Exercises in EM

Bernard Flury; Alice Zoppe

The American Statistician, Vol. 54, No. 3. (Aug., 2000), pp. 207-209.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28200008%2954%3A3%3C207%3AEIE%3E2.0.CO%3B2-D

The American Statistician is currently published by American Statistical Association.

Exercises in EM

Bernard FLURY and Alice ZOPPÈ

Suppose survival times follow an exponential distribution, and some observations are right-censored: in this situation the EM algorithm gives a straightforward solution to the problem of maximum likelihood estimation. But what happens if survival times are also left-censored, or if they follow a uniform distribution? The EM algorithm is a generic device useful in a variety of problems with incomplete data, and it appears more and more often in statistical textbooks. This article presents two exercises, which are extensions of a well-known example used in introductions to the EM algorithm. They focus on two points: the applicability of the algorithm and its self-consistency property.

KEY WORDS: EM-algorithm; Exponential distribution; Incomplete data; Maximum likelihood estimation; Uniform distribution.

1. INTRODUCTION

The EM algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997) is a powerful tool for computing maximum likelihood estimates with incomplete data. "Incomplete" is a generic word that, according to the situation, can assume different meanings: missing values, unknown components, censored observations, latent variables, and so on. Its appearance in modern textbooks gives the teacher of medium to advanced level statistics courses excellent opportunities to challenge students with exercises. Two such exercises are described in this note.

A brief (and incomplete) description of the EM algorithm follows.

Let $\mathcal{Y}$ denote the observed data and $\mathcal{X}$ the unknown data, θ the parameter of interest and $\ell_c(\theta; \mathcal{Y}, \mathcal{X})$ the (hypothetical) complete-data log-likelihood, defined for all θ in a parameter space Ω. Starting with an initial parameter value $\theta^{(0)} \in \Omega$, the EM algorithm repeats the following two steps until convergence.

- E-step: compute $\ell^{(j)}(\theta) = E_{\mathcal{X}|\mathcal{Y},\theta^{(j-1)}}[\ell_c(\theta; \mathcal{Y}, \mathcal{X})]$, where the expectation is taken with respect to the conditional distribution of the missing data $\mathcal{X}$ given the observed data $\mathcal{Y}$, and the current numerical value $\theta^{(j-1)}$ is used in evaluating the expected value.
- M-step : find $\theta^{(j)} \in \Omega$ that maximizes $\ell^{(j)}(\theta)$.

Iterating for $j = 1, 2, \ldots$ between the two steps leads to a sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ that converges to a local maximum of the observed-data log-likelihood, if it exists, under fairly general conditions (for details see Wu 1983).

Prominent applications of the EM algorithm include missing values situations, search of the mode of posterior distribution in Bayesian framework (Tanner 1996), applications to grouped, censored, or truncated data, and finite mixture models; see Dempster et al. (1977) or McLachlan and Krishnan (1997).

2. THE FIRST EXERCISE

Generalizing one of the standard examples of censored data, suppose that the lifetime of litebulbs follows an exponential distribution with unknown mean θ. A total of $M+N$ litebulbs are tested in two independent experiments. In the first experiment, with N bulbs, the exact lifetimes $y_1, \ldots, y_N$ are recorded. In the second experiment, the experimenter enters the laboratory at some time $t > 0$, and all she registers is that some of the M litebulbs are still burning, while the others have expired. Thus, the results from the second experiment are right- or left-censored, and the available data are indicators $E_1, \ldots, E_M$, where $E_i = 1$ if the bulb is still burning, and $E_i = 0$ if the light is out.

Having this data, which is the MLE $\hat{\theta}$?

Let $X_1, \ldots, X_M$ be the (unobserved) lifetimes associated with the second experiment, and $Z = \sum_{i=1}^{M} E_i$ the number of litebulbs in the second experiment that are still alive at time t. Thus, the observed data from both the experiments combined is

$$\mathcal{Y} = (Y_1, Y_2, \ldots, Y_N, E_1, E_2, \ldots, E_M),$$

and the unobserved data is

$$\mathcal{X} = (X_1, \ldots, X_M).$$

The complete-data log-likelihood is

$$\ell_c(\theta; \mathcal{Y}, \mathcal{X}) = -N(\log\theta + \bar{Y}/\theta) - \sum_{i=1}^{M}(\log\theta + X_i/\theta), \quad (1)$$

which is linear in the unobserved X_i. But

$$E[X_i|\mathcal{Y}] = E[X_i|E_i] = \begin{cases} t+\theta & \text{if} \quad E_i = 1 \\ \theta - \frac{te^{-t/\theta}}{1-e^{-t/\theta}} & \text{if} \quad E_i = 0, \end{cases} \quad (2)$$

and therefore the jth step consists of replacing X_i in (1) by its expected value (2), using the current numerical parame-

Bernard Flury was Professor of Statistics, Groupe de Statistique, University of Neuchatel (CH). Professor Flury left this life in a tragic accident in July 1999. A close friend and a great teacher, Dr. Flury touched all those around him and is profoundly missed. Alice Zoppè is researcher, Department of Management and Computer Sciences, University of Trento, Via Inama 5, 38100 Trento, Italy (E-mail: azoppe@gelso.unitn.it). The authors thank Beat Neuenschwander for insightful comments on an early draft of this article.

ter value $\theta^{(j-1)}$. The result is

$$\ell^{(j)}(\theta) = -(N+M)\log\theta - \frac{1}{\theta}[N\bar{Y} + Z(t+\theta^{(j-1)}) + (M-Z)(\theta^{(j-1)} - tp^{(j-1)})], \quad (3)$$

where

$$p^{(j)} = \frac{e^{-t/\theta^{(j)}}}{1 - e^{-t/\theta^{(j)}}}.$$

The jth M-step maximizes (3), yielding

$$\theta^{(j)} = f(\theta^{(j-1)}) \equiv \frac{N\bar{Y} + Z(t+\theta^{(j-1)}) + (M-Z)\left(\theta^{(j-1)} - tp^{(j-1)}\right)}{N+M}. \quad (4)$$

Thus, one can simply iterate Equation (4), starting with an arbitrary positive $\theta^{(0)}$, until convergence.

The self-consistency equation $\theta = f(\theta)$ has no explicit solution unless $Z = M$, (i.e., all litebulbs in experiment 2 are still on at time t); in this case, we obtain the well-known solution $\hat{\theta} = (N\bar{Y} + Mt)/N$.

This exercise may also be tackled directly without using the EM algorithm. The joint log-likelihood of both experiments, using observed data only, is

$$\ell(\theta) = -N(\log\theta + \bar{Y}/\theta) - Zt/\theta + (M-Z)\log(1 - e^{-t/\theta}). \quad (5)$$

The log-likelihood (5) can then be maximized with standard numerical methods. If $Z = M$, the self-consistency equation and the maximum of the (5) can be solved analytically, and in this case ML and EM give the same answer (McLachlan and Krishnan 1997, p. 24).

3. THE SECOND EXERCISE

Contrary to litebulbs, lifetimes of heavybulbs follow a uniform distribution in the interval $(0, \theta]$, where $\theta > 0$ is unknown. Suppose the same experiments are performed as in the first exercise, and again the second experimenter registers only that Z out of M heavybulbs are still burning at time t, while $M - Z$ have expired.

Using again the EM algorithm, the solution to the problem of maximum likelihood estimation is fairly straightforward. We know that for (hypothetical) complete data, the MLE would be $\max\{X_{\max}, Y_{\max}\}$, where $Y_{\max}$ is the largest of the observed lifetimes, and $X_{\max}$ is the largest of the unobserved lifetimes.

Assume for simplicity that $Z \geq 1$, so that we are sure that $\theta \geq t$. Then

$$E[X_i|E_i] = \begin{cases} \frac{1}{2}(t+\theta) & \text{if} \quad E_i = 1 \\ \frac{1}{2}t & \text{if} \quad E_i = 0, \end{cases}$$

and the EM algorithm consists simply of iterations of the equation:

$$\theta^{(j)} = f(\theta^{(j-1)}) \equiv \max\{Y_{\max}, \frac{1}{2}(t + \theta^{(j-1)})\}. \quad (6)$$

Starting with any $\theta^{(0)} > 0$, iterations for $j = 1, 2, \ldots$ will converge to the solution $\hat{\theta} = \max\{Y_{\max}, t\}$, and this conclusion may be obtained easily by noticing that the self-consistency equation $\theta = f(\theta)$ is solved by $\hat{\theta}$.

The main advantage of this solution is its simplicity. Its main disadvantage is that it is wrong. Why it is wrong we will explain later, after sketching a correct solution.

The joint likelihood function of both experiments is

$$L(\theta) = \theta^{-N} I_{[Y_{\max},\infty)}(\theta) \times \left(\frac{t}{\max(t,\theta)}\right)^{M-Z} \left(1 - \frac{t}{\max(t,\theta)}\right)^{Z}. \quad (7)$$

Consider first the case $Z = 0$. Then

$$L(\theta) = \theta^{-N} I_{[Y_{\max},\infty)}(\theta) \left(\frac{t}{\max(t,\theta)}\right)^{M},$$

which is decreasing for $\theta \geq Y_{\max}$, and therefore the maximum likelihood estimator is $\hat{\theta} = Y_{\max}$.

Next consider the case $Z \geq 1$, which implies $\theta \geq t$. For $\theta > t$, the function $H(\theta) = \theta^{-(N+M)}(\theta - t)^{Z}$ has a unique maximum in $\dot{\theta} = \frac{N+M}{N+M-Z}\, t$ and is monotonically decreasing for $\theta > \dot{\theta}$. Thus, the likelihood function (7) takes its maximum at $\dot{\theta}$ if $\dot{\theta} > Y_{\max}$, and at $Y_{\max}$ if $\dot{\theta} < Y_{\max}$.

Summarizing these results we obtain the maximum likelihood estimate as

$$\hat{\theta} = \begin{cases} \dot{\theta} & \text{if} \quad \dot{\theta} > Y_{\max} \quad \text{and} \quad Z \geq 1 \\ Y_{\max} & \text{otherwise.} \end{cases}$$

Why is the solution given by the EM algorithm wrong? The answer is simple: the EM algorithm is not applicable because the log-likelihood function does not exist for all $\theta > 0$, which means that its expected value is not defined. To see this, assume that one heavybulb has survived time t, and let X_m be its (unobserved) lifetime. The unconditional pdf of X is

$$f_X(x_m;\theta) = \begin{cases} 1/\theta & \text{if} \quad 0 \leq x_m \leq \theta \\ 0 & \text{elsewhere.} \end{cases}$$

In the jth E-step we need to find $\ell^{(j)}(\theta) = E_{X|\mathcal{Y},\theta^{(j-1)}}[\ell_c(\theta;\mathcal{X},\mathcal{Y}]$. Conditionally on $X_m|Y_m$, which means conditionally on $X_m > t$, and using $\theta^{(j-1)}$ as the parameter, X_m follows a uniform distribution in $[t, \theta^{(j-1)}]$. Now, for all $\theta < \theta^{(j-1)}$, $f(x_m;\theta)$ takes value zero with positive probability, and hence $\ell^{(j)}(\theta)$ does not exist for $\theta < \theta^{(j-1)}$. This could be seen from Equation (7), but in the rush of applying the EM algorithm, it is easy to skip this check.

4. CONCLUSIONS

This note originated in exercises given to students in graduate level mathematical statistics courses, following an introduction to the EM algorithm. The first exercise emphasizes the fact that the self-consistency property cannot be derived for every situation, opposite to what the students might think. Even the best students failed to solve the second exercise correctly, falling into the well-prepared trap

(provided by the first exercise) of simply replacing unobserved data by their conditional expectation. The lesson to be learned is this: it can not be stressed enough that the E-step does not simply involve replacing missing data by their conditional expectation (although this is true for many important applications of the algorithm). Rather, the E-step takes the expected value of the complete-data log-likelihood function, conditional on the observed data. If the likelihood function takes value zero in a subset of the parameter space, then the log-likelihood function does not exist, and the EM algorithm is not applicable.

[Received April 1999. Revised October 1999.]

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

McLachlan, G., and Krishnan, T. (1997), *The EM-algorithm and Extensions*, New York: Wiley.

Tanner, M. A. (1996), *Tools for Statistical Inference*, (3rd ed.), New York: Springer-Verlag.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.