Module 7: Part IV: Gaussian Mixture Models, Latent Variables, and Gibbs Sampling

Rebecca C. Steorts

This goes with Lab 8 (https://github.com/resteorts/modern-bayes/blob/master/labs/08-gibbs-augmentation/lab-8-partial-solutions.pdf), which has been prepared by Olivier Binette and Rebecca C. Steorts

Agenda

- ▶ Goals
- ▶ Background of Multinomial-Dirichlet conjugacy
- ▶ This corresponds with Lab 8 (and your ungraded homework this week)
- ▶ A three component mixture model
- ▶ Likelihood for the three component mixture model
- ▶ Building the model specification
- ▶ Moving to a latent variable approach
- ▶ Re-deriving the conditional distributions
- ▶ Coding up the Gibbs sampler
- ▶ Understanding the output and diagnostics

You can find Olivier's outline for Lab 8 here:
https://github.com/OlivierBinette/Labs-STA-360

Goal

The goal of this lecture, which corresponds with lab 8, is to introduce you to the **three component mixture model**.

This easily extends to any type of mixture model.

Background

Suppose that
$$X \sim \text{InverseGamma}(a, b).$$

Then
$$Y = 1/X \sim \text{Gamma}(a, b).$$

Proof: Similar proof here http://www.math.wm.edu/~leemis/chart/UDR/PDFs/GammaInvertedgamma.pdf.

Background

In order to work with this module, we will need to know to work with the Multinomial and Dirichlet distributions.

Background on the Mulinomial-Dirichlet

Before going through the lecture, we will first go over background material on the

- ▶ Multinomial distribution
- ▶ Dirichlet distribution

which can be found here https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-7/exercise-multinomial-dirichlet-with-blanks.pdf

The document contains blank pages, so you'll need to fill this in through the lecture today.

Big picture

▶ We will build a model specification based upon a **three component mixture model**
▶ What does this model look like?
▶ The posterior will be intractable? Why?
▶ What should we do in order to help us solve the problem?
▶ This will be your homework for this coming week, and it will be solved and worked in lab 8 with your TA's this week.
▶ You should have read through the lecture notes for class and attempted to work through the problems on your own before attending Lab 8.

Likelihood (three component mixture model)

For $i = 1, \ldots, n$

$$p(Y_i | \mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2)$$
$$= \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$
$$= w_1\, N(\mu_1, \varepsilon^2) + w_2\, N(\mu_2, \varepsilon^2) + w_3\, N(\mu_3, \varepsilon^2)$$

▶ 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.

Prior specification on likelihood terms

Let's specify the priors on

- 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.

Prior specification on likelihood terms

For $i = 1, \ldots, n$

$$p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$

$$\mu_j|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0) \qquad (1)$$
$$\varepsilon^2 \sim \text{InverseGamma}(2, 2) \qquad (2)$$
$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1) \qquad (3)$$

What is the Dirichlet$(1, 1, 1)$? This is the multivariate distribution of the Beta distribution.

Complete the model specification

Let's specify the priors on

- μ_0
- σ_0^2

Finalizing model specification

For $i = 1, \ldots, n$

$$p(Y_i | \mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$

$$\mu_j | \mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0) \qquad (4)$$
$$\varepsilon^2 \sim \text{InverseGamma}(2, 2) \qquad (5)$$
$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1) \qquad (6)$$

$$\mu_0 \sim N(0, 3) \qquad (7)$$
$$\sigma_0^2 \sim \text{InverseGamma}(2, 2) \qquad (8)$$

Transformed model

Let $\tau = \dfrac{1}{\epsilon^2}$ and $\phi_o = \dfrac{1}{\sigma_o^2}$.

$$p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2) \qquad (9)$$

$$\mu_j|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0) \qquad (10)$$

$$\tau = (1/\varepsilon^2) \sim \text{Gamma}(2, 2) \qquad (11)$$

$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1) \qquad (12)$$

$$\mu_0 \sim N(0, 3) \qquad (13)$$

$$\phi_o = (1/\sigma_0^2) \sim \text{Gamma}(2, 2) \qquad (14)$$

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.

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.

$$p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) = \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2)$$

$$\mu_j|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0^2)$$

$$\mu_0 \sim N(0, 3)$$

$$\tau = (1/\varepsilon^2) \sim \text{Gamma}(2, 2)$$

$$(w_1, w_2, w_3) \sim \text{Dirichlet}(1, 1, 1)$$

$$\mu_0 \sim N(0, 3)$$

$$\phi_o = (1/\sigma_0^2) \sim \text{Gamma}(2, 2)$$

for $i = 1, \ldots n.$

Task 1

Derive the joint posterior up to a normalizing constant. What do you observe?

Specifically, derive

$$p(w_1, w_2, w_3, \mu_1, \mu_2, \mu_3, \epsilon^2, \mu_o, \sigma_o^2 \mid y_{1:n})$$

up to a normalizing constant, where it may be helpful to let $\tau = \frac{1}{\epsilon^2}$, $\phi_o = \frac{1}{\sigma_o^2}$.

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) \right) \left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0) \right) p(\mu_0) p(\phi_0) p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right) \left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0) p(\phi_0) p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right) \left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau)\right) \left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0)\right) p(\mu_0)p(\phi_0)p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

$$p(\phi_0) = \text{Gamma}(\phi_0; 2, 2),$$

Task 1

Show that the full joint distribution can be written as follows:

$$\left(\prod_{i=1}^{n} p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) \right) \left(\prod_{j=1}^{3} p(\mu_j \mid \mu_0, \phi_0) \right) p(\mu_0) p(\phi_0) p(\tau);$$

$$p(Y_i \mid \mu_{1:3}, w_{1:3}, \tau) = \sum_{j=1}^{3} w_j N(Y_i; \mu_j, \tau),$$

$$p(\mu_j \mid \mu_0, \phi_0) = N(\mu_j; \mu_0, \phi_0^{-1}),$$

$$p(\mu_0) = N(\mu_0; 0, 3),$$

$$p(\phi_0) = \text{Gamma}(\phi_0; 2, 2),$$

$$p(\tau) = \text{Gamma}(\tau; 2, 2).$$

Task 2

Using Task 1, derive the full conditionals below up to a normalizing constant. What do you observe?

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

Task 2

Observe that the likelihood is difficult to work with. In order to more clearly see this, we will derive the full conditionals under our current specification so that we can see this more clearly.

In addition, for those that asked earlier in the course about more realistic models, this is an example of these.

Think on your own now about the tradeoffs regarding fully conjugate models versus these more realistic models that we are now working with.

Using latent variables

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

Using latent variables

Our data points $Y_1, \ldots, Y_n$ come from a mixture model:

$$Y_i \sim \sum_{j=1}^{3} w_j N(\mu_j, \varepsilon^2) \tag{15}$$

For each data point, define Z_i to be a latent variable such that

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

for $j = 1, 2, 3$. This implies that

$$Y_i \mid Z_i \sim N(\mu_{Z_i}, \varepsilon^2). \tag{16}$$

Latent variables

▶ 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

$$Y_i \mid Z_i, \mu_1, \mu_2, \mu_3, \epsilon^2 \sim N(\mu_{Z_i}, \epsilon^2)$$
$$\mu_j \mid \mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0^2)$$
$$Z_i \mid w_1, w_2, w_3 \sim \mathsf{Cat}(3, \mathbf{w})$$
$$\mathbf{w} = (w_1, w_2, w_3) \sim \mathsf{Dirichlet}(1, 1, 1)$$
$$\mu_0 \sim N(0, 3)$$
$$\sigma_0^2 \sim IG(2, 2)$$
$$\epsilon^2 \sim IG(2, 2)$$

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

Task 3

Using the latent variable mode, when re-derive the full conditional when necessary.

What do you observe?

Hint: You should be able to sample from all your full conditional distributions!

Task 4

In task 3, you should have found full conditionals that you can easily sample from.

Use the full conditionals from task 3 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 at
https://github.com/resteorts/modern-bayes/tree/master/labs/08-gibbs-augmentation

Before the next class, work through lab 8 on your own (or in groups).

On Friday, your TA will go through the lab and answer questions that you have. Please post to Piazza if you have specific questions that you'd like each TA to go through in more detail.

Detailed Summary

- Relationship between aamma and inverse gamma distributions
- Multinomial/Categorical distribution
- Dirichlet distribution
- Multinomial-Dirichelt conjugacy
- Three component mixture model
- Model specification (meaning prior specification)
- Full conditional derivations
- Using a latent variable model
- Re-deriving conditional derivations
- Coding a Gibbs sampler
- Diagnostics
- Analyzing the output (posterior densities and credible intervals)