The EM Algorithm — Example

BS2 Statistical Inference, Lecture 10
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 12, 2004

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 of each observation being proportional to the probability that this observation is contaminated.

We now just need to know how to calculate the weights x_i^* needed in the E-step. But

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

so this is not difficult.

And now a live software-demonstration...

Exponential families

This example is typical for exponential families. For if the complete data likelihood is

$$l_{x,y}(\theta) = a(\theta)^\top t(x,y) - c(\theta)$$

then the E-step is determined as

$$q(\theta \mid \theta^*) = \theta^\top t^* - c(\theta)$$

where

$$t^* = \mathbf{E}_{\theta^*}\{t(X,Y) \mid Y = y\}.$$

The M-step will then be made in the same way as the maximization of the complete data likelihood, just with $t(x,y)$ replaced by t^*.

When the family is a linear canonical family, so that we can take $a(\theta) = \theta$, the M-step solves the equation

$$\mathbf{E}_\theta\{t(X,Y)\} = t^*.$$

If we let $T = t(X,Y)$ this equation can also be written as

$$\mathbf{E}_\theta(T) = \mathbf{E}_{\theta^*}(T \mid Y = y).$$

The E-step then calculates the right-hand side and the M-step solves the equation for θ.