EM Algorithm Extensions

ECM (Meng and Rubin, 1993)

(Expectation Conditional Maximization)

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

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

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 1983)

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. $S_i Y_i$ is the vector of observed components)

E-step:

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

and

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

M-step: maximize

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

$$-\frac{1}{2}\sum_i \left(\hat{Y}_i^{(n)} - X_i\beta \right)^T V^{-1} \left(\hat{Y}_i^{(n)} - X_i\beta \right)$$

M1:

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

M2:

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

$$+ X_i \beta^{(n+1)} \beta^{(n+1)^T} X_i^T)$$

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 $\Theta_{-j} = \{ \theta_i : i \neq j \}$

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

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

The second form holds since as shown last time

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

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

$$\pi(\theta|\mathrm{Y}) \propto g(Y|\theta)\pi(\theta)$$

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:

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

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: $X \sim \mathrm{Bin}(n, p)$

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

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

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

Posterior:

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

$$p|x \sim \mathrm{Beta}(x+\alpha, n-x+\beta)$$

E-step:

$$\begin{aligned} Q(p|p_n) = E\Big[&(x+a-1)\log p \\ &+(n-x+\beta-1)\log(1-p)\Big|Y, p_n\Big] \end{aligned}$$

So we need $E\left[X|Y, p_n\right]$, 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 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 $D^{10}Q\left(\theta_{n+1}\middle|\theta_n\right)=0$, 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

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

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,

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

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

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	$\left\|\lambda_n - \hat{\lambda}\right\|$	$\left\|\lambda_{n+1} - \hat{\lambda}\right\| / \left\|\lambda_n - \hat{\lambda}\right\|$
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

Louis' formula (Louis, 1982)

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

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.

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

SEM algorithm (Meng and Rubin, 1991)

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

Their idea is based on the missing information principle and

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

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.