Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

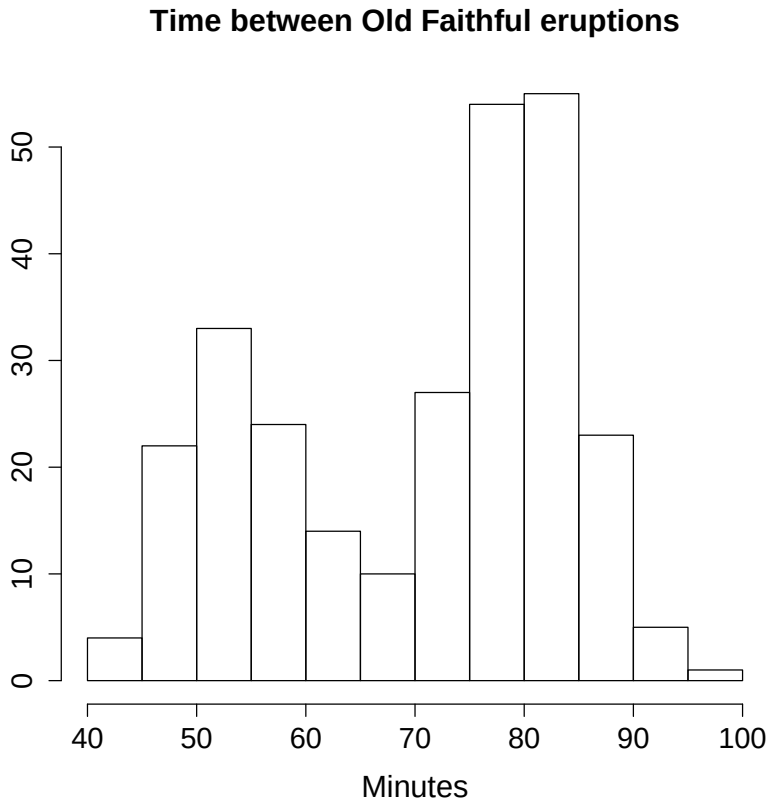
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

EM algorithm

Mixture models

↪ Let us consider the popular old faithful data. The data consists of 272 waiting times between eruptions for the Old Faithful geyser in Yellowstone National park, Wyoming, USA.



EM algorithm

Mixture models

↪ For this dataset we posit as a model a mixture model with two normal components, i.e.,

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} f(y; \theta),$$

where

$$f(y; \boldsymbol{\theta}) = p\phi(y; \mu_1, \sigma_1^2) + (1-p)\phi(y; \mu_2, \sigma_2^2), \qquad \boldsymbol{\theta} = (p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

↪ The observed data likelihood is

$$L(\boldsymbol{\theta}; y) = \prod_{i=1}^{n} \{p\phi(y_i; \mu_1, \sigma_1^2) + (1-p)\phi(y_i; \mu_2, \sigma_2^2)\},$$

with corresponding log likelihood given by

$$\log L(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \log \left\{ p\phi(y_i; \mu_1, \sigma_1^2) + (1-p)\phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ This log likelihood is difficult to maximise due to the sum inside the logarithm.

EM algorithm

Mixture models

- **Idea**: If we knew the group each observation belongs to, we could simply fit a normal distribution to each group.

- We define an augmented complete dataset where $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_n)$ and $\mathbf{y}_{\text{mis}} = z = (z_1, \ldots, z_n)$ is a vector of unobserved/latent group data indicator, such that

$$z_i = \begin{cases} 1, & \text{if } y_i \text{ belongs to the first component (short waiting times),} \\ 0 & \text{if } y_i \text{ belongs to the second component (long waiting times).} \end{cases}$$

- Note that $\Pr(Z_i = 1) = p$ or, equivalently stated, $Z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

- Then, the complete data likelihood is

$$L(\theta \mid y, z) = \prod_{i=1}^{n} \left\{ [p\phi(y_i; \mu_1, \sigma_1^2)]^{z_i} [(1-p)\phi(y_i; \mu_2, \sigma_2^2)]^{1-z_i} \right\}.$$

EM algorithm

Mixture models

↪ Therefore,

$$\log L(\theta \mid y, z) = \sum_{i=1}^{n} z_i \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} + \sum_{i=1}^{n} (1 - z_i) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ For the E-step we would need to compute

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_Z[\log L(\theta \mid y, z) \mid y, \theta^{(t)}] \\ &= \sum_{i=1}^{n} E[Z_i \mid y, \theta^{(t)}] \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} \\ &+ \sum_{i=1}^{n} \left(1 - E[Z_i \mid y, \theta^{(t)}]\right) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right\}. \end{aligned}$$

↪ Now,

$$\begin{aligned} E[Z_i \mid y, \theta^{(t)}] &= E[Z_i \mid y_i, \theta^{(t)}] \\ &= 1 \times \Pr(Z_i = 1 \mid y_i, \theta^{(t)}) + 0 \times \Pr(Z_i = 0 \mid y_i, \theta^{(t)}) \\ &= \frac{p^{(t)} \phi\left(y_i; \mu_1^{(t)}, (\sigma_1^{(t)})^2\right)}{p^{(t)} \phi\left(y_i; \mu_1^{(t)}, (\sigma_1^{(t)})^2\right) + (1 - p^{(t)}) \phi\left(y_i; \mu_2^{(t)}, (\sigma_2^{(t)})^2\right)} \\ &= \widetilde{p}_i^{(t)} \end{aligned}$$

EM algorithm

Mixture models

↪ Thus,

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \widetilde{p}_i^{(t)} \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} + \sum_{i=1}^{n} \left(1 - \widetilde{p}_i^{(t)}\right) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ For the M-step,

$$\frac{\partial}{\partial p} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow p^{(t+1)} = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}{n}$$

$$\frac{\partial}{\partial \mu_1} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)} y_i}{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}$$

$$\frac{\partial}{\partial \sigma_1^2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (\sigma_1^{(t+1)})^2 = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)} (y_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}$$

EM algorithm

Mixture models

↪ Continuing the M-step:

$$\frac{\partial}{\partial \mu_2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \mu_2^{(t+1)} = \frac{\sum_{i=1}^{n}(1-\widetilde{p}_i^{(t)})y_i}{\sum_{i=1}^{n}(1-\widetilde{p}_i^{(t)})}$$

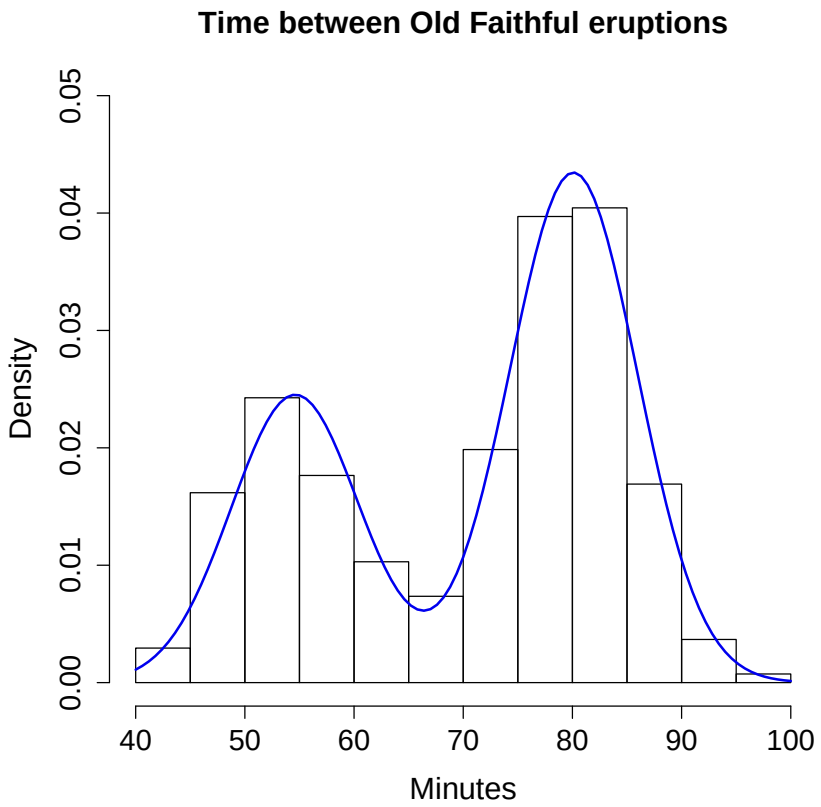
$$\frac{\partial}{\partial \sigma_2^2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (\sigma_2^{(t+1)})^2 = \frac{\sum_{i=1}^{n}(1-\widetilde{p}_i^{(t)})(y_i-\mu_2^{(t+1)})^2}{\sum_{i=1}^{n}(1-\widetilde{p}_i^{(t)})},$$

which can be solved iteratively.

EM algorithm

Mixture models

↪ The plot below depicts the fit of two-component Gaussian mixture model to the observed data.



EM algorithm

Mixture models

↪ More generally, we may have a K-component mixture model

$$f(y) = \sum_{k=1}^{K} p_k f(y; \theta_k), \qquad \sum_{k=1}^{K} p_k = 1.$$

↪ In the particular case of normal components, we have

$$f(y) = \sum_{k=1}^{K} p_k \phi(y; \mu_k, \sigma_k^2).$$

EM algorithm

Mixture models and identifiability issues

- ↪ Due to identifiability issues, such as the so-called label switching problem, it makes difference whether there is interest in making inferences about the mixture component-specific parameters and clustering.

- ↪ The label switching problem (also known as label ambiguity) refers to the fact that there is nothing in the likelihood to distinguish mixture component k from mixture component k'.

- ↪ Permuting the K labels in any of $K!$ ways results in the same model for the data.

EM algorithm

Mixture models and identifiability issues

- As a concrete example, in the $K = 2$ case, consider

$$p_1 = 0.3, \quad \mu_1 = 1, \quad p_2 = 0.7, \quad \mu_2 = 1.5, \quad \sigma_1^2 = \sigma_2^2 = 1, \qquad \text{(Scenario A)}.$$

- Then, the model is equivalent to one with

$$p_1 = 0.7, \quad \mu_1 = 1.5, \quad p_2 = 0.3, \quad \mu_2 = 1, \quad \sigma_1^2 = \sigma_2^2 = 1, \qquad \text{(Scenario B)}.$$

- If one is only interested in density estimation, then everything is fine, because as illustrated in the example

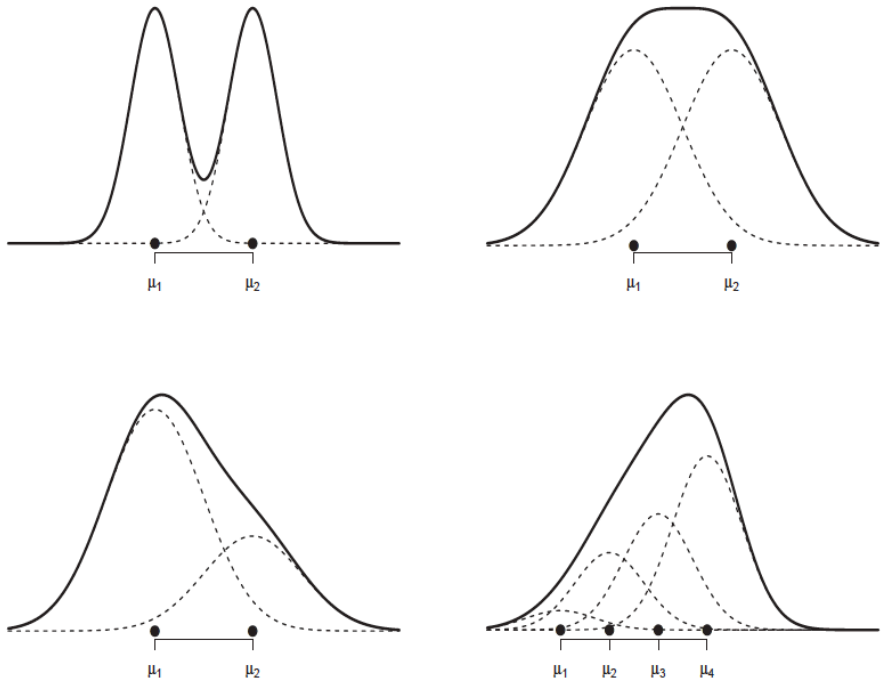
$$\begin{aligned} f_{\text{A}}(y) &= 0.3\phi(y \mid 1, 1) + 0.7\phi(y \mid 1.5, 1) \\ &= 0.7\phi(y \mid 1.5, 1) + 0.3\phi(y \mid 1, 1) = f_{\text{B}}(y). \end{aligned}$$

- Of course, if we are 'only' interested in estimating the density, the label switching poses no problem.

EM algorithm

Mixture models

↪ It is worth noting that multimodality is not the sole motivation for the use of mixture models. For instance, skewed data can also be handled by mixtures.



source: Komarek, A., 2006, PhD thesis

EM algorithm

- ↪ The following paper, available on Learn, is an interesting reading (I think!). It advocates the use of direct (numerical) maximisation of the likelihood, in several well-known problems, where the EM algorithm is the 'gold standard' solution.
- ↪ We will explore the first example mentioned in the paper (mixture of Poisson distributions), from the EM perspective, in Workshop 4.

International Statistical Review (2014), 82, 2, 296–308 doi:10.1111/insr.12041

Numerical Maximisation of Likelihood: A Neglected Alternative to EM?

Iain L. MacDonald

Actuarial Science, University of Cape Town, 7701 Rondebosch, South Africa
Email: iain.macdonald@uct.ac.za