The expectation maximization (EM) algorithm[†]

TMA4300: Computer Intensive Statistical Methods

(Spring 2015)

Andrea Riebler

[†]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.

Introduction

- 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.

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

- No closed form solution, but MLEs can be found through suitable choice of the design matrix and response vector to which the standard least squares equations are applied

Idea of the EM algorithm

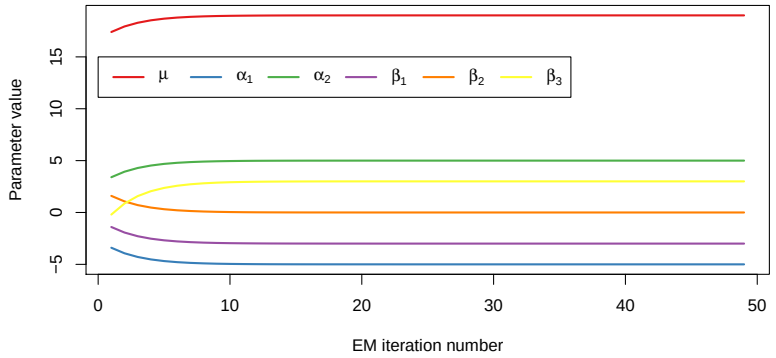
- Through knowledge of y_{23} the table becomes balanced and hence the estimates are easy to calculate

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ \hat{\alpha}_i &= \bar{y}_{i\cdot} - \bar{y} \\ \hat{\beta}_j &= \bar{y}_{\cdot j} - \bar{y} \end{aligned}$$

- 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 $\mathbf{X}^\top\mathbf{X}$ matrix. The entry y_{23} is estimated to be 27.

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

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

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

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

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

Input: Function $l(\boldsymbol{\theta}; \boldsymbol{x})$ and start value $\boldsymbol{\theta}^{(0)}$
$i \leftarrow 0$;
while *not converged* **do**
  E-Step: Compute the conditional expectation

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

  where $l(\boldsymbol{\theta}; \boldsymbol{x})$ is the complete data loglikelihood ;
  M-Step: Determine $\boldsymbol{\theta}^{(i+1)} = \text{argmax}_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$
  Update iteration number: $i \leftarrow i + 1$;
end

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

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

be multinomial distributed with probabilities

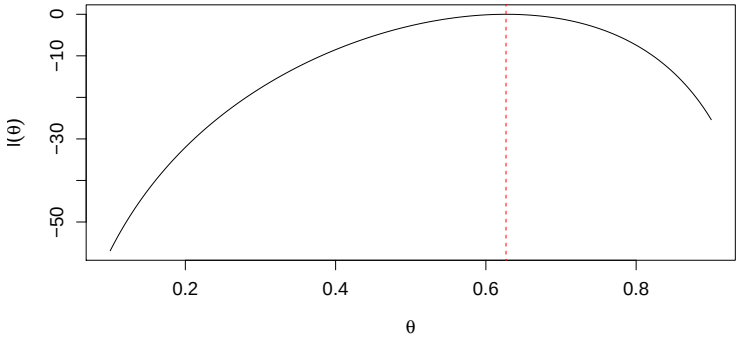
$$\boldsymbol{p}(\theta) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}.\right)^\top$$

- The loglikelihood function based on $\boldsymbol{y}$

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

We can solve it using the quadratic formula or maximise it numerically.

Numerical optimisation



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

Application of the EM algorithm

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

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

- Now, the loglikelihood $l(\theta; \boldsymbol{x})$ 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*?

The E-step

- The loglikelihood for the complete data is

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

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

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

Properties of the EM algorithm

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

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

+ 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!

Frequent applications of the EM algorithm

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