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

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

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

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

Another way of getting this is based on

$$X_1+X_4\left|X_1+X_2+X_3+X_4=n\right.\sim\mathrm{Bin}\left(n,\lambda\right)$$

E-step:

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

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

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

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

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

$$\begin{aligned}\hat{X}_4 &= E\left[X_4 \middle| Y_4, \lambda_n\right] \\ &= Y_4 \frac{\lambda_n}{\lambda_n + 2}\end{aligned}$$

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:

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

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

$$\begin{aligned}\log f\left(X_i \middle| \theta\right) &= -\frac{1}{2}\log\det V - \frac{1}{2}\left(X_i - \mu\right)^T V^{-1}\left(X_i - \mu\right) \\ &= -\frac{1}{2}\log\det V \\ &\qquad -\frac{1}{2}\text{trace}\left[V^{-1}\left(X_i - \mu\right)^T\left(X_i - \mu\right)\right] \\ &= -\frac{1}{2}\log\det V \\ &\qquad -\frac{1}{2}\text{trace}\left[V^{-1}\left(X_i X_i^T - 2\mu X^T + \mu\mu^T\right)\right]\end{aligned}$$

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

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

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, $Y_1 = X_1$, $Y_2 = (X_{21}, X_{23}, X_{25}, \ldots X_{2k})^T$ (with $Z_2 = (X_{22}, X_{24})^T$) 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

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

If the observations are independent, the problem reduces to calculating

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

and

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

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

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

$$E\left[Y_i \left| Y_i\right.\right] = Y_i$$

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

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

First

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

$$\text{Var}\left(Y_i \left| Y_i\right.\right) = 0$$

$$\text{Cov}\left(Z_i, Y_i \left| Y_i\right.\right) = 0$$

Then

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

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,

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

Instead, when calculating $Q\left(\theta \middle| \theta_n\right)$ 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

$$\begin{aligned} Q\left(\theta \middle| \theta_n\right) &= E\left[\beta(\theta) + h(X)^T \gamma(\theta) \middle| Y, \theta_n\right] \\ &= \beta(\theta) + E\left[h(X)^T \middle| Y, \theta_n\right] \gamma(\theta) \end{aligned}$$

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

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

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 $\left(Y, Z\right)$, 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).

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

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

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

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

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

where

$$\begin{aligned}H\left(\theta\middle|\theta_n\right)&=\int\log h\left(Z\middle|Y,\theta\right)h\left(Z\middle|Y,\theta_n\right)dZ\\&=E\left[\log h\left(Z\middle|Y,\theta\right)\middle|Y,\theta_n\right]\end{aligned}$$

Then

$$\begin{aligned}&\log g\left(Y\middle|\theta_{n+1}\right)-\log g\left(Y\middle|\theta_n\right)\\&\quad=\underbrace{\left[Q\left(\theta_{n+1}\middle|\theta_n\right)-Q\left(\theta_n\middle|\theta_n\right)\right]}_{\geq 0}\\&\qquad\qquad-\underbrace{\left[H\left(\theta_{n+1}\middle|\theta_n\right)-H\left(\theta_n\middle|\theta_n\right)\right]}_{\leq 0}\\&\quad\geq 0\end{aligned}$$

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

Jensen's inequality:

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

$$E\left[h(W)\right] \geq h\left(E[W]\right)$$

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

Lemma (Prop 10.3.2):

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

Proof

$$\begin{aligned}
&H\left(\theta|\theta\right) - H\left(\theta'|\theta\right) \\
&\quad = \int \left[\log h\left(Z|\mathrm{Y},\theta\right) - \log h\left(Z|Y,\theta'\right)\right] \\
&\qquad\qquad\qquad \times h\left(Z|\mathrm{Y},\theta\right) dZ \\
&\quad = -\int \left[\log \frac{h\left(Z|\mathrm{Y},\theta'\right)}{h\left(Z|\mathrm{Y},\theta\right)}\right] h\left(Z|\mathrm{Y},\theta\right) dZ \\
&\quad \geq -\log\left[\int \frac{h\left(Z|\mathrm{Y},\theta'\right)}{h\left(Z|\mathrm{Y},\theta\right)} h\left(Z|\mathrm{Y},\theta\right) dZ\right] \\
&\quad = -\log\left[\int h\left(Z|Y,\theta'\right) dZ\right] = 0
\end{aligned}$$

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

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

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 $\left\{g\left(Y\left|\theta_n\right.\right)\right\}$ 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 $g\left(Y\left|\theta_n\right.\right)$.

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 $\{\theta_n\}$,

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

if

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

Proof:

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

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

This comes from

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

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

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

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

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 $\log g\left(Y\left|\theta_n\right.\right)$.

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

EM Algorithm Extensions

ECM (Meng and Rubin, 1993)

(Expectation Conditional Maximization)

Idea: Suppose that $\theta = \left(\theta_1, \theta_2, \ldots, \theta_k\right)$ and that optimizing $Q\left(\theta \middle| \theta^{(n)}\right)$ isn't easy. However suppose that

$$
\begin{gathered}
Q\left(\theta_1, \theta_2^{(n)}, \theta_3^{(n)}, \ldots, \theta_k^{(n)} \middle| \theta^{(n)}\right) \\
Q\left(\theta_1^{(n)}, \theta_2, \theta_3^{(n)}, \ldots, \theta_k^{(n)} \middle| \theta^{(n)}\right) \\
\vdots \\
Q\left(\theta_1^{(n)}, \theta_2^{(n)}, \theta_3^{(n)}, \ldots, \theta_k \middle| \theta^{(n)}\right)
\end{gathered}
$$

are all easy to maximize.

Note in the above θ_j may be a vector of parameters.

Then the basic ECM algorithm modifies the M-step as follows

M$_1$: Given $\theta_2 = \theta_2^{(n)}$, $\theta_3 = \theta_3^{(n)}$, … , $\theta_k = \theta_k^{(n)}$ find the value of θ_1, $\theta_1^{(n+1)}$, that maximizes $Q\left(\theta_1, \theta_2^{(n)}, \theta_3^{(n)}, \ldots, \theta_k^{(n)} \middle| \theta^{(n)}\right)$

M$_2$: Given $\theta_1 = \theta_1^{(n+1)}$, $\theta_3 = \theta_3^{(n)}$, … , $\theta_k = \theta_k^{(n)}$ find the value of θ_2, $\theta_2^{(n+1)}$, that maximizes $Q\left(\theta_1^{(n+1)}, \theta_2, \theta_3^{(n)}, \ldots, \theta_k^{(n)} \middle| \theta^{(n)}\right)$

…

M$_k$: Given $\theta_1 = \theta_1^{(n+1)}$, $\theta_2 = \theta_2^{(n+1)}$, … , $\theta_{k-1} = \theta_{k-1}^{(n+1)}$ find the value of θ_k, $\theta_k^{(n+1)}$, that maximizes $Q\left(\theta_1^{(n+1)}, \theta_2^{(n+1)}, \ldots, \theta_{k-1}^{(n+1)}, \theta_k \middle| \theta^{(n)}\right)$

So step through and maximize each piece separately.

This procedure is a GEM since

$$\begin{aligned} Q\left(\theta^{(n+1)}\,\middle|\,\theta^{(n)}\right) &\geq Q\left(\theta_1^{(n+1)},\theta_2^{(n+1)},\ldots,\theta_{k-1}^{(n+1)},\theta_k^{(n)}\,\middle|\,\theta^{(n)}\right) \\ &\geq Q\left(\theta_1^{(n+1)},\theta_2^{(n+1)},\ldots,\theta_{k-1}^{(n)},\theta_k^{(n)}\,\middle|\,\theta^{(n)}\right) \\ &\geq \ldots \geq Q\left(\theta_1^{(n+1)},\theta_2^{(n)},\ldots,\theta_k^{(n)}\,\middle|\,\theta^{(n)}\right) \\ &\geq Q\left(\theta^{(n)}\,\middle|\,\theta^{(n)}\right) \end{aligned}$$

So all the nice properties I talked about last time go through, (though you need to be slightly careful with the regularity conditions showing that ECM converges to a stationary point of the likelihood surface – see Meng and Rubin 1993)

Example: Multivariate normal regression with incomplete response data

Complete Data Model:

$$Y_i \sim N\left(X_i\beta, V\right); \quad i = 1,\ldots,m$$

where X_i is a $k \times p$ matrix of covariates, β is a $p \times 1$ vector of unknown parameters, and V is a positive definite covariance matrix ($k\left(k+1\right)/2$ unknown parameters)

Missing Data:

Components of Y_i are missing at random (similar to example from last time)

Let S_i be a matrix on ones and zeros which indicates which observations have been observed (e.g. S_iY_i is the vector of observed components)

E-step:

$$\hat{Y}_i^{(n)} = E\left[Y_i \left| S_i Y_i, \beta_n, V_n \right. \right]$$

and

$$\hat{W}_i^{(n)} = E\left[Y_i Y_i^T \left| S_i Y_i, \beta_n, V_n \right. \right]$$

M-step: maximize

$$-\frac{m}{2}\log|V| - \frac{1}{2}\text{trace}\left[V^{-1} \sum_i \left(\hat{W}_i^{(n)} - \hat{Y}_i^{(n)} \hat{Y}_i^{(n)^T} \right) \right]$$
$$-\frac{1}{2}\sum_i \left(\hat{Y}_i^{(n)} - X_i\beta \right)^T V^{-1} \left(\hat{Y}_i^{(n)} - X_i\beta \right)$$

M_1:

$$\beta^{(n+1)} = \left(\sum_i X_i^T V^{(n)^{-1}} X_i \right)^T \left(\sum_i X_i^T V^{(n)^{-1}} \hat{Y}_i^{(n)} \right)$$

M_2:

$$V^{(n+1)} = \frac{1}{m} \sum_i (\hat{W}_i^{(n)} - \hat{Y}_i^{(n)} \left(X_i \beta^{(n+1)} \right)^T - X_i \beta^{(n+1)} \hat{Y}_i^{(n)^T} \\ + X_i \beta^{(n+1)} \beta^{(n+1)^T} X_i^T)$$

Analogies with other procedures:

Iterative Proportional Fitting (IPF):

> Approach for fitting log linear models for contingency tables when there are no closed form solutions. Actually this is a special case of ECM (Lange, section 12.2).

Gibbs sampler:

> Draw θ_j from $\left[\theta_j \middle| \Theta_{-j} \right]$ where $\Theta_{-j} = \left\{ \theta_i : i \neq j \right\}$

Iterative Conditional Modes (ICM) (Besag, 1986):

> Iteratively maximize components of the posterior distribution (or the likelihood function)

Variations:

Additional E-steps can be mixed into the series of M-steps. For example, if $k = 2$, a modified ECM scheme could be

$$(E - M_1 - E - M_2) - (E - M_1 - E - M_2)$$

instead of

$$(E - M_1 - M_2) - (E - M_1 - M_2)$$

Another modification is to skip E-steps, giving for example,

$$(E - M_1 - M_2 - M_1 - M_2) - (E - M_1 - M_2 - M_1 - M_2)$$

Note that this sort of scheme usually isn’t particularly advantageous, though if calculating the E-step is slow, this can lead to speed ups.

EM Gradient Algorithm

Even with careful thinking, the M-step may not be feasible, even with extensions like ECM.

As all that is really needed is a GEM, what we really need is an approximation to the maximizer.

One approach for doing this is one Newton-Raphson step on Q. This given

Gradient M-step: Set

$$\begin{aligned}\theta_{n+1} &= \theta_n - d^{20}Q\left(\theta_n \left|\theta_n\right.\right)^{-1} d^{10}Q\left(\theta_n \left|\theta_n\right.\right)^T \\ &= \theta_n - d^{20}Q\left(\theta_n \left|\theta_n\right.\right)^{-1} dL\left(\theta_n\right)^T\end{aligned}$$

The second form holds since as shown last time

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

Since NR isn't a ascent algorithm, you need to watch things a bit, but it is possible to show than when you get close to $\hat{\theta}$, the EM gradient algorithm satisfies the ascent condition $L\left(\theta_{n+1}\right) \geq L\left(\theta_n\right)$.

This idea can also be combined with ECM, e.g., run EM Gradient on a couple of the $\theta_j^{(n)}$'s and regular ECM on the rest.

Another advantage to this combination, is that NR often works better on smaller parameter spaces (more likely to have an ascent algorithm)

Note that this idea can be used with regular NR. There is nothing special about doing it on the Q function.

Bayesian EM

Let $\pi(\theta)$ be the prior distribution on the parameter θ. Then the posterior density

$$\pi(\theta | Y) \propto g(Y | \theta)\pi(\theta)$$

So finding the posterior mode is equivalent to maximizing

$$\log g(Y | \theta) + \log \pi(\theta)$$

Assuming that a nice complete data model $f(X | \theta)$ can be found, the Bayesian version of EM involves

Bayesian E-step:

$$\begin{aligned} Q\left(\theta\middle|\theta_n\right) &= E\left[\log f\left(X\middle|\theta\right)+\log\pi\left(\theta\right)\middle|Y,\theta_n\right] \\ &= E\left[\log f\left(X\middle|\theta\right)\middle|Y,\theta_n\right]+\log\pi\left(\theta\right) \end{aligned}$$

Bayesian M-step: Set

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

By similar arguments as for basic EM, the sequence $\left\{\theta_n\right\}$ leads to an increasing sequence of the log posteriors, converges to a stationary point of the log posterior, etc.

One potential problem is that the log prior often complicates the M-step.

Usually things only work nicely when the prior is conjugate to the complete data model.

A prior is conjugate if the posterior distribution is a member of the same family of distributions as the prior

Example:

Complete data: $X \sim \text{Bin}(n, p)$

$$f\left(x \middle| p\right) = \binom{n}{x} p^{x} \left(1-p\right)^{n-x}$$

Prior: $p \sim \text{Beta}(\alpha, \beta)$

$$\pi\left(p\right) \propto p^{\alpha-1}\left(1-p\right)^{\beta-1}$$

Posterior:

$$\pi\left(p \middle| x\right) \propto p^{x+\alpha-1}\left(1-p\right)^{n-x+\beta-1}$$

$$p \middle| x \sim \text{Beta}\left(x+\alpha, n-x+\beta\right)$$

E-step:

$$Q\left(p \middle| p_n\right) = E\Big[\left(x+a-1\right)\log p \\ +\left(n-x+\beta-1\right)\log\left(1-p\right)\Big|Y, p_n\Big]$$

So we need $E\left[X \middle| Y, p_n\right]$, where Y is the observed data.

M-step:

$$p_{n+1} = \frac{E\left[X\middle|Y, p_n\right] + \alpha - 1}{N + \alpha + \beta - 2}$$

Missing Information Principle

Remember from last time

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

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

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

This implies

$$-D^2 \log g\left(Y\middle|\theta\right) = -D^2 \log f\left(X\middle|\theta\right) - \left(-D^2 \log h\left(Z\middle|Y,\theta\right)\right)$$

Taking conditional expectations gives

$$I_O\left(\theta\middle|Y\right) = I_{OC}\left(\theta\middle|Y\right) - I_{OM}\left(\theta\middle|Y\right)$$

Observed Information

= Complete Information – Missing Information

$$I_{OC}\left(\theta \middle| Y\right) = -D^{20} Q\left(\theta \middle| \theta\right)$$

$$I_{OM}\left(\theta \middle| Y\right) = -D^{20} H\left(\theta \middle| \theta\right)$$

Convergence of EM

EM can be considered as an iterative update scheme where

$$\theta_{n+1} = M\left(\theta_n\right)$$

It has been shown (Dempster, Laird, and Rubin, 1977) that EM has linear convergence and that

$$\lim_{n \to \infty} \frac{\left\| \theta_{n+1} - \hat{\theta} \right\|_2}{\left\| \theta_n - \hat{\theta} \right\|_2} = \lambda$$

where λ is the largest eigenvalue of $DM\left(\hat{\theta}\right)$.

Note that the mapping $\theta_{n+1} = M\left(\theta_n\right)$ may be difficult to determine in a nice form so the Jacobian can be calculated. However, $DM\left(\hat{\theta}\right)$ can be tied in with the missing information principle as follows.

Theorem:

If $D^{10}Q\left(\theta_{n+1}\left|\theta_n\right.\right)=0$, then

$$DM\left(\hat{\theta}\right)=I_{OM}\left(\hat{\theta}\middle|Y\right)I_{OC}^{-1}\left(\hat{\theta}\middle|Y\right)$$

Proof:

$D^{10}Q\left(M\left(\theta\right)\middle|\theta\right)=0$

Applying the chain rule

$$DM\left(\theta\right)D^{20}Q\left(M\left(\theta\right)\middle|\theta\right)+D^{11}Q\left(M\left(\theta\right)\middle|\theta\right)=0$$

which implies

$$DM\left(\hat{\theta}\right)D^{20}Q\left(\hat{\theta}\middle|\hat{\theta}\right)+D^{11}Q\left(\hat{\theta}\middle|\hat{\theta}\right)=0 \qquad (*)$$

Then

$$\log g\left(Y\left|\theta\right.\right)=Q\left(\theta\middle|\theta'\right)-H\left(\theta\middle|\theta'\right)$$

implies

$$D^{11}Q\left(\theta\middle|\theta\right)=D^{11}H\left(\theta\middle|\theta\right)=-D^{20}H\left(\theta\middle|\theta\right)$$

So plugging this into (*) gives

$$DM\left(\hat{\theta}\right)D^{20}Q\left(\hat{\theta}\middle|\hat{\theta}\right)-D^{20}H\left(\hat{\theta}\middle|\hat{\theta}\right)=0$$

which then gives the result.

One way of thinking of this, particularly for the scalar parameter case, is the rate of convergence is the fraction of information that is missing.

This implies for fast convergence, you want $I_{OM}\left(\theta|Y\right)$ to be “small” and $I_{OC}\left(\theta|Y\right)$ to be “big”

So for the genetics example,

$$D^{20}Q\left(\lambda|\lambda'\right) = -\frac{E\left[X_4|y_4,\lambda'\right]+y_1}{\lambda'^2} - \frac{y_2+y_3}{\left(1-\lambda'\right)^2}$$

$$D^{20}H\left(\lambda|\lambda'\right) = -\frac{E\left[X_4|y_4,\lambda'\right]}{\lambda'^2} + \frac{y_4}{\left(2+\lambda'\right)^2}$$

Plugging in $\hat{\lambda}$ = 0.626821 gives

$$DM\left(\hat{\lambda}\right) = I_{OM}\left(\hat{\lambda}|Y\right) I_{OC}^{-1}\left(\hat{\lambda}|Y\right) = 0.132778$$

If we look at the sequence of iterations

Iteration	λ_n	$\lvert\lambda_n - \hat{\lambda}\rvert$	$\lvert\lambda_{n+1} - \hat{\lambda}\rvert / \lvert\lambda_n - \hat{\lambda}\rvert$
0	0.5	0.126821	0.1465
1	0.608247423	0.018574	0.1346
2	0.624321050	0.002500	0.1330
3	0.626488879	0.000333	0.1327
4	0.626777322	0.000044	0.1322
5	0.626815632	0.000006	0.1287
6	0.626820719		-
7	0.626821394		-

This doesn't quite match the table in DLR.

In the scalar parameter case

$$\begin{aligned}\mathrm{Var}\left(\hat{\theta}\middle|Y\right) &= \frac{\mathrm{Var}\left(\hat{\theta}\middle|X\right)}{1-\lambda} \\ &= \mathrm{Var}\left(\hat{\theta}\middle|X\right) + \frac{1}{1-\lambda}\mathrm{Var}\left(\hat{\theta}\middle|X\right)\end{aligned}$$

Calculating the information matrix

1) Calculate directly

$$I_O(\theta) = -D^2 \log g(Y|\theta)$$

Usually not feasible if you're forced to run EM instead of, say, Newton-Raphson.

2) Use the fact

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

and differentiate this.

3) Use the fact

$$\begin{aligned} I_O(\theta) &= -D^2 \log g(Y|\theta) \\ &= -D^{20} Q(\theta|\theta_0) + D^{20} H(\theta|\theta_0) \end{aligned}$$

Usually not useful for calculation purposes. When this can be used for getting $I_O(\theta)$, you can probably do 1) directly. This fact is more useful for doing proofs about EM.

4) Louis' formula (Louis, 1982)

$$
\begin{aligned}
I_O(\theta) &= -D^2 \log g(Y|\theta) \\
&= E\left[-D^2 \log f(X|\theta)|Y\right] \\
&\qquad - E\left[D\log f(X|\theta)\left(D\log f(X|\theta)\right)^T|Y\right] \\
&\qquad + E\left[D\log f(X|\theta)|Y\right]E\left[D\log f(X|\theta)|Y\right]^T
\end{aligned}
$$

This can be though of in terms of the missing information principle. The first term in the sum is the complete information and the last two terms are the missing information.

The second term in the sum might be a bit difficult as it will involve products of the sufficient statistics.

Note that the third term is 0 when evaluated at the MLE.

There is a simplification which sometimes helps as the last two terms are just

$$\begin{aligned}\mathrm{Var}\left[D\log f\left(X\middle|\theta\right)\middle|Y\right] &= E\left[D\log f\left(X\middle|\theta\right)\left(D\log f\left(X\middle|\theta\right)\right)^T\middle|Y\right] \\ &\quad - E\left[D\log f\left(X\middle|\theta\right)\middle|Y\right]E\left[D\log f\left(X\middle|\theta\right)\middle|Y\right]^T\end{aligned}$$

so Louis' formula is sometimes presented as

$$\begin{aligned}I_O\left(\theta\right) &= -D^2\log g\left(Y\middle|\theta\right) \\ &= E\left[-D^2\log f\left(X\middle|\theta\right)\middle|Y,\theta\right] - \mathrm{Var}\left(D\log f\left(X\middle|\theta\right)\middle|Y,\theta\right)\end{aligned}$$

Example: Genetics

$$\begin{aligned}\log f\left(X\middle|\lambda\right) &= \left(X_1+X_4\right)\log\lambda + \left(X_2+X_3\right)\log\left(1-\lambda\right) \\ &\quad + Y_5\log 2 - 197\log 4\end{aligned}$$

$$\begin{aligned}D\log f\left(X\middle|\lambda\right) &= \frac{X_1+X_4}{\lambda} - \frac{X_2+X_3}{1-\lambda} \\ D^2\log f\left(X\middle|\lambda\right) &= -\frac{X_1+X_4}{\lambda^2} - \frac{X_2+X_3}{\left(1-\lambda\right)^2}\end{aligned}$$

So

$$E\left[D^2 \log f\left(X|\lambda\right)|Y,\lambda\right] = -\frac{E\left[X_4\,|y_4,\lambda\right]+y_1}{\lambda^2} - \frac{y_2+y_3}{\left(1-\lambda\right)^2}$$

$$= \frac{y_4\,\frac{\lambda}{2+\lambda}+y_1}{\lambda^2} - \frac{y_2+y_3}{\left(1-\lambda\right)^2}$$

$$\mathrm{Var}\left(D\log f\left(X|\lambda\right)|Y,\lambda\right) = \mathrm{Var}\left(\frac{X_1+X_4}{\lambda} - \frac{X_2+X_3}{1-\lambda}|Y,\lambda\right)$$

$$= \mathrm{Var}\left(\frac{X_4}{\lambda}|Y,\lambda\right)$$

$$= \frac{y_4}{\lambda^2}\frac{\lambda}{2+\lambda}\frac{2}{2+\lambda}$$

Plugging in gives

$$E\left[D^2 \log f\left(X|\lambda\right)|Y,\lambda\right] = 435.3$$

$$\mathrm{Var}\left(D\log f\left(X|\lambda\right)|Y,\lambda\right) = 57.8$$

so

$$I\left(\hat{\lambda}\right) = 435.3 - 57.8 = 377.5$$

5) SEM algorithm (Meng and Rubin, 1991)

Their idea is based on the missing information principle and the fact

$$DM\left(\hat{\theta}\right) = I_{OM}\left(\hat{\theta}\middle|Y\right) I_{OC}^{-1}\left(\hat{\theta}\middle|Y\right)$$

Combining the two gives

$$\begin{aligned} I_O &= I_{OC} - I_{OM} \\ &= \left(I - I_{OM} I_{OC}^{-1}\right) I_{OC} \\ &= \left(I - DM\right) I_{OC} \end{aligned}$$

Thus, if we can figure out *DM* and I_{OC}, we can get the observed information in the data.

In the genetics example discussed last time, we saw that using iterates from EM we could get a reasonable guess for *DM*, at least in a single parameter problem.

Instead of calculating the matrices above exactly, the idea is to use the iterates of the EM sequences to approximately numerically, *DM*.

$$
\begin{aligned}
r_{ij}\left(\hat{\theta}\right) &= \frac{\partial M_j\left(\theta\right)}{\partial \theta_i}\Big|_{\theta_i=\hat{\theta}_i} \\
&= \lim_{\theta_i \to \hat{\theta}_i} \frac{M_j\left(\hat{\theta}_1,\ldots,\theta_i,\ldots,\hat{\theta}_k\right) - M_j\left(\hat{\theta}\right)}{\theta_i - \hat{\theta}_i} \\
&= \lim_{t \to \infty} \frac{M_j\left(\hat{\theta}_1,\ldots,\theta_i^t,\ldots,\hat{\theta}_k\right) - M_j\left(\hat{\theta}\right)}{\theta_i^t - \hat{\theta}_i} \\
&= \lim_{t \to \infty} r_{ij}^t
\end{aligned}
$$

So the following scheme can be used to get r_{ij}^t.

1) Fix i = 1 and set $\theta^t\left(i\right) = \left(\hat{\theta}_1,\ldots,\theta_i^t,\ldots,\hat{\theta}_k\right)$

   Evaluate $\tilde{\theta}^{t+1}\left(i\right) = M\left(\theta^t\left(i\right)\right)$

2) Form

$$
r_{ij}^t = \frac{\tilde{\theta}_j^{t+1}\left(i\right) - \hat{\theta}_j}{\theta_i^t - \hat{\theta}_i}
$$

   for j = 1, … , k.

3) Repeat steps 1 and 2 for i = 2, … , k.

To implement this algorithm, k evaluations of the mapping M are required.

Doing this for each *EM* iteration leads to the sequence $\{r_{ij}^{1}, r_{ij}^{2}, \ldots\}$,. which can be stopped at t^* when the sequence stablizes. Note that t^* may not be the same for each (i, j) combination.

Also for numerical reasons the sequence may appear to become unstablized at some point. We saw this last time with the genetics example

Iteration	λ_n	$\lvert\lambda_n - \hat{\lambda}\rvert$	$(\lambda_{n+1} - \hat{\lambda})/(\lambda_n - \hat{\lambda})$
0	0.5	0.126821	0.1465
1	0.608247423	0.018574	0.1346
2	0.624321050	0.002500	0.1330
3	0.626488879	0.000333	0.1327
4	0.626777322	0.000044	0.1322
5	0.626815632	0.000006	0.1287
6	0.626820719		0.1009
7	0.626821394		-0.1831

This is an artifact of the numerical precision of computer code. When calculating the errors at each iteration you lose significant digits.

However if you had infinite precision, the sequence would converge.

So for deciding when the r_{ij}^{t} have converged, you need a different convergence criterion.

One suggestion I've seen (though I can't remember where) is if your convergence criterion for EM is stop when

$$\left\| \theta_{n+1} - \theta_{n} \right\| < TOL$$

then use the SEM stopping criterion

$$\left\| r_{ij}^{t+1} - r_{ij}^{t} \right\| < \sqrt{TOL}$$

One way to think of this is to go for only half as many digits of accuracy.

Also look to see when the sequence $\left\| r_{ij}^{t+1} - r_{ij}^{t} \right\|$ starts to increase (as it probably will).

One potential problem with this algorithm, is that this estimate I_O is not guaranteed to be symmetric and thus $V = I_O^{-1}$ will not be either.

Meng and Rubin suggest replacing V with $\frac{1}{2}\left(V + V^{T}\right)$.

Another idea would be to replace I_O with $\frac{1}{2}\left(I_O + I_O^{T}\right)$.

Asymmetry in I_O and V can be used to look for problems in SEM.

Note that you do not need to iterate the SEM algorithm as I've described, You can run through steps 1) through 3) only once. However you need to think about the values θ_i^t you use for each i.

SEM when θ is a single value

You do not need to run the extra EM steps to get $\tilde{\theta}^{t+1}(i)$ as $\tilde{\theta}^{t+1}(i) = \theta^{t+1}$.

So for this case, you get SEM for free. However for the multiparameter case, you do need to run the extra EM steps.

Genetics Example:

As shown last time the true value of *DM* = 0.1327798. Plugging into

$$\begin{aligned} I_O &= \left(I - DM\right) I_{OC} \\ &= \left(1 - 0.1327798\right) 435.3 \\ &= 377.501 \end{aligned}$$

The same answer as Louis' method.

If we estimate *DM* with 0.132739278 (where $\left\| r_{ij}^{t+1} - r_{ij}^{t} \right\|$ starts to increase, we get

$$\begin{aligned} I_O &= \left(I - DM\right) I_{OC} \\ &= \left(1 - 0.1327392\right) 435.3 \\ &= 377.519 \end{aligned}$$

If we look at the standard error of $\hat{\lambda}$, we get

$$SE\left(\hat{\lambda}\right) = 0.0514684$$

Iteration	$SE_{SEM}\left(\hat{\lambda}\right)$
0	0.0518792
1	0.0515231
2	0.0514753
3	0.0514672
4	0.0514524
5	0.0513471
6	0.0505479

Cyclic Coordinate Ascent:

Last time I briefly discussed optimizing along each coordinate in turn (ICM, ECM). In general the algorithm can be thought of as

Step 1): Find t_1 which maximizes

$$L\left(\theta + te_1\right)$$

where e_i is the vector whose i^{th} coordinate is 1 and the rest are 0.

Step *i*): Find t_i which maximizes

$$L\left(\theta + \sum_{j=1}^{i-1} t_j e_j + t e_i\right)$$

When all *k* coordinates have been updated, one iteration is complete.

It can be shown that cyclic coordinate schemes will converge as long as a maximum is determined in each step.

However that convergence might be to saddle point, instead of a local maximum.

None of the schemes I've discussed so far are guaranteed to converge to the global maximum, unless strong assumptions can be made of the function being optimized, such as the function is convex over its parameter space

Problem 13.7 in Lange discusses a multivariate normal case where the Likelihood has two modes and a saddle point.