

The expectation maximization (EM) algorithm[†]

TMA4300: Computer Intensive Statistical Methods

(Spring 2014)

Andrea Riebler

- The expectation maximization (EM) algorithm (Dempster et al., 2007) is an **alternative procedure for the computation of maximum likelihood estimators**.
- In certain models – particularly **missing data and data augmentation problems** – the EM algorithm appears naturally and simplifies the maximum likelihood problem.

[†]Slides are based on slides from the CISM2014 class at Stockholm University by Michael Höhle, which represent a translated and much extended version of ancient slides in German of the Cim2004@LMU class by Leonhard Held.

Example: Linear model with missing observation

- Consider a 2×3 table with one missing observation:

10	15	17
22	23	NA

We assume the following linear model

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

with $\sum_{i=1}^2 \alpha_i = 0$, $\sum_{j=1}^3 \beta_j = 0$ and e_{ij} independent and identically $\mathcal{N}(0, \sigma^2)$ distributed

- Estimation only numerically possible through suitable choice of the design matrix and response vector to which the standard least squares equations are applied

Idea of the EM algorithm

- Through knowledge of y_{23} the table becomes balanced and hence the estimates are easy to calculate

$$\hat{\mu} = \bar{y}$$

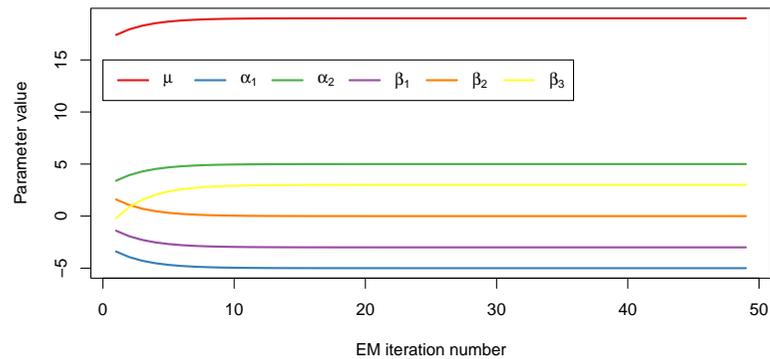
$$\hat{\alpha}_i = \bar{y}_i - \bar{y}$$

$$\hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}$$

- Idea: Use **an iterative imputation** of the missing value y_{23} by choosing a “plausible” value (start with $\hat{y}_{23} = \hat{\mu}$):

$$\hat{y}_{23} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_3$$

Parameter estimates versus iteration number



The algorithm arrives at the MLE solution without inverting the $\mathbf{X}^T \mathbf{X}$ matrix. The entry y_{23} is estimated to be 27.

The EM algorithm to maximize $l(\boldsymbol{\theta}; \mathbf{x})$

We would like to maximize $L(\boldsymbol{\theta}; \mathbf{y})$ regarding $\boldsymbol{\theta}$, but we use $L(\boldsymbol{\theta}; \mathbf{x})$ or rather $l(\boldsymbol{\theta}; \mathbf{x})$.

Input: Function $l(\boldsymbol{\theta}; \mathbf{x})$ and start value $\boldsymbol{\theta}^{(0)}$

$i \leftarrow 0$;

while *not converged* **do**

E-Step: Compute the conditional expectation

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(i)}) = E(l(\boldsymbol{\theta}; \mathbf{x}) | \mathbf{y}, \boldsymbol{\theta}^{(i)}),$$

where $l(\boldsymbol{\theta}; \mathbf{x})$ is the complete data loglikelihood ;

M-Step: Determine

$$\boldsymbol{\theta}^{(i+1)} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$$

Update iteration number: $i \leftarrow i + 1$;

end

Missing Data Setup and the EM Algorithm

The previous example has all the ingredients of an EM algorithm.

In a data setup with missing data, the key notions are

- The **incomplete** (observed) data \mathbf{y}
- The **complete** (but partially unobserved) data \mathbf{x} (= \mathbf{y}, y_{23} above)
- One has the following property: $\mathbf{y} = h(\mathbf{x})$, but the inverse does not exist. Example: $\mathbf{x} = (x_1, x_2, x_3)$ and

$$\mathbf{y} = \begin{cases} (x_1, x_2) \text{ or} \\ (x_1 + x_2, x_2 + x_3) \text{ or} \\ (x_1, x_2 + 2x_3) \end{cases}$$

\Rightarrow Some information is lost by going from \mathbf{x} to \mathbf{y} .

Back to table with missing entry

How does our motivating example fit into this framework?

See blackboard

Genetic example of Rao (1973, page 369)

- Let the vector

$$\mathbf{y} = (y_1, y_2, y_3, y_4)^\top = (125, 18, 20, 34)^\top \sim \text{Mult} \left(\sum_{i=1}^4 y_i, \mathbf{p}(\theta) \right)$$

be multinomial distributed with probabilities

$$\mathbf{p}(\theta) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4} \right)^\top$$

- The loglikelihood function based on \mathbf{y}

$$l(\theta; \mathbf{y}) = y_1 \log(2 + \theta) + (y_2 + y_3) \log(1 - \theta) + y_4 \log \theta$$

can be maximized analytically or numerically.

Application of the EM algorithm

- Idea: Assume complete data $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^\top$ while the incomplete data are $\mathbf{y} = h(\mathbf{x}) = (x_1 + x_2, x_3, x_4, x_5)^\top$. Then,

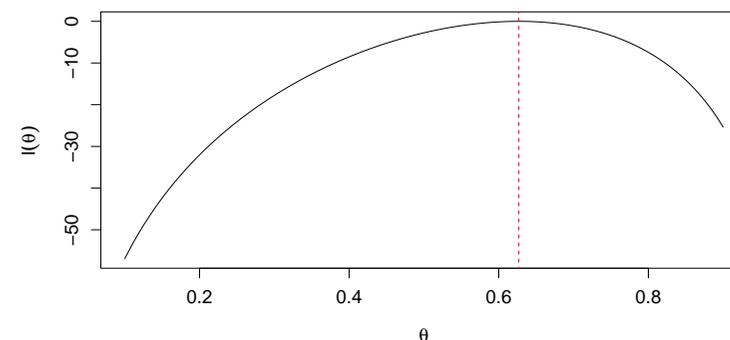
$$\mathbf{x} \sim \text{Mult} \left(\sum_{i=1}^5 x_i, \left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4} \right)^\top \right).$$

- Now, the loglikelihood $l(\theta; \mathbf{x})$ is easy to maximize analytically

$$\hat{\theta} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5}.$$

- What is the *E-step*?

Numerical optimisation



One obtains $\hat{\theta}_{ML} = 0.627$ and $\text{se}(\hat{\theta}_{ML}) = 0.051$.

The E-step

- The loglikelihood for the complete data is

$$l(\theta; \mathbf{x}) = (x_2 + x_5) \log \theta + (x_3 + x_4) \log(1 - \theta).$$

- For the calculation of $E(l(\theta, \mathbf{x}) | \mathbf{y}, \theta)$ one uses

$$x_2 | y_1, \theta \sim \text{Bin}(y_1, \frac{\theta}{2 + \theta})$$

- This yields the *E-Step* of the EM algorithm

$$E(x_2 | y_1, \theta) = y_1 \frac{\theta}{2 + \theta}.$$

Altogether ...

- ...one iterates between

$$\hat{x}_2 = y_1 \frac{\hat{\theta}}{2 + \hat{\theta}}$$

and

$$\hat{\theta} = \frac{\hat{x}_2 + x_5}{\hat{x}_2 + x_3 + x_4 + x_5}$$

until convergence.

- This is equivalent to the one-step update

$$\theta^{(i+1)} = \frac{y_1 \theta^{(i)} + x_5 (2 + \theta^{(i)})}{y_1 \theta^{(i)} + (x_3 + x_4 + x_5) (2 + \theta^{(i)})}$$

Frequent applications of the EM algorithm

- Mixture models, cluster analysis
- Hidden Markov models
- Likelihood-based parameter estimation with missing data

Properties of the EM algorithm

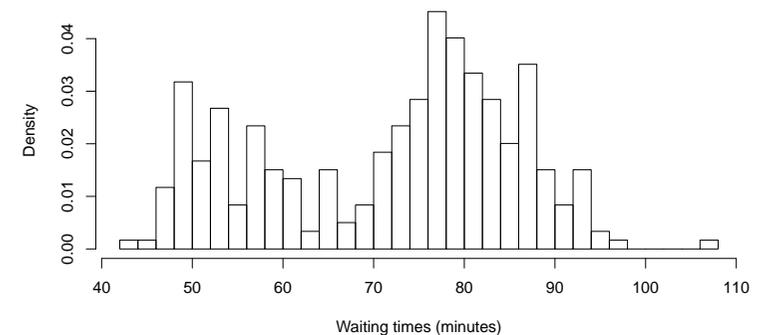
- + In each iteration step of the EM algorithm the (incomplete) likelihood is increased:

$$L(\theta^{(i+1)}; \mathbf{y}) \geq L(\theta^{(i)}; \mathbf{y})$$

- + Parameter restrictions are (mostly) automatically fulfilled
- Convergence can be very slow – this especially depends on the “amount” of missing data
- Standard errors are not directly available. Some methods exist to try to approximate it, but they are not so easy to use in practice. Much easier just to bootstrap your data!

Application to mixture models

Histogram of the waiting time between eruptions of the Old Faithful Geyser in Yellowstone National Park.



→ Mixture of two univariate normal distributions

Statistical model

- We model the data as coming from a normal mixture

$$y_i \sim \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2),$$

$$i = 1, \dots, n \text{ with } \pi_1 \in (0, 1), \pi_1 + \pi_2 = 1$$

- The vector of unknowns is $\theta = (\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^\top$
- Idea: Had the class $z_i = 1, 2$ of each observation (or the probability for being in class 1 or 2) be known then ML estimation would have been easy.

Application of the EM algorithm (2)

1. Take initial guesses for $\hat{\pi}_j, \hat{\mu}_j, \hat{\sigma}_j^2$.

2. **E-Step:** Compute

$$\hat{p}_{ij} = E\{I(z_i = j) | y_i, \theta\} = \frac{\hat{\pi}_j f_j(y_i | \hat{\mu}_j, \hat{\sigma}_j^2)}{\sum_{k=1}^2 \hat{\pi}_k f_k(y_i | \hat{\mu}_k, \hat{\sigma}_k^2)}$$

3. **M-Step:** Compute the mixing parameters, weighted means and variances:

$$\begin{aligned} \hat{\pi}_j &= \frac{\sum_i \hat{p}_{ij}}{n} \\ \hat{\mu}_j &= \frac{\sum_i \hat{p}_{ij} y_i}{\sum_i \hat{p}_{ij}} \\ \hat{\sigma}_j^2 &= \frac{\sum_i \hat{p}_{ij} (y_i - \hat{\mu}_j)^2}{\sum_i \hat{p}_{ij}} \end{aligned}$$

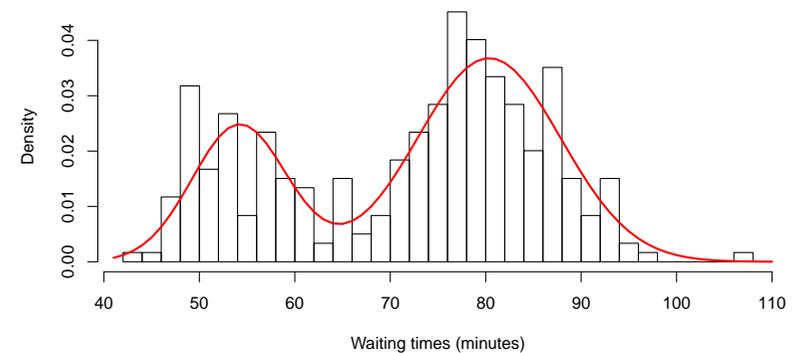
4. Iterate steps 2 and 3 until convergence.

Application of the EM algorithm (1)

- Complete data:** $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ with $\mathbf{x}_i = (y_i, z_i)$. However values of z_i are unknown.
- The marginal probability of Z_i is $P(Z_i = j) = \pi_j$
- Let $f_j(y_i | \mu_j, \sigma_j^2)$ be the density of the normal distribution for observation i which belongs to class $j = 1, 2$
- Loglikelihood

$$l(\theta; \mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^2 I(z_i = j) (\log f_j(y_i | \mu_j, \sigma_j^2) + \log \pi_j)$$

Application of the EM algorithm (3)



We obtain

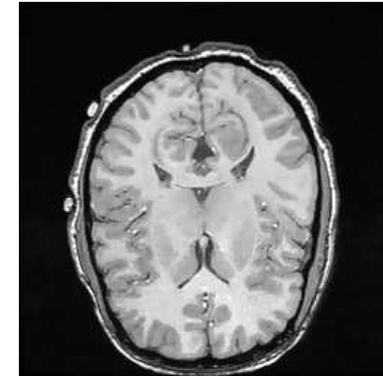
$$\hat{\pi}_1 = 0.308, \hat{\mu}_1 = 54.203, \hat{\sigma}_1^2 = 4.952, \hat{\mu}_2 = 80.360, \hat{\sigma}_2^2 = 7.508.$$

Direct Maximization

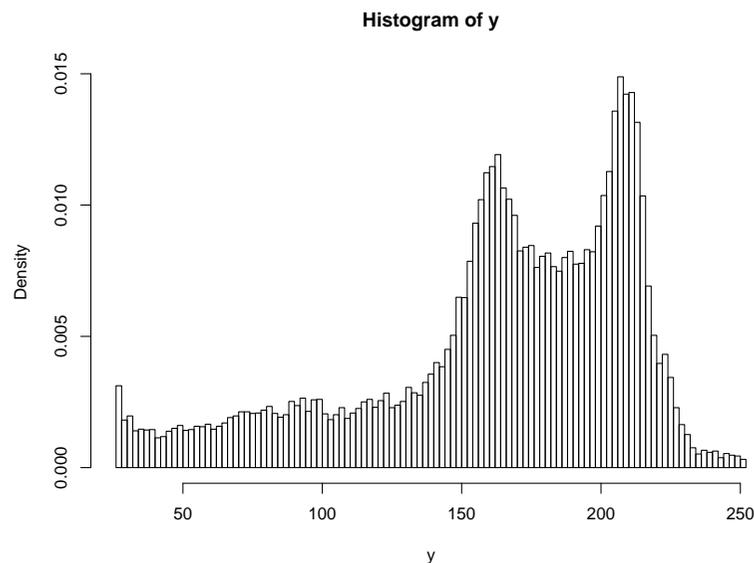
- Possible with the R function `optim()`
- Advantage: Standard errors are directly available
- Important: Choose good starting values

Example: Image Segmentation

- The above mixture modelling using the EM algorithm can be used for the segmentation of raster images based on their intensity.



Histogram of image intensity



Mixture model

- We shall use a mixture of 3 normals to describe the intensity of each pixel

$$y_i \sim \sum_{i=1}^3 \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$$

with $\pi_1 + \pi_2 + \pi_3 = 1$.

- Once the parameters are estimated we can for each of the 3 components/distributions infer the probability that a pixel belongs to it.
- We segment the image by assigning each pixel to the component which is most likely (i.e. classes 1-3).

Classification

