the marginalized estimator is $\mathcal{Z}$, which is smaller than the sample space of the original estimator $\mathcal{X}\times\mathcal{Z}$, thus the marginalized estimator should be more reliable (better accuracy) and converge faster (better efficiency). Especially we can often integrate out the model parameters in a hierarchical Bayes model by using conjugate prior, which makes the marginalized estimator feasible. Furthermore, the variance reduction guaranteed by the Rao-Blackwell theorem. All these reasons justify the collapsed Gibbs sampler.

Consider the FMM, assume $F(x_i|\theta_k)$ belongs to an exponential family and $G_0(\theta_k|\lambda)$ is a conjugate prior for θ_k. Integrating out π and $\{\theta_k\}_{k=1}^{K}$ means we need to draw samples from $p(z_i|\mathbf{z}_{-i},\mathbf{x},\alpha_0,\lambda)$. Factorize this distribution, we have:

$$\begin{aligned}& p(z_i=k|\mathbf{z}_{-i},\mathbf{x},\alpha_0,\lambda)\\ =\;& p(z_i=k|x_i,\mathbf{z}_{-i},\mathbf{x}_{-i},\alpha_0,\lambda) &(4.30)\\ \propto\;& p(z_i=k|\mathbf{z}_{-i},\mathbf{x}_{-i},\alpha_0,\lambda)p(x_i|z_i=k,\mathbf{z}_{-i},\mathbf{x}_{-i},\alpha_0,\lambda) &(4.31)\\ =\;& p(z_i=k|\mathbf{z}_{-i},\alpha_0)p(x_i|\mathbf{x}_{k,-i},\lambda) &(4.32)\end{aligned}$$

Here we use the Bayesian rule in (4.31) and apply the Markov property of the FMM graphical model in (4.32). The first term in (4.32) is due to the marginalization of π, whose details can be found in my technical report "Derivation of Gibbs Sampling for Finite Gaussian Mixture Model". The result of this term has been given by (3.7). The second term in (4.32) can be considered as a predictive likelihood of x_i given $\mathbf{x}_{k,-i}$, i.e., the other data currently assigned

to cluster k. It is due to the marginalization of θ_k, as we will see next. This distribution has analytic forms if $F(x_i|\theta_k)$ belongs to an exponential family and $G_0(\theta_k|\lambda)$ is a conjugate prior for θ_k. The derivations are as follows.

An exponential family of distribution is parameterized as:

$$p(x|\theta) = \exp\left(t(\theta)^T s(x) - \phi(x) - \psi(\theta)\right), \tag{4.33}$$

where $s(x)$ is the sufficient statistics vector, $t(\theta)$ the natural parameter vector, $\psi(\theta)$ is the log normalization quantity. The conjugate prior of θ is an exponential family distribution over θ with hyperparameter ν and η:

$$p(\theta|\nu,\eta) = \exp\left(t(\theta)^T \nu - \eta\psi(\theta) - \xi(\nu,\eta)\right). \tag{4.34}$$

The posterior given observation $\mathbf{x} = \{x_j\}_{j=1}^K$ is in the same form of $p(\theta)$ with updated hyperparameter $\tilde{\nu} = \nu + \sum_i s(x_i)$ and $\tilde{\eta} = \eta + n$:

$$\begin{aligned} & p(\theta|\mathbf{x},\nu,\eta) \\ = \;& \exp\left(t(\theta)^T(\nu + \sum_j s(x_j)) - (\eta+n)\psi(\theta) - \xi(\nu + \sum_i s(x_j), \eta+n)\right) \\ = \;& p(\theta|\tilde{\nu},\tilde{\eta}). \end{aligned} \tag{4.35}$$

The marginal probability can be obtained by simply apply the Bayes rule:

$$\begin{aligned} p(\mathbf{x}) &= \frac{p(\theta)p(\mathbf{x}|\theta)}{p(\theta|\mathbf{x})} \\ &= \exp\left(\xi(\nu + \sum_j s(x_j), \eta+n) - \xi(\nu,\eta) - \sum_j \phi(x_j)\right). \end{aligned} \tag{4.36}$$

Now go back to the predictive likelihood of x_i. This distribution can be obtained by marginalizing θ_k:

$$p(x_i|\mathbf{x}_{k,-i},\lambda) = \int p(x_i|\theta_k)p(\theta_k|\mathbf{x}_{k,-i},\lambda)d\theta_k. \tag{4.37}$$

This integration can be obtained directly by applying the results in (4.35) and (4.36). First, $p(\theta_k|\mathbf{x}_{k,-i},\lambda)$ is the posterior of θ_k after observing the data $\mathbf{x}_{k,-i}$, where θ_k is given a conjugate prior parameterized by $\lambda = (\nu,\eta)$, use (4.35) we get:

$$p(\theta_k|\mathbf{x}_{k,-i},\nu,\eta) = p(\theta_k|\tilde{\nu},\tilde{\eta}), \tag{4.38}$$

where $\tilde{\nu} = \nu + \sum_{\mathbf{x}_{k,-i}} s(x_j)$ and $\tilde{\eta} = \eta + n_{k,-i}$. Second, we take (4.38) as a new prior of θ with hyperparameter $\tilde{\nu}$ and $\tilde{\eta}$, and x_i as the observation. Notice $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$, thus (4.37) is actually the marginal probability of x_i

with the new prior of θ_k as in (4.38). Use the result in (4.36) and get

$$\begin{aligned} & p(x_i|\mathbf{x}_k, \lambda) \\ = \; & \exp\left(\xi(\tilde{\nu} + s(x_i), \tilde{\eta} + 1) - \xi(\tilde{\nu}, \tilde{\eta}) - \phi(x_i)\right) \\ = \; & \exp\left(\xi(\nu + \sum_{\mathbf{x}_{k,-i}} s(x_j) + s(x_i), \eta + n_{k,-i} + 1) - \xi(\nu + \sum_{\mathbf{x}_{k,-i}} s(x_j), \eta + n_{k,-i}) - \phi(x_i)\right) \\ \equiv \; & f_k(x_i; \mathcal{S}_k, n_k), \end{aligned} \tag{4.39}$$

where $\mathcal{S}_k \equiv \{s(x_j)\}_{\mathbf{x}_k}$ is the set of sufficient statistics for data in $\mathbf{x}_k$, and $f_k(x_i; \mathcal{S}_k, n_k)$ is defined as the predictive likelihood of x_i based on the observations in $\mathbf{x}_{k,-i}$. Notice that in case we need $\mathcal{S}_{k,-i} \equiv \{s(x_j)\}_{\mathbf{x}_{k,-i}}$, $\mathcal{S}_{k,-i}$ can be routinely obtained by excluding $s(x_i)$ from $\mathcal{S}_k$. Similarly, $n_{k,-i} = n_k - 1$ can be used when we need $n_{k,-i}$.

Substitute the results in (4.39) and (3.7) into (4.32), we get

$$p(z_i = k|\mathbf{z}_{(-i)}, \mathbf{x}, \alpha_0, \lambda) = \frac{n_{k,-i} + \alpha_0/K}{n + \alpha_0 - 1} \times f_k(x_i; \mathcal{S}_k, n_k) \tag{4.40}$$

We notice that all the information we need to compute (4.40) is the number and sufficient statistics of data points in each cluster, i.e., n_k and $\mathcal{S}_k$. So we only need to update these values when a new sample of z_i is drawn.

In summary, the collapsed Gibbs sampling for FMM as illustrated in Algorithm (5).

Similar to Algorithm 4, we can generalize FMM to DPMM by letting K go to infinity. By doing so, the conditional prior of z_i evolves from (3.7) to (3.8). When z_i is assigned to one of the current K clusters, the conditional posterior of z_i can be obtained by replacing the conditional prior $p(z_i = k|\mathbf{z}_{-i}, \alpha_0)$ in (4.32) by $\frac{n_{k,-i}}{n+\alpha_0-1}$, and consequently the term $\frac{n_{k,-i}+\alpha_0/K}{n+\alpha_0-1}$ in (4.40) is then replaced by $\frac{n_{k,-i}}{n+\alpha_0-1}$. If z_i is assigned to a new cluster index, which we denote as $K+1$ without loss generality, z_i's conditional posterior in this case is the same as (4.24).

In summary, the collapsed Gibbs sampling for DPMM as illustrated in Algorithm (6).

References

[1] M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 1995.

[2] B. J. Frey. Extending factor graphs so as to unify directed and undirected graphical models. In *UAI2003*, 2003.

[3] S. N. MacEarchern. Computational Methods for Mixture of Dirichlet Process Models. *Practical nonparametric and semiparametric Bayesian statistics*, 2:23–44, 1998.

Algorithm 5 Collapsed Gibbs sampling for FMM

Given $\{z_i^{(t-1)}\}_{i=1}^n$ from the previous iteration, sample a new set of $\{z_i^{(t)}\}_{i=1}^n$ as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers {1,...,n }.
2. Set $z = z^{(t-1)}$
3. For $i \in \tau(1), ..., \tau(n)$
   (a) Remove data item x_i from the cluster z_i, since we are going to sample a new z_i for x_i. This is done by updating $\mathcal{S}_{z_i}$ and n_{z_i}.
   (b) For each of the K clusters, compute the predictive likelihood $f_k(x_i; \mathcal{S}_k, n_k)$ using the information in $\{\mathcal{S}_k\}_{k=1}^K$ and $\{n_k\}_{k=1}^K$.
   (c) Draw a new sample for z_i from the following multinomial probabilities:
   $$p(z_i = k) \propto \frac{n_{k,-i} + \alpha_0/K}{n + \alpha_0 - 1} f_k(x_i; \mathcal{S}_k, n_k)$$
   (d) Update $\{\mathcal{S}_k\}_{k=1}^K$ and $\{n_k\}_{k=1}^K$ to reflect the new value of z_i
4. Set $z^{(t)} = z$
5. After the burn-in period, optionally, we can draw samples for $\pi^{(t)}$ and $\{\theta_k^{(t)}\}_{k=1}^K$ via Step 2 and 3 in Algorithm 2 respectively.

Algorithm 6 Collapsed Gibbs sampling for DPMM

Given $\alpha_0^{(t-1)}$ and $\{z_i^{(t-1)}\}_{i=1}^n$ from the previous iteration, sample a new set of $\{z_i^{(t)}\}_{i=1}^n$ as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers {1,...,n }.
2. Set $z = z^{(t-1)}$, $\alpha_0 = \alpha_0^{(t-1)}$
3. For $i \in \tau(1), ..., \tau(n)$
   (a) Remove data item x_i from the cluster z_i, since we are going to sample a new z_i for x_i. This is done by updating $\mathcal{S}_{z_i}$ and n_{z_i}.
   (b) If x_i is the only data in its current cluster, this cluster becomes empty after Step (3.a). This cluster is then removed, together with its parameter. This is done by updating $\mathcal{S}_{z_i}$ and n_{z_i}, and K is decreased by 1.
   (c) Re-arrange cluster indices so that $1, ..., K$ are active (i.e., non-empty)
   (d) For each of the K active clusters, compute the predictive likelihood $f_k(x_i; \mathcal{S}_k, n_k)$ using the information in $\{\mathcal{S}_k\}_{k=1}^K$ and $\{n_k\}_{k=1}^K$ as in (4.39). Also compute the predictive likelihood of the potential new cluster $f_{K+1}(x_i) \equiv \int F(x_i|\theta)G_0(\theta)d\theta$.
   (e) Draw a new sample for z_i from the following (K+1) multinomial probabilities:

$$\begin{aligned} p(z_i = k, k \leq K) &\propto \frac{n_{k,-i}}{n + \alpha_0 - 1} f_k(x_i; \mathcal{S}_k, n_k) \\ p(z_i = K + 1) &\propto \frac{\alpha_0}{n + \alpha_0 - 1} f_{K+1}(x_i). \end{aligned}$$

   (f) If $z_i = K + 1$, we get a new cluster. Index this cluster as $K + 1$, sample a new cluster parameter ϕ_i from $H(\phi_i|x_i)$ as defined in (4.2), assign it to θ_{K+1}, and increase K by 1
   (g) Update $\{\mathcal{S}_k\}_{k=1}^K$ and $\{n_k\}_{k=1}^K$ to reflect the new value of z_i
4. Set $z^{(t)} = z$
5. After the burn-in period, optionally, we can draw samples for $\{\theta_k^{(t)}\}_{k=1}^K$ via Step 3 in Algorithm 2.
6. If $\alpha_0 \sim \text{Gamma}(a, b)$, sample $\alpha_0^{(t)} \sim p(\alpha_0|K, n, a, b)$ via auxiliary variable method [1].

[4] R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[5] A. Ranganathan. *Probabilistic Topological Maps*. PhD thesis, Georgia Institute of Technology, 2008.

[6] C. E. Rasmussen. The Infinite Gaussian Mixture Model. 12:554–560, 2000.

[7] S.M.Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[8] E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, MIT, 2006.

[9] Y. W. Teh. Dirichlet Processes: Tutorial and Practical Course. Machine Learning Summer School, 2007.

[10] 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.

[11] M. West, P. Muller, and M. Escobar. Hierarchical priors and mixture models with applications in regression and density estimation. In P. R. Freeman and A. F. Smith, editors, *Aspects of Uncertainty*, pages 363– 386. John Wiley, 1994.