# A Lecture Note on The EM Algorithm

Xia Shen*

November 28, 2012

## Introduction

This lecture note discusses the *Expectation Maximization (EM) algorithm* of Dempster et al. (1977). They pointed out that the method had been "proposed many times in special circumstances" by other authors, but the 1977 paper generalized the method and developed the theory behind it. The EM algorithm has become a popular tool in statistical estimation problems involving incomplete data, or in problems which can be posed in a similar form, such as mixture estimation (McLachlan and Krishnan, 1996; McLachlan and Peel, 2000). The EM algorithm has also been used in various motion estimation frameworks (Weiss, 1998) and variants of it have been used in multiframe superresolution restoration methods which combine motion estimation along the lines of Hardie et al. (1997).

Generally in statistics, the EM algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E-step. The parameters found on the M-step are then used to begin another E step, and the process is repeated.

To illustrate the EM algorithm, the rest of this note will first review the properties of the relative convex functions, and after that, the derivation, convergence, standard errors and some examples of the algorithm will be introduced.

## Review of Convexity

**Definition 1** *Let $f$ be a real valued function defined on an interval $I = [a, b]$. $f$ is said to be convex on $I$ if $\forall x_1, x_2 \in I$ and $\lambda \in [0, 1]$,*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leqslant \lambda f(x_1) + (1 - \lambda)f(x_2)$$

*$f$ is said to be strictly convex if the inequality is strict.*

**Definition 2** *$f$ is concave (strictly concave) if $-f$ is convex (strictly convex).*

**Theorem 3** *If $f(x)$ is twice differentiable on $[a, b]$ and $f''(x) \geqslant 0$ on $[a, b]$ then $f(x)$ is convex on $[a, b]$.*

*Postdoctoral researcher, Division of Computational Genetics, Department of Clinical Sciences, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden; Lecturer, School of Technology and Business Studies/Statistics, Dalarna University, Borlänge, Sweden. Homepage: http://www.19850911.com

***Proof.*** *For $x \leqslant y \in [a,b]$ and $\lambda \in [0,1]$ let $z = \lambda y + (1-\lambda)x$. By definition, $f$ is convex if $f(\lambda y + (1-\lambda)x) \leqslant \lambda f(y) + (1-\lambda)f(x)$. Noting that $f(z) = \lambda f(z) + (1-\lambda)f(z)$ we have that $f(z) = \lambda f(z) + (1-\lambda)f(z) \leqslant \lambda f(y) + (1-\lambda)f(x)$. Rearranging terms gives an equivalent definition for convexity that $f$ is convex if*

$$\lambda[f(y) - f(z)] \geqslant (1-\lambda)[f(z) - f(x)] \tag{i}$$

*By the mean value theorem, $\exists s$, $x \leqslant s \leqslant z$ s.t.*

$$f(z) - f(x) = f'(s)(z-x)$$

*Similarly, $\exists t$, $z \leqslant t \leqslant y$ s.t.*

$$f(y) - f(z) = f'(t)(y-z)$$

*Thus we have the situation $x \leqslant s \leqslant z \leqslant t \leqslant y$. By assumption, $f''(x) \geqslant 0$ on $[a,b]$ so*

$$f'(s) \leqslant f'(t)$$

*Rewriting $z = \lambda y + (1-\lambda)x$ gives*

$$(1-\lambda)(z-x) = \lambda(y-z)$$

*Combining the above we have*

$$\begin{aligned}(1-\lambda)[f(z) - f(x)] &= (1-\lambda)f'(s)(z-x) \\ &\leqslant f'(t)(1-\lambda)(z-x) \\ &= f'(t)\lambda(y-z) \\ &= \lambda[f(y) - f(z)]\end{aligned}$$

*i.e. (i) holds, therefore $f(x)$ is convex on $[a,b]$.* ■

**Corollary 4** *$-\log(x)$ is strictly convex on $(0,\infty)$.*
***Proof.*** *With $f(x) = -\log(x)$, we have $f''(x) = x^{-2} > 0$ for $x \in (0,\infty)$. By Theorem 3, $-\log(x)$ is strictly convex on $(0,\infty)$. Also, by Definition 2, $\log(x)$ is strictly concave on $(0,\infty)$.* ■

Extending the notion of convexity to apply to $n$ points, we have the result known as Jensen's inequality.

**Theorem 5 (Jensen's Inequality)** *Let $f$ be a convex function defined on an interval $I$. If $x_1, x_2, \ldots, x_n \in I$ and $\lambda_1, \lambda_2, \ldots, \lambda_n \geqslant 0$ with $\sum_{i=1}^n \lambda_i = 1$, then*

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leqslant \sum_{i=1}^n \lambda_i f(x_i)$$

***Proof.*** *For $n = 1$, trivial. The case $n = 2$ is just Definition 1. To show the inequality is true for all natural numbers, we proceed by induction. Assume the theorem is true for some natural number $k$, then for the case*

$n = k + 1$,

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_{k+1}x_{k+1} + \sum_{i=1}^{k}\lambda_i x_i\right) \\
&= f\left(\lambda_{k+1}x_{k+1} + (1-\lambda_{k+1})\frac{1}{1-\lambda_{k+1}}\sum_{i=1}^{k}\lambda_i x_i\right) \\
&\leqslant \lambda_{k+1}f(x_{k+1}) + (1-\lambda_{k+1})f\left(\frac{1}{1-\lambda_{k+1}}\sum_{i=1}^{k}\lambda_i x_i\right) \\
&= \lambda_{k+1}f(x_{k+1}) + (1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right) \\
&\leqslant \lambda_{k+1}f(x_{k+1}) + (1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}f(x_i) \\
&= \lambda_{k+1}f(x_{k+1}) + \sum_{i=1}^{k}\lambda_i f(x_i) \\
&= \sum_{i=1}^{k+1}\lambda_i f(x_i)
\end{aligned}$$

*therefore the theorem holds for all natural numbers.* ■

From Corollary 4 and Theorem 5, we have the useful result

$$\log\sum_{i=1}^{n}\lambda_i x_i \geqslant \sum_{i=1}^{n}\lambda_i \log(x_i) \tag{ii}$$

which will be used in the derivation of the EM algorithm in the next section.

**Exercise 6** *Prove that for a positive-valued sample $x_1, x_2, \ldots, x_n$, the arithmetic mean is greater than or equal to the geometric mean, i.e.*

$$\frac{1}{n}\sum_{i=1}^{n}x_i \geqslant \sqrt[n]{x_1 x_2 \ldots x_n}$$

## The EM Algorithm

The EM algorithm is an efficient iterative procedure to compute the *Maximum Likelihood (ML)* estimate in the presence of missing or hidden data. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely. Each iteration of the EM algorithm consists of two processes: the **E-step (expectation)**, and the **M-step (maximization)**. In the E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

## Derivation of the EM Algorithm

Let $\mathbf{X}$ be a random vector which results from a parameterized family. We wish to find $\theta$ such that $\mathcal{P}(\mathbf{X}|\theta)$ is a maximum, where $\mathcal{P}$ is a *probability measure* which usually may be a *probability mass function (pmf)* or a *probability density function (pdf)*. This is known as the ML estimate for $\theta$. In order to estimate $\theta$, it is typical to introduce the *log-likelihood function* defined as,

$$\ell(\theta) = \log \mathcal{P}(\mathbf{X}|\theta)$$

The likelihood function is considered to be a function of the parameter $\theta$ given the data $\mathbf{X}$. Since $\log(x)$ is a strictly increasing function, the value of $\theta$ which maximizes $\mathcal{P}(\mathbf{X}|\theta)$ also maximizes $\ell(\theta)$. The EM algorithm is an iterative procedure for maximizing $\ell(\theta)$. Assume that after the $n^{th}$ iteration, the current estimate for $\theta$ is given by $\theta_n$. Since the objective is to maximize $\ell(\theta)$, we wish to compute an updated estimate $\theta$ such that

$$\ell(\theta) > \ell(\theta_n)$$

Equivalently, we want to maximize the difference

$$\ell(\theta) - \ell(\theta_n) = \log \mathcal{P}(\mathbf{X}|\theta) - \log \mathcal{P}(\mathbf{X}|\theta_n) \tag{iii}$$

So far, we have not considered any unobserved or missing variables. In problems where such data exist, the EM algorithm provides a natural framework for their inclusion. Alternately, hidden variables may be introduced purely as an artifice for making the maximum likelihood estimation of $\theta$ tractable. In this case, it is assumed that knowledge of the hidden variables will make the maximization of the likelihood function easier. Either way, denote the hidden random vector by $\mathbf{Z}$ and a given realization by $\mathbf{z}$. Using the *law of total probability*, $\mathcal{P}(\mathbf{X}|\theta)$ may be written in terms of the hidden variable $\mathbf{Z}$ as

$$\mathcal{P}(\mathbf{X}|\theta) = \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)$$

We may then rewrite Equation (iii) as

$$\ell(\theta) - \ell(\theta_n) = \log \left( \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \log \mathcal{P}(\mathbf{X}|\theta_n) \tag{iv}$$

Notice that this expression involves the logarithm of a sum. Using the inequality (ii), consider letting the constants be of the form $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$. Since $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$ is a probability measure, we have that $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \geqslant 0$ and that $\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) = 1$ as required. Then introducing constants $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$ and starting from Equation (iv), we have

$$\begin{aligned}
\ell(\theta) - \ell(\theta_n) &= \log \left( \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \log \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \log \left( \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \frac{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right) - \log \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \log \left( \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right) - \log \mathcal{P}(\mathbf{X}|\theta_n) \\
&\geqslant \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \log \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} - \log \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \log \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \\
&= \Delta(\theta|\theta_n)
\end{aligned}$$

Then,

$$\ell(\theta) \geqslant \ell(\theta_n) + \Delta(\theta|\theta_n) \tag{v}$$

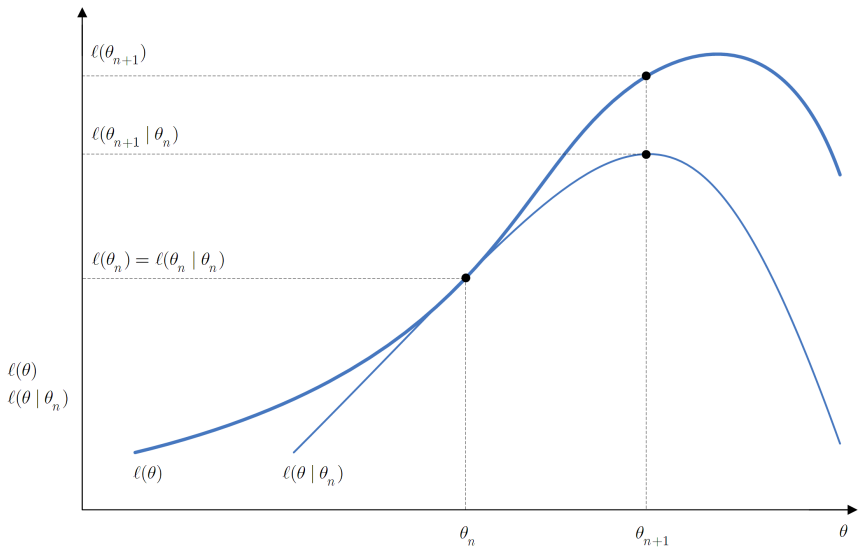


Figure 1: Graphical interpretation of a single iteration of the EM algorithm: The function $\ell(\theta|\theta_n)$ is bounded above by the likelihood function $\ell(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses $\theta_{n+1}$ as the value of $\theta$ for which $\ell(\theta|\theta_n)$ is a maximum. Since $\ell(\theta) \geqslant \ell(\theta|\theta_n)$, increasing $\ell(\theta|\theta_n)$ ensures that the value of the likelihood function $\ell(\theta)$ is increased at each step.

For convenience, define

$$\ell(\theta|\theta_n) = \ell(\theta_n) + \Delta(\theta|\theta_n)$$

so that the relationship in (v) becomes

$$\ell(\theta) \geqslant \ell(\theta|\theta_n)$$

We have now a function, $\ell(\theta|\theta_n)$ which is bounded above by the likelihood function $\ell(\theta)$. Additionally, observe that

$$\begin{aligned}
\ell(\theta_n|\theta_n) &= \ell(\theta_n) + \Delta(\theta_n|\theta_n) \\
&= \ell(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \log \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta_n)\mathcal{P}(\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \\
&= \ell(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \log \frac{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)} \\
&= \ell(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \log 1 \\
&= \ell(\theta_n)
\end{aligned}$$

so for $\theta = \theta_n$, the functions $\ell(\theta|\theta_n)$ and $\ell(\theta)$ are equal.

Our objective is to choose a values of $\theta$ so that $\ell(\theta)$ is maximized. We have shown that the function $\ell(\theta|\theta_n)$ is bounded above by the likelihood function $\ell(\theta)$ and that the value of the functions $\ell(\theta|\theta_n)$ and $\ell(\theta)$ are equal at the current estimate for $\theta = \theta_n$. Therefore, any $\theta$ which increases $\ell(\theta|\theta_n)$ in turn increase the $\ell(\theta)$. In order to achieve the greatest possible increase in the value of $\ell(\theta)$, the EM algorithm calls for selecting $\theta$ such that $\ell(\theta|\theta_n)$ is maximized. We denote this updated value as $\theta_{n+1}$. This process is illustrated

in Figure 1. Formally, we have

$$\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta}\{\ell(\theta|\theta_n)\} \\
&= \arg\max_{\theta}\{\ell(\theta_n) + \Delta(\theta|\theta_n)\} \\
&= \arg\max_{\theta}\left\{\ell(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\log\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)}\right\} \\
&= \arg\max_{\theta}\left\{\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\log(\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta))\right\} \\
&= \arg\max_{\theta}\left\{\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\log\mathcal{P}(\mathbf{X},\mathbf{z}|\theta)\right\}
\end{aligned}$$

i.e., according to the definition of mathematical expectation,

$$\theta_{n+1} = \arg\max_{\theta}\{E_{\mathbf{Z}|\mathbf{X},\theta_n}[\log\mathcal{P}(\mathbf{X},\mathbf{Z}|\theta)]\} \tag{vi}$$

Equation (vi) actually contains the expectation and maximization steps, thus the EM algorithm consists of iterating the

**Algorithm 7 (The EM Algorithm)**

1. ***E-step:*** *Determine the conditional expectation* $Q(\theta,\theta_n) = E_{\mathbf{Z}|\mathbf{X},\theta_n}[\log\mathcal{P}(\mathbf{X},\mathbf{Z}|\theta)]$*;*
2. ***M-step:*** *Maximize* $Q(\theta,\theta_n)$ *with respect to* $\theta$*.*

At this point it is fair to ask what has been gained given that we have simply traded the maximization of $\ell(\theta)$ for the maximization of $\ell(\theta|\theta_n)$. The answer lies in the fact that $\ell(\theta|\theta_n)$ takes into account the unobserved or missing data $\mathbf{Z}$. In the case where we wish to estimate these variables, the EM algorithms provides a framework for doing so. Also, as alluded to earlier, it may be convenient to introduce such hidden variables so that the maximization of $\ell(\theta|\theta_n)$ is simplified given knowledge of the hidden variables (as compared with a direct maximization of $\ell(\theta)$).

## Convergence of the EM Algorithm

The convergence properties of the EM algorithm are discussed in detail by McLachlan and Krishnan (1996). In this section, we discuss the general convergence of the algorithm. Recall that $\theta_{n+1}$ is the estimate for $\theta$ which maximizes the difference $\Delta(\theta|\theta_n)$. Starting with the current estimate for $\theta$, i.e., $\theta_n$, we had that $\Delta(\theta_n|\theta_n) = 0$. Since $\theta_{n+1}$ is chosen to maximize $\Delta(\theta|\theta_n)$, we then have that $\Delta(\theta_{n+1}|\theta_n) \geqslant \Delta(\theta_n|\theta_n) = 0$, so for each iteration the likelihood $\ell(\theta)$ is nondecreasing.

**Exercise 8** *By Jensen's inequality, if* $\mathbf{Y}$ *is a random variable, the following inequality holds for all densities* $f$ *and* $g$*,*

$$E_g\left[\log\frac{f(\mathbf{Y})}{g(\mathbf{Y})}\right] \leqslant 0$$

*Use this result to show that the EM algorithm always increases the likelihood* $\ell(\theta)$*.*

When the algorithm reaches a fixed point for some $\theta_n$, the value $\theta_n$ maximizes $\ell(\theta|\theta_n)$. Since $\ell(\theta)$ and $\ell(\theta|\theta_n)$ are equal at $\theta_n$ if $\ell(\theta)$ and $\ell(\theta|\theta_n)$ are differentiable at $\theta_n$, then $\theta_n$ must be a stationary point of $\ell(\theta)$. The stationary point need not, however, be a local maximum. McLachlan and Krishnan (1996) showed that it

is possible for the algorithm to converge to local minima or saddle points in unusual cases. Furthermore, the rate of convergence can be very slow, suggesting that alternative procedures, or techniques which accelerate the convergence, may often be appropriate.

### Standard Errors

As usual, calculation of standard errors requires calculation of the inverse Hessian:

$$\left[-\nabla_\theta^2 \log \mathcal{P}(\mathbf{X}|\theta)\right]^{-1}$$

evaluated at the mode $\widehat{\theta}$. In missing data problems this may be difficult to evaluate directly. It is possible however to exploit the structure of the EM algorithm in simplifying this calculation. This uses what is known as the *missing information principle*:

$$\textit{Observed Information} = \textit{Complete Information} - \textit{Missing Information}$$

Explicitly, this takes the form

$$-\nabla_\theta^2 \log \mathcal{P}(\mathbf{X}|\theta) = [-\nabla_\theta^2 Q(\theta,\phi)]_{\phi=\theta} - [-\nabla_\theta^2 H(\theta,\phi)]_{\phi=\theta} \tag{vii}$$

where

$$H(\theta,\phi) = \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\phi) \log \mathcal{P}(\mathbf{z}|\mathbf{X},\theta)$$

Without the missing data, the first term on the right-hand side of (vii) would be the Hessian; the second term compensates for the missing information. The proof of (vii) is almost immediate from the representation

$$\log \mathcal{P}(\mathbf{X}|\theta) = \log \mathcal{P}(\mathbf{X},\mathbf{Z}|\theta) - \log \mathcal{P}(\mathbf{Z}|\mathbf{X},\theta) + C$$

which is essentially a statement of *Bayes' theorem*. Application of the missing information principle is simplified by using the result

$$-\nabla_\theta^2 H(\theta,\phi) = Var[\nabla_\theta \log \mathcal{P}(\mathbf{X},\mathbf{Z}|\theta)]$$

a result which mirrors the corresponding result in the classical (full data) theory.

## Examples*

### A Toy Example

Assume that we observe a sample $\mathbf{x} = (9, 11, ?)$ from the univariate normal distribution $\mathcal{N}(\theta, \sigma^2)$. Our objective is to estimate the mean parameter $\theta$. The third observation of $\mathbf{x}$ is missing, thus we may give an initial guess of it as $\theta_0 = 0$. Then the maximum likelihood estimate of $\theta$ will be $\theta_1 = (9 + 11 + 0)/3 = 20/3$. We fill in the missing observation with the estimate $\theta_1$ since the conditional expectation of the missing observation is the current estimate of $\theta$. Generally, we have

$$\theta_{n+1} = \frac{9 + 11 + \theta_n}{3}$$

We may iterate this to obtain the ML estimate of $\theta$. Applying some limit theory, we see that $\theta_n \to 10$ as $n \to \infty$. For this simple model (a univariate normal distribution with unit variance), it can be seen that substituting the average of the known values is the best answer. For more complex models, there is no easy way to find the best answer, and the EM algorithm is a very popular approach for estimating the answer.

*Two computer exercises can be found at http://www.du.se/~lrn/StatMod12.

## Poisson Process Recording

Assume a radioactive material is known to emit $N \sim Possion(100)$ particles in unit time. A measurement equipment records each particle with probability $\theta$. If the recorded value was $y = 84$, what is the maximum likelihood estimate of $\theta$? Here we failed to observe $Z$, the total number of particles emitted. Conditionally on $Z$, $y$ is an observation of a $Binomial(Z, \theta)$ variable. The augmented likelihood is

$$\begin{aligned}\mathcal{L}(\theta|y,z) &= f(y,z|\theta)\\ &= f(y|z,\theta)f(z|\theta)\\ &= \binom{z}{y}\theta^y(1-\theta)^{z-y}\frac{100^z\exp(-100)}{z!}\\ &\propto \theta^y(1-\theta)^{z-y}\end{aligned}$$

and

$$\begin{aligned}f(z|y,\theta) &\propto \frac{z!}{(z-y)!}(1-\theta)^{z-y}\frac{100^z\exp(-100)}{z!}\\ &\propto \frac{(1-\theta)^{z-y}100^z}{(z-y)!}\\ &\propto \frac{(100(1-\theta))^{z-y}}{(z-y)!}\end{aligned}$$

which can be recognized as the pmf of $y + W$, where $W \sim Possion(100(1-\theta))$. Hence,

$$\begin{aligned}Q(\theta,\theta_n) &= E_{Z|y,\theta_n}[\log\mathcal{L}(\theta|y,Z)]\\ &= E_{Z|y,\theta_n}[y\log\theta + (Z-y)\log(1-\theta)]\\ &= y\log\theta + (E_{Z|y,\theta_n}[Z]-y)\log(1-\theta)\\ &= y\log\theta + (E_{W|y,\theta_n}[y+W]-y)\log(1-\theta)\\ &= y\log\theta + E_{W|y,\theta_n}[W]\log(1-\theta)\\ &= y\log\theta + 100(1-\theta_n)\log(1-\theta)\end{aligned}$$

which is maximized by

$$\theta_{n+1} = \frac{y}{y + 100(1-\theta_n)}$$

We iterate this to obtain the ML estimate of $\theta$. Applying some limit theory, we see that $\theta_n \to 0.84$ as $n \to \infty$.

## A Classic Genetic Example

This is a famous example from Rao (1973). We consider the genetic linkage of 197 animals, in which the phenotypes are distributed into 4 categories:

$$\mathbf{Y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with cell probabilities

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

Though it is by no means impossible to maximize this multinomial likelihood directly, we illustrate how the EM algorithm brings a substantial simplification, by using the augmentation method. Specifically, we augment the observed data $\mathbf{Y}$ by dividing the first cell into two, with respective cell probabilities $1/2$ and $\theta/4$. This gives an augmented data set $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)$, where $x_1 + x_2 = y_1$, and $x_3 = y_2$, $x_4 = y_3$, $x_5 = y_4$. Now we have

$$\mathcal{L}(\theta|\mathbf{Y}) \propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$$

whereas

$$\mathcal{L}(\theta|\mathbf{X}) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}$$

Thus we obtain

$$\begin{aligned} Q(\theta,\theta_n) &= E_{\mathbf{X}|\mathbf{Y},\theta_n}[\log \mathcal{L}(\theta|\mathbf{X})] \\ &= E_{\mathbf{X}|\mathbf{Y},\theta_n}[(X_2+X_5)\log(\theta)+(X_3+X_4)\log(1-\theta)] \\ &= (E_{\mathbf{X}|\mathbf{Y},\theta_n}[X_2]+x_5)\log(\theta)+(x_3+x_4)\log(1-\theta) \end{aligned}$$

where

$$X_2|\mathbf{Y},\theta_n \sim Binomial\left(125, \frac{\theta_n}{\theta_n+2}\right)$$

Thus

$$Q(\theta,\theta_n) = \left(\frac{125\theta_n}{\theta_n+2}+34\right)\log(\theta)+38\log(1-\theta)$$

which is maximized by

$$\theta_{n+1} = \frac{159\theta_n+68}{197\theta_n+144}$$

The alternation between estimation and maximization is clearly seen in this iteration formula. Starting with $\theta_0 = 0.5$ we obtain the sequence as follows. Hence the ML estimate is $\widehat{\theta} = 0.6268$.

| $n$ | $\theta_n$ |
|---|---|
| 0 | 0.5 |
| 1 | 0.6082 |
| 2 | 0.6243 |
| 3 | 0.6265 |
| 4 | 0.6268 |
| 5 | 0.6268 |

**Exercise 9** *For this genetic example, using the missing information principle introduced in the previous section, calculate the standard error of the ML estimate obtained by the EM algorithm.*

# References

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, Nov 1977.

R. Hardie, K. Barnard, and E. Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, Dec 1997.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 1996.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

C. Rao. *Linear Statistical Inference and Its Applications (2nd edn.)*. New York: Wiley, 1973.

Y. Weiss. *Bayesian Motion Estimation and Segmentation*. PhD thesis, Massachusetts Institute of Technology, May 1998.