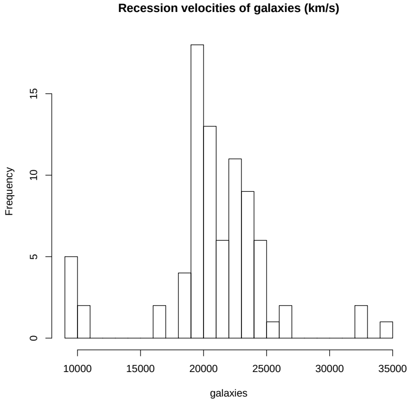
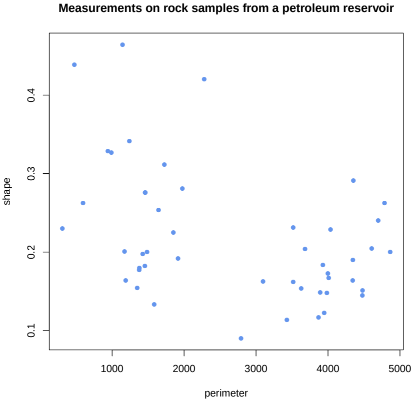
The EM algorithm for finite Gaussian mixtures

Jochen Einbeck

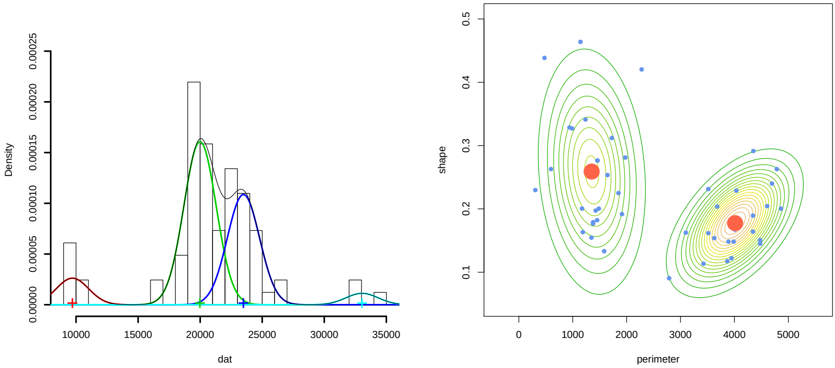
January 18, 2019

Motivation: Data with unobserved heterogeneity





Aim: Fit 'mixture' distribution



- multivariate data set $Y = (y_1, \ldots, y_n) \in \mathbb{R}^p$
- unobserved heterogeneity ("clustering")
- represented by mixture components $k = 1, \ldots, K$
- Finite Gaussian mixture model: $f(y_i) = \sum_{k=1}^{K} \pi_k f(y_i|\mu_k, \Sigma_k)$, where

$$\begin{aligned} & f(y_i|\mu_k, \Sigma_k) = \\ & \quad = \quad (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\right\}, \end{aligned}$$

- Parameters: $\{\pi_k, \mu_k, \Sigma_k\}_{1 \le k \le K}$; restriction $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

Why do this?

- 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
- ...

- Need to estimate $\mu_k, \pi_k, \Sigma_k,\ k = 1, \ldots, K$ from data $y_i,\ i = 1, \ldots, n$.
- Idea: If, for each y_i, we knew to which class k it belonged, then estimation straightforward.
- However, we do not know this. But, assuming given 'current' values of $\mu_k, \pi_k, \Sigma_k,\ k = 1, \ldots, K$, we can compute the probability that case i belongs to class k via Bayes' theorem as

$$w_{ik} \equiv P(k|y_i) = \frac{P(y_i|k)P(k)}{\sum_\ell P(y_i|\ell)P(\ell)} = \frac{f(y_i|\mu_k, \Sigma_k)\pi_k}{\sum_{\ell=1}^K f(y_i|\mu_\ell, \Sigma_\ell)\pi_\ell}.$$

Estimation via EM algorithm

- Fix K and choose starting values for μ_k, π_k, Σ_k, $k = 1, \dots, K$. Then, iterate between...

- E-step: Update posterior probabilities of class membership,

$$w_{ik} = \frac{\pi_k f(y_i|\mu_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell f(y_i|\mu_\ell, \Sigma_\ell)}.$$

- M-step: Update parameter estimates via

$$\begin{aligned} \hat{\pi}_k &= \frac{1}{n}\sum_{i=1}^{n} w_{ik}; \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{n} w_{ik} y_i}{\sum_{i=1}^{n} w_{ik}}; \\ \hat{\Sigma}_k &= \frac{\sum_{i=1}^{n} w_{ik}(y_i - \mu_k)(y_i - \mu_k)^T}{\sum_{i=1}^{n} w_{ik}}. \end{aligned}$$

- ... until convergence is reached (convergence proven in Dempster et al., 1997, Wu, 1983).

Estimates for rock data

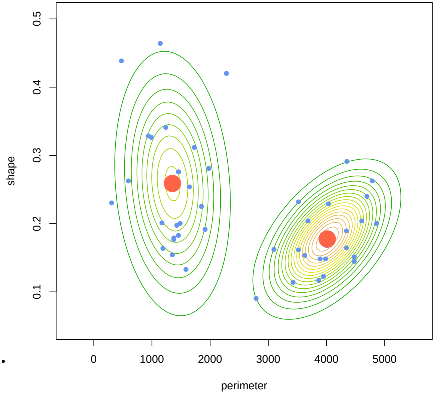
$$\hat{\pi}_1 = 0.5;\ \hat{\pi}_2 = 0.5;$$

$$\hat{\mu}_1 = \begin{pmatrix} 4014.2 \\ 0.1773 \end{pmatrix};$$

$$\hat{\mu}_2 = \begin{pmatrix} 1353.9 \\ 0.2588 \end{pmatrix};$$

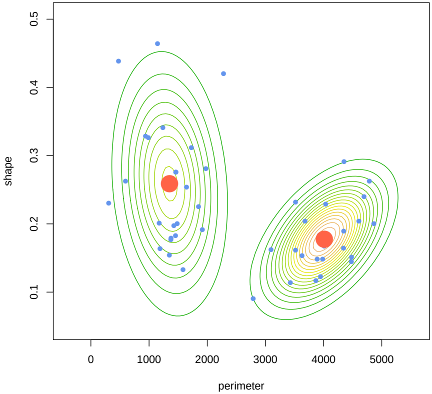
$$\hat{\Sigma}_1 = \begin{pmatrix} 275805 & 13.3909 \\ 13.3909 & 0.00238 \end{pmatrix};$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 220173 & -6.06322 \\ -6.06322 & 0.00839 \end{pmatrix}.$$



Posterior probabilities w_{ik}:

	k=1	k=2
1	0.969	0.031
2	1.000	0.000
3	1.000	0.000
4	1.000	0.000
5	1.000	0.000
6	1.000	0.000
7	1.000	0.000
8	1.000	0.000
9	1.000	0.000
10	0.997	0.003
11	1.000	0.000
12	1.000	0.000
	....	
45	0.000	1.000
46	0.000	1.000
47	0.000	1.000
48	0.000	1.000



Consider the model

$$f(y|\theta) = \sum_{k=1}^{K} \pi_k \phi_{\mu_k, \sigma_k^2}(y) \tag{1}$$

where $\theta = \{\pi_k, \mu_k, \sigma_k\}_{1 \le k \le K}$, and

$$\phi_{\mu_k, \sigma_k^2}(y) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu_k}{\sigma_k} \right)^2 \right\}$$

is the density of a (one-dimensional) normal distribution $N(\mu_k,\, \sigma_k^2)$, evaluated at y. Note that $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

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

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

  and the corresponding log-likelihood

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

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

- 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

$$\begin{aligned} P(G_{ik}=1) &= \pi_k \qquad (\text{“prior”}) \\ P(y_i, G_{ik}=1) &= P(y_i|G_{ik}=1)P(G_{ik}=1) = f_{ik}\pi_k \end{aligned} \tag{2}$$

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

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

- The corresponding likelihood function, called complete likelihood, is

$$L^*(\theta|y_1,\ldots,y_n) = \prod_{i=1}^{n}\prod_{k=1}^{K}(\pi_k f_{ik})^{G_{ik}}. \tag{3}$$

- One obtains the complete log-likelihood

$$\ell^* = \log L^* = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \pi_k + G_{ik}\log f_{ik} \tag{4}$$

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

$$w_{ik} \equiv E(G_{ik}|y_i) = P(G_{ik}=1|y_i) = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}$$

  This corresponds to the **E-Step** as explained earlier.

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

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

yielding

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

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^n w_{ik} (y_i - \hat{\mu}_k)^2}{\sum_{i=1}^n w_{ik}}; \tag{6}$$

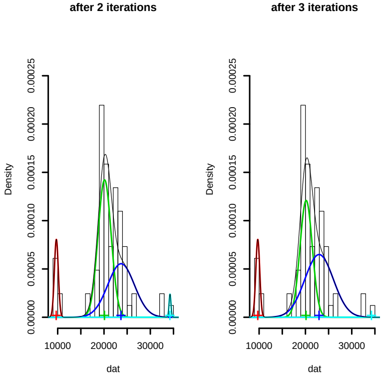
$$\hat{\pi}_k = \frac{\sum_{i=1}^n w_{ik}}{n}. \tag{7}$$

- 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.

- Most simple solution: Set all $\sigma_k \equiv \sigma$. In this case, expression (6) becomes

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

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



- In the practical part, we will implement the EM algorithm for univariate Gaussian mixtures, for equal component variances $\sigma_k^2 = \sigma^2$.
- We will use the statistical programming language R, which is freely available from `https://cran.r-project.org/`.
- R works best in conjunction with the (free) software RStudio, which includes an Editor.
- Please follow the instructions on the R source code file that you have been given; and make use of the Handout.