Bayesian inference for mixture models

M. Concepción Ausín
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative Methods
Master in Mathematical Engineering

Contents

Finite mixtures

A finite mixture of k densities of the same distribution is a convex combination,

$$f\left(x \mid k, \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \sum_{i=1}^{k} \rho_i f\left(x \mid \boldsymbol{\theta}_i\right),$$

of densities $f\left(x \mid \boldsymbol{\theta}_i\right)$, where $\boldsymbol{\rho} = (\rho_1, ..., \rho_k)$ such that $\sum_{i=1}^{k} \rho_i = 1$.

Remarks

- Mixture models are frequently referred as semi-parametric models as their flexibility allow to approximate non-parametric problems.
- Mixture component do not always have a physical meaning, they can describe complex behaviour of data in different research areas: biology, astronomy, engineering...

Finite mixtures

- Note that $E\left[X^r\right] = \sum_{i=1}^{k} \rho_i E\left[X^r \mid \boldsymbol{\theta}_i\right]$
- Computationally intensive methods must be considered for inference in mixture models: MCMC methods, EM algorithm,...
- The Bayesian approach using MCMC methods allows us to transform the complex structure of a mixture model in a set of simple structures using latent variables.

Finite mixtures

Example (Gaussian mixtures)

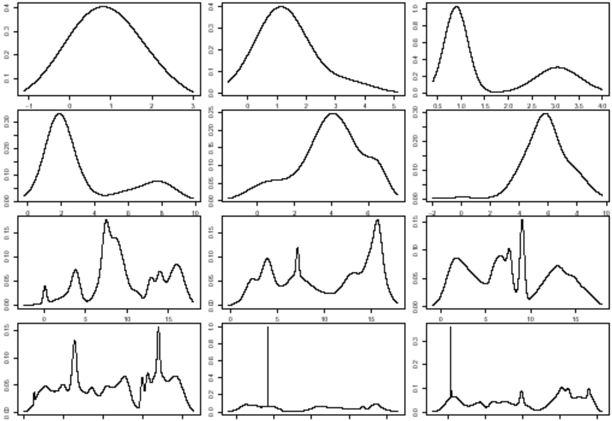
A finite Gaussian mixture of size k has the following density:

$$f\,(x \mid k, \boldsymbol{\rho}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \rho_i f_N(x \mid \mu_i, \phi_i),$$

where $\boldsymbol{\theta} = (\mu_1, \phi_1, \ldots, \mu_k, \phi_k)$ and $f_N(x \mid \mu_i, \phi_i)$ is the density of a Gaussian distribution with mean μ_i and precision ϕ_i.

Finite mixtures

The following figure shows various density functions of Gaussian mixtures with $k = 2$ components (first row), $k = 5$ components (second row), $k = 25$ components (third row) and $k = 50$ components (fourth row):



Bayesian inference for finite mixtures

Assume we have n observations $\mathbf{x} = (x_1, \ldots, x_n)$ sampled i.i.d. from a finite mixture distribution with density,

$$f(x \mid \boldsymbol{\rho}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \rho_i f(x \mid \boldsymbol{\theta}_i),$$

where k is finite and known.

We wish to make Bayesian inference for the model parameters $(\boldsymbol{\rho}, \boldsymbol{\theta})$. The likelihood is,

$$l(\boldsymbol{\rho}, \boldsymbol{\theta} \mid \mathbf{x}) = \prod_{j=1}^{n} \sum_{i=1}^{k} \rho_i f(x_j \mid \boldsymbol{\theta}_i),$$

which is given by k^n terms, which implies a large computational cost for a not very large sample size, n.

Bayesian inference for finite mixtures

In order to simplify the likelihood, we can introduce latent variables Z_j such that:

$$X_j \mid Z_j = i \sim f(x \mid \boldsymbol{\theta}_i) \qquad \text{and} \qquad P(Z_j = i) = \rho_i.$$

These auxiliary variables allows us to identify the mixture component each observation has been generated from.

Therefore, for each sample of data $\mathbf{x} = (x_1, \ldots, x_n)$, we assume a missing data set $\mathbf{z} = (z_1, \ldots, z_n)$, which provide the labels indicating the mixture components from which the observations have been generated.

Bayesian inference for finite mixtures

Using this missing data set, the likelihod simplifies to:

$$l(\boldsymbol{\rho}, \boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z}) = \prod_{j=1}^{n} \rho_{z_j} f\left(x_j \mid \boldsymbol{\theta}_{z_j}\right)$$
$$= \prod_{i=1}^{k} \rho_i^{n_i} \left[\prod_{j:z_j=i} f\left(x_j \mid \boldsymbol{\theta}_i\right)\right],$$

where $n_i = \#\{z_j = i\}$ and $\Sigma n_i = n$.

Then, the posterior probability that the observation x_j has been generated from the i-th component is:

$$P\left(z_j = i \mid x_j, \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \frac{\rho_i f\left(x_j \mid \boldsymbol{\theta}_i\right)}{\sum_{i=1}^{k} \rho_i f\left(x_j \mid \boldsymbol{\theta}_i\right)}.$$

Bayesian inference for finite mixtures

Example (Bayesian inference for Gaussian mixtures)

Using the missing data, $\mathbf{z} = (z_1, ..., z_n)$ the likelihood simplifies to:

$$l\left(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z}\right) \propto \prod_{i=1}^{k} \left(\rho_i \phi_i\right)^{n_i} \exp\left(-\frac{\phi_i}{2} \sum_{j:z_j=i} \left(x_j - \mu_i\right)^2\right),$$

where $n_i = \#\{z_j = i\}$.

And we have that:

$$P\left(z_j = i \mid x_j, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi}\right) = \frac{\rho_i \phi_i \exp\{-\frac{\phi_i}{2}\left(x_j - \mu_i\right)^2\}}{\sum_{i=1}^{k} \rho_i \phi_i \exp\{-\frac{\phi_i}{2}\left(x_j - \mu_i\right)^2\}}$$

Bayesian inference for finite mixtures

For the model parameters, $\boldsymbol{\rho}$, $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$, we assume conjugate priors:

Prior	Posterior
$\boldsymbol{\rho} \sim D(\delta_1, \ldots, \delta_k)$	$\boldsymbol{\rho} \mid \mathbf{x}, \mathbf{z} \sim D(\delta_1^*, \ldots, \delta_k^*)$
$\phi_i \sim G(a/2, b/2)$	$\phi_i \mid \mathbf{x}, \mathbf{z} \sim G(a_i^*/2, b_i^*/2)$
$\mu_i \mid \phi_i \sim N\left(m_i, \frac{1}{\alpha_i \phi_i}\right)$	$\mu_i \mid \mathbf{x}, \mathbf{z}, \phi_i \sim N\left(m_i^*, \frac{1}{\alpha_i^* \phi_i}\right)$

where

$$\delta_i^* = \delta_i + n_i, \qquad a_i^* = a + n_i,$$

$$b_i^* = b + \sum_{j:z_j=i} (x_j - \mu_i)^2, \qquad \alpha_i^* = \alpha_i + n_i,$$

$$m_i^* = \frac{\alpha_i m_i + n_i \bar{x}_i}{\alpha_i + n_i}, \qquad \text{where } \bar{x}_i = \frac{1}{n_i} \sum_{j:z_j=i} x_j.$$

For identifiability reasons, we assume that $\mu_1 < \ldots < \mu_k$.

Bayesian inference for finite mixtures

Note that $D(\delta_1, \ldots, \delta_k)$ a Dirichlet distribution with density:

$$f(\rho_1, \ldots, \rho_k) \propto \prod_{i=1}^{k} \rho_i^{\delta_i - 1}.$$

The usual prior choice is to take $(\delta_1, \ldots, \delta_k) = (1, \ldots, 1)$ to impose a uniform prior over the mixture weights.

Note that this prior choice is equivalent to use the following reparameterization:

$$\begin{aligned} \rho_1 &= \eta_1, \\ \rho_i &= (1 - \eta_1) \ldots (1 - \eta_{i-1}) \, \eta_i \end{aligned}$$

assuming that $\eta_i \sim \mathcal{B}(1, k - i + 1)$.

Bayesian inference for finite mixtures

MCMC algorithm

1. Set initial values $\boldsymbol{\eta}^{(0)}, \boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\phi}^{(0)}$.
2. Update $\mathbf{z}$ sampling from $\mathbf{z}^{(j+1)} \sim \mathbf{z}|\mathbf{x}, \boldsymbol{\rho}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\phi}^{(j)}$.
3. Update $\boldsymbol{\eta}$ sampling from $\boldsymbol{\eta}^{(j+1)} \sim \boldsymbol{\eta}|\mathbf{x}, \mathbf{z}^{(j+1)}$.
4. Update ϕ_i sampling from $\phi_i^{(j+1)} \sim \phi_i|\ \mathbf{x}, \mathbf{z}^{(j+1)}$.
5. Update μ_i sampling from $\mu_i^{(j+1)} \sim \mu_i|\mathbf{x}, \mathbf{z}^{(j+1)}, \phi_i^{(j+1)}$.
6. Order $\boldsymbol{\mu}^{(j+1)}$ and arrange $\boldsymbol{\rho}^{(j+1)}$ y $\boldsymbol{\phi}^{(j+1)}$ with this order.
7. $j = j + 1$. Go to 2.

Infinite mixtures

Now, consider an infinite mixture of densities of the same distribution,

$$f\left(x \mid \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \sum_{i=1}^{\infty} \rho_i f\left(x \mid \boldsymbol{\theta}_i\right),$$

of densities $f\left(x \mid \boldsymbol{\theta}_i\right)$, where $\boldsymbol{\rho} = (\rho_1, \rho_2, ...)$ such that $\sum_{i=1}^{\infty} \rho_i = 1$.

Suppose that we reparametrize the weights such that:

$$\begin{aligned} \rho_1 &= \eta_1, \\ \rho_i &= (1 - \eta_1) \ldots (1 - \eta_{i-1})\, \eta_i \end{aligned}$$

and assume a priori that:

$$\begin{aligned} \eta_i &\sim \mathcal{B}(1, \alpha), \\ \theta_i &\sim P_0, \end{aligned}$$

for $i = 1, 2, \ldots$.

Infinite mixtures

Note that this infinite mixture model with the considered prior choice corresponds to a Dirichlet process mixture model (DPM model) given by,

$$\begin{aligned} X_i \mid \boldsymbol{\theta}_i &\sim f(x \mid \boldsymbol{\theta}_i), \\ \boldsymbol{\theta}_i \mid P &\sim P(\boldsymbol{\theta}) \\ P &\sim DP(\alpha, P_0) \end{aligned}$$

or equivalently, using the stick-breaking representation,

$$\begin{aligned} x_j|z_j &\sim f\left(x \mid \theta_{z_j}\right) \\ \Pr\left(z_j = i\right) &= \rho_i, \\ \theta &\sim P_0 \\ \rho_1 = \eta_1, \qquad \rho_i &= \left(1 - \eta_1\right) \ldots \left(1 - \eta_{i-1}\right) \eta_s \\ \eta_i &\sim \mathcal{B}\left(1, \alpha\right) \end{aligned}$$

Bayesian inference for infinite mixtures

Observe that even using the latent variables, z_j, the likelihood is complicated:

$$l(\boldsymbol{\rho}, \boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{\infty} \rho_i^{n_i} \left[\prod_{j:z_j=i} f(x_j \mid \boldsymbol{\theta}_i) \right],$$

where $n_i = \#\{z_j = i\}$ and $\Sigma n_i = n$.

And the posterior probability that the observation x_j has been generated from the i-th component is difficult to evaluate:

$$P(z_j = i \mid x_j, \boldsymbol{\rho}, \boldsymbol{\theta}) = \frac{\rho_i f(x_j \mid \boldsymbol{\theta}_i)}{\sum_{i=1}^{\infty} \rho_i f(x_j \mid \boldsymbol{\theta}_i)}.$$

Bayesian inference for infinite mixtures

To solve this problem, Walker (2007) proposes to introduce a new set of latent variables, $\mathbf{u} = (u_1, \ldots, u_n)$ such that,

$$f\left(x_j, u_j | \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \sum_{i=1}^{\infty} I\left(u_j < \rho_i\right) f\left(x_j \mid \theta_i\right),$$

where I is the indicator function. Observe that integrating over u_j the marginal density is $f\left(x \mid \boldsymbol{\rho}, \boldsymbol{\theta}\right)$. Also note that we can write,

$$f\left(x_j, u_j | \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \sum_{i=1}^{\infty} \rho_i f_U\left(u_j \mid 0, \rho_i\right) f\left(x_j \mid \theta_i\right),$$

where f_U is the density of a uniform $U(0, \rho_i)$. Then, with probability ρ_i, the auxiliary variable u_j follows a uniform distribution in $(0, \rho_i)$ and the variable x_j follows the density $f\left(x_j \mid \theta_i\right)$.

Bayesian inference for infinite mixtures

With this new set of latent variables, the complete likelihood function is,

$$l(\boldsymbol{\rho}, \boldsymbol{\theta} \mid \mathbf{x}, \mathbf{u}, \mathbf{z}) \propto \prod_{j=1}^{n} I\left(u_j < \rho_{z_j}\right) f\left(x_j \mid \theta_{z_j}\right).$$

And the posterior probability that the observation x_j has been generated from the i-th component is:

$$P\left(z_j = i \mid x_j, u_j, \boldsymbol{\rho}, \boldsymbol{\theta}\right) = \frac{f\left(x_j \mid \boldsymbol{\theta}_i\right)}{\sum_{i:\rho_i > u_j} f\left(x_j \mid \boldsymbol{\theta}_i\right)}.$$

Bayesian inference for infinite mixtures

Given $\boldsymbol{\rho}$, the posterior distribution of u_j is:

$$u_j \sim U(0, \rho_{z_j}),$$

for $j = 1, \ldots, n$, where $\rho_{z_j} = (1-\eta_1)\ldots\left(1-\eta_{z_j-1}\right)\eta_{z_j}$.

Given $\mathbf{z}$, the posterior distribution of $\boldsymbol{\eta}$ is:

$$\eta_j|\mathbf{z} \sim Beta\left(n_s + 1, n - \sum_{l=1}^{s} n_l + \alpha\right)$$

where $n_i = \sum_{j=1}^{n} I\left(z_j = i\right)$.

Clearly, assuming a conjugate prior, P_0, for all $\boldsymbol{\theta}_i$, the conditional posterior distribution of $\boldsymbol{\theta}_i$ given $\mathbf{z}$ is straightforward to obtain.

Bayesian inference for infinite mixtures

MCMC algorithm

1. Set an initial allocation $\mathbf{z} = \{z_1, \ldots, z_n\}$.
2. Update η_i by simulating from the beta distribution for $i = 1, \ldots, z^*$, where $z^* = \max\{z_j\}_{j=1}^n$.
3. Update u_j by simulating from $u_j \sim U\left(0, \rho_{z_j}\right)$ for $j = 1, \ldots, n$.
4. Update η_i by simulating from $\eta_i \sim Beta\left(1, \alpha\right)$ for $i = z^* + 1, \ldots, s^*$, where s^* is the smallest value such that: $\sum_{i=1}^{s^*} \rho_i > 1 - u^*$ where $u^* = \min\{u_1, \ldots, u_n\}$.
5. Update $\boldsymbol{\theta}_i$ by simulating from its conditional posterior distribution for $i = 1, \ldots, s^*$.
6. Update z_j by simulating from $Z_j \mid x_j, u_j, \boldsymbol{\rho}, \boldsymbol{\theta}$ for $j = 1, \ldots, n$.