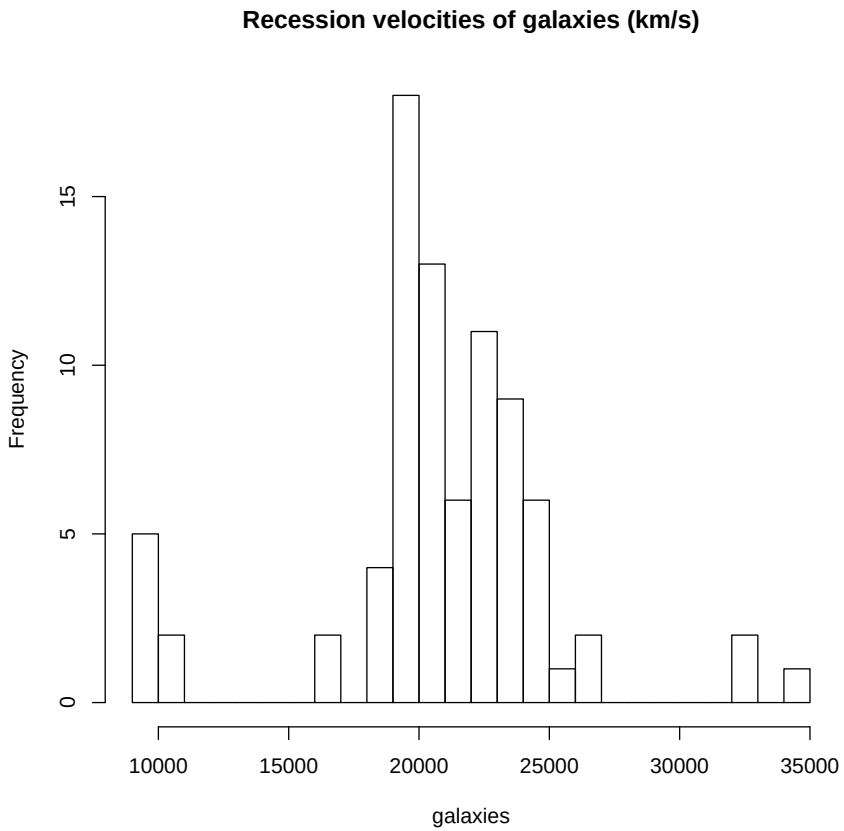
The EM algorithm for finite Gaussian mixtures

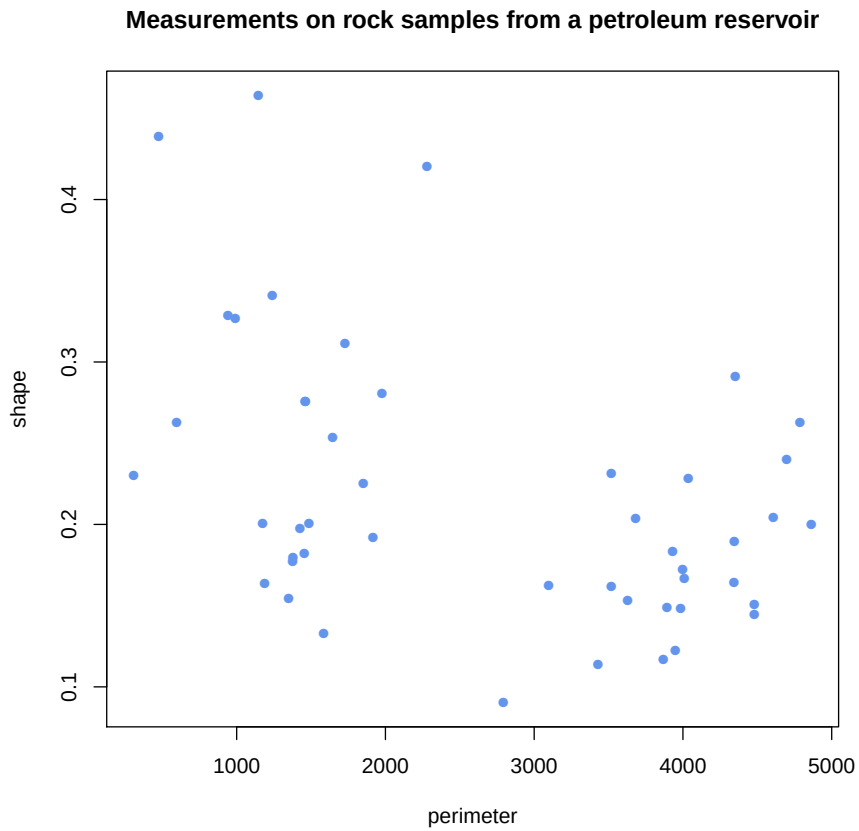
Jochen Einbeck

CMStatistics, December 13, 2019

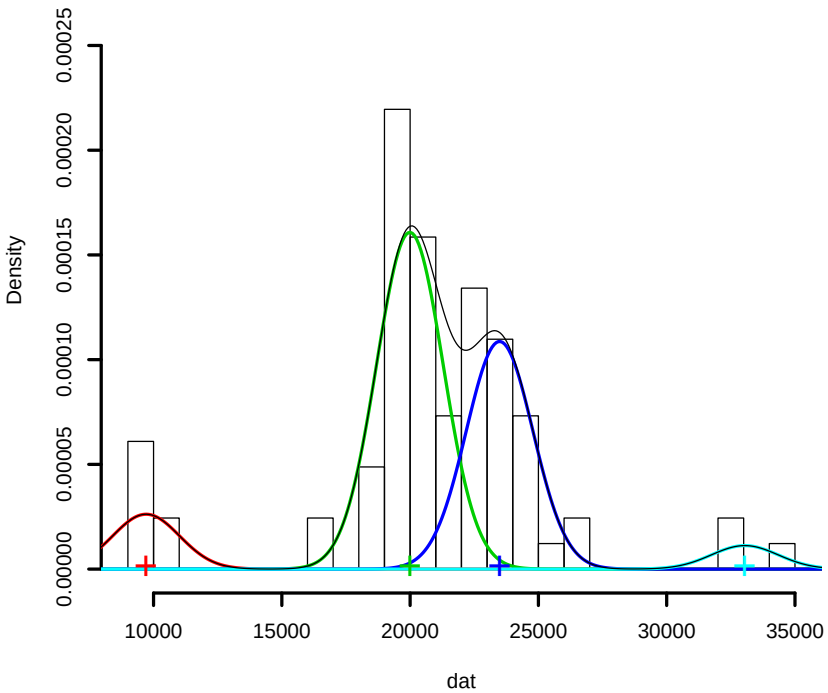
Durham University

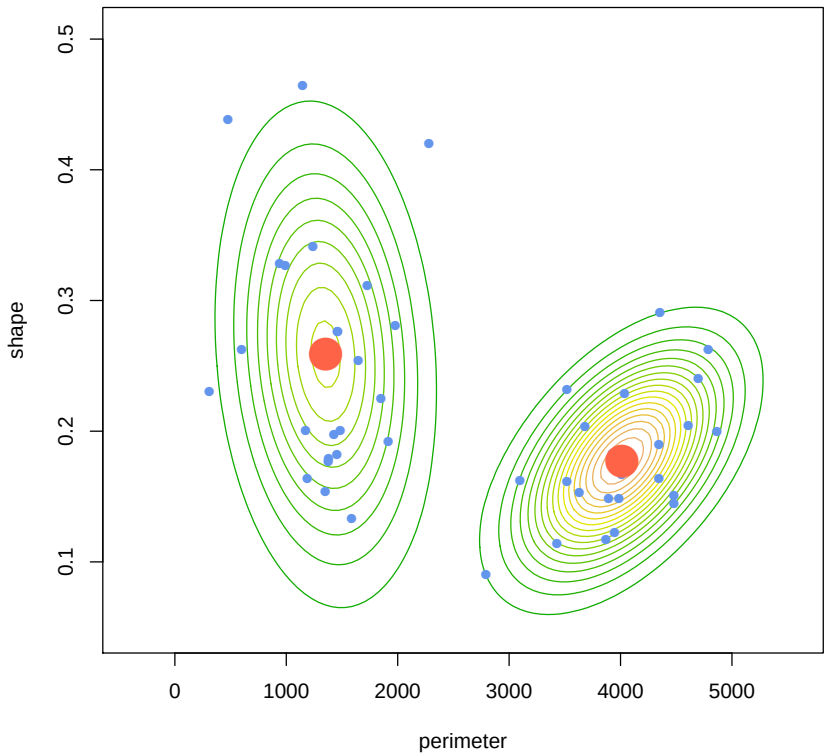
Motivation: Data with unobserved heterogeneity





Aim: Fit 'mixture' distribution





- This course: Focus on univariate mixtures (left).

Gaussian mixture models

- univariate data set $Y = (y_1, \ldots, y_n)$
- unobserved heterogeneity ("clustering")
- represented by mixture components $k = 1, \ldots, K$
- Finite Gaussian mixture model:

$$f(y_i|\theta) = \sum_{k=1}^{K} p_k \phi(y_i|\mu_k, \sigma_k^2)$$

where $\phi(y_i|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2\right)$,
and $\theta = \{p_k, \mu_k, \sigma_k^2\}_{1 \leq k \leq K}$; with constraint $p_K = 1 - \sum_{k=1}^{K-1} p_k$.

Why is this model interesting?

- Visualization
- Ability to simulate new data (evolutionary algorithms, etc)
- Correct representation of heterogeneity in further inference, for instance regression models
- Identification of subpopulations/clusters
- Classification of new observations
- ...

Estimation via EM algorithm: Derivation

- Given data $y_i, i = 1, \ldots, n$, we wish to obtain an estimator, $\hat{\theta}$, of θ.
- Define $f_{ik} = \phi(y_i|\mu_k, \sigma_k^2)$, so $f(y_i|\theta) = \sum_k p_k f_{ik}$.
- Then one has the Likelihood function

$$L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} p_k f_{ik} \right)$$

  and the corresponding log-likelihood

$$\ell(\theta|y_1, \ldots, y_n) = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} p_k f_{ik})$$

- However, $\frac{\partial \ell}{\partial \theta} = 0$ has no (analytic) solution!

Estimation via EM algorithm: Derivation

- Idea: Give the likelihood some more 'information.' Assume that, for an observation y_i, we know to which of the K components it belongs; i.e. we assume we know

$$G_{ik} = \begin{cases} 1 & \text{if observation } \quad i \quad \text{belongs to component } \quad k \\ 0 & \text{otherwise.} \end{cases}$$

- Then we also know

$$\begin{aligned} P(G_{ik} = 1) &= p_k \qquad (\text{"prior"}) \\ P(y_i, G_{ik} = 1) &= P(y_i | G_{ik} = 1) P(G_{ik} = 1) = f_{ik} p_k \end{aligned} \tag{1}$$

- This gives complete data $(y_i, G_{i1}, \ldots G_{iK})$, $i = 1, \ldots, n$, with

$$P(y_i, G_{i1}, \ldots, G_{iK}) = \prod_{k=1}^{K} (f_{ik} p_k)^{G_{ik}}.$$

Estimation via EM algorithm: Derivation

- The corresponding likelihood function, called complete likelihood, is

$$L^*(\theta|y_1,\ldots,y_n) = \prod_{i=1}^{n}\prod_{k=1}^{K}(p_k f_{ik})^{G_{ik}}. \tag{2}$$

- One obtains the complete log-likelihood

$$\ell^* = \log L^* = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log p_k + G_{ik}\log f_{ik} \tag{3}$$

Estimation via EM algorithm: Derivation

- As the G_{ik} are unknown, we replace them by their expectations

$$\begin{aligned} w_{ik} &\equiv E(G_{ik}|y_i) = P(G_{ik}=1|y_i) \\ &= \frac{p_k P(y_i|G_{ik}=1)}{\sum_\ell p_\ell P(y_i|G_{i\ell}=1)} = \frac{p_k f_{ik}}{\sum_\ell p_\ell f_{i\ell}} \end{aligned}$$

This corresponds to the **E-Step**.

Estimation via EM algorithm: Derivation

- For the **M-step**, set

$$\frac{\partial \ell^*}{\partial \mu_k} = 0; \quad \frac{\partial \ell^*}{\partial \sigma_k} = 0; \quad \frac{\partial \left(\ell^* - \lambda(\sum_{k=1}^K p_k - 1) \right)}{\partial p_k} = 0;$$

yielding

$$\hat{\mu}_k = \frac{\sum_{i=1}^n w_{ik} y_i}{\sum_{i=1}^n w_{ik}}; \tag{4}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^n w_{ik} (y_i - \hat{\mu}_k)^2}{\sum_{i=1}^n w_{ik}}; \tag{5}$$

$$\hat{p}_k = \frac{\sum_{i=1}^n w_{ik}}{n}. \tag{6}$$

EM Output

Cycling between the E-step and M-step until convergence, leads to two sorts of outputs:

- Obviously, $\hat{\mu}_k$, $\hat{p}_k$, $\hat{\sigma}_k^2$, $k = 1, \ldots, K$.
- But, also a matrix

$$W = (w_{ik})_{1 \leq i \leq n, 1 \leq j \leq K}$$

of posterior probabilities of component membership.

Useful for clustering and classification!

EM Output

For instance, galaxies data:

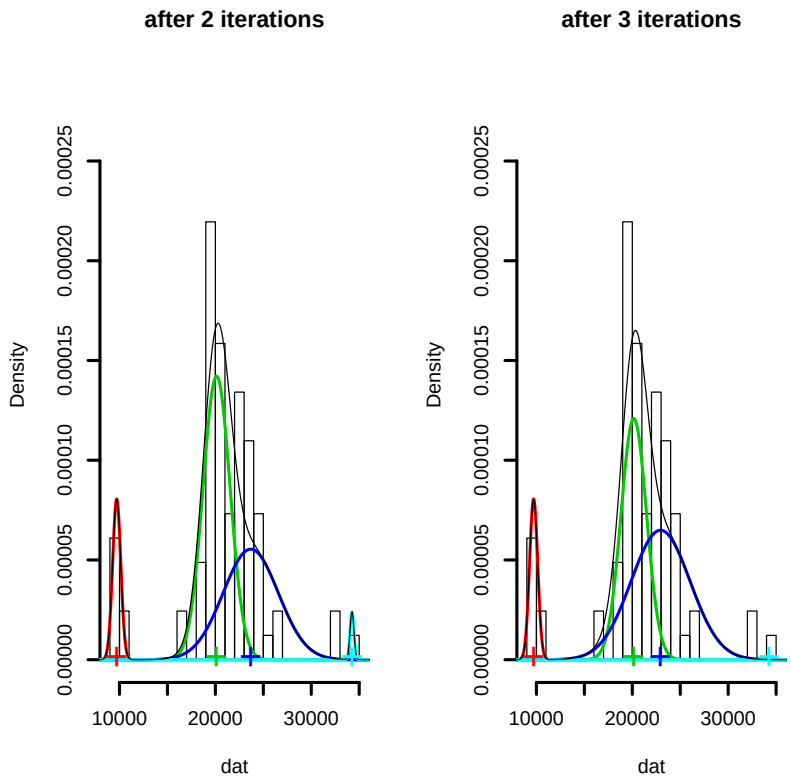
```
...
 [6,]    1 0.000 0.000    0
 [7,]    1 0.000 0.000    0
 [8,]    0 1.000 0.000    0
 [9,]    0 1.000 0.000    0
[10,]    0 0.999 0.001    0
[11,]    0 0.999 0.001    0
[12,]    0 0.999 0.001    0
[13,]    0 0.998 0.002    0
[14,]    0 0.997 0.003    0
...
```

Likelihood spikes

- Problem: If an individual data point, say x_0, 'captures' a mixture component (*i.e.*, $\mu_k = x_0$ and $\sigma_k^2 \longrightarrow 0$), one obtains a spurious solution with infinite likelihood.
- Simplistic solution: Set all $\sigma_k \equiv \sigma$. In this case, expression (5) becomes

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} (y_i - \hat{\mu}_k)^2.$$

$\hat{\sigma}_4 = 207.4$ $\hat{\sigma}_4 = 9.1 \times 10^{-3}$



Implementation

- In the practical part, we will implement the EM algorithm for univariate Gaussian mixtures, for general (unequal) component variances σ_k^2.

- Continue to use the provided R Notebook that you have used in Part I, and consult the Handout for theoretical support.