Multivariate Mixture of Normals

by Hedibert Freitas Lopes
Graduate School of Business
University of Chicago

Model and prior setting

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample from a (multivariate) mixture of normals, i.e.

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{S}) = \sum_{k=1}^{K} \pi_k dN(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{S}) \tag{1}$$

where $dN(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{S})$ denotes the probability density function of a (multivariate) normal with mean (vector) $\boldsymbol{\mu}$ and variance(-covariance matrix) $\boldsymbol{S}$, evaluated at $\boldsymbol{x}$, with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K)$.

Traditionally, latent indicator variables z_i are included in the model for clarity in such a way that

$$Pr(z_i = j) = \pi_j \tag{2}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, K$. Therefore, the likelihood in (1) can be rewritten as,

$$p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{S}) = \prod_{j=1}^{K} \prod_{i \in I_j} dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S}) \tag{3}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{z} = (z_1, \ldots, z_n)$, $I_j = \{i : z_i = j, i = 1, \ldots, n\}$ and $n_j = card(I_j)$. Consequently, when combining the likelihood in (3) with prior distributions for $\boldsymbol{z}$ and $\boldsymbol{\mu}_k$,

$$\boldsymbol{z}|\boldsymbol{\pi} \sim M(K, \boldsymbol{\pi}) \quad \text{and} \quad \boldsymbol{\mu}_k \sim N(\boldsymbol{\mu}_{k0}, \boldsymbol{V}_{k0})$$

respectively, the joint posterior for $\boldsymbol{\mu}$ and $\boldsymbol{z}$ is,

$$p(\boldsymbol{\mu}, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{S}) \propto \left[\prod_{j=1}^{K} \prod_{i \in I_j} dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S})\right] \left[\prod_{j=1}^{K} dN(\boldsymbol{\mu}_j|\boldsymbol{\mu}_{j0}, \boldsymbol{V}_{j0})\right] \tag{4}$$

Obviously, analytically tratable posterior inference is impossible. Next, we present the full conditional distributions $\boldsymbol{z}$, $\boldsymbol{\mu}$, $\boldsymbol{\pi}$ and $\boldsymbol{S}$, to be used in the Gibbs sampler algorithm.

Full conditionals

Full conditionals for $\boldsymbol{\mu}$ and $\boldsymbol{z}$ are given below. Notice that the full conditional for the elements of $\boldsymbol{z}$ are not conditional on $\boldsymbol{\mu}$. We believe that this strategy improves the Markov chain mixing since it eliminates, at least partially, high dependence of $\boldsymbol{\mu}$ and $\boldsymbol{z}$, usually observed in practical applications.

Full conditionals $\boldsymbol{\eta}$

If the prior distribution of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is a Dirichlet with parameter $\boldsymbol{\eta}_0$, ie. $\boldsymbol{\pi} \sim Dir(\boldsymbol{\eta}_0)$, then

$$\boldsymbol{\pi}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{S} \sim Dir(\boldsymbol{\eta}_0 + \boldsymbol{\eta})$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ and $\eta_j = \sum_{i=1}^n 1(z_i = j)$.

Full conditionals of $\boldsymbol{\Phi} = \boldsymbol{S}^{-1}$

If the prior distribution of $\boldsymbol{\Phi}$ is a Wishart with parameters ν_0 and $\nu_0^{-1}\boldsymbol{\Phi}_0$, ie. $\boldsymbol{\Phi} \sim W(\nu_0, \nu_0^{-1}\boldsymbol{\Phi}_0)$, then

$$\boldsymbol{\Phi}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\pi} \sim W(\nu_1, \nu_1^{-1}\boldsymbol{\Phi}_1)$$

where

$$\nu_1 = \nu_0 + n \quad \text{and} \quad \nu_1^{-1}\boldsymbol{\Phi}_1 = \nu_0^{-1}\boldsymbol{\Phi}_0 + \sum_{j=1}^{K} \sum_{i \in I_j} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)'$$

Full conditionals of $\boldsymbol{\mu}$

Sampling $\boldsymbol{\mu}$ given $\boldsymbol{z}$ and $\boldsymbol{x}$ is straightforward. It is easily shown that

$$p(\boldsymbol{\mu}|\boldsymbol{z}, \boldsymbol{x}) \;\propto\; \prod_{j=1}^{K} \left\{ \left[\prod_{i \in I_j} dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S}) \right] dN(\boldsymbol{\mu}_j|\boldsymbol{\mu}_{j0}, \boldsymbol{V}_{j0}) \right\} \tag{5}$$

which has the kernel of a multivariate normal distribution with mean vector and covariance matrix given by

$$\boldsymbol{\mu}_j = \boldsymbol{V}_{j1} \left(n_j \boldsymbol{S}^{-1} \tilde{\boldsymbol{x}}_j + \boldsymbol{V}_{j0}^{-1} \boldsymbol{\mu}_{j0} \right) \quad \text{and} \quad \boldsymbol{V}_{j1} = \left(n_j \boldsymbol{S}^{-1} + \boldsymbol{V}_{j0}^{-1} \right)^{-1}$$

respectively, for $n_j \tilde{\boldsymbol{x}}_j = \sum_{i \in I_j} \boldsymbol{x}_i$.

Full conditionals of $\boldsymbol{z}$: case I

Sampling z_i given $\boldsymbol{z}_{(i)} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, $\boldsymbol{\mu}$ and $\boldsymbol{x}$ is also relatively simple:

$$p(z_i = j|\boldsymbol{z}_{(i)}, \boldsymbol{\mu}, \boldsymbol{x}) \propto Pr(z_i = j) dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S}) \equiv q(z_i)$$

such that

$$p(z_i|\boldsymbol{z}_{(i)}, \boldsymbol{\mu}, \boldsymbol{x}) = \frac{\pi_j dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S})}{\sum_{l=1}^{K} \pi_l dN(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{S})}$$

The main drawback of previous sampling scheme is that $z_1, z_2, \ldots, z_n$ are highly correlated and that can significantly affect the performance of the MCMC algorithm.

Full conditionals of $\boldsymbol{z}$: case II

An alternative is to sample z_i given $\boldsymbol{z}_{(i)}$ and $\boldsymbol{x}$, ie. by integrating out $\boldsymbol{\mu}$. Initially,

$$\begin{aligned} Pr(z_i = j|\boldsymbol{z}_{(i)}, \boldsymbol{x}) &\propto Pr(z_i = j|\boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) p(\boldsymbol{x}_i|z_i = j, \boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) \\ &\propto Pr(z_i = j) p(\boldsymbol{x}_i|z_i = j, \boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) \end{aligned}$$

where $\boldsymbol{x}_{(i)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n)$ and $Pr(z_i = j) = \pi_j$. Also,

$$\begin{aligned} p(\boldsymbol{x}_i|z_i = j, \boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) &= \int p(\boldsymbol{x}_i|z_i = j, \boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}, \boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) d\boldsymbol{\mu} \\ &= \int dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S}) p(\boldsymbol{\mu}_j|\boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) d\boldsymbol{\mu}_j \\ &\times \prod_{l \neq j} \int p(\boldsymbol{\mu}_l|\boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) d\boldsymbol{\mu}_l \end{aligned}$$

where the last product of integrals is equal to one, following the conditional independence of $\boldsymbol{\mu}$'s given $\boldsymbol{x}$ and $\boldsymbol{z}$ shown above. It is easy to show that

$$p(\boldsymbol{\mu}_j|\boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) = dN(\boldsymbol{\mu}_j|\boldsymbol{\mu}_{j1,i}, \boldsymbol{V}_{j1,i})$$

with $\boldsymbol{V}_{j1,i} = \left(n_{j,i}\boldsymbol{S}^{-1} + \boldsymbol{V}_{j0}^{-1}\right)^{-1}$, $\boldsymbol{\mu}_{j1,i} = \boldsymbol{V}_{j1,i}\left(n_{j,i}\boldsymbol{S}^{-1}\tilde{\boldsymbol{x}}_{j,i} + \boldsymbol{V}_{j0}^{-1}\boldsymbol{\mu}_{j0}\right)$, $I_{j,i} = \{l : z_l = j, l = 1, \ldots, n \text{ and } l \neq i\}$, $n_{j,i} = card(I_{j,i})$, and $n_{j,i}\tilde{\boldsymbol{x}}_{j,i} = \sum_{l \in I_{j,i}} \boldsymbol{x}_l$. Therefore,

$$\begin{aligned} p(\boldsymbol{x}_i|z_i = j, \boldsymbol{z}_{(i)}, \boldsymbol{x}_{(i)}) &\propto \int dN(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{S}) dN(\boldsymbol{\mu}_j|\boldsymbol{\mu}_{j1,i}, \boldsymbol{V}_{j1,i}) d\boldsymbol{\mu}_j \\ &\propto dN(\boldsymbol{x}_i|\boldsymbol{\mu}_{j1,i}, \boldsymbol{V}_{j1,i} + \boldsymbol{S}) \end{aligned}$$

and

$$Pr(z_i = j|\boldsymbol{z}_{(i)}, \boldsymbol{x}) \propto \pi_j dN(\boldsymbol{x}_i|\boldsymbol{\mu}_{j1,i}, \boldsymbol{V}_{j1,i} + \boldsymbol{S}) \tag{6}$$

which is easy to sample from.