EM Algorithm

(Dempster, Laird, and Rubin, 1977)

An approach for finding MLEs and posterior modes.

Based on decomposing data into observed and missing parts.

The missing data might be real, a theoretical construct, or both.

Let Y be the observed data and X be the unobserved, complete data.

In general there is a function $t(X) = Y$ that collapses the complete data X onto Y.

Often $X = (Y, Z)$, where Z is the missing data

Assume that X has density $f(X|\theta)$, and Y has density $g(Y|\theta)$. When choosing X you need to set it up such that

$$g(Y|\theta) = \int_{t(X)=Y} f(X|\theta)dX$$

Problem: Find

$$\hat{\theta} = \arg \sup g\left(y \,|\, \theta\right).$$

Assume that this is tough to do.

Idea: Pick X such that $f\left(X \,|\, \theta\right)$ is easy to maximize.

Can't deal with $f\left(X \,|\, \theta\right)$ exactly, since X can't be known with certainty.

Instead we want to deal with an expectation involving it.

The EM algorithm gives a sequence of estimates $\theta_0, \theta_1, \theta_2, \ldots$ by iterating the following 2 steps.

E-step: Calculate

$$Q\left(\theta \,|\, \theta_n\right) = E\left[\log f\left(X \,|\, \theta\right) \,|\, Y, \theta_n\right],$$

the conditional expectation of the complete data log likelihood.

M-step: Set

$$\theta_{n+1} = \arg \sup Q\left(\theta \,|\, \theta_n\right)$$

This scheme has the property that the sequence of estimators increases the observed data likelihood $g(Y|\theta)$. (To be made more precise later)

Example: Linkage Analysis (Rao, 1973, pp 368-369, Feb 4th lecture))

Phenotype	Probability	Counts	Y
ab	$\lambda/4$	34	y_1
Ab	$(1-\lambda)/4$	18	y_2
aB	$(1-\lambda)/4$	20	y_3
AB	$(2+\lambda)/4$	125	y_4

$$(Y_1, Y_2, Y_3, Y_4) \sim \text{Multi}\left(197, \left(\frac{\lambda}{4}, \frac{1-\lambda}{4}, \frac{1-\lambda}{4}, \frac{2+\lambda}{4}\right)\right)$$

The likelihood and log likelihood functions are

$$g(Y|\lambda) = \left(\frac{\lambda}{4}\right)^{Y_1} \left(\frac{1-\lambda}{4}\right)^{Y_2+Y_3} \left(\frac{2+\lambda}{4}\right)^{Y_4}$$

$$\log g(Y|\lambda) = Y_1 \log \lambda + (Y_2 + Y_3) \log(1-\lambda)$$
$$+ Y_4 \log(2+\lambda) - 197 \log 4$$

As we've seen, this needs some work to maximize.

Let $X = (X_1, X_2, X_3, X_4, X_5)$ such that

$$(X_1, X_2, X_3, X_4, X_5)$$
$$\sim \text{Multi}\left(197, \left(\frac{\lambda}{4}, \frac{1-\lambda}{4}, \frac{1-\lambda}{4}, \frac{\lambda}{4}, \frac{1}{2}\right)\right)$$

and $X_1 = Y_1, X_2 = Y_2, X_3 = Y_3$.

So Y_4 is being split into 2 groups.

Notice that for this problem X_4 and X_5 don't have any particular meaning. It's a theoretical construct set up to make things easy to deal with.

Its also a situation where X isn't of the form (Y, Z), though it could be extended to that setup.

With X, it is easy to solve for λ. With this data

$$\hat{\lambda} = \frac{X_1 + X_4}{X_1 + X_2 + X_3 + X_4}$$

as

$$\log f(X|\lambda) = (X_1 + X_4)\log \lambda + (X_2 + X_3)\log(1-\lambda)$$
$$- X_5 \log 2 - 197 \log 4$$

Another way of getting this is based on

$$X_1 + X_4 \,|\, X_1 + X_2 + X_3 + X_4 = n \sim \text{Bin}(n, \lambda)$$

E-step:

$$Q(\lambda|\lambda_n)$$
$$= E\left[(X_1 + X_4)\log \lambda + (X_2 + X_3)\log(1-\lambda)\,|\,Y, \lambda_n\right]$$

Since most of the components of X are fixed given Y, this reduces to

$$Q(\lambda|\lambda_n) = Y_1 \log \lambda + (Y_2 + Y_3)\log(1-\lambda)$$
$$+ E\left[X_4 \log \lambda \,|\, Y, \lambda_n\right]$$
$$= Y_1 \log \lambda + (Y_2 + Y_3)\log(1-\lambda)$$
$$+ E\left[X_4 \log \lambda \,|\, Y_4, \lambda_n\right]$$
$$= \left(Y_1 + \hat{X}_4\right)\log \lambda + (Y_2 + Y_3)\log(1-\lambda)$$

where $\hat{X}_4 = E\left[X_4 \,|\, Y_4, \lambda_n\right]$.

Now $X_4 \,|\, Y_4 \sim \text{Bin}\left(Y_4, \dfrac{\lambda}{\lambda+2}\right)$ so

$$\hat{X}_4 = E\left[X_4 \,|\, Y_4, \lambda_n\right]$$

$$= Y_4 \frac{\lambda_n}{\lambda_n + 2}$$

M-step:

$$\hat{\lambda}_{n+1} = \frac{Y_1 + \hat{X}_4}{Y_1 + Y_2 + Y_3 + \hat{X}_4}$$

Iteration	λ_n	$\log g(\lambda_n)$
0	0.5	64.6297445
1	0.608247423	67.3201705
2	0.624321050	67.3829250
3	0.626488879	67.3840812
4	0.626777322	67.3841017
5	0.626815632	67.3841021
6	0.626820719	67.3841021
7	0.626821394	67.3841021

Notice that the observed data log likelihood increases at each step.

The above run was based on the convergence criteria of $\left|\lambda_{n+1} - \lambda_n\right| < 10^{-6}$

Example: Multivariate Normal with missing data

Complete Data:

$$X_i \sim N_k(\mu, V); \quad i = 1, \ldots, n$$

$$X_i^T = (X_{i1}, \ldots, X_{ik})$$

$$\log f(X_i|\theta) = -\frac{1}{2}\log\det V - \frac{1}{2}(X_i - \mu)^T V^{-1}(X_i - \mu)$$

$$= -\frac{1}{2}\log\det V$$

$$\quad -\frac{1}{2}\operatorname{trace}\left[V^{-1}(X_i - \mu)^T(X_i - \mu)\right]$$

$$= -\frac{1}{2}\log\det V$$

$$\quad -\frac{1}{2}\operatorname{trace}\left[V^{-1}(X_i X_i^T - 2\mu X^T + \mu\mu^T)\right]$$

So a set of sufficient statistics for μ and V are

$$\sum_{i=1}^n X_i \text{ and } \sum_{i=1}^n X_i X_i^T.$$

For the complete data set up

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$$

and

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - \hat{\mu}\hat{\mu}^T$$

Missing Data:

Assume that components of X_i are missing at random. So the missing data pattern for each vector could be arbitrary.

For example, $Y_1 = X_1$, $Y_2 = \left(X_{21}, X_{23}, X_{25}, \ldots X_{2k} \right)^T$ (with $Z_2 = \left(X_{22}, X_{24} \right)^T$) and so on.

While each Y_i is multivariate normal, the parameterization is potentially different for each observation, so you can't directly get the MLE.

However it can be done quite easily with EM

E-step:

As the complete data log likelihood is a linear function of the sufficient statistics, the E-step involves calculating

$$E\left[\sum_{i=1}^{n} X_i \middle| Y, \mu_n, V_n\right] \text{ and } E\left[\sum_{i=1}^{n} X_i X_i^T \middle| Y, \mu_n, V_n\right]$$

If the observations are independent, the problem reduces to calculating

$$\hat{X}_i^{(n)} = E\left[X_i \middle| Y_i, \mu_n, V_n\right]$$

and

$$\hat{S}_i^{(n)} = E\left[X_i X_i^T \middle| Y_i, \mu_n, V_n\right]$$

for each observation. (How to do it to come)

M-step:

$$\hat{\mu}_{n+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i^{(n)}$$

and

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i^{(n)} - \hat{\mu}_{n+1} \hat{\mu}_{n+1}^T$$

How to get $X_i^{(n)}$ and $\hat{S}_i^{(n)}$

$$X_i = P_i^T \begin{bmatrix} Z_i \\ Y_i \end{bmatrix}$$

where P_i^T is a square matrix which permutes the rows into the correct order. ($P_i^T = P_i^{-1}$).

For multivariate normals

$$E\left[Z_i \middle| Y_i\right] = \mu_z + V_{ZY}V_Y^{-1}\left(Y_i - \mu_Y\right) = \mu_{Z|Y,i}$$
$$E\left[Y_i \middle| Y_i\right] = Y_i$$

So

$$X_i^{(n)} = P_i^T \begin{bmatrix} \mu_{Z|Y,i} \\ Y_i \end{bmatrix}$$

To get $\hat{S}_i^{(n)}$, we'll use the fact that

$$E\left[XX^T\right] = \mathrm{Var}\left(X\right) + \mu_X \mu_X^T$$

First

$$\mathrm{Var}\left(Z_i \middle| Y_i\right) = V_Z - V_{ZY}V_Y^{-1}V_{YZ} = V_{Z|Y,i}$$
$$\mathrm{Var}\left(Y_i \middle| Y_i\right) = 0$$
$$\mathrm{Cov}\left(Z_i, Y_i \middle| Y_i\right) = 0$$

Then

$$\text{Var}\left(X_i \mid Y_i\right) = P_i^T \begin{bmatrix} V_{Z|Y,i} & 0 \\ 0 & 0 \end{bmatrix} P_i = V_{X|Y,i}$$

So

$$S_i^{(n)} = V_{X|Y,i} + X_i^{(n)} X_i^{(n)^T}$$

As can be seen from this example, EM doesn't just fill in missing parts of X with their expectation, i.e,

$$Q\left(\theta \mid \theta_n\right) \neq \log f\left(E\left[X \mid Y, \theta_n\right] \mid \theta\right)$$

Instead, when calculating $Q\left(\theta \mid \theta_n\right)$ you need to calculate expectations of functions of the sufficient statistics.

When the distribution of X comes from the exponential family, the problem reduces calculating the conditional expectation of the sufficient statistics since

$$Q\left(\theta \mid \theta_n\right) = E\left[\beta(\theta) + h(X)^T \gamma(\theta) \mid Y, \theta_n\right]$$

$$= \beta(\theta) + E\left[h(X)^T \mid Y, \theta_n\right]\gamma(\theta)$$

up to an additive constant (which doesn't affect the optimization.

This exactly what was done in the multivariate normal example ($h(X)^T = \left[\sum X_i \quad \sum X_i X_i^T \right]$)

So in addition, when choosing X, the complete data, you also need to think of situations where you can calculate the conditional expectations in addition to whether the likelihood is easy to optimize.

Optimality properties of EM

Theorem

$$g\left(Y \middle| \theta_{n+1}\right) \geq g\left(Y \middle| \theta_n\right)$$

or equivalently

$$\log g\left(Y \middle| \theta_{n+1}\right) \geq \log g\left(Y \middle| \theta_n\right)$$

Proof:

For simplicity, lets assume that X can be decomposed into (Y, Z), the observed and missing parts. The proofs go through without this assumption, but they aren't quite as intuitive (technical note, in Sec 10.3.1).

$$f\left(X|\theta\right) = g\left(Y|\theta\right)h\left(Z|Y,\theta\right)$$

$$\log f\left(X|\theta\right) = \log g\left(Y|\theta\right) + \log h\left(Z|Y,\theta\right)$$

$$\log g\left(Y|\theta\right) = \log f\left(X|\theta\right) - \log h\left(Z|Y,\theta\right)$$

by taking expectations of both sides of the third line with respect to Y and θ_n, we get

$$\log g\left(Y|\theta\right) = Q\left(\theta|\theta_n\right) - H\left(\theta|\theta_n\right)$$

where

$$H\left(\theta|\theta_n\right) = \int \log h\left(Z|Y,\theta\right)h\left(Z|Y,\theta_n\right)dZ$$

$$= E\left[\log h\left(Z|Y,\theta\right)|Y,\theta_n\right]$$

Then

$$\log g\left(Y|\theta_{n+1}\right) - \log g\left(Y|\theta_n\right)$$

$$= \underbrace{\left[Q\left(\theta_{n+1}|\theta_n\right) - Q\left(\theta_n|\theta_n\right)\right]}_{\geq 0}$$

$$\underbrace{-\left[H\left(\theta_{n+1}|\theta_n\right) - H\left(\theta_n|\theta_n\right)\right]}_{\leq 0}$$

$$\geq 0$$

Thus $\log g\left(Y|\theta_{n+1}\right) \geq \log g\left(Y|\theta_n\right)$

Jensen's inequality:

Let W be a random variable. If $h(w)$ is a convex function on the range of W, then

$$E\big[h(W)\big] \geq h\big(E[W]\big)$$

assuming both expectations exist. For a strictly convex function, equality holds *iff* $W = E[W]$ almost surely.

Lemma (Prop 10.3.2):

$$H(\theta'|\theta) \leq H(\theta|\theta)$$

Proof

$$H(\theta|\theta) - H(\theta'|\theta)$$

$$= \int \Big[\log h(Z|Y,\theta) - \log h(Z|Y,\theta')\Big]$$

$$\times h(Z|Y,\theta)\,dZ$$

$$= -\int \left[\log \frac{h(Z|Y,\theta')}{h(Z|Y,\theta)}\right] h(Z|Y,\theta)\,dZ$$

$$\geq -\log\left[\int \frac{h(Z|Y,\theta')}{h(Z|Y,\theta)} h(Z|Y,\theta)\,dZ\right]$$

$$= -\log\left[\int h(Z|Y,\theta')\,dZ\right] = 0$$

Generalized EM (GEM):

In the M-step, you don't actually have to maximize the Q function at each step.

What is needed is to choose a value θ_{n+1} such that

$$Q\left(\theta_{n+1}\,|\,\theta_n\right) \geq Q\left(\theta_n\,|\,\theta_n\right).$$

Since this relationship was all that was used in the earlier proof, any GEM will increase the likelihood.

So the assumption that X has to be easy to maximize can be relaxed and leads to extensions to EM, some of which are discussed in Chapter 12 of Lange.

Corollary to increasing likelihood theorem

If the sequence $\left\{g\left(Y\,|\,\theta_n\right)\right\}$ is bounded above then it will converge to some value g^*.

So this implies that EM (or a GEM) converges to something.

It doesn't imply that θ_n to an optima of $g\left(Y\,|\,\theta_n\right)$.

You need a bit more to do that.

Note that the proof of this in Dempster, Laird and Rubin was in error. Wu (1983) finds conditions which do imply what θ_n converges to.

Theorem: Under some regularity conditions (see Wu, 1983), for any EM sequence $\{\theta_n\}$,

$$\log g\left(Y\,|\,\theta_{n+1}\right) > \log g\left(Y\,|\,\theta_n\right)$$

if

$$\theta_n \notin \Gamma = \left\{\theta : D\log g\left(Y\,|\,\theta\right) = 0\right\}$$

Proof:

$$D\log g\left(Y\,|\,\theta\right) = D^{10}Q\left(\theta\,|\,\theta\right)$$

where D^{10} indicates taking partial derivatives with respect to the first θ.

This comes from

$$\log g\left(Y\,|\,\theta\right) = Q\left(\theta\,|\,\theta\right) - H\left(\theta\,|\,\theta\right)$$

and $D^{10}H\left(\theta\,|\,\theta\right) = 0$ since $H\left(\theta\,|\,\theta\right) \geq H\left(\theta'\,|\,\theta\right)$

So if $Q\left(\theta_{n+1}\,|\,\theta_n\right) > Q\left(\theta_n\,|\,\theta_n\right)$, then

$$\log g\left(Y\,|\,\theta_{n+1}\right) > \log g\left(Y\,|\,\theta_n\right)$$

which holds for points in Γ^c.

This theorem then implies that any limit point of an EM sequence must be a stationary point of $\log g\left(Y\,|\,\theta_n\right)$.

Thus a sequence $\left\{\theta_n\right\}$ must converge to a local maximum or saddle point of $\log g\left(Y\,|\,\theta_n\right)$.