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)

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

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

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

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

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

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

So

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

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

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

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

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

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.