## EM algorithm

Our discussion of the EM algorithm will follow chapter 4 of Martin Tanner's *Tools for Statistical Inference : Methods for the Exploration of Posterior Distributions and Likelihood Functions*.

Let $\boldsymbol{\beta}$ be a parameter vector of a parametric model and $\boldsymbol{Y}$ be the observed data. The EM algorithm is useful for simplifying the search for the mode of the posterior $p[\boldsymbol{\beta} \mid \boldsymbol{Y}]$, in situations when it might otherwise be a difficult search.

It works by incorporating the use of latent data $\boldsymbol{Z}$ such that $p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]$ is a simple expression.

It exploits the simplicity of
$p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]$ in order to maximize $p[\boldsymbol{\beta} \mid \boldsymbol{Y}]$.

The EM algorithm involves iteratively performing an **E**xpectation step and a **M**aximization step.

**E-step**: Let $\boldsymbol{\beta}^{(m)}$ denote the current guess of the mode of $p[\boldsymbol{\beta} \mid \boldsymbol{Y}]$, the **observed posterior**.

In the general setting we refer to $p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]$ as the **augmented posterior** and we call $p[\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y}]$ the **conditional predictive distribution** of the latent data $\boldsymbol{Z}$, conditional on $\boldsymbol{\beta}^{(m)}$.

The E-step consists of computing $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)})$, the expectation of $log(p[\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y}])$ with respect to the density $p[\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y}]$ .

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)}) = \int log(p[\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y}])p[\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y}]d\boldsymbol{Z}$$

**M-step**: In the M-step, $Q$ is maximized with respect to $\boldsymbol{\beta}$ to obtain $\boldsymbol{\beta}^{(m+1)}$. Depending on the problem, this step might require techniques such as Newton-Raphson or numerical integration if the integral in the E-step is difficult.

In any case, the method is only useful if maximizing $Q$ with respect to $\boldsymbol{\beta}$ is much easier than directly maximizing $p[\boldsymbol{\beta} \mid \boldsymbol{Y}]$.

The algorithm is iterated until
$\mid Q(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) - Q(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)}) \mid$ or
$\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|$ is sufficiently small.

**Trival example:**

$Y_1, Y_2 \sim$ i.i.d. Exp($\theta$) with $y_1 = 5$ observed but $y_2$ missing.

The complete data log likelihood function is

$\log\{L(\theta|\mathbf{y})\} = \log\{f_{\mathbf{Y}}(\mathbf{y}|\theta)\} = 2\log\{\theta\} - \theta y_1 - \theta y_2.$

Thus

$$Q(\theta, \theta^{(t)}) = 2\log\{\theta\} - 5\theta - \theta/\theta^{(t)}$$

since $\mathrm{E}\{Y_2|y_1, \theta^{(t)}\} = \mathrm{E}\{Y_2|\theta^{(t)}\} = 1/\theta^{(t)}$ follows from independence.

The maximizer of $Q(\theta, \theta^{(t)})$ is the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$. Thus $\theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)}+1}$. Converges quickly to $\widehat{\theta} = 0.2$.

Note: The E step and M step do not need to be re-derived at each iteration

This example is not realistic because of an easy analytic solution. Taking the required expectation is tricker in real applications because one needs to know the conditional distribution of the complete data given the missing data.

Note that we can write

$$log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}]) = log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]) - log(p[\boldsymbol{Z} \mid \boldsymbol{\beta}, \boldsymbol{Y}])$$

$$+ log(p[\boldsymbol{Z} \mid \boldsymbol{Y}]).$$

Now integrate both sides of this equation with respect to the conditional predictive density $p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*]$ and we obtain,

$$log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}]) = \int log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]) p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*] d\boldsymbol{Z}$$

$$- \int log(p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}]) p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*] d\boldsymbol{Z}$$

$$+ \int log(p[\boldsymbol{Z} | \boldsymbol{Y}]) p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*] d\boldsymbol{Z}.$$

We have already discussed the first term on the right side,

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \int log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}])p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*]d\boldsymbol{Z}.$$

Let $H$ denote the term that is subtracted.

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \int log(p[\boldsymbol{Z} \mid \boldsymbol{\beta}, \boldsymbol{Y}])p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*]d\boldsymbol{Z}$$

Finally let $K$ denote the third function,

$$K(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \int log(p[\boldsymbol{Z} \mid \boldsymbol{Y}])p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*]d\boldsymbol{Z}.$$

We can see how the iterations of the EM-algorithm relate to the observed posterior through these functions.

$$log[p(\boldsymbol{\beta}^{m+1} \mid \boldsymbol{Y})] - log[p(\boldsymbol{\beta}^{m} \mid \boldsymbol{Y})]$$

$$= Q(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) - Q(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)})$$
$$- \left[H(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) - H(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)})\right]$$

$$+ \left[K(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) - K(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)})\right]$$

Dempster, Laird and Rubin (1977) define the EM algorithm to maximize $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)})$ at every iteration.

They also defined the **generalized EM algorithm** (GEM) to select $\boldsymbol{\beta}^{(m+1)}$ to satisfy

$$Q(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) > Q(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)})$$

at every iteration.

In general, it can be shown that

$$H(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) \leq H(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)}).$$

Also, we clearly have
$K(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) - K(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)}) = 0.$

From these facts we have the following theorem.

**Theorem EM1** Every EM or GEM algorithm increases the posterior at each iteration. That is,

$$p[\boldsymbol{\beta}^{m+1} \mid \boldsymbol{Y}] \geq p[\boldsymbol{\beta}^{m} \mid \boldsymbol{Y}]$$

with equality holding iff
$Q(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\beta}^{(m)}) = Q(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(m)})$.

Wu (1983) proved a stronger theorem.

**Theorem EM2** Suppose a sequence of EM iterates $\boldsymbol{\beta}^{(m)}$ satisfies the three conditions:

1. $\frac{\partial Q(\boldsymbol{\beta},\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m+1)}} = \mathbf{0}$

2. The sequence $\boldsymbol{\beta}^{(m)}$ converges to some value $\boldsymbol{\beta}^*$.

3. The predictive distribution $p[\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}]$ is sufficiently smooth.

Then it follows that

$$\frac{\partial p[\boldsymbol{\beta} \mid \boldsymbol{Y}]}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}.$$

Thus, if the sequence converges, it converges to a stationary point of $p[\boldsymbol{\beta} \mid \boldsymbol{Y}]$. When there are multiple stationary points it may not converge to a global maximum.

## Exponential Family Case

Let $\boldsymbol{X}$ denoted the augmented data $(\boldsymbol{Y}, \boldsymbol{Z})$. And suppose that

$$f(\boldsymbol{X} \mid \boldsymbol{\beta}) = b(\boldsymbol{X}) exp[\boldsymbol{\beta}' s(\boldsymbol{X})] a(\boldsymbol{\beta})$$

That is, it belongs to an exponential family where $\boldsymbol{\beta}$ is the parameter vector and $s(\boldsymbol{X})$ is a sufficient statistic of the same dimension.

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)}) = \int log[b(\boldsymbol{X})] p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(m)}) d\boldsymbol{Z}$$

$$+\boldsymbol{\beta}' \int s(\boldsymbol{X}) p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(m)}) d\boldsymbol{Z} - log[a(\boldsymbol{\beta})].$$

Since the first term does not depend on $\boldsymbol{\beta}$, we can see that the E-step consists of computing

$$E[s(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(m)}] = \int s(\boldsymbol{X}) p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(m)}) d\boldsymbol{Z} = s^{(m)}$$

The M-step consists of maximizing

$$-log[a(\boldsymbol{\beta})] + \boldsymbol{\beta}' s^{(m)}$$

Using a well known result for exponential families, we have

$$\frac{\partial log[a(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = E[s(\boldsymbol{X}) \mid \boldsymbol{\beta}].$$

Thus the M-step amounts to solving

$$E[s(\boldsymbol{X}) \mid \boldsymbol{\beta}] = s^{(m)}$$

for $\boldsymbol{\beta}$.

## Monte Carlo Implementation of the E-step

Given the current guess of the posterior mode $\boldsymbol{\beta}^{(m)}$, the E-step requires computing

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)}) = \int log[p(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z})] p(\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y}) d\boldsymbol{Z}.$$

In cases where this integration is difficult or must be done numerically in high dimensions, the result might be an expression that is difficult to maximize in the M-step.

In such cases, if it is possible to draw directly from $p(\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y})$ a useful alternative is a Monte Carlo approximation of $Q$. The steps are

a. Draw $z_1, z_2, ..., z_K \sim p(\boldsymbol{Z} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{Y})$.

b. Let $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(m)}) = \frac{1}{K} \sum_{k=1}^{K} log[p(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{z}_k)]$

A suggestion in Tanner's book is to let $K$ increase as the iterations increase. It is inefficient to spend too much on $K$ early, when little is known about $\boldsymbol{\beta}$, but near the end the precision of $\hat{Q}$ must be very high.

## Standard Errors

The output of the EM algorithm is the mode of the observed posterior $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$. The normal approximation to $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$, requires finding the Hessian of $log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})$.

Let $\boldsymbol{\beta}^*$ denote the posterior mode computed using the EM algorithm. It might be possible to directly evaluate the **observed information**,

$$-\frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{\beta}^*}.$$

However, the very fact that the EM algorithm was used might imply that $log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]$ is a difficult object to work with, and the computation above may not be feasible.

One alternative is to numerically differentiate the score function, $\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]/\partial\boldsymbol{\beta}$.

To do this, one can perturb $\boldsymbol{\beta}^*$ by adding a small amount $\epsilon$ to one coordinate while leaving the other coordinates unperturbed to obtain $\tilde{\boldsymbol{\beta}}$ so that the row of the Hessian corresponding to the perturbed coordinate becomes

$$\frac{1}{\epsilon}\left[\frac{\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial\boldsymbol{\beta}}|_{\tilde{\boldsymbol{\beta}}} - \frac{\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}^*}\right]$$

Dealing with rounding errors due to small $\epsilon$ can be a problem

## Missing Information Principle

The result that

$$log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}]) = log(p[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{Z}]) - log(p[\boldsymbol{Z} \mid \boldsymbol{\beta}, \boldsymbol{Y}]) + C$$

implies

$$-\frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2} = -\frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2} + \frac{\partial^2 log[p(\boldsymbol{Z} \mid \boldsymbol{\beta}, \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2}.$$

If we integrate both sides of this equation with respect to $p(\boldsymbol{Z} \mid \boldsymbol{\beta}, \boldsymbol{Y})$ we get

$$-\frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2} = \frac{-\partial^2 Q(\boldsymbol{\beta}, \boldsymbol{\omega})}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{\omega}=\boldsymbol{\beta}} - \frac{-\partial^2 H(\boldsymbol{\beta}, \boldsymbol{\omega})}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{\omega}=\boldsymbol{\beta}}$$

Louis (1982) referred to the terms in this equations by

observed information=complete information - missing information

Louis (1982) showed that $-\partial^2 H/\partial\boldsymbol{\beta}^2 = var(\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})]/\partial\boldsymbol{\beta})$. Thus, we have

$$-\frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Y})]}{\partial\boldsymbol{\beta}^2}$$

$$= \int \frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})]}{\partial\boldsymbol{\beta}^2} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}) d\boldsymbol{Z}$$

$$-var[\frac{\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})}{\partial\boldsymbol{\beta}}]$$

where the variance is taken with respect to $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta})$.

This result can often be valuable in finding the observed information from the two terms on the right side.

In some cases it might be difficult to compute

$$\int \frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*) d\boldsymbol{Z}$$

Suppose that it is possible to draw from $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*)$. Then one can obtain a sample $z_1, z_2, ..., z_M$ from $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*)$, and approximate the integral by

$$\frac{1}{M} \sum_{m=1}^{M} \frac{\partial^2 log[p(\boldsymbol{\beta} \mid \boldsymbol{z}_m, \boldsymbol{Y})]}{\partial \boldsymbol{\beta}^2}.$$

Also, one can compute

$$var[\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y})]/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}^*}]$$

by

$$\frac{1}{M}\sum_{m=1}^{M}(\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{z}_m, \boldsymbol{Y})]/\partial\boldsymbol{\beta})^2$$

$$-\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\partial log[p(\boldsymbol{\beta} \mid \boldsymbol{z}_m, \boldsymbol{Y})]}{\partial\boldsymbol{\beta}}\right]^2$$

where $z_1, z_2, ..., z_M$ are drawn from $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}^*)$.