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:

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

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:

$$f(y \mid \mathbf{\Psi}) = \sum_{j=1}^{k} \lambda_j f_j(y \mid \boldsymbol{\theta}_j) \ .$$

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:

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

The mixture parameter vector is $\boldsymbol{\Psi} = (\lambda_1, \lambda_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$; the number of components is $k = 2$; the component density parameters are $\boldsymbol{\theta_1} = (\mu_1, \sigma_1^2)$ and $\boldsymbol{\theta_2} = (\mu_2, \sigma_2^2)$; 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)

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

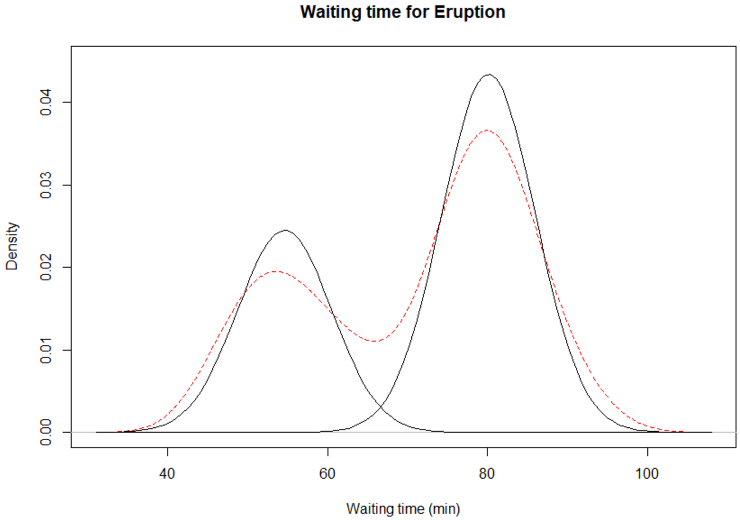
By looking at the distribution of the variable 'waiting' using kernel density estimator, we clearly see that this distribution is bimodal. We will therefore model the distribution using a Gaussian Mixture model with $k = 2$ components.

Summary

In order to perform a Gaussian Mixture model, we call the function 'normalmixEM()' from the package 'mixtools', applying an EM algorithm and containing two essential arguments: the variable on which we want to perform GMM, the number of components k that we have to specify beforehand. The result we obtains are the estimates of the mixing proportions, the means and standard deviations of the two Gaussian components.

```
> # perform GMM
> set.seed(2024)
> gmm = normalmixEM(faithful$waiting, k = 2)
number of iterations= 28
> summary(gmm)
summary of normalmixEM object:
          comp 1    comp 2
lambda  0.360887  0.639113
mu     54.614897 80.091095
sigma   5.871248  5.867713
loglik at estimate:  -1034.002
```

Visualizing GMM



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.

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