

STAT 238 - Bayesian Statistics

Lecture Thirty

Spring 2026, UC Berkeley

Aditya Guntuboyina

13 April 2026

1 The Gibbs Sampler for Gaussian Mixture Models

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

$$y_i \stackrel{\text{i.i.d.}}{\sim} \sum_{j=1}^k w_j N(\mu_j, \sigma_j^2)$$

with unknown parameters (w_1, \dots, w_k) and (μ_j, σ_j^2) for $j = 1, \dots, k$. (w_1, \dots, w_k) is a probability vector (i.e., $w_j \geq 0$ and $\sum_j w_j = 1$).

We use the following priors:

$$\begin{aligned} (w_1, \dots, w_k) &\sim \text{Dirichlet}(a_1, \dots, a_k) \\ \mu_1, \dots, \mu_k &\stackrel{\text{i.i.d.}}{\sim} N(m, s^2) \\ \sigma_1^2, \dots, \sigma_k^2 &\stackrel{\text{i.i.d.}}{\sim} IG(\alpha, \beta). \end{aligned}$$

The default choices are $a_j = 1$ for all j , $m = 0$ and s^2 to be large, and α, β to be near zero. Note that the density of the Inverse Gamma distribution $IG(\alpha, \beta)$ is given by $x^{-\alpha-1} \exp(-\beta/x) I\{x > 0\}$. It is easy to check that this is simply the density of X^{-1} where $X \sim \text{Gamma}(\alpha, \beta)$.

To use the Gibbs sampler, we introduce latent variables z_1, \dots, z_n which represent group memberships. Specifically, z_i represents which of the k Gaussians the observation y_i is coming from. More precisely, z_i takes the values $1, \dots, k$ with probabilities w_1, \dots, w_k (i.e., $z_i \sim \text{Categorical}(w)$), and

$$y_i \mid z_i = j \sim N(\mu_j, \sigma_j^2).$$

The Gibbs sampler will then jointly sample from all the unknown variables:

$$w_1, \dots, w_k, \mu_1, \sigma_1^2, \dots, \mu_k, \sigma_k^2, z_1, \dots, z_n \mid y_1, \dots, y_n$$

The full conditionals corresponding to all the variables can be written in closed form as described below. We will use the notation $w = (w_1, \dots, w_k)$, $\mu = (\mu_1, \dots, \mu_k)$ and $\sigma = (\sigma_1, \dots, \sigma_k)$, also $y = (y_1, \dots, y_n)$ and $z = (z_1, \dots, z_n)$.

For the full conditional corresponding to z_1, \dots, z_n (i.e., the conditional distribution of z_1, \dots, z_n given y_1, \dots, y_n as well as $w_1, \dots, w_k, \mu_1, \sigma_1^2, \dots, \mu_k, \sigma_k^2$) is given by:

$$\mathbb{P}\{z_i = j \mid y_i, w, \mu, \sigma\} = r_{ij} := \frac{w_j \phi(y_i, \mu_j, \sigma_j^2)}{\sum_{j=1}^k w_j \phi(y_i, \mu_j, \sigma_j^2)}.$$

In other words,

$$z_i \mid y_i, w, \mu, \sigma \sim \text{Categorical}(r_i) \quad \text{where } r_i = (r_{i1}, \dots, r_{ik}).$$

The full conditional of $w = (w_1, \dots, w_k)$ (i.e., the conditional distribution of w given z, y, μ, σ) is:

$$w \mid z, y, \mu, \sigma \sim \text{Dirichlet}(a_1 + n_1, \dots, a_k + n_k) \quad \text{where } n_j := \sum_{i=1}^n I\{z_i = j\}.$$

The full conditional of μ_j (i.e., the conditional distribution of μ_j given z, y, σ) is:

$$\mu_j \mid z, y, \sigma \sim N\left(\frac{m/s^2 + \sum_{i:z_i=j} y_i/\sigma_j^2}{1/s^2 + n_j/\sigma_j^2}, \frac{1}{1/s^2 + n_j/\sigma_j^2}\right).$$

Note that we are also conditioning on σ_j above. If we do not condition on σ_j , then the conditional distribution of μ_j will be t -distributed and not normal distributed.

The full conditional of σ_j^2 (i.e., the conditional distribution of σ_j^2 given z, y, μ) is:

$$\sigma_j^2 \mid z, y, \mu \sim IG\left(\alpha + \frac{n_j}{2}, \beta + \frac{1}{2} \sum_{i:z_i=j} (y_i - \mu_j)^2\right),$$

where again $n_j = \sum_{i=1}^n I\{z_i = j\}$.

So the Gibbs sampler algorithm is given by:

1. Initialize the parameters $w^{(0)}, \mu^{(0)}, \sigma^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$:

a) Generate $z_1^{(t)}, \dots, z_n^{(t)}$ via:

$$z_i^{(t)} \sim \text{Categorical}\left(\frac{w_j^{(t)} \phi(y_i, \mu_j^{(t)}, (\sigma_j^{(t)})^2)}{\sum_{j=1}^k w_j^{(t)} \phi(y_i, \mu_j^{(t)}, (\sigma_j^{(t)})^2)}, j = 1, \dots, k\right)$$

b) Calculate

$$n_j^{(t)} := \sum_{i=1}^n I\{z_i^{(t)} = j\} \quad \text{for } j = 1, \dots, k.$$

c) Generate $w^{(t+1)} = (w_1^{(t+1)}, \dots, w_k^{(t+1)})$ via

$$w^{(t+1)} \sim \text{Dirichlet}(a_1 + n_1^{(t)}, \dots, a_k + n_k^{(t)}).$$

d) Generate $\mu_1^{(t+1)}, \dots, \mu_k^{(t+1)}$ via

$$\mu_j^{(t+1)} \sim N \left(\frac{m/s^2 + \sum_{i:z_i^{(t)}=j} y_i / (\sigma_j^{(t)})^2}{1/s^2 + n_j^{(t)} / (\sigma_j^{(t)})^2}, \frac{1}{1/s^2 + n_j^{(t)} / (\sigma_j^{(t)})^2} \right).$$

e) Generate $\sigma_1^{(t+1)}, \dots, \sigma_k^{(t+1)}$ via

$$(\sigma_j^2)^{(t+1)} \sim IG \left(\alpha + \frac{n_j^{(t)}}{2}, \beta + \frac{1}{2} \sum_{i:z_i^{(t)}=j} (y_i - \mu_j^{(t+1)})^2 \right),$$

2 The EM Algorithm

The EM algorithm can be seen as a deterministic variant of the Gibbs sampler. We shall look at the EM in more generality later. In the case of fitting finite normal mixtures, the EM algorithm is given by the following:

1. Initialize the parameters $w^{(0)}, \mu^{(0)}, \sigma^{(0)}$.
2. Repeat the following for $t = 0, 1, 2, \dots$ until convergence:
 - a) **E-step.** Compute the responsibility of component j for observation i :

$$r_{ij}^{(t)} := \frac{w_j^{(t)} \phi(y_i, \mu_j^{(t)}, (\sigma_j^{(t)})^2)}{\sum_{l=1}^k w_l^{(t)} \phi(y_i, \mu_l^{(t)}, (\sigma_l^{(t)})^2)}, \quad j = 1, \dots, k.$$

b) Compute the effective counts

$$n_j^{(t)} := \sum_{i=1}^n r_{ij}^{(t)}, \quad j = 1, \dots, k.$$

c) **M-step.** Update the weights:

$$w_j^{(t+1)} := \frac{n_j^{(t)}}{n}, \quad j = 1, \dots, k.$$

d) Update the means:

$$\mu_j^{(t+1)} := \frac{\sum_{i=1}^n r_{ij}^{(t)} y_i}{n_j^{(t)}}, \quad j = 1, \dots, k.$$

e) Update the variances:

$$(\sigma_j^2)^{(t+1)} := \frac{\sum_{i=1}^n r_{ij}^{(t)} (y_i - \mu_j^{(t+1)})^2}{n_j^{(t)}}, \quad j = 1, \dots, k.$$

There are the following close connections between the EM algorithm and the Gibbs sampler:

- The **E-step** replaces sampling $z_i \sim \text{Categorical}(\dots)$ with computing its expectation

$$r_{ij}^{(t)} = \mathbb{E} \left[I\{z_i = j\} \mid y_i, w^{(t)}, \mu^{(t)}, \sigma^{(t)} \right],$$

which are the same normalised weights used in the Categorical draw.

- The **effective count** $n_j^{(t)} = \sum_{i=1}^n r_{ij}^{(t)}$ replaces the integer count $\sum_{i=1}^n I\{z_i^{(t)} = j\}$.
- The **weight update** replaces the Dirichlet draw and takes the maximum likelihood estimate

$$w_j^{(t+1)} = \frac{n_j^{(t)}}{n},$$

which is equivalent to the posterior mean of the Dirichlet as the prior becomes uninformative.

- The **mean update** replaces the Gaussian conjugate draw and is the responsibility-weighted sample mean

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n r_{ij}^{(t)} y_i}{n_j^{(t)}},$$

the maximum likelihood estimate with no prior shrinkage toward m .

- The **variance update** replaces the inverse-Gamma draw and is the responsibility-weighted sample variance

$$(\sigma_j^2)^{(t+1)} = \frac{\sum_{i=1}^n r_{ij}^{(t)} \left(y_i - \mu_j^{(t+1)} \right)^2}{n_j^{(t)}},$$

the maximum likelihood estimate with no α, β regularisation terms.