Module 7: Part IV: Gibbs Sampling, Data Augmentation, Mixture Models

Rebecca C. Steorts

Agenda

- Review of data augmentation
- A three component mixture model
- Dirichlet distribution
- The Dirichlet-Multinomial
- Return to the three component mixture model problem

Data augmentation for auxiliary variables

- A commonly-used technique for designing MCMC samplers is to use *data augmentation*, also known as *auxiliary variables*.
- Introduce variable(s) Z that depends on the distribution of the existing variables in such a way that the resulting conditional distributions, with Z included, are easier to sample from and/or result in better mixing.
- Z's are latent/hidden variables that are introduced for the purpose of simplifying/improving the sampler.

Idea: Create Z's and throw them away at the end!

- Suppose we want to sample from $p(x, y)$, but $p(x|y)$ and/or $p(y|x)$ are complicated.
- Choose

$$p(z|x, y)$$

  such that $p(x|y, z)$, $p(y|x, z)$, and $p(z|x, y)$ are easy to sample from.
- Then construct a Gibbs sampler to sample all three variables (X, Y, Z) from $p(x, y, z)$.
- Then we just throw away the Z's and we will have samples (X, Y) from $p(x, y)$.

Three component mixture model (Lab 8)

- Consider a three component mixture of normal distribution with a common prior on the mixture component means, the error variance and the variance within mixture component means.
- The prior on the mixture weights w is a three component Dirichlet distribution.

$$
\begin{aligned}
p(Y_i|\mu_1,\mu_2,\mu_3,w_1,w_2,w_3,\varepsilon^2) &= \sum_{j=1}^{3} w_j N(\mu_j,\varepsilon^2) \\
\mu_j|\mu_0,\sigma_0^2 &\sim N(\mu_0,\sigma_0^2) \\
\mu_0 &\sim N(0,3) \\
\sigma_0^2 &\sim \text{InverseGamma}(2,2) \\
(w_1,w_2,w_3) &\sim \text{Dirichlet}(1,1,1) \\
\varepsilon^2 &\sim \text{InverseGamma}(2,2),
\end{aligned}
$$

for $i = 1, \ldots n$.

Three component mixture model (Lab 8)

Specifically,

- w_1, w_2 and w_3 are the mixture weight of mixture components 1,2 and 3 respectively
- μ_1, μ_2 and μ_3 are the means of the mixture components
- ε^2 is the variance parameter of the error term around the mixture components.

Three component mixture model (Lab 8)

In order to be able to work on this problem, we need to:

1. We need to realize that the full conditionals as written cannot be easily sampled from. (Lab 8).
2. Next, we want to re-write the model using latent allocation variables to make it easier to work with.
3. Finally, in order to work with this model, we need to know about two distributions — the Dirichlet and the Multinomial. It's also essential to note that the Dirichlet is the conjugate prior for the Multinomial.

We will start by learning about the Dirichlet and Multinomial distributions and then come back to the three component mixture model problem.

Dirichlet

A Dirichlet distribution[1] is a distribution of the K-dimensional probability simplex[2]

$$\triangle_K = \{(\pi_1, \ldots, \pi_k) : \pi_k \geq 0, \sum_k \pi_k = 1\}.$$

We say that $(\pi_1, \ldots, \pi_k)$ is Dirichlet distributed:

$$(\pi_1, \ldots, \pi_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$$

if

$$p(\pi_1, \ldots, \pi_k) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_{k-1}}.$$

[1] This is the multivariate version of the Beta distribution.

[2] In geometry, a simplex is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions.

Dirichlet distribution

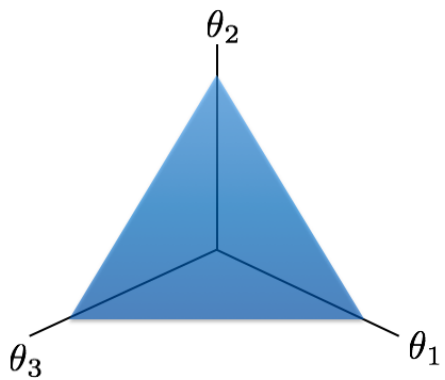
Let

$$\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$

where the probability density function is

$$p(\theta \mid \alpha) \propto \prod_{k=1}^{m} \theta_k^{\alpha_k - 1},$$

where $\sum_k \theta_k = 1, \theta_i \geq 0$ for all i



Dirichlet distribution

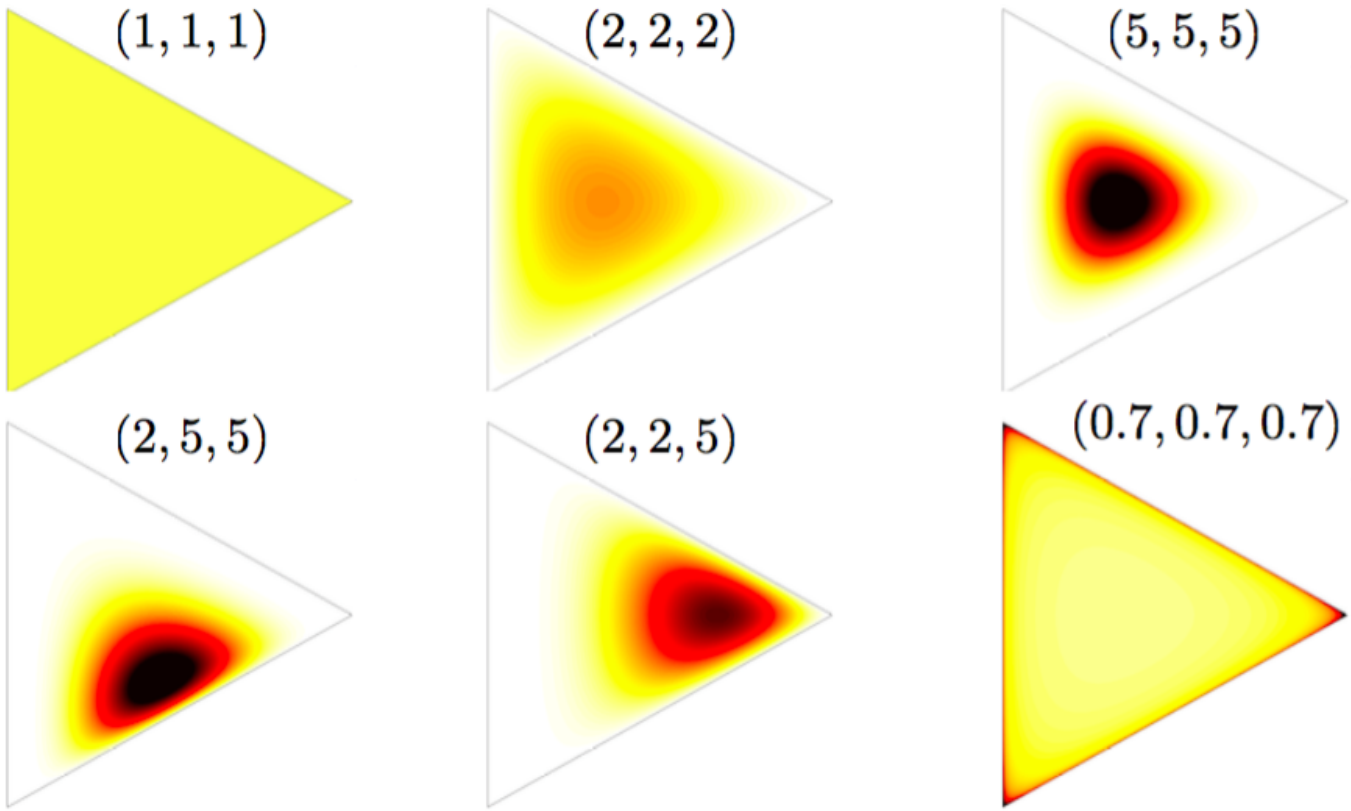


Figure 1: Far left: We get a uniform prior on the simplex. Moving to the right we get things unimodal. On the bottom, we get distributions that are multimodal at the corners.

Multinomial-Dirichlet

In order to proceed with the lab, we'll need to learn about the Multinomial or Categorical distribution.[3]

[3] This is the multivariate generalization of of the Binomial distribution.

Multinomial or Categorical distribution

Assume $X = (x_1, x_2, \ldots x_n)$ where $x_i \in \{1, \ldots, m\}$. Assume $\theta = (\theta_1, \ldots, \theta_m)$, where $\sum_i \theta_i = 1$.

Assume that

$$X \mid \theta \overset{ind}{\sim} \text{Multinomial}(\theta)$$

or

$$X \mid \theta \overset{ind}{\sim} \text{Categorical}(\theta)$$

$$P(X_i = j \mid \theta) = \theta_j$$

Conjugate prior (Dirichlet)

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Recall the density of the Dirichlet is the following:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1},$$

where $\sum_j \theta_j = 1, \theta_i \geq 0$ for all i

Likelihood

Define the data as $\boldsymbol{X} = (x_1, \ldots, x_n)$, $x_i \in \{1, \ldots m\}$. Consider

$$p(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{n} P(X_i = x_i \mid \theta) \tag{1}$$

$$= \prod_{i=1}^{n} \theta_{x_i} = \theta_{x_1} \times \theta_{x_2} \times \theta_{x_n} \tag{2}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{m} \theta_j^{I(x_i=j)} = \prod_{j=1}^{m} \prod_{i=1}^{n} \theta_j^{I(x_i=j)} \tag{3}$$

$$= \prod_{j=1}^{m} \theta_j^{\sum_i I(x_i=j)} \tag{4}$$

$$= \prod_{j=1}^{m} \theta_j^{c_j} \tag{5}$$

where $c = (c_1, \ldots c_m)$, $c_j = \#\{i : x_i = j\}$.

Likelihood, Prior, and Posterior

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{j=1}^{m} \theta_j^{c_j}$$

$$P(\boldsymbol{\theta}) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i)$$

Then

$$P(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto \prod_{j=1}^{m} \theta_j^{c_j} \times \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i) \tag{6}$$

$$\propto \prod_{j=1}^{m} \theta_j^{c_j + \alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i) \tag{7}$$

This implies

$$\boldsymbol{\theta} \mid \boldsymbol{X} \sim \text{Dirichlet}(\boldsymbol{c} + \boldsymbol{\alpha}).$$

Takeaways

1. Dirichlet is conjugate for Categorical or Multinomial.[4]
2. Useful formula:

$$\prod_i \text{Multinomial}(x_i \mid \theta) \times \text{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \propto \text{Dir}(\boldsymbol{\theta} \mid \boldsymbol{c} + \boldsymbol{\alpha}).$$

[4] The word Categorical seems to be used in CS and ML. The word Multinomial seems to be used in Statistics and Mathematics.

Three component mixture model

- Recall the three component mixture of normal distribution with a common prior on the mixture component means, the error variance and the variance within mixture component means.
- The prior on the mixture weights w is a three component Dirichlet distribution.

$$
\begin{aligned}
p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) &= \sum_{j=1}^{3} w_i N(\mu_j, \varepsilon^2) \\
\mu_j|\mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim \text{InverseGamma}(2, 2) \\
(w_1, w_2, w_3) &\sim \text{Dirichlet}(1, 1, 1) \\
\varepsilon^2 &\sim \text{InverseGamma}(2, 2),
\end{aligned}
$$

for $i = 1, \ldots n$.

Three component mixture model (Task 1 and 2)

Derive the full conditionals for all the parameters up to a normalizing constant and see that three of the conditional distributions are very difficult to sample from.

Task 1 and 2

Specifically, you should derive the following conditional distributions below:

- $p(w_1, w_2, w_3 | \mu_1, \mu_2, \mu_3, \varepsilon^2, Y_1, ..., Y_N) \propto$
- $p(\mu_1 | \mu_2, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_2 | \mu_1, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_3 | \mu_1, \mu_2, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\varepsilon^2 | \mu_1, \mu_2, \mu_3, Y_1, ..., Y_N) \propto$
- $p(\mu_0 | \mu_1, \mu_2, \mu_3, \sigma_0^2) \propto$
- $p(\sigma_0^2 | \mu_0, \mu_1, \mu_2, \mu_3) \propto$

Task 1 (Solution)

We start by deriving the full conditional kernels.

$$p(\mu_0|\mu_1, \mu_2, \mu_3, \varepsilon^2, \sigma_0^2) \propto \text{Normal-Normal mean update} \tag{8}$$

$$p(\sigma_0^2|\mu_1, \mu_2, \mu_3, \mu_0) \propto \text{Normal-InverseGamma variance update} \tag{9}$$

Task 1 (Solution)

$$p(\mu_k | Y_1, ..., Y_N, \sigma_0^2, \varepsilon^2, w_1, w_2, w_3) \tag{10}$$

$$\propto \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{\sigma_0^2}(\mu_k - \mu_0)^2} \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{11}$$

$$\propto? \tag{12}$$

Task 1 (Solution)

$$p(\varepsilon^2 | Y_1, ..., Y_N, \mu_1, \mu_2, \mu_3, w_1, w_2, w_3) \tag{13}$$

$$\propto (\varepsilon^2)^{-3} e^{-\frac{2}{\varepsilon^2}} \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{14}$$

$$\propto? \tag{15}$$

Task 1 (Solution)

$$p(w_1, w_2, w_3 | Y_1, ..., Y_N, \mu_1, \mu_2, \mu_3, \varepsilon^2) \tag{16}$$

$$\propto \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{17}$$

$$\propto ? \tag{18}$$

Note that everything that involves the likelihood includes the products of sums, and becomes exceedingly painful. Thus, let us look at the full conditionals under data augmentation.

Data augumentation scheme

- Neither the joint posterior nor any of the full conditionals involving the likelihood are of a form that's easy to sample from.

Solution: introduce an additional set of random variables $\{Z_i\}_{i=1}^N$ that assign each observation to one of the mixture components with the proabilitiy of assignment being the respective mixture weight.

If we condition on Z_i we can then write the likelihood of Y_i as

$$p(Y_i|Z_i, \mu_1, \mu_2, \mu_3, \varepsilon^2) = \sum_{j=1}^{3} N(\mu_j, \varepsilon^2)\delta_j(Z_i) = \sum_{j=1}^{3} N(\mu_{Z_i}, \varepsilon^2)$$

$$P(Z_i = j) = w_j.$$

Data augmentation (continued)

- Conditional on Z_i we no longer have a sum of Normal pdfs in our likelihood, resulting in a significant simplification.
- Conditional on the $\{Z_i\}$ updates will be straightforward, only depending on the mixture component that any given Y_i is currently assigned to.
- The drawback is that we also have to update $\{Z_i\}_{i=1}^N$ as well, introducing extra steps into our sampler.

The updated model

The model is now

$$
\begin{aligned}
Y_i \mid Z_i, \mu_1, \mu_2, \mu_3, \epsilon^2 &\sim \sum_{i=1}^{3} N(\mu_{Z_i}, \epsilon^2) \\
\mu_j \mid \mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
Z_i \mid w_1, w_2, w_3 &\sim \text{Cat}(3, \boldsymbol{w}) \\
\boldsymbol{w} = (w_1, w_2, w_3) &\sim \text{Dirichlet}(1, 1, 1) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim IG(2, 2) \\
\epsilon^2 &\sim IG(2, 2)
\end{aligned}
$$

$i = 1, \ldots, n$ $j = 1, \ldots, 3$

Task 3

Where necessary, (re)derive the full conditionals under the data augmentation scheme.

(See the lab solutions).

Task 4

In task 3 you derived all the full conditionals, and due to data augmentation scheme they are all in a form that is easy to sample. Use these full conditionals to implement Gibbs sampling using the data from "Lab8Mixture.csv".

Task 5

- Show traceplots for all estimated parameters
- Show means and 95% credible intervals for the marginal posterior distributions of all the parameters

Now suppose you re-run the sampler using 3 different starting values, are your results in a,b the same? Justify your reasoning with visualizations.

Sample code

Partial code for this problem can be found on github under lab 8.

Recap of Module 8 (Part I – Part IV)

1. We introduced the two-stage Gibbs sampler.
2. You should be able to derive conditional distributions. for two-stage Gibbs samplers. (See Part I, Module 8 for examples).
3. Be familar with diagnostic plots.
4. We then looked at a three-stage sampler and generalized to the multi-stage Gibbs sampler.
5. We looked at an application to censoring (a type of missing data here).
6. Why would we use latent variables in a Gibbs sampler? (We looked at these for Gaussian mixture models). Notice that the hierarhical modeling setup was more complicated here, which is why we used this trick.
7. In short, we saw many ways to use Gibbs sampling in many applications and various tricks that one needs to use in order to derive the full conditionals in closed form. This is always driven by the data and will vary by the model specified.

Exercise

Consider the following Exponential model for an observation x:

$$p(x|a,b) = ab\exp(-abx)\mathbb{1}(x > 0)$$

and suppose the prior is

$$p(a,b) = \exp(-a-b)\mathbb{1}(a,b > 0).$$

You want to sample from the posterior $p(a,b|x)$. Find the conditional distributions needed for implementing a Gibbs sampler.

Note: you did a generalization of this problem in lab.

Solution

The Gibbs sampler consists of alternately sampling from $a|b, x$ and $b|a, x$.

First note that the joint p.d.f. is

$$p(x, a, b) = ab \exp(-abx - a - b)\mathbb{1}(a, b, x > 0).$$

Thus,

$$\begin{aligned} p(a|b, x) &\underset{a}{\propto} p(x, a, b) \underset{a}{\propto} a \exp(-abx - a)\mathbb{1}(a > 0) \\ &= a \exp(-(bx + 1)a)\mathbb{1}(a > 0) \underset{a}{\propto} \text{Gamma}(a \mid 2,\ bx + 1). \end{aligned}$$

Therefore, $p(a|b, x) = \text{Gamma}(a \mid 2,\ bx + 1)$ and by symmetry, $p(b|a, x) = \text{Gamma}(b \mid 2,\ ax + 1)$.