

STAT 238 - Bayesian Statistics

Lecture Twenty Nine

Spring 2026, UC Berkeley

Aditya Guntuboyina

10 April 2026

1 Gibbs Sampler for Mixture Models

1.1 Known variances and proportion

We observe real-valued data y_1, \dots, y_n and consider the model:

$$y_i \stackrel{\text{i.i.d.}}{\sim} (1-w)N(\mu_0, 1) + wN(\mu_1, 1) \quad (1)$$

with w known and fixed (e.g., $w = 0.3$ or 0.7 as in the simulations) and μ_0, μ_1 being unknown. We will describe the Gibbs sampler algorithm in this setting first, and then generalize it to the case where w is unknown and we have variance parameters σ_0^2 and σ_1^2 .

The log-likelihood corresponding to (1) is:

$$\sum_{i=1}^n \log(p\phi(y_i, \mu_1, 1) + (1-p)\phi(y_i, \mu_2, 1)),$$

where $\phi(x, \mu, \sigma^2)$ is the normal density with mean μ , variance σ^2 evaluated at x .

We will take a standard prior for μ_0, μ_1 e.g., $\mu_0, \mu_1 \stackrel{\text{i.i.d.}}{\sim} N(0, C)$ for a large C . The posterior of μ_0, μ_1 is then:

$$\pi(\mu_0, \mu_1 \mid \text{data}) \propto \phi(\mu_0, 0, C)\phi(\mu_1, 0, C) \prod_{i=1}^n ((1-w)\phi(y_i, \mu_0, 1) + w\phi(y_i, \mu_1, 1)).$$

This posterior cannot be evaluated in closed form and numerical methods need to be used. A standard approach is to use the Gibbs sampler with augmentation. First observe that the model (1) can be rewritten in the following way:

$$z_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(w) \quad \text{and} \quad y_i \mid z_i = 1 \sim N(\mu_1, 1) \quad \text{and} \quad y_i \mid z_i = 0 \sim N(\mu_0, 1).$$

It should be clear that, under the above model, the marginal distribution of y_i coincides with (1). z_1, \dots, z_n can be thought of as unobserved latent variables which represent which of the two populations (corresponding to the distributions $N(\mu_0, 1)$ and $N(\mu_1, 1)$ respectively) the observation y_i comes from.

Gibbs sampler is implemented for jointly sampling from the posterior of $\mu_0, \mu_1, z_1, \dots, z_n$ given the data. This requires being able to sample from the full conditionals

$$z \mid \mu_0, \mu_1, y \quad \text{and} \quad (\mu_0, \mu_1) \mid z, y.$$

where $y = (y_1, \dots, y_n)$ is the data and $z = (z_1, \dots, z_n)$. It is easy to see that these full conditionals can be written in closed form as follows. Given $\mu_0, \mu_1, y_1, \dots, y_n$, the variables z_1, \dots, z_n are independent with

$$z_i \mid \mu_0, \mu_1, y \sim \text{Bernoulli} \left(\frac{w\phi(y_i, \mu_1, 1)}{w\phi(y_i, \mu_1, 1) + (1-w)\phi(y_i, \mu_0, 1)} \right).$$

On the other hand, given $z_1, \dots, z_n, y_1, \dots, y_n$, the variables μ_0, μ_1 are independent with

$$\mu_1 \mid z, y \sim N \left(\frac{\sum_{i=1}^n y_i z_i}{n_1 + (1/C)}, \frac{1}{n_1 + (1/C)} \right) \quad \text{and} \quad \mu_0 \mid z, y \sim N \left(\frac{\sum_{i=1}^n y_i (1 - z_i)}{n_0 + (1/C)}, \frac{1}{n_0 + (1/C)} \right)$$

where n_0 is the number of z_i 's that are equal to 0, and n_1 is the number of z_i 's that are equal to 1. Note that in the limit as $C \rightarrow \infty$, we can write

$$\mu_1 \mid z, y \sim N \left(\frac{\sum_{i=1}^n y_i z_i}{\sum_{i=1}^n z_i}, \frac{1}{\sum_{i=1}^n z_i} \right) \quad \text{and} \quad \mu_0 \mid z, y \sim N \left(\frac{\sum_{i=1}^n y_i (1 - z_i)}{\sum_{i=1}^n (1 - z_i)}, \frac{1}{\sum_{i=1}^n (1 - z_i)} \right)$$

Based on these full conditional distributions, the Gibbs sampler algorithm takes the following form:

1. Initialize $\mu_0^{(0)}, \mu_1^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$:
 - a) Generate $z_1^{(t)}, \dots, z_n^{(t)}$ via:

$$z_i^{(t)} \sim \text{Bernoulli} \left(\frac{w\phi(y_i, \mu_1^{(t)}, 1)}{w\phi(y_i, \mu_1^{(t)}, 1) + (1-w)\phi(y_i, \mu_0^{(t)}, 1)} \right).$$

- b) Generate $\mu_0^{(t+1)}, \mu_1^{(t+1)}$ via:

$$\mu_1^{(t+1)} \sim N \left(\frac{\sum_{i=1}^n y_i z_i^{(t)}}{\sum_{i=1}^n z_i^{(t)}}, \frac{1}{\sum_{i=1}^n z_i^{(t)}} \right)$$

$$\mu_0^{(t+1)} \sim N \left(\frac{\sum_{i=1}^n y_i (1 - z_i^{(t)})}{\sum_{i=1}^n (1 - z_i^{(t)})}, \frac{1}{\sum_{i=1}^n (1 - z_i^{(t)})} \right).$$

As we saw in the simulations in the last couple of lectures, initialization is very important. The log-likelihood can have multiple modes only one of which is the correct mode (in the sense of having large likelihood). If the Gibbs sampler is not properly initialized, the algorithm would sample from a spurious peak.

1.2 Unknown variances and proportion

Now consider the model:

$$y_i \stackrel{\text{i.i.d}}{\sim} (1-w)N(\mu_0, \sigma_0^2) + wN(\mu_1, \sigma_1^2).$$

The unknown parameters are now $w, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2$. We place the following independent priors on these variables:

$$w \sim \text{Beta}(a, b) \quad \text{and} \quad \mu_0, \mu_1 \sim N(m, s^2) \quad \text{and} \quad \sigma_0^2, \sigma_1^2 \sim IG(\alpha, \beta). \quad (2)$$

The proportion parameter w is constrained to lie in $[0, 1]$ so it is natural to take the Beta prior for it. A standard choice here is $a = b = 1$ (which corresponds to the uniform prior). For μ_0, μ_1 , we are using the $N(m, s^2)$ prior. Standard choice is $m = 0$ and s very large. The Inverse Gamma prior $IG(\alpha, \beta)$ corresponds to the density:

$$\propto x^{-\alpha-1} \exp(-\beta/x) I\{x > 0\}.$$

The standard uninformative prior for σ is $\propto \sigma^{-1} I\{\sigma > 0\}$. The corresponding prior density for σ^2 is also $\propto x^{-1} I\{x > 0\}$. This is a special case of $IG(\alpha, \beta)$ corresponding to $\alpha = \beta = 0$.

The prior (2) is therefore uses conjugate families while including the standard uninformative choices as special cases. The reason for conjugacy is that the full conditional distributions corresponding to the posterior distribution can be written in closed form. We shall look at the formulae in the next lecture.