Using EM to Obtain Asymptotic Variance Matrices: The SEM Algorithm

Hung Chen
Department of Mathematics
National Taiwan University

OUTLINE

1. Motivated Examples
2. Quick Review on EM
3. Rao's Multinomial Example
4. Review on Standard Asymptotic Analysis
5. Louis's Suggestion
6. The SEM Algorithm
7. References:
   - Meng and Rubin (JASA, pp899-909, 1991) Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm.
   - Louis (JRSSB, pp226-233, 1982) Finding the Observed Information Matrix when Using the EM Algorithm.
   - Dempster, Laird and Rubin (JRSSB, pp1-37, 1977) Maximum Likelihood rom Incomplete Data via the EM Algorithm.

Motivated Examples

Incomplete Covariate Data in Linear Regression Models:

- complete-data score

$$S(\beta) = \sum_i X_i(Y_i - X_i^T\beta)$$

- observed data: (Y_i, Z_i)
- score for the observed data

$$E[S(\beta)|Y,Z]$$
$$= \sum_i E_{\beta_1}(X_i|Y_i,Z_i)Y_i - \beta \sum_i E_{\beta_1}(X_iX_i^T|Y_i,Z_i)$$

- implementation:
  Need $P_\beta(Y|X,Z)$ and $P_\gamma(X|Z)$.

$$P_{\beta,gamma}(X|Y,Z) = \frac{P_\beta(Y|X,Z)P_\gamma(X|Z)}{\int P_\beta(Y|X,Z)P_\gamma(X|Z)dX}.$$

- Surrogate Condition: $P_\beta(Y|X,Z) = P_\beta(Y|X)$.

Cancer Epidemiological Studies

Fact: The log-log plot of the cumulative incidence against age in many human adult carcinomas is almost a straight line.

Armitage-Doll Model (1961):

- Cancer arises from a single cell undergoing a series of p distinct heritable mutations in a particular sequence.
- The waiting time between any age and the subsequent stage is exponentially distributed.

$$E_1 \stackrel{\lambda_1}{\to} E_2 \stackrel{\lambda_2}{\to} E_3 \stackrel{\lambda_3}{\to} \cdots \stackrel{\lambda_p}{\to} E_{p+1}.$$

- E_1: a normal cell, E_{p+1}: a malignant cell
- The hazard of developing a cancer at age t, $h(t) \approx ct^{p-1}$.

Pike (Nature, 1983)

The Occurrence of Breast Cancer

Data:

The log-log plot of the cumulative incidence against age is a broken line and the slope before age 45 is much larger than after it.

Explanation:

The Breast Tissue Age

The mitotic activity begins at **menarche** at age t_0 and proceeds at a constant rate until the **first full-term pregnancy**, then proceeds at reduced rate until **menopause**, and thereafter continues at a still lower rate.

- Several to-be-determined change points
- Constant rate between consecutive change points

EM Algorithm
Dempster, Laird and Rubin, JRSSB(1977)

A general approach to iterative computation of MLEs when the observations can be viewed as incomplete data; each iteration of the algorithm consists of an expectation step followed by a maximization step.

- Complete data:
  $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{X}$
  $\mathcal{X}$: sample space
  Model: density $f(\mathbf{x}|\theta)$
  $\theta \in \Theta \subset R^p$: unknown parameter

- Incomplete data:
  $Y(\mathbf{x}) = \mathbf{x}_{obs} = y \in \mathcal{Y}$
  Many-One mapping: $\mathbf{x} \to Y(\mathbf{x})$
  $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis})$, where

$\mathbf{x}_{obs}$ represents the observed part of $\mathbf{x}$ and

$\mathbf{x}_{mis}$ denotes the missing values.

- Objective of the EM algorithm: Find the MLE for θ, $\hat{\theta}$, based on $\mathbf{x}_{obs}$.
- Model:

$$g(Y(\mathbf{x})|\theta) = \int_{\mathcal{X}(y)} f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}$$

An Illustrative Example

Rao (1965, pp. 368-369)

- 197 animals are distributed multinomially into four categories
- Observed data: $\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$
- $\mathbf{y}$ can be modeled as a four-category multinomial population with cell probabilities

$(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$,

and $0 \le \theta \le 1$.

- View $\mathbf{y}$ as incomplete data of $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$. Here $\mathbf{X}$ is a multinomial distribution with cell probabilities

$$(1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4).$$

- Observe that

$$g(\mathbf{y}|\theta) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!} \cdot (\frac{1}{2}+\frac{\theta}{4})^{y_1}(\frac{1}{4}-\frac{\theta}{4})^{y_2}(\frac{1}{4}-\frac{\theta}{4})^{y_3}(\frac{\theta}{4})^{y_4}$$

and

$$f(\mathbf{x}|\theta) = \frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!} \cdot \left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2+x_5}\left(\frac{1}{4}-\frac{\theta}{4}\right)^{x_3+x_4}.$$

- The MLE of θ based on $\mathbf{y}$ is the solution of the following:

$$(y_1+\cdots+y_4)\theta^2+(-y_1+2y_2+2y_3+y_4)\theta-y_4=0.$$

- The MLE of θ based on $\mathbf{x}$ is

  $\hat{\theta}^* = (x_2 + x_5)/(x_2 + x_3 + x_4 + x_5)$.

- $\mathbf{X}|\mathbf{Y}$ or $X_2|y_1$ can be modeled as $Bin(y_1, \frac{\theta/4}{1/2+\theta/4})$.

What is the EM algorithm?

- Find a value of θ maximizing $g(\mathbf{y}|\theta)$ by making essential use of $f(\mathbf{x}|\theta)$.
- The E step:
  Estimate the sufficient statistics of the complete data $\mathbf{x}$ given $\mathbf{y}$.
  What is the **sufficient statistics**?

Rao's Example

  - The log-likelihood $L(\theta|\mathbf{x})$ is linear in $\mathbf{x}$.
  - Estimate x_1 and x_2
    when $x_1 + x_2 = y_1 = 125$.

– At the $(t+1)$th iteration,

$$x_2^{(t)} = 125\frac{\theta^{(t)}/4}{1/2+\theta^{(t)}/4}.$$

Exponential Family:

– $f(\mathbf{x}|\theta) = b(\mathbf{x})\exp(\theta\mathbf{t}(\mathbf{x})^T)/a(\theta)$,

$$E(\mathbf{t}(\mathbf{x})|\theta) = a^{(1)}(\theta)/a(\theta).$$

– Conditional density of $\mathbf{x}$ given $\mathbf{y}$ and θ:

$k(\mathbf{x}|\mathbf{y},\theta) = b(\mathbf{x})\exp(\theta\mathbf{t}(\mathbf{x})^T)/a(\theta|\mathbf{y})$

where

$a(\theta|\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} b(\mathbf{x})\exp(\theta\mathbf{t}(\mathbf{x})^T)d\mathbf{x}.$

Note that

$$a(\theta) = \int_{\mathcal{X}} b(\mathbf{x})\exp(\theta\mathbf{t}(\mathbf{x})^T)d\mathbf{x}.$$

– At the $(t+1)$th iteration, estimate the complete-data sufficient statistics $\mathbf{t}(\mathbf{x})$ by

$$\mathbf{t}^{(t)} = E(\mathbf{t}(\mathbf{x})|\mathbf{y},\theta^{(t)}).$$

In General:

– Conditional density of $\mathbf{x}$ given $\mathbf{y}$ and θ:

$$k(\mathbf{x}|\mathbf{y},\theta) = f(\mathbf{x}|\theta)/g(\mathbf{y}|\theta).$$

– Log-likelihood function of $\mathbf{y}$:

$$\begin{aligned} L(\theta) &= \log g(\mathbf{y}|\theta) \\ &= \log f(\mathbf{x}|\theta) - \log k(\mathbf{x}|\mathbf{y},\theta). \end{aligned}$$

For exponential family,

$$L(\theta) = -\log a(\theta) + \log a(\theta|\mathbf{y}).$$

– At the $(t+1)$th iteration, estimate the complete-data sufficient statistics $\log f(\mathbf{x}|\theta)$ by

$$Q(\theta|\theta^{(t)}) = E(\log f(\mathbf{x}|\theta)|\mathbf{y},\theta^{(t)}),$$

which is the expected pseudo-complete data log-likelihood if θ were $\theta^{(t)}$.

- The M step:
  Maximize the expected log-likelihood

Rao's Example

$$\theta^{(t+1)} = (x_2^{(t)} + 34)/(x_2^{(t)} + 72).$$

Exponential Family:

Determine $\theta^{(t+1)}$ as the solution of the equations

$$E(\mathbf{t}(\mathbf{x})|\theta) = \mathbf{t}^{(t)} (= E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \theta^{(t)})).$$

In General:

Choose $\theta^{(t+1)}$ to be the one maximizing $Q(\theta|\theta^{(t)})$.

Review on Asymptotic Variance:

Consider $X_1, \ldots, X_n \sim N(\theta_0, \sigma^2)$.

- Score Equation:

$$\begin{aligned} S(\theta) &= \sum_{i=1}^{n} (X_i - \theta)/\sigma^2, \\ S(\hat{\theta}) &= 0. \end{aligned}$$

- $S(\theta_0) \sim N(0, n/\sigma^2)$.
- $-S^{(1)}(\theta) = n/\sigma^2$.
- $\theta - \theta_0 \sim N(0, \sigma^2/n)$.

In General:

Under regularity condition,

$$\begin{aligned} 0 &= S(\hat{\theta}) \\ 0 &\approx S(\theta_0) + (\hat{\theta} - \theta_0) S^{(1)}(\theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= \frac{\frac{1}{\sqrt{n}} S(\theta_0)}{\frac{1}{n} S^{(1)}(\theta_0)} \\ -\frac{1}{n} S^{(1)}(\theta_0) &\approx AVar\left(\frac{1}{\sqrt{n}} S(\theta_0)\right). \end{aligned}$$

Recall that

$$[\log g(\mathbf{y}|\theta)]^{(1)} = [\log f(\mathbf{x}|\theta)]^{(1)} - [\log k(\mathbf{x}|\mathbf{y},\theta)]^{(1)}.$$

Rao's Example:

Complete-Data: $\mathbf{Y}$

- Score Equation: g

$$\begin{aligned} S(\theta) &= \frac{Y_1}{2+\theta} - \frac{Y_2+Y_3}{1-\theta} + \frac{Y_4}{\theta}, \\ \sqrt{n}S(\theta_0) &\sim N(0,?) \\ -\frac{1}{n}S^{(1)}(\theta) &\xrightarrow{P} -\frac{2\theta+1}{2\theta(2+\theta)(1-\theta)}. \end{aligned}$$

- $\sqrt{n}(\theta-\theta_0) \sim N(0,?)$.

Pseudo-Complete Data: $\mathbf{X}$

Score Equation: f

$$\begin{aligned} S(\theta) &= \frac{X_2+X_5}{\theta} - \frac{X_3+X_4}{1-\theta}, \\ -S^{(1)}(\theta) &= \frac{x_2+x_5}{\theta^2} + \frac{x_3+x_4}{(1-\theta)^2}. \end{aligned}$$

$$X_2|Y_1$$

$$X_2|Y_1 \sim Bin(Y_1, \theta/(2+\theta)).$$

- Score Equation: k

$$S(\theta) = \frac{X_2}{\theta} - \frac{Y_1}{2+\theta},$$
$$-S^{(1)}(\theta) = \frac{X_2}{\theta^2} - \frac{Y_1}{(2+\theta^2)^2}.$$

$$\frac{1}{\sqrt{Y_1}} S(\theta_0) \sim N(0, 2/[\theta(2+\theta)^2]),$$
$$-Y_1^{-1} S^{(1)}(\theta) \xrightarrow{P} 2/[\theta(2+\theta)^2].$$

- $\sqrt{Y_1}(\theta - \theta_0) \sim N(0, \theta_0(2+\theta_0)^2/2)$.

Remark: $Y = X_1 + X_2,\ X_1, X_2 \sim Exp(\theta)$
Complete Data: Y
Pseudo-Complete Data: X_1, X_2
The distribution of $X_1|Y$ is independent of θ.

The EM algorithm implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space of θ to itself such that

$$\theta^{(t+1)} = M(\theta^{(t)}), \quad t = 0, 1, \ldots.$$

Suppose that $\theta^{(t)}$ converges to θ^*. Then

$$\theta^* = M(\theta^*)$$

and in the neighborhood of θ^*

$$\theta^{(t+1)} - \theta^* \approx (\theta^{(t)} - \theta^*) M^{(1)}(\theta^*).$$

Here $M^{(1)}$ defines the rate of convergence.

How can we figure out $M^{(1)}(\theta^*) = I_{om} I_{oc}^{-1}$?

Large sample variance of $\hat{\theta} - \theta_0$ based on $\mathbf{y}$:

Denote the negative second derivative of the log-likelihood of θ given $\mathbf{y}$ by $I_o(\hat{\theta}|\mathbf{y})$ which is

$$I_o(\theta|\mathbf{y}) = -\frac{d^2 \log g(\mathbf{y}|\theta)}{d\theta^2}.$$

Assume that $\theta^* = \hat{\theta}$.
Then $V = AVar(\sqrt{n}(\theta^* - \theta_0))$ is the inverse of $I_o^{-1}(\theta^*|\mathbf{y})$.

A Scenario: V can be difficult to evaluate directly but $I_O(\theta^*|\mathbf{x})$ is easy to compute.
Here

$$I_O(\theta|\mathbf{x}) = -\frac{d^2 \log f(\mathbf{x}|\theta)}{d\theta^2}.$$

Question: Can we find V by using the complete-data variance matrix and successive iterates of the EM algorithm?

Smith's Comment on DLR (1977):

- V: the observed-data asymptotic variance
- V_C: the complete-data asymptotic variance
- Claim:

$$V = V_C/(1-r) = V_C + \frac{r}{1-r}V_C.$$

  r: the rate of convergence of EM

- r can be well approximated by

$$(\theta^{(t+1)} - \theta^*)/(\theta^{(t)} - \theta^*).$$

Rao's Example:

- $\hat{\theta} = 0.6268215\ldots$ seems to be the limit of $0.6(= \theta^{(0)}), 0.623188, 0.626338, 0.626757, 0.626812, \cdots$.

- $r = 0.1328$ seems to be the limit of

  $0.1465, 0.1346, 0.1330, 0.1328, 0.1328,$

  $0.1328, 0.1328, \ldots.$

- According to Smith's claim,

$$\begin{aligned} V^{-1} &= I_{oc}(1-r) = 435.3 \times (1 - 0.1328) \\ &= 377.49. \end{aligned}$$

- Pseudo-complete data:

$$X_2 + X_5 \sim Bin(X_2 + X_3 + X_4 + X_5, \theta)$$

$$\begin{aligned} &\hat{X}_2 + X_3 + X_4 + X_5 \\ &= 125 \times \frac{\theta^*/4}{1/2 + \theta^*/4} + 72 = n^* \end{aligned}$$

$$\sqrt{V} = \sqrt{\frac{\theta^*(1-\theta^*)}{n^*(1-r)}} = 0.051$$

Suggestion in Louis (1982)

Missing Information Principle (Orchard and Woodbury, 1972):
observed information = complete information - missing information

Why?

Recall that

$$[\log g(\mathbf{y}|\theta)]^{(1)} = [\log f(\mathbf{x}|\theta)]^{(1)} - [\log k(\mathbf{x}|\mathbf{y},\theta)]^{(1)}.$$

Hence,

$$\begin{aligned} V^{-1} &= I_{oc} - I_{om} \\ &= E\left[\frac{x_2 + x_5}{\theta^{*2}} + \frac{x_3 + x_4}{(1-\theta)^{*2}}\middle|\mathbf{y}, \theta^*\right] \\ &- \left[\frac{E(x_2|\theta^*, \mathbf{y})}{\theta^{*2}} - \frac{y_1}{(2+\theta^*)^2}\right] \\ &= 435.3 - 57.8 = 377.5. \end{aligned}$$

Claim: $M^{(1)}(\theta^*) = I_{om} I_{oc}^{-1}$

Why?

From now on, we use Meng and Rubin's notations.

- The Factorization:

$$f(\mathbf{x}|\theta) = f(\mathbf{y}|\theta)f(\mathbf{x}_{mis}|\mathbf{y},\theta)$$

  Note that $f(\mathbf{y}|\theta)$ is called $g(\mathbf{y}|\theta)$ and $f(\mathbf{x}_{mis}|\mathbf{y},\theta)$ is called $k(\mathbf{x}|\mathbf{y},\theta)$.

- log-likelihhod:

$$L(\theta|\mathbf{y}) = L(\theta|\mathbf{x}) - \log f(\mathbf{x}_{mis}|\mathbf{y},\theta).$$

  The following result is obtained by taking second derivatives, averaging over $f(\mathbf{x}_{mis}|\mathbf{y},\theta)$, and evaluating at $\theta = \theta^*$.

- missing information principle:

$$I_o(\theta^*|\mathbf{y}) = I_{oc} - I_{om},$$

  where

$$I_{oc} = E\big[I_o(\theta|\mathbf{x})|\mathbf{y},\theta\big]\big|_{\theta=\theta^*}$$

and

$$I_{om} = E\left[-\frac{d^2 \log f(\mathbf{x}_{mis}|\mathbf{y},\theta)}{d\theta^2}\middle|\mathbf{y},\theta\right]\Bigg|_{\theta=\theta^*}.$$

- Write $I_o(\theta^*|\mathbf{y})$ as

$$(1 - I_{om}I_{oc}^{-1})I_{oc}.$$

- Dempster et al.(1977) showed that

$$M^{(1)}(\theta^*) = I_{om}I_{oc}^{-1}.$$

- Conclusion:

$$V = I_{oc}^{-1}(1 - M^{(1)}(\theta^*))^{-1}.$$

Proof of $M^{(1)}(\theta^*) = I_{om} I_{oc}^{-1}$

Refer to Theorem 4 of DLR (1977).

- Score Equation:

$$Q(\theta_1|\theta) = L(\theta_1) + H(\theta_1|\theta)$$

where

$$\begin{aligned} Q(\theta_1|\theta) &= E(\log f(\mathbf{x}|\theta_1)|\mathbf{y}, \theta) \\ L(\theta_1) &= \log f(\mathbf{y}|\theta_1) \\ H(\theta_1|\theta) &= E(\log f(\mathbf{x}_{mis}|\mathbf{y}, \theta_1)|\mathbf{y}, \theta). \end{aligned}$$

- Assumptions:
  $\theta^{(t)}$ converges to θ^*.
  $D^{10} Q(\theta^{(t+1)}|\theta^{(t)}) = 0$.
  $D^{20} Q(\theta^{(t+1)}|\theta^{(t)})$ is negative definite with eigenvalues bounded away from 0.

- Some Facts:

$$\begin{aligned} D^{11}H(\theta|\theta) &= -D^{20}H(\theta|\theta) \\ D^{11}Q(\theta|\theta) &= D^{11}H(\theta|\theta) \\ \theta^{(t+1)} &= M(\theta^{(t)}) \\ \theta^* &= M(\theta^*). \end{aligned}$$

- Expansion at θ^*:

$$\begin{aligned} D^{10}Q(\theta_1|\theta) &= D^{10}Q(\theta^*|\theta^*) \\ &+(\theta_1 - \theta^*)D^{20}Q(\theta^*|\theta^*) \\ &+(\theta - \theta^*)D^{11}Q(\theta^*|\theta^*) + \cdots. \end{aligned}$$

Substituting $\theta_1 = \theta^{(t+1)}$ and $\theta = \theta^{(t)}$, we have

$$\begin{aligned} 0 &= (\theta^{(t+1)} - \theta^*)D^{20}Q(\theta^*|\theta^*) \\ &+ (\theta^{(t)} - \theta^*)D^{11}Q(\theta^*|\theta^*) + \cdots \end{aligned}$$

and

$$0 = M^{(1)}(\theta^*)D^{20}Q(\theta^*|\theta^*)+D^{11}Q(\theta^*|\theta^*).$$

- Conclusion:

  $M^{(1)}(\theta^*) = D^{20}H(\theta^*|\theta^*)[D^{20}Q(\theta^*|\theta^*)]^{-1}.$

  $M^{(1)}(\theta^*) = I_{om}I_{oc}^{-1} = 0.1327778$

  which is close to 0.1328 derived before.

SEM:

INPUT: θ^* and $\theta^{(t)}$.

- Run the usual E and M steps to obtain $\theta^{(t+1)}$.
- Repeat the following two steps for $i = 1, \ldots, p$.
- Calculate $\theta^{(t)}(i)$ by

  $$\theta^{(t)}(i) = (\theta_1^*, \cdots, \theta_{i-1}^*, \theta_i, \theta_{i+1}^*, \cdots, \theta_p^*).$$

  Treat $\theta^{(t)}(i)$ as current estimate and run one iteration of EM to obtain $\tilde{\theta}^{(t+1)}(i)$.
- Obtain the ratio

  $$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \theta_j^*}{\theta_i^{(t)} - \theta_i^*}, \quad \text{for } j = 1, \ldots, p.$$

OUTPUT: $\theta^{(t+1)}$ and $\{r_{ij}^{(t)}, i, j = 1, \ldots, p\}$.

Question: How do we handle the case that some components have no missing information?

Fact: When there is no missing information on a particular component of θ,

1. EM will converge in one step for that component from any starting value, and

2. the corresponding component of $\mathbf{M}(\theta)$ will be a constant with zero derivative.

Recall the Gamma distribution example.

Assume $\{M_j(\theta), j = 1, \ldots, p_1\}$ constant.

Write

$$DM = \begin{array}{c} \\ p_1 \\ p_2 \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p_2 \end{array} \\ \begin{pmatrix} 0 & A \\ 0 & DM^* \end{pmatrix} \end{array}.$$

where $p_1 + p_2 = p$.

The SEM algorithm can be used to compute DM. Write

$$I_{oc}^{-1} = \begin{array}{c} \\ p_1 \\ p_2 \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p_2 \end{array} \\ \begin{pmatrix} G_1 & G_2 \\ G_2^t & G_3 \end{pmatrix} \end{array}.$$

Then we have

$$V = I_{oc}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \triangle V^* \end{pmatrix},$$

where

$$\triangle V^* = (G_3 - G_2^T G_1^{-1} G_2) DM^* (I - DM^*)^{-1}.$$