The Method of Scoring.
The EM Algorithm

BS2 Statistical Inference, Lecture 9
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 8, 2004

The method of scoring

The iteration

$$\theta \leftarrow \theta + j_n(\theta)^{-1} S(\theta)$$

has a tendency to be unstable for many reasons, one of them being that $j_n(\theta)$ may be negative unless θ already is very close to to the MLE $\hat{\theta}$. In addition, $j(\theta)$ might sometimes be hard to calculate.

R. A. Fisher introduced the *method of scoring* which simply replaces the observed second derivative with its expectation to yield the iteration

$$\theta \leftarrow \theta + i_n(\theta)^{-1} S(\theta)$$

which in the case of independent and identically distributed observations gives

$$\theta \leftarrow \theta + i(\theta)^{-1} S(\theta)/n.$$

In many cases, $i(\theta)$ is easier to calculate and $i(\theta)$ is always positive.

In canonical exponential families we get

$$j(\theta) = \frac{\partial^2}{\partial \theta^2}\{c(\theta) - \theta t(X)\} = c''(\theta) = i(\theta)$$

so *for canonical exponential families the method of scoring and the method of Newton–Raphson coincide.*

If we let $v(\theta) = c''(\theta)$ the iteration becomes

$$\theta \leftarrow \theta + v(\theta)^{-1} S(\theta)/n.$$

The identity of Newton–Raphson and the method of scoring *only holds for the canonical parameter.* If $\theta = g(\mu)$

$$\begin{aligned} j(\mu) &= \frac{\partial^2}{\partial \mu^2}\{c(g(\mu)) - g(\mu)t(X)\} \\ &= \frac{\partial}{\partial \mu}\left[g'(\mu)\tau\{g(\mu)\} - g'(\mu)t(X)\right] \\ &= v\{g(\mu)\}\{g'(\mu)\}^2 + g''(\mu)\left[\tau\{g(\mu)\} - t(X)\right]. \end{aligned}$$

The method of scoring is simpler because the last term has expectation equal to 0:

$$i(\mu) = \mathbf{E}\{j(\mu)\} = v\{g(\mu)\}\{g'(\mu)\}^2.$$

The method of scoring is used in the `glim` procedure for estimation in so-called *generalised linear models*.

The EM algorithm

The EM algorithm is a supplement or alternative to Newton–Raphson in cases where the complications in calculating the MLE are due to *incomplete observation.*

Data (X, Y) are the *complete data* whereas only *incomplete data* $Y = y$ are observed.

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 $q(\theta \mid \theta^*)$:

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

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

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

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

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

Then

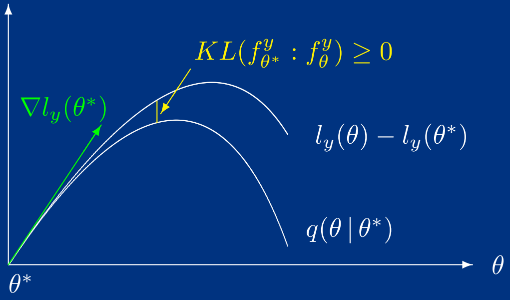
$$\begin{aligned} l_y(\theta_{n+1}) &= l_y(\theta_n) + q(\theta_{n+1} \,|\, \theta_n) + KL(f^y_{\theta_{n+1}} : f^y_{\theta_n}) \\ &\geq l_y(\theta_n) + 0 + 0. \end{aligned}$$

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



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

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