The EM Algorithm

Further Statistical Methods, Lecture 6
Hilary Term 2004

Steffen Lauritzen, University of Oxford; February 7, 2005

The EM algorithm

The EM algorithm is an alternative to Newton–Raphson or the method of scoring for computing MLE in cases where the complications in calculating the MLE are due to *incomplete observation* and data are *MAR*, missing at random, with *separate parameters* for observation and the missing data mechanism, so the missing data mechanism can be ignored.

Data (X, Y) are the *complete data* whereas only *incomplete data* $Y = y$ are observed. (Rubin uses $Y = Y_{\mathrm{obs}}$ and $X = Y_{\mathrm{mis}}$).

The *complete data log-likelihood* is:

$$l(\theta) = \log L(\theta; x, y) = \log f(x, y; \theta).$$

The *marginal log-likelihood* or *incomplete data log-likelihood* is based on y alone and is equal to

$$l_y(\theta) = \log L(\theta; y) = \log f(y; \theta).$$

We wish to maximize l_y in θ but l_y is typically quite unpleasant:

$$l_y(\theta) = \log \int f(x, y; \theta)\, dx.$$

The EM algorithm is a method of maximizing the latter iteratively and alternates between two steps, one known as the *E-step* and one as the *M-step*, to be detailed below.

We let θ^* be and arbitrary but fixed value, typically the value of θ at the current iteration.

The E-step calculates the expected complete data log-likelihood ratio $q(\theta \mid \theta^*)$:

$$
\begin{aligned}
q(\theta \,|\, \theta^*) &= \mathbf{E}_{\theta^*}\left[\log \frac{f(X,y;\theta)}{f(X,y;\theta^*)} \,|\, Y = y\right] \\
&= \int \log \frac{f(x,y;\theta)}{f(x,y;\theta^*)} f(x \,|\, y;\theta^*)\, dx.
\end{aligned}
$$

The M-step maximizes $q(\theta \,|\, \theta^*)$ in θ for for fixed θ^*, i.e. calculates

$$\theta^{**} = \arg\max_{\theta} q(\theta \,|\, \theta^*).$$

We will show that *after an E-step and subsequent M-step, the likelihood function has never decreased.*

Kullback-Leibler divergence

The *KL divergence* between f and g is

$$KL(f:g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx.$$

Also known as *relative entropy* of g with respect to f.

Since $-\log x$ is a convex function, Jensen's inequality gives

$KL(f:g) \geq 0$ *and* $KL(f:g) = 0$ *if and only if* $f = g$,
since

$$KL(f:g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \geq -\log \int f(x) \frac{g(x)}{f(x)}\, dx = 0,$$

so KL divergence defines an (asymmetric) distance measure between probability distributions.

Expected and marginal log-likelihood

Since $f(x \mid y; \theta) = f\{(x, y); \theta\}/f(y; \theta)$ we have

$$\begin{aligned}
q(\theta \mid \theta^*) &= \int \log \frac{f(y;\theta) f(x \mid y; \theta)}{f(y;\theta^*) f(x \mid y; \theta^*)} f(x \mid y; \theta^*)\, dx \\
&= \log f(y;\theta) - \log f(y;\theta^*) \\
&\quad + \int \log \frac{f(x \mid y; \theta)}{f(x \mid y; \theta^*)} f(x \mid y; \theta^*)\, dx \\
&= l_y(\theta) - l_y(\theta^*) - KL(f^y_{\theta^*} : f^y_\theta).
\end{aligned}$$

Since the KL-divergence is minimized for $\theta = \theta^*$, differentiation of the above expression yields

$$\frac{\partial}{\partial \theta} q(\theta \mid \theta^*)\Big|_{\theta=\theta^*} = \frac{\partial}{\partial \theta} l_y(\theta)\Big|_{\theta=\theta^*}.$$

Let now $\theta_0 = \theta^*$ and define the iteration

$$\theta_{n+1} = \arg\max_{\theta} q(\theta \mid \theta_n).$$

Then

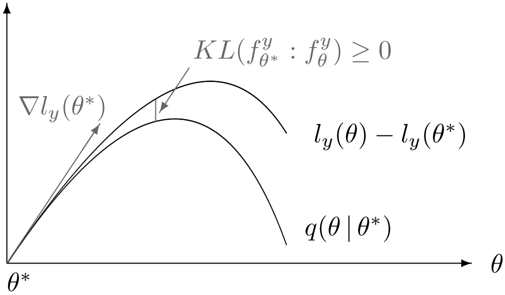
$$\begin{aligned} l_y(\theta_{n+1}) &= l_y(\theta_n) + q(\theta_{n+1} \mid \theta_n) + KL(f^y_{\theta_{n+1}} : f^y_{\theta_n}) \\ &\geq l_y(\theta_n) + 0 + 0. \end{aligned}$$

So the log-likelihood never decreases after a combined E-step and M-step.

It follows that any limit point must be a saddle point or a local maximum of the likelihood function.

The picture on the next overhead should show it all.

Expected and complete data likelihood


$KL(f^y_{\theta^*} : f^y_\theta) \geq 0$
$\nabla l_y(\theta^*)$
$l_y(\theta) - l_y(\theta^*)$
$q(\theta \,|\, \theta^*)$
θ
θ^*

$$l_y(\theta) - l_y(\theta^*) = q(\theta \,|\, \theta^*) + KL(f^y_{\theta^*} : f^y_\theta)$$

$$\nabla l_y(\theta^*) = \left. \frac{\partial}{\partial\theta} l_y(\theta) \right|_{\theta=\theta^*} = \left. \frac{\partial}{\partial\theta} q(\theta \,|\, \theta^*) \right|_{\theta=\theta^*}.$$

Mixtures

Consider a sample $Y = (Y_1, \dots, Y_n)$ from individual densities

$$f(y; \alpha, \mu) = \{\alpha\phi(y - \mu) + (1 - \alpha)\phi(y)\}$$

where ϕ is the normal density

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

and α and μ are both unknown, $0 < \alpha < 1$.

This corresponds to a fraction α of the observations being contaminated, or originating from a different population.

Incomplete observation

The likelihood function becomes

$$L_y(\alpha, \mu) = \prod_i \{\alpha\phi(y_i - \mu) + (1 - \alpha)\phi(y_i)\}$$

is quite unpleasant, although both Newton–Raphson and the method of scoring can be used.

But suppose we knew which observations came from which population?

In other words, let $X = (X_1, \ldots, X_n)$ be i.i.d. with $P(X_i = 1) = \alpha$ and suppose that the conditional distribution of Y_i given $X_i = 1$ was $\mathcal{N}(\mu, 1)$ whereas given $X_i = 0$ it was $\mathcal{N}(0, 1)$, i.e. that X_i was indicating whether Y_i was contaminated or not.

Then the marginal distribution of Y is precisely the mixture distribution and the 'complete data likelihood' is

$$\begin{aligned} L_{x,y}(\alpha,\mu) &= \prod_i \alpha^{x_i}\phi(y_i-\mu)^{x_i}(1-\alpha)^{1-x_i}\phi(y_i)^{1-x_i} \\ &\propto \alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}\prod_i \phi(y_i-\mu)^{x_i} \end{aligned}$$

so taking logarithms we get (ignoring a constant) that

$$\begin{aligned} l_{x,y}(\alpha,\mu) &= \sum x_i \log\alpha + \left(n-\sum x_i\right)\log(1-\alpha) \\ &\quad -\sum_i x_i(y_i-\mu)^2/2. \end{aligned}$$

If we did not know how to maximize this explicitly, differentiation easily leads to:

$$\hat{\alpha} = \sum x_i/n, \quad \hat{\mu} = \sum x_i y_i / \sum x_i.$$

Thus, when complete data are available the frequency of contaminated observations is estimated by the observed frequency and the mean μ of these is estimated by the average among the contaminated observations.

E-step and M-step

By taking expectations, we get the E-step as

$$
\begin{aligned}
q(\alpha, \mu \,|\, \alpha^*, \mu^*) &= \mathbf{E}_{\alpha^*,\mu^*}\{l_{X,y}(\alpha, \mu) \,|\, Y = y\} \\
&= \sum x_i^* \log \alpha + \left(n - \sum x_i^*\right) \log(1 - \alpha) \\
&\quad - \sum_i x_i^* (y_i - \mu)^2/2
\end{aligned}
$$

where

$$x_i^* = \mathbf{E}_{\alpha^*,\mu^*}(X_i \,|\, Y_i = y_i) = P_{\alpha^*,\mu^*}(X_i = 1 \,|\, Y_i = y_i).$$

Since this has the same form as the complete data likelihood, just with x_i^* replacing x_i, the M-step simply becomes

$$\alpha^{**} = \sum x_i^*/n, \quad \mu^{**} = \sum x_i^* y_i / \sum x_i^*,$$

i.e. here the mean of the contaminated observations is estimated by a weighted average of all the observations, the weight being proportional to the probability that this observation is contaminated. In effect, x_i^* act as *imputed values* of x_i.

The imputed values x_i^* needed in the E-step are calculated as follows:

$$\begin{aligned} x_i^* &= \mathbf{E}(X_i \,|\, Y_i = y_i) = P(X_i = 1 \,|\, Y_i = y_i) \\ &= \frac{\alpha^* \phi(y_i - \mu^*)}{\alpha^* \phi(y_i - \mu^*) + (1 - \alpha^*)\phi(y_i)}. \end{aligned}$$