

Mixture models: introduction

Mixtures are complicated distributions build from simpler ones. In this respect, these distributions can be viewed as a weighted combination of densities. Let Y be a random variable (or a d -dimensional random vector in the multivariate case) and y be any observed values of this random variable. Then Y obeys a finite mixture distribution if its density can be written as:

$$f(y) = \lambda_1 f_1(y) + \dots + \lambda_k f_k(y) = \sum_{j=1}^k \lambda_j f_j(y),$$

provided that $\lambda_j > 0$ and $\sum_{j=1}^k \lambda_j = 1$. The weights λ_j are called the *mixing proportions* and $f_j(y)$ are called the *component densities*. Further, a k -component parametric finite mixture model has the form:

$$f(y \mid \Psi) = \sum_{j=1}^k \lambda_j f_j(y \mid \theta_j) .$$

Gaussian Mixture models

We are concerned with the particular case of univariate gaussian mixture models. The simplest case of a two-component model, parametrized by μ_j and σ_j^2 , for $j = 1, 2$, decomposes as follows:

$$\begin{aligned} f(y \mid \Psi) &= \sum_{j=1}^2 \lambda_j f_j(y \mid \theta_j) \\ &= \lambda_1 N_1(y \mid \theta_1) + \lambda_2 N_2(y \mid \theta_2) \\ &= \lambda_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(y - \mu_1)^2}{2\sigma_1^2}\right) + \lambda_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y - \mu_2)^2}{2\sigma_2^2}\right). \end{aligned}$$

The mixture parameter vector is $\Psi = (\lambda_1, \lambda_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$; the number of components is $k = 2$; the component density parameters are $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_2 = (\mu_2, \sigma_2^2)$; the mixing proportions are λ_1 and $\lambda_2 = (1 - \lambda_1)$.

'faithful' dataset

faithful: A data frame with 272 observations on 2 variables.

eruptions: (numeric) Eruption time in mins

waiting: (numeric) Waiting time to next eruption (in mins)

```
1 > data(faithful)
2 > head(faithful)
3   eruptions waiting
4 1      3.600      79
5 2      1.800      54
6 3      3.333      74
7 4      2.283      62
8 5      4.533      85
9 6      2.883      55
```

By looking at the distribution of the variable 'waiting' using kernel density estimator, we clearly see that this distribution is bimodal. We will therefore model the distribution using a Gaussian Mixture model with $k = 2$ components.

Choice of prior structure

Several kind of prior structures for θ have proved to be appropriate for normal mixtures. We first mention the *independence prior* (IP) having the form:

$$p(\theta) = \prod_{j=1}^k p(\mu_j) p(\sigma_j^2) ,$$

where μ_j is modeled using the conjugate normal distribution prior $N(m_0, v_0)$ and σ_j^2 the conjugate inverse gamma prior $IG(sh_0, ra_0)$. Such prior structure reflects the belief that a priori no particular reason justifies to introduce any kind of dependence whatsoever. Another common type of prior structure is the *conditionnally conjugate prior* (CCP) which has been used for example in Diebolt and Robert (1994), and takes the form:

$$p(\theta) = \prod_{j=1}^k p(\mu_j \mid \sigma_j^2) p(\sigma_j^2) ,$$

where $\mu_j \mid \sigma_j^2$ comes from a normal distribution $N(m_0, p_0^{-1} \sigma_j^2)$ and σ_j^2 comes from an inverse gamma distribution $IG(sh_0, ra_0)$. This prior assumes that there is a dependance between μ_j and σ_j^2 within a given subpopulation j . Such a prior for θ has also been studied in Raftery (1996). To remain weakly informative, hyperparameter values are chosen so that their influence on the posterior is limited.

Gibbs Sampler - Independent Priors (1/3)

Under the independent priors, a possible implementation of a Gibbs sampler could be as follows :

1. Specify a number of simulations T .
2. At iteration $t = 0$, start with some preliminary parameter estimation and classification.
3. At iteration $t \geq 1$, compute the ratio \hat{w}_{ij}

$$\hat{w}_{ij}^{(t)} = P(Z_{ij} = 1 \mid y_i) = \frac{\lambda_j f_j(y_i \mid \boldsymbol{\theta}_j)}{\sum_{j'=1}^k \lambda_{j'} f_{j'}(y_i \mid \boldsymbol{\theta}_{j'})}$$

Gibbs Sampler - Independent Priors (2/3)

4. Sample the indicator variable Z from a multinomial distribution

$$z_i^{(t)} \sim \text{Multinomial}(1, \hat{w}_{i1}, \dots, \hat{w}_{ik})$$

5. (a) Sample the mixing proportions from the full conditional Dirichlet distribution

$$\hat{\boldsymbol{\lambda}}^{(t)} \mid \mathbf{z} \sim \text{Dirichlet}(\delta_1 + n_1, \dots, \delta_k + n_k)$$

(b) Sample the mean of the component j from the full conditional normal distribution

$$\hat{\mu}_j^{(t)} \mid \mathbf{y}, \mathbf{z}, \boldsymbol{\lambda}, \sigma_j^2 \sim N\left(\frac{\frac{m_0}{v_0} + \frac{y_j}{\sigma_j^2}}{\frac{1}{v_0} + \frac{n_j}{\sigma_j^2}}, \frac{1}{\frac{1}{v_0} + \frac{n_j}{\sigma_j^2}} \right)$$

Gibbs Sampler - Independent Priors (3/3)

(c) Sample the variance of the component j from the full conditional inverse gamma distribution

$$\hat{\sigma}_j^{2(t)} \mid \mathbf{y}, \mathbf{z}, \boldsymbol{\lambda}, \mu_j \sim IG\left(sh_0 + \frac{n_j}{2}, \quad ra_0 + \frac{s_j}{2} \right)$$

6. Return to step 3 and loop through step 5 until the specified number of simulations T is achieved.

R code and Results summary

The R code to perform Bayesian Gaussian Mixture Modelling using Gibbs sampling with independent prior is accessible here:

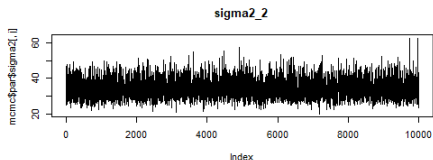
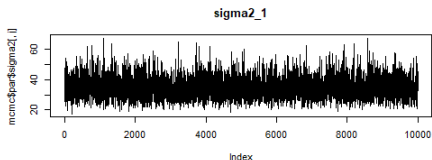
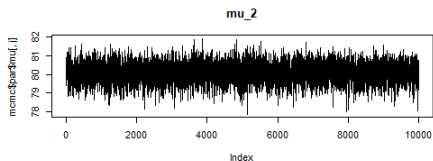
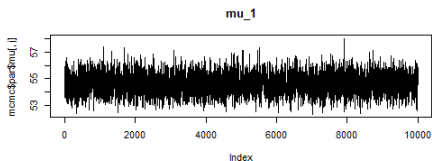
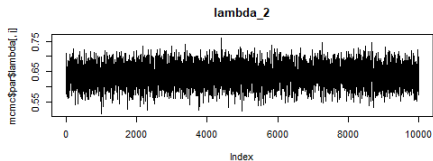
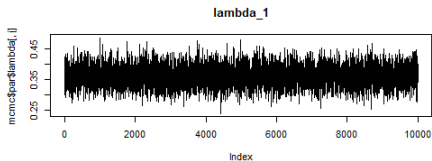
https://github.com/JRigh/Short-Analyses-in-R-and-Python/blob/main/Mixture%20Models/R/Bayesian%20Mixture%20Modelling/GMMb_IP.R

Bay IP	<i>mean</i>	<i>sd</i>	2.5 %	97.5 %
λ_j	0.3614	0.0316	0.3000	0.4241
	0.6386	0.0316	0.5759	0.7000
μ_j	54.6229	0.7046	53.2617	56.0589
	80.0849	0.5088	79.0683	81.0526
σ_j	5.8281	0.538	4.8624	6.9860
	5.8595	0.4016	5.1376	6.7172

Visualizing GMM



Traceplots (Convergence of the chains)



Main observations

- The GMM distinctly separates the waiting times into two clusters, corresponding to shorter (around 55 minutes) and longer (around 80 minutes) eruption intervals.
- The mixing proportions typically show a near 36-64 split, indicating that long eruptions are more frequent than short ones.
- The standard deviation within each cluster suggests that the longer eruptions exhibit more variability in waiting times compared to the shorter eruptions.
- Despite clear separation, there is moderate overlap between the two clusters in the 65-75 minute range, indicating some ambiguity in classifying waiting times within this interval.

References

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. 0471006262, John Wiley & Sons.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* 56: 363-375.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. 13978038732909 New York / Berlin / Heidelberg: Springer.

Raftery, A.E. (1996). Hypothesis testing and model selection. In Markov Chain Monte Carlo in Practice(W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), *London: Chapman and Hall*, pp. 163–188.

Walsh, D., course notes. (2016). available for download at this address :
<http://www.massey.ac.nz/~dcwalsh/161304/CourseMaterials/LectureNotes>

Walsh, D., R Code to implement Gibbs sampling for component univariate normal mixture. available for download at this address :
<http://www.massey.ac.nz/~dcwalsh/161304/Code/MCMC.R>

The R Project for Statistical Computing:
<https://www.r-project.org/>