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

↪ The EM algorithm was proposed by Dempster, Laird, and Rubin in their seminal paper published in *JRSS B* in 1977 (although similar ideas had appeared earlier) and it is one of the great success stories of statistics over the past 40 years.

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

Harvard University and Educational Testing Service

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

Keywords: MAXIMUM LIKELIHOOD; INCOMPLETE DATA; EM ALGORITHM; POSTERIOR MODE

EM algorithm

Context

- ↪ The Expectation-Maximisation (EM) algorithm is a general iterative algorithm for maximum likelihood estimation in incomplete data problems.
- ↪ The range of problems that can be attacked by the EM algorithm is remarkably broad and include problems that would not usually be considered as, or are not obvious, incomplete data problems, such as mixture models and random effects models.
- ↪ The EM algorithm has been the first choice of most researchers seeking maximum likelihood estimates in a model that involves incomplete data or that can be structured in such a way (e.g., through latent variables).
- ↪ At the time of writing these notes (21/10/2020), the number of citations of the article on Google Scholar exceeds 60 000.

EM algorithm

Context

- ↪ The EM algorithm is most useful when maximisation from the complete data likelihood is easy wile maximisation based on the observed data likelihood is difficult.
- ↪ Starting from an initial value $\boldsymbol{\theta}^{(0)}$ inside the parameter space $\boldsymbol{\Theta}$, the EM algorithm consists of an expectation (E) step and a maximisation (M) step within each iteration.
- ↪ The condition for the EM algorithm to be valid, in its basic form, is ignorability and hence MAR data.

EM algorithm

Expectation step

↪ We denote the observed data log likelihood by $\log L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})$ and the complete data log likelihood by $\log L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$.

↪ At iteration $(t+1)$ of the iterative procedure, the E-step calculates the conditional expectation, with respect to the missing observations, of the complete data log-likelihood given the observed data and the estimate of $\boldsymbol{\theta}$ from iteration t, $\boldsymbol{\theta}^{(t)}$, that is,

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= E_{Y_{\text{mis}}}[\log L(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}] \\ &= E_{Y_{\text{mis}}}[\log f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}] \\ &= \int \log f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}) \mathrm{d}\mathbf{y}_{\text{mis}}. \end{aligned}$$

EM algorithm

Maximisation step

↪ In the M-step, we obtain $\boldsymbol{\theta}^{(t+1)}$, the value of $\boldsymbol{\theta}$ that maximises $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. That is,

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$

this implies that $Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

↪ The E and M steps are repeated until some convergence criterion is met.

↪ One possible criterion is

$$D(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) < \varepsilon, \qquad D(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{p} |u_i - v_i|.$$

EM algorithm

Toy example

- To understand the EM basic setup, let us consider a trivial example (Givens and Hoeting, 2012, p. 99).
- Let $Y_1, Y_2 \overset{\text{iid}}{\sim} \text{Exp}(\theta)$. Remember that $f(y;\theta) = \theta e^{-\theta y}$, $y \geq 0$, and $\theta > 0$.
- Suppose that $y_1 = 5$ and y_2 is missing. That is, $\mathbf{y}_{\text{obs}} = \{y_1\}$ and $\mathbf{y}_{\text{mis}} = \{y_2\}$.
- The likelihood of the complete data is

$$\begin{aligned} L(\theta \mid y_1, y_2) &= f(y_1;\theta) f(y_2;\theta) \\ &= \theta^2 e^{-\theta(y_1+y_2)}. \end{aligned}$$

- The complete log likelihood is then

$$\log L(\theta \mid y_1, y_2) = 2 \log \theta - \theta y_1 - \theta y_2.$$

EM algorithm

Toy example

- For the E-step at iteration $t+1$ we need to calculate the expectation, with respect to what is missing, Y_2, of the above complete data log likelihood, given what is observed Y_1 and the current estimate of θ, $\theta^{(t)}$, that is

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_{Y_2}[\log L(\theta \mid y_1, y_2) \mid y_1, \theta^{(t)}] \\ &= E[2\log\theta - \theta y_1 - \theta Y_2 \mid y_1, \theta^{(t)}] \\ &= 2\log\theta - 5\theta - \theta E[Y_2 \mid y_1, \theta^{(t)}]. \end{aligned}$$

- Now, since $Y_2 \sim \text{Exp}(\theta)$, then $E[Y_2] = \frac{1}{\theta}$, and so $E[Y_2 \mid y_1, \theta^{(t)}] = E[Y_2 \mid \theta^{(t)}] = \frac{1}{\theta^{(t)}}$.

- Replacing, we get

$$Q(\theta \mid \theta^{(t)}) = 2\log\theta - 5\theta - \frac{\theta}{\theta^{(t)}}.$$

EM algorithm

Toy example

- For the M-step, we maximise $Q(\theta \mid \theta^{(t)})$ with respect to θ.
- We have

$$\frac{\mathrm{d}}{\mathrm{d}\theta} Q(\theta \mid \theta^{(t)}) = \frac{2}{\theta} - 5 - \frac{1}{\theta^{(t)}}.$$

- So,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \frac{2}{\theta} - 5 - \frac{1}{\theta^{(t)}} = 0$$

$$\Rightarrow \theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)} + 1},$$

  which can be solved iteratively.

- Note that here the E and M steps do not need to be re-derived at each iteration: iterative application of the updating formula starting from some initial value provides estimates that converge to $\widehat{\theta} = 0.2$.

EM algorithm

Properties of the EM algorithm

↪ Each iteration of the EM algorithm, leads to an observed likelihood that is greater than or equal to the previous observed likelihood

$$\log L(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{y}_{\text{obs}}) \geq \log L(\boldsymbol{\theta}^{(t)} \mid \mathbf{y}_{\text{obs}}),$$

that is, EM iterates always improve the estimate in the sense that each iterate is more likely than its predecessors.

↪ It is possible for the iterates to converge to a local mode and it is always wise to start the algorithm with several very different initial values, especially if we do not have a clear idea about the behaviour (e.g., the number of modes) of the likelihood being maximised.

EM algorithm

Generalised EM

- ↪ It is worth remarking that Dempster, Laird, and Rubin (1977) have also defined a Generalised EM (GEM) algorithm that differs from the EM algorithm in the M step.

- ↪ With the GEM algorithm, $\boldsymbol{\theta}^{(t+1)}$ is chosen not to globally maximise $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ but rather to ensure that $Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$.

EM algorithm

Genetic linkage model: Rao (1973), Dempster, Laird, and Rubin (1977)

- This example has been used on numerous occasions to illustrate the EM algorithm, including the original article by Dempster, Laird, and Rubin (1977).

- Suppose that 197 animals are distributed into 4 categories, so that the observed data are

$$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34),$$

and y is postulated to have arisen from a multinomial distribution with cell probabilities

$$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right),$$

for some $\theta \in (0, 1)$ unknown.

- The goal is to estimate θ.

EM algorithm

Genetic linkage model: Rao (1973), Dempster, Laird, and Rubin (1977)

- The observed data likelihood is

$$\begin{aligned} L(\theta \mid y) &= \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} \\ &\propto \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y1} \left(\frac{1}{4}(1-\theta)\right)^{y_2+y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}. \end{aligned}$$

- To maximise this likelihood there is no need to use iterative methods. We can find the mle analytically. **Suggestion**: try it by yourself!

- However, for illustration purposes, we will use the EM algorithm.

EM algorithm

Genetic linkage model: Rao (1973), Dempster, Laird, and Rubin (1977)

- To this end, let us suppose that the first cell is divided into tow sub cells with probabilities $1/2$ and $\theta/4$, respectively.

- Let z and $y_1 - z$ be the number of observations (=animals) that fall into the first and second sub cells, respectively. Note that z is unobserved/missing/latent.

- The random vector $(Z, Y_1 - Z, Y_2, Y_3, Y_4)$ has multinomial distribution with probabilities

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- Let (y, z) form the hypothetical complete data. The likelihood of the complete data is

$$L(\theta \mid y, z) \propto \left(\frac{1}{2}\right)^{z} \left(\frac{\theta}{4}\right)^{y_1 - z + y_4} \left(\frac{1}{4}(1-\theta)\right)^{y_2 + y_3}.$$

EM algorithm

Genetic linkage model: Rao (1973), Dempster, Laird, and Rubin (1977)

↪ The complete data log likelihood is thus

$$\log L(\theta \mid y, z) = (y_1 - z + y_4) \log \theta + (y_2 + y_3) \log(1 - \theta).$$

↪ Let $\theta^{(t)}$ be the current estimate of θ, in the E-step we compute

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_Z[\log L(\theta \mid y, z) \mid y, \theta^{(t)}] \\ &= (y_1 - E[Z \mid y, \theta^{(t)}] + y_4) \log \theta + (y_2 + y_3) \log(1 - \theta). \end{aligned}$$

↪ Now, by construction

$$Z \mid Y_1 = y_1 \sim \text{Binomial}\left(y_1, \frac{1/2}{1/2 + \theta/4}\right),$$

implying that

$$E[Z] = y_1 \times \frac{1/2}{1/2 + \theta/4}.$$

EM algorithm

Genetic linkage model: Rao (1973), Dempster, Laird, and Rubin (1977)

↪ Hence,

$$E[Z \mid y, \theta^{(t)}] = E[Z \mid y_1, \theta^{(t)}] = y_1 \times \frac{1/2}{1/2 + \theta^{(t)}/4} = z^{(t)}.$$

↪ So,

$$Q(\theta \mid \theta^{(t)}) = (y_1 - z^{(t)} + y_4) \log \theta + (y_2 + y_3) \log(1 - \theta).$$

↪ For the M-step,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (y_1 - z^{(t)} + y_4) \frac{1}{\theta} - (y_2 + y_3) \frac{1}{1 - \theta} = 0$$

$$\Rightarrow \theta^{(t+1)} = \frac{y_1 - z^{(t)} + y_4}{n - z^{(t)}},$$

$n = y_1 + y_2 + y_3 + y_4$, which can be solved iteratively.

EM algorithm

Incomplete univariate (normal) data

- Let us assume that the data $\mathbf{y} = (y_1, \ldots, y_n)$ comes from a normal random sample with mean μ (unknown) and variance σ^2 (known).

- Further suppose that, possibly after reordering, only the first m observations $(y_1, \ldots, y_m)$ are observed, with the remainder $n - m$ observations $(y_{m+1}, \ldots, y_n)$ being missing. Assume also that the missing data mechanism is ignorable.

- We have already worked with this example in week 5, and it was a very simple one. With the only purpose of getting used to the calculations, we will now do again this example under the perspective of the EM algorithm.

- We have that $\mathbf{y}_{\text{obs}} = \{y_1, \ldots, y_m\}$ and $\mathbf{y}_{\text{mis}} = \{y_{m+1}, \ldots, y_n\}$.

EM algorithm

Incomplete univariate (normal) data

↪ The log likelihood of the complete data is

$$\log L(\mu \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\mu)^2 - \frac{1}{2\sigma^2}\sum_{i=m+1}^{n}(y_i-\mu)^2.$$

↪ At iteration $t+1$, the E step is given by

$$\begin{aligned} Q(\mu \mid \mu^{(t)}) &= E_{\mathbf{Y}_{\text{mis}}}\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\mu)^2 - \frac{1}{2\sigma^2}\sum_{i=m+1}^{n}(y_i-\mu)^2 \mid \mathbf{y}_{\text{obs}}, \mu^{(t)}\right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\mu)^2 - \frac{1}{2\sigma^2}E_{\mathbf{Y}_{\text{mis}}}\left[\sum_{i=m+1}^{n}(y_i-\mu)^2 \mid \mu^{(t)}\right] \end{aligned}$$

EM algorithm

Incomplete univariate (normal) data

- Let us work with the expectation term and bear in mind that $E[Y] = \mu$ and $E[Y^2] = \mu^2 + \sigma^2$.

- We have that

$$E_{\mathbf{Y}_{\text{mis}}}\left[\sum_{i=m+1}^{n}(y_i-\mu)^2 \mid \mu^{(t)}\right] = (n-m)\left\{(\mu^{(t)})^2 + (\sigma^{(t)})^2\right\} - 2(n-m)\mu\mu^{(t)} + (n-m)\mu^2.$$

- Therefore

$$\begin{aligned} Q(\mu \mid \mu^{(t)}) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\mu)^2 \\ &- \frac{1}{2\sigma^2}\left[(n-m)\left\{(\mu^{(t)})^2 + \sigma^2\right\} - 2(n-m)\mu\mu^{(t)} + (n-m)\mu^2\right] \end{aligned}$$

EM algorithm

Incomplete univariate (normal) data

- We can now proceed to the M step and we have that

$$\frac{\mathrm{d}}{\mathrm{d}\mu} Q(\mu \mid \mu^{(t)}) = \frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - \mu) + \frac{1}{\sigma^2}(n-m)\mu^{(t)} - \frac{1}{\sigma^2}\mu(n-m).$$

- Then

$$\frac{\mathrm{d}}{\mathrm{d}\mu} Q(\mu \mid \mu^{(t)}) = 0 \Rightarrow \mu^{(t+1)} = \frac{\sum_{i=1}^{m} y_i + (n-m)\mu^{(t)}}{n}.$$

- As for the toy example, the E and M steps fo not need to be re-derived at each iteration.