Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

EM algorithm

- The Expectation-Maximisation (EM) algorithm is a very general algorithm for maximum likelihood estimation in incomplete data problems.
- The problems that can be tackled by the EM algorithm are very broad and include problems that would not usually be considered as or are not obvious incomplete data problems, such as random effects models and mixture models.
- The EM algorithm was coined by Dempster, Laird, and Rubin in their seminal paper published in JRSS B in 1977 and it is one of the great success stories of statistics over the past 40 years.
- It has been the first choice of most researchers seeking maximum likelihood estimates in a model that involves incomplete data or can be structured in such a way to have that form.
- At the time of writing these notes, the number of citations of the article on Google Scholar exceeds 50 000, and many of these citations come from fields other than statistics.

EM algorithm

↪ The EM algorithm is closely related to the idea of:

1. filling in missing values by estimated values,
2. estimating the parameters given the filled in values,
3. re-estimating the missing values given the new parameter estimates,
4. re-estimating the parameters given the new filled in values, and

continuing along these lines, iterating until convergence of the parameter estimates.

↪ In a strong sense, the EM algorithm formalised this *ad hoc* idea of handling missing data.

↪ In fact, the *ad hoc* approach mentioned above is an EM algorithm under the condition that the complete log likelihood $\log L(\theta \mid \mathbf{y}_{obs}, \mathbf{y}_{mis})$ is linear in the missing values $\mathbf{y}_{mis}$, otherwise the approach may lead to biased estimates.

EM algorithm

- ↪ Each iteration of the EM algorithm consists of an E-step (expectation step) and an M-step (maximisation step).
- ↪ These steps are often easy to construct conceptually and to program for calculation.
- ↪ An additional advantage of the EM algorithm is that it can be shown to converge reliably, in the sense that under general conditions, each iteration increases the log likelihood of the observed data $\log L(\theta \mid \mathbf{y}_{\text{obs}})$.
- ↪ A disadvantage of the EM algorithm is that its rate of convergence can be too slow when there is a large fraction of missing information.
- ↪ Dempster, Laird, and Rubin (1977) show that convergence is linear with rate proportional to the fraction of information about θ in the observed data likelihood (for more details see Little and Rubin, 2002, Section 8.4.3).
- ↪ The condition for the EM algorithm to be valid, in its basic form, is ignorability and hence MAR.

EM algorithm

The E step and the M step of the EM

- The E-step calculates the conditional expectation of the complete data log-likelihood given the observed data and the current estimates of the parameters.
- The M-step requires performing maximum likelihood estimation for the parameters on the (conditional) expected log likelihood found in the E-step.
- The observed log likelihood is $\log L(\theta \mid \mathbf{y}_{\text{obs}})$, while the complete data log likelihood is $\log L(\theta \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$.
- At the $(t+1)$th iteration, let $\theta^{(t)}$ denote the current estimate of the parameter θ.
- In the E-step, compute the expectation of the complete log likelihood, with respect to $\mathbf{y}_{\text{mis}}$, given the observed data $\mathbf{y}_{\text{obs}}$ and the current estimate $\theta^{(t)}$. That is,

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

EM algorithm

The E step and the M step of the EM

↪ In the M-step, choose $\theta^{(t+1)}$ to be any value of θ that maximises $Q(\theta \mid \theta^{(t)})$. That is,

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

↪ The E- and M-steps are repeated until either

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

changes by at most a specified small amount, or if

$$||\theta^{(t+1)} - \theta^{(t)}||,$$

is sufficiently small, or if

$$\left| Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) \right|,$$

is sufficiently small.

EM algorithm

The E step and the M step of the EM

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

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

EM algorithm

Properties of the EM algorithm

↪ **Stability/monotonicity**: Each iteration of the EM algorithm, leads to an observed likelihood that is greater than or equal to the previous observed likelihood. That is,

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

↪ Under regularity conditions, if $\theta^{(t)}$'s converge then they converge to a stationary point of $\log L(\theta \mid \mathbf{y}_{\text{obs}})$. If there are multiple stationary points (local maxima or saddle points) then the algorithm may not converge to the global maximum. Therefore, it is good practice to try different starting values.

EM algorithm

Toy example

- Let $Y_1, Y_2 \overset{\text{iid}}{\sim} \text{Exp}(\theta)$. Remember that $f(y;\theta) = \theta e^{-\theta y}$, $y > 0$, and $\theta > 0$.
- Suppose that $y_1 = 5$ and y_2 is missing.
- 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 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.

- Let $\theta^{(t)}$ be the current estimate of θ, then

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

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

- 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.

EM algorithm

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

- This example has been used on numerous occasions to demonstrate the EM algorithm, including Dempster, Laird, and Rubin to introduce it.

- 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,

- 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,

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

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,

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

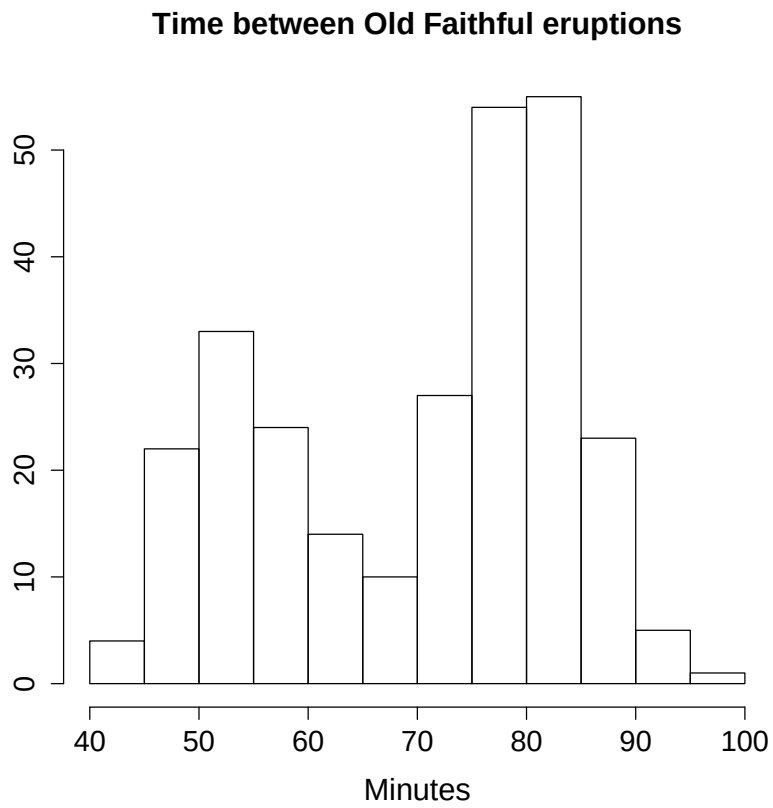
$$\Rightarrow \theta^{(t+1)} = \frac{y_1 - z^{(t)} + y_4}{n - z^{(t)}},$$

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

EM algorithm

Mixtures

↪ Let us consider the popular old faithful data. The data consists of 272 waiting times between eruption for the Old Faithful geyser in Yellowston National park, Wyoming, USA.



EM algorithm

Mixtures

↪ For this dataset we posit as a model a mixture model with two normal components, i.e.,

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} f(y; \theta),$$

where

$$f(y; \theta) = p\phi(y; \mu_1, \sigma_1^2) + (1 - p)\phi(y; \mu_2, \sigma_2^2), \qquad \theta = (p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

↪ The observed data likelihood is

$$L(\theta; y) = \prod_{i=1}^{n} \{p\phi(y_i; \mu_1, \sigma_1^2) + (1 - p)\phi(y_i; \mu_2, \sigma_2^2)\},$$

with corresponding log likelihood given by

$$\log L(\theta; y) = \sum_{i=1}^{n} \log \left\{ p\phi(y_i; \mu_1, \sigma_1^2) + (1 - p)\phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ This log likelihood is difficult to maximise due to the sum inside the logarithm.

EM algorithm

Mixtures

- **Idea**: If we knew the group which observation belongs to, we could simply fit a normal distribution to each group.

- We define an augmented complete dataset where $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_n)$ and $\mathbf{y}_{\text{mis}} = z = (z_1, \ldots, z_n)$ is a vector of unobserved/latent group data indicator, such that

$$z_i = \begin{cases} 1, & \text{if } y_i \text{ belongs to the first component (short waiting times),} \\ 0 & \text{if } y_i \text{ belongs to the second component (long waiting times).} \end{cases}$$

- Note that $\Pr(Z_i = 1) = p$ or, equivalently stated, $Z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

- Then, the complete data likelihood is

$$L(\theta \mid y, z) = \prod_{i=1}^{n} \left\{ [p\phi(y_i; \mu_1, \sigma_1^2)]^{z_i} [(1-p)\phi(y_i; \mu_2, \sigma_2^2)]^{1-z_i} \right\}$$

EM algorithm

Mixtures

↪ Therefore,

$$\log L(\theta \mid y, z) = \sum_{i=1}^{n} z_i \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} + \sum_{i=1}^{n} (1 - z_i) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ For the E-step we would need to compute

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_Z[\log L(\theta \mid y, z) \mid y, \theta^{(t)}] \\ &= \sum_{i=1}^{n} E[z_i \mid y, \theta^{(t)}] \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} + \sum_{i=1}^{n} \left(1 - E[z_i \mid y, \theta^{(t)}]\right) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right. \end{aligned}$$

↪ Now,

$$\begin{aligned} E[Z_i \mid y, \theta^{(t)}] &= E[Z_i \mid y_i, \theta^{(t)}] \\ &= 1 \times \Pr(Z_i = 1 \mid y_i, \theta^{(t)}) + 0 \times \Pr(Z_i = 0 \mid y_i, \theta^{(t)}) \\ &= \frac{p^{(t)} \phi\left(y_i; \mu_1^{(t)}, (\sigma_1^{(t)})^2\right)}{p^{(t)} \phi\left(y_i; \mu_1^{(t)}, (\sigma_1^{(t)})^2\right) + (1 - p^{(t)}) \phi\left(y_i; \mu_2^{(t)}, (\sigma_2^{(t)})^2\right)} \\ &= \widetilde{p}_i^{(t)} \end{aligned}$$

EM algorithm

Mixtures

↪ Thus,

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \widetilde{p}_i^{(t)} \left\{ \log p + \log \phi(y_i; \mu_1, \sigma_1^2) \right\} + \sum_{i=1}^{n} \left(1 - \widetilde{p}_i^{(t)}\right) \left\{ \log(1-p) + \log \phi(y_i; \mu_2, \sigma_2^2) \right\}.$$

↪ For the M-step,

$$\frac{\partial}{\partial p} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow p^{(t+1)} = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}{n}$$

$$\frac{\partial}{\partial \mu_1} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)} y_i}{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}$$

$$\frac{\partial}{\partial \sigma_1^2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (\sigma_1^{(t+1)})^2 = \frac{\sum_{i=1}^{n} \widetilde{p}_i^{(t)} (y_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^{n} \widetilde{p}_i^{(t)}}$$

EM algorithm

Mixtures

↪ Continuing:

$$\frac{\partial}{\partial \mu_2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \mu_2^{(t+1)} = \frac{\sum_{i=1}^{n} (1 - \widetilde{p}_i^{(t)}) y_i}{\sum_{i=1}^{n} (1 - \widetilde{p}_i^{(t)})}$$

$$\frac{\partial}{\partial \sigma_2^2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (\sigma_2^{(t+1)})^2 = \frac{\sum_{i=1}^{n} (1 - \widetilde{p}_i^{(t)})(y_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^{n} (1 - \widetilde{p}_i^{(t)})},$$

which can be solved iteratively.

EM algorithm

Mixtures

↪ Let us apply this algorithm to the old faithful data. The plot below depicts the fit of the model to the observed data.