# STA6171: Statistical Computing for DS 1

# EM Algorithm

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

2020.10.07

# Introduction

- The expectation-maximization (EM) algorithm is an iterative optimization strategy motivated by a notion of missingness and by consideration of the conditional distribution of what is missing given what is observed.
- Popularity of the EM algorithm
  - Simple to implement
  - Reliable to find the global optimum.

# Introduction

- Frequentist Setting
  - Observed data from $X$ along with missing data from $Z$.
  - Complete data $Y = (X, Z)$.
  - Given observed data $x$, maximize a likelihood $L(\theta|x)$. Difficult to work with this likelihood
  - A easier way is working with the density $Y|\theta$ and $Z|x, \theta$.
- Bayesian Setting: Interest Often focuses on estimating the mode of a posterior distribution $f(\theta|x)$.
- Missing data may not truly be missing: they may be only a conceptual ploy that simplifies the problem. In this case, $Z$ is often referred to as *latent*.

# Missing Data and Marginalization - Frequentist

- In the presence of missing data, only a function of the complete-data $y$ is observed.
- Assume that the missing data are random, so that

$$f(y|\theta) = f(x, z|\theta) = f_X(x|\theta) \cdot f_{Z|X}(z|x, \theta).$$

  Thus, it follows that

$$l_X(\theta) = l_Y(\theta) - \log f_{Z|X}(Z|X, \theta).$$

- Useful when maximizing $l_X(\theta)$ can be difficult but maximizing the complete log-likelihood $l$ is simple.

# Missing Data and Marginalization - Bayesian

- View the likelihood $L(\theta|x)$ as a marginalization of the complete-data likelihood $L(\theta|y) = L(\theta|x, z)$.
- Consider there to be missing parameter $\psi$, whose inclusion simplifies Bayesian calculations even though $\psi$ is of no interest itself. Since $Z$ and $\psi$ are both missing random quantities, it matters little whether we use notation that suggests the missing variables to be unobserved data or parameters.

# EM Algorithm

- EM algorithm iteratively seeks to maximize $L(\theta|x)$ with respect to $\theta$.
- Let $\theta^{(t)}$ denote the estimated maximizer at iteration $t$, for $t = 0, 1, \cdots$.
- Define $Q(\theta|\theta^{(t)})$ to be the expectation of the joint log-likelihood for the complete data, conditional on the observed data $X = x$.

$$Q\left(\theta|\theta^{(t)}\right) = E\left\{\log L(\theta|Y)\middle|x, \theta^{(t)}\right\}.$$

- Then, $Q\left(\theta|\theta^{(t)}\right)$ is maximized w.r.t $\theta$, that is $\theta^{(t+1)}$ is found such that

$$Q\left(\theta^{(t+1)}|\theta^{(t)}\right) \geq Q\left(\theta|\theta^{(t)}\right)$$

for all $\theta \in \Theta$.

# EM Algorithm

- EM is initiated from $\theta^{(0)}$ then alternates between two steps: E is for expectation and M for maximization.
- The algorithm is summarized as:
  1. **E Step**: Compute $Q\left(\theta|\theta^{(t)}\right)$.
  2. **M Step**: Maximize $Q\left(\theta|\theta^{(t)}\right)$ with respect to $\theta$. Set $\theta^{(t+1)}$ equal to the maximizer of $Q$.
  3. Return to the E step unless a stopping criterion has been met.

# EM Algorithm - Exponential Families

- The computation of these two steps simplify a great deal when it can be shown that the log-likelihood is linear in the sufficient statistic for $\theta$ (Exponential Family).
- The E-step reduces to computing the expectation of the complete-data sufficient statistic given the observed data.
- In M-step, the conditional expectations of the sufficient statistics computed in the E-step can be directly substituted for the sufficient statistics that occur in the expressions obtained for the complete-data maximum likelihood estimators of $\theta$, to obtain the next iterate.

# Appealing Properties of the EM Algorithm

- It is typically easily implemented because it relies on complete-data computations
  1. The E-step of each iteration only involves taking expectations over complete-data conditional distributions.
  2. The M-step of each iteration only requires complete-data maximum likelihood estimation, for which simple closed form expressions are already available.
- It is numerically stable: each iteration is required to increase the log-likelihood $l_X(\theta)$ in each iteration, and if $l_X(\theta)$ is bounded, the sequence $l_X(\theta^{(t)})$ converges to a stationery value.

# Convergence

The log of the observed-data density can be re-expressed as

$$\log f_X(x \mid \theta) = \log f_Y(y \mid \theta) - \log f_{Z|X}(z \mid x, \theta).$$

Therefore,

$$E\left\{\log f_X(x \mid \theta) \mid x, \theta^{(t)}\right\} = E\left\{\log f_Y(y \mid \theta) \mid x, \theta^{(t)}\right\} - E\left\{\log f_{Z|X}(z \mid x, \theta) \mid x, \theta^{(t)}\right\},$$

where the expectations are taken with respect to the distribution of $Z \mid (x, \theta^{(t)})$. Thus,

$$\log f_X(x \mid \theta) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}),$$

where

$$H(\theta \mid \theta^{(t)}) = E\left\{\log f_{Z|X}(Z \mid x, \theta) \mid x, \theta^{(t)}\right\}.$$

# Convergence

Show that $H(\theta \mid \theta^{(t)})$ is maximized with respect to $\theta$ when $\theta = \theta^{(t)}$. Write

$$\begin{aligned} H\left(\theta^{(t)} \mid \theta^{(t)}\right) - H\left(\theta \mid \theta^{(t)}\right) &= E\left\{\log f_{Z|X}\left(Z \mid x, \theta^{(t)}\right) - \log f_{Z|X}\left(Z \mid x, \theta\right) \mid x, \theta^{(t)}\right\} \\ &= \int -\log\left[\frac{f_{Z|X}(z \mid x, \theta)}{f_{Z|X}(z \mid x, \theta^{(t)})}\right] f_{Z|X}(z \mid x, \theta^{(t)}) dz \\ &\geq -\log \int f_{Z|X}(z \mid x, \theta) dz = 0 \end{aligned}$$

Thus, any $\theta \neq \theta^{(t)}$ makes $H(\theta \mid \theta^{(t)})$ smaller than $H(\theta^{(t)} \mid \theta^{(t)})$.

If we choose $\theta^{(t+1)}$ to maximize $Q(\theta \mid \theta^{(t)})$ with respect to $\theta$, then

$$\log f_X(x \mid \theta^{(t+1)}) - \log f_X(x \mid \theta^{(t)}) \geq 0,$$

since $Q$ increases and $H$ decreases, with restrict inequality when $Q(\theta^{(t+1)} \mid \theta^{(t)}) > Q(\theta^{(t)} \mid \theta^{(t)})$.

# Convergence

- Choosing $\theta^{(t+1)}$ at each iteration to maximize $Q(\theta \mid \theta^{(t)})$ with respect to $\theta$ constitutes the standard EM algorithm.
- If instead we simply select any $\theta^{(t+1)}$ for which

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) > Q(\theta^{(t)} \mid \theta^{(t)}),$$

  then resulting algorithm is called generalized EM.
- A step that increase $Q$ increases the log-likelihood.
- The global rate of EM convergence is defined as

$$\rho = \lim_{n \to \infty} \frac{||\theta^{(t+1)} - \hat{\theta}||}{||\theta^{(t)} - \hat{\theta}||}.$$

  $\rho$ effectively serves as a scalar summary of overall proportion of missing information.

# Convergence

- Conceptually, the proportion of missing information equals one minus the ratio of the observed information to the information that would be contained in the complete data.
- EM suffers slower convergence when the proportion of missing information is larger.
- The linear convergence of EM can be extremely slow compared to the quadratic convergence of, say, Newton's method, particularly when the fraction of missing information is large.
- However, the ease of implementation and the stable ascent of EM are often very attractive despite its slow convergence.

# One-dimensional Illustration of EM Algorithm

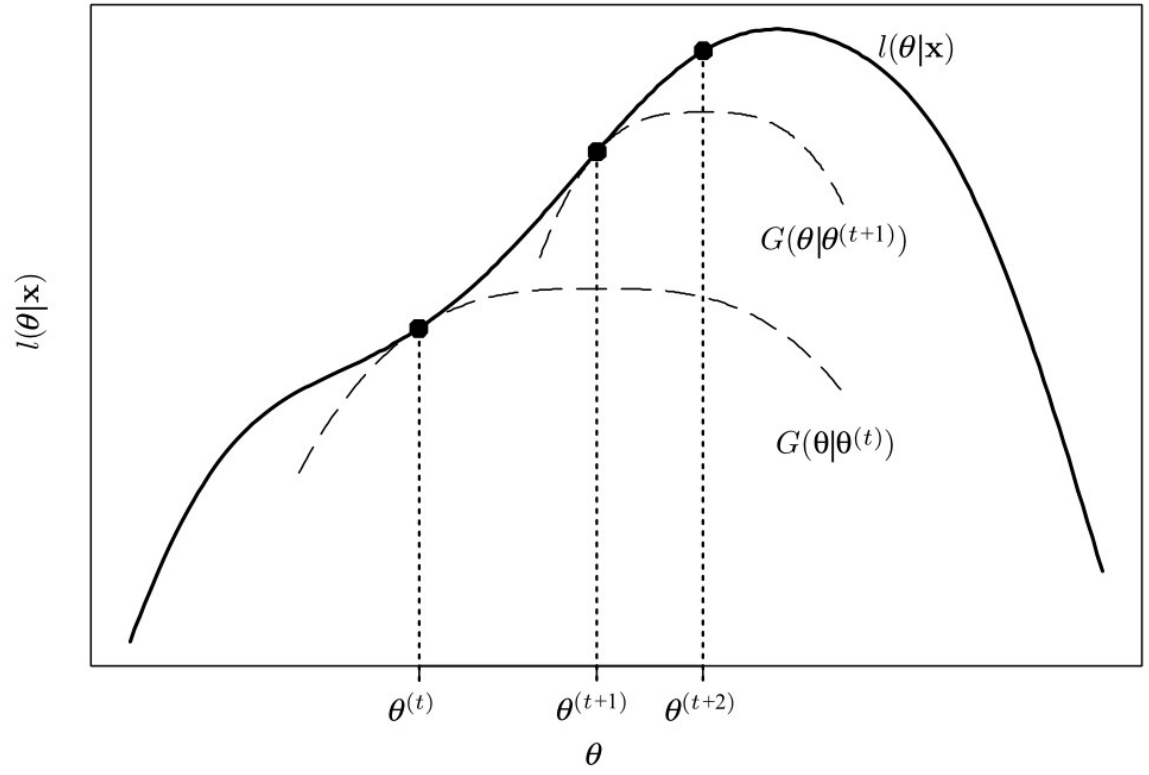


**FIGURE 4.1** One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy.

# Illustration of EM Algorithm

- To further understand how EM works, note that

$$l(\theta|x) \geq Q\left(\theta \mid \theta^{(t)}\right) + l\left(\theta^{(t)} \mid x\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right) = G\left(\theta \mid \theta^{(t)}\right).$$

- Since the last two terms in $G\left(\theta \mid \theta^{(t)}\right)$ are independent of $\theta$, the functions $Q$ and $G$ are maximized at the same $\theta$.
- $G$ is tangent to $l$ at $\theta^{(t)}$ and lies everywhere below $l$. We say that $G$ is a minorizing function for $l$.
- The EM strategy transfers optimization from $l$ to the surrogate function $G$ (effectively to $Q$), which is more convenient to maximize. The maximizer of $G$ provides an increase in $l$.
- Each E step amounts to forming the minorizing function G, and each M step amounts to maximizing it to provide an uphill step.

# Example: Simple Exponential Distribution

- Suppose $Y_1, Y_2 \sim exp(\theta)$ and $y_1 = 5$ is observed but the value $y_2$ is missing.
- The complete-data log likelihood function is

$$\log L(\theta|y) = 2\log\theta - \theta y_1 - \theta y_2.$$

- Because

$$E(Y_2|y_1, \theta^{(t)}) = E(Y_2|\theta^{(t)}) = \frac{1}{\theta^{(t)}},$$

the conditional expectation of $\log L(\theta|Y)$ yields

$$Q(\theta|\theta^{(t)}) = 2\log\theta - 5\theta - \theta/\theta^{(t)}.$$

- The maximizer of $Q(\theta|\theta^{(t)})$ with respect to $\theta$ is easily to found to be the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$.

# Example: Univariate Normal Distribution

Let the complete-data vector $y = (y_1, \cdots, y_n)$ be a random sample from $N(\mu, \sigma^2)$. Then,

$$f(y|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum y_i^2 - 2\mu\sum y_i + n\mu^2\right)\right\}.$$

which implies that $\left(\sum y_i, \sum y_i^2\right)$ are sufficient statistics for $\theta = (\mu, \sigma^2)$. The complete-data log-likelihood function is

$$l(\mu.\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum y_i^2 + \frac{\mu}{\sigma^2}\sum y_i - \frac{n\mu^2}{\sigma^2} + \text{constant}.$$

# Example: Univariate Normal Distribution

Suppose $y_i, i = 1, \cdots, m$ are observed and $y_i, i = m+1, \cdots, n$ are missing. Denote the observed data by $y_{obs} = (y_1, \cdots, y_m)$.

- The E-step requires the computation of

$$E_\theta\left(\sum y_i | y_{obs}\right) \quad \text{and} \quad E_\theta\left(\sum y_i^2 | y_{obs}\right).$$

- At the $t$-th iteration of the E-step, compute

$$s_1^{(t)} = E_\theta\left(\sum y_i | y_{obs}\right) = \sum_{i=1}^{m} y_i + (n-m)\mu^{(t)}$$

$$s_2^{(t)} = E_\theta\left(\sum y_i^2 | y_{obs}\right) = \sum_{i=1}^{m} y_i^2 + (n-m)\left[\sigma^{2(t)} + \mu^{(t)2}\right]$$

# Example: Univariate Normal Distribution

- The complete-data maximum likelihood estimation of $\mu$ and $\sigma^2$ are

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \hat{\sigma^2} = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)^2.$$

- At the $t$-th iteration of the M-step, compute

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{n}$$
$$\sigma^{2(t+1)} = \frac{s_2^{(t)}}{n} - \mu^{(t+1)2}.$$

# Example: Multinomial Distribution

- Define $y = (y_1, y_2, y_3, y_4)$ with multinomial probability $\left(\frac{1}{2} - \frac{1}{2}\theta, \frac{1}{4}\theta, \frac{1}{4}\theta, \frac{1}{2}\right) = (p_1, p_2, p_3, p_4)$.
- Let $y_{obs} = (y_1, y_2, y_3 + y_4) = (38, 34, 125)$ be observed counts from a multinomial population, and then, observed data vector is a function of the complete-data vector.
- Since only $y_3 + y_4$ is observed and $y_3$ and $y_4$ are not, the observed data is considered incomplete.
- The complete-data log-likelihood is

$$l(\theta) = y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3 + y_4 \log p_4 + \text{constant}.$$

where $y_1, y_2, y_3, y_4$ are sufficient statistics.

# Example: Multinomial Distribution

- E-step:

$$E_\theta(y_1|y_{obs}) = y_1 = 38, \quad \text{and} \quad E_\theta(y_2|y_{obs}) = y_2 = 34,$$

$$E_\theta(y_3|y_{obs}) = E_\theta(y_3|y_3 + y_4) = 125\frac{\frac{1}{4}\theta}{\frac{1}{2} + \frac{1}{4}\theta}, \quad y_3^{(t)} = 125\frac{\frac{1}{4}\theta^{(t)}}{\frac{1}{2} + \frac{1}{4}\theta^{(t)}}$$

- M-step: The complete-data maximum likelihood estimate of $\theta$ is

$$\frac{y_2 + y_3}{y_1 + y_2 + y_3}.$$

Then,

$$\theta^{(t+1)} = \frac{34 + y_3^{(t)}}{72 + y_3^{(t)}}.$$

# Example: Peppered Moths

- Three possible alleles: C, I, and T.
- C is dominant to I and T is recessive to I
  - Genotype CC, CI, and CT $\Rightarrow$ carbonaria phenotype.
  - Genotype TT $\Rightarrow$ typica phenotype.
  - Genotype II and TI $\Rightarrow$ intermediate phenotype called insularia.
- If the allele frequencies in the population are $p_C$, $p_I$, and $p_T$, then the genotype frequencies should be $p_C^2$, $2p_Cp_1$, $2p_Cp_T$, $p_I^2$, $2p_Ip_T$, and $p_T^2$ for genotypes CC, CI, CT, II, IT, and TT, respectively. Note that $p_C + p_I + p_T = 1$.
- Capture $n$ moths, of which there are $n_C$, $n_I$, and $n_T$ of the carbonaria, insularia, and typica phenotypes, respectively. Thus, $n = n_C + n_I + n_T$.

# Example: Peppered Moths

- The observed data are $x = (n_C, n_I, n_T)$ and the complete data are $y = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$. The mapping from the complete data to the observed data is $x = M(y) = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$.
- Wish to estimate the allele probabilities, $p_C$, $p_I$, and $p_T$.
- The complete data log likelihood function is multinomial

$$\begin{aligned}\log L(p|y) = {} & n_{CC} \log\left\{p_C^2\right\} + n_{CI} \log\left\{2p_C p_I\right\} + n_{CT} \log\left\{2p_C p_T\right\} \\ & + n_{II} \log\left\{p_I^2\right\} + n_{IT} \log\left\{2p_I p_T\right\} + n_{TT} \log\left\{p_T^2\right\} \\ & + \log \binom{n}{n_{CC}\ n_{CI}\ n_{CT}\ n_{II}\ n_{IT}\ n_{TT}}.\end{aligned}$$

# Example: Peppered Moths

- E-Step

$$E\left\{N_{CC}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CC}^{(t)} = \frac{n_C\left(p_C^{(t)}\right)^2}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\left\{N_{CI}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CI}^{(t)} = \frac{2n_C p_C^{(t)} p_I^{(t)}}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\left\{N_{CT}|n_C, n_I, n_T, p^{(t)}\right\} = n_{CT}^{(t)} = \frac{2n_C p_C^{(t)} p_T^{(t)}}{\left(p_C^{(t)}\right)^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

# Example: Peppered Moths

- E-Step

$$E\left\{N_{II}|n_C, n_I, n_T, p^{(t)}\right\} = n_{II}^{(t)} = \frac{n_I\left(p_I^{(t)}\right)^2}{\left(p_I^{(t)}\right)^2 + 2p_I^{(t)}p_T^{(t)}}$$

$$E\left\{N_{IT}|n_C, n_I, n_T, p^{(t)}\right\} = n_{IT}^{(t)} = \frac{2n_I p_I^{(t)} p_T^{(t)}}{\left(p_I^{(t)}\right)^2 + 2p_I^{(t)}p_T^{(t)}}$$

Then,

$$\begin{aligned} Q(p|p^{(t)}) = & n_{CC}^{(t)}\log\{p_C^2\} + n_{CI}^{(t)}\log\{2p_Cp_I\} + n_{CT}^{(t)}\log\{2p_Cp_T\} \\ & + n_{II}^{(t)}\log\{p_I^2\} + n_{IT}^{(t)}\log\{2p_Ip_T\} + n_{TT}\log\{p_T^2\} + \kappa(n_C, n_I, n_T, p^{(t)}). \end{aligned}$$

# Example: Peppered Moths

- M-Step: Recall $p_T = 1 - p_C - p_I$ and differentiate w.r.t $p_C$ and $p_I$.

$$\frac{dQ(p|p^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} + \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}$$
$$\frac{dQ(p|p^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_C} + \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}$$

  Then,

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}, \qquad p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}$$

- Suppose the observed phenotpye counts are $n_C = 85$, $n_I = 196$, and $n_T = 341$.

# Example: Bayesian Posterior Mode

- Consider a Bayesian problem with likeihood $L(\theta|x)$, prior $\pi(\theta)$, and missing data or parameters $Z$. To find the posterior mode, the E-step requires

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\left[\log\left\{L(\theta|y)\pi(\theta)\kappa(y)\right\}|x,\theta^{(t)}\right] \\ &= E\left\{\log L(\theta|y)|x,\theta^{(t)}\right\} + \log\pi(\theta) + E\left\{\log\kappa(y)|x,\theta^{(t)}\right\}, \end{aligned}$$

  where the final term is a normalizing constant that can be ignored.

- This function $Q$ is obtained by simply adding the log prior to the $Q$ function that would be used in a maximum likelihood setting.
- Unfortunately, the addition of the log prior often makes it more difficult to maximize $Q$ during the M step.

# Variance Estimation

- In maximum likelihood settings, the EM algorithm is used to find an MLE but does not automatically produce an estimate of the covariance matrix of the MLEs.
- Use the asymptotic normality of the MLEs to justify seeking an estimate of the Fisher information matrix. One way to estimate the covariance matrix is to compute the observed information, $-l''(\hat{\theta}|x)$.
- In a Bayesian setting, an estimate of the posterior covariance matrix for $\theta$ can be motivated by noting the asymptotic normality of the posterior.

# Louis's Method

- Taking second partial derivatives of

$$\log f_X(x \mid \theta) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}),$$

and negating both sides yields

$$-l''(\theta \mid x) = -Q''(\theta \mid \omega)\Big|_{\omega=\theta} + H''(\theta \mid \omega)\Big|_{\omega=\theta},$$
$$\hat{i}_X(\theta) = \hat{i}_Y(\theta) - \hat{i}_{Z|X}(\theta),$$

where $\hat{i}_X(\theta) = -l''(\theta \mid x)$ is the observed information, and $\hat{i}_Y(\theta)$ and $\hat{i}_{Z|X}(\theta)$ will be called the complete information and the missing information, respectively.

# Louis's Method

- Interchanging integration and differentiation (when possible), we have

$$\hat{i}_Y(\theta) = -Q''(\theta \mid \omega)\Big|_{\omega=\theta} = -E\left\{l''(\theta \mid Y) \mid x, \theta\right\},$$

  which is reminiscent of the Fisher information.

- The missing-information principle can be used to obtain an estimated covariance matrix for $\hat{\theta}$. It can be shown that

$$\hat{i}_{Z|X}(\theta) = \text{var}\left\{\frac{d \log f_{Z|X}(Z \mid x, \theta)}{d\theta}\right\}$$

  where the variance is taken with respect to $f_{Z|X}$.

# Louis's Method

- Since the expected score is zero at $\hat{\theta}$,

$$\hat{i}_{Z|X}(\hat{\theta}) = \int S_{Z|X}(\hat{\theta}) S_{Z|X}(\hat{\theta})^T f_{Z|X}(z \mid X, \hat{\theta}) dz,$$

  where

$$S_{Z|X}(\theta) = \frac{d \log f_{Z|X}(z \mid x, \theta)}{d\theta}.$$

- The missing-information principle enables us to express $\hat{i}_X(\theta)$ in terms of the complete-data likelihood and the conditional density of the missing data given the observed data, while avoiding calculations involving the complicated marginal likelihood of the observed data.
- If $\hat{i}_Y(\theta)$ or $\hat{i}_{Z|X}(\theta)$ is difficult to compute analytically, it may be estimated via the Monte Carlo methods.

# SEM Algorithm

- The EM algorithm defines a mapping $\theta^{(t+1)} = \Psi(\theta^{(t)})$ where the function $\Psi(\theta) = (\Psi_1(\theta), \cdots, \Psi_p(\theta))$ and $\theta = (\theta_1, \cdots, \theta_p)$. When EM converges, it converges to a fixed point of this mapping, so $\hat{\theta} = \Psi(\hat{\theta})$ with Jacobian matrix $\Psi'(\theta)$ with $(i, j)$-th element equaling $d\Psi_i(\theta)/d\theta_j$. Then,

$$\Psi'\left(\hat{\theta}\right)^T = \hat{i}_{Z|X}\left(\hat{\theta}\right)\hat{i}_Y\left(\hat{\theta}\right)^{-1}.$$

- If we reexpress the missing information principle as

$$\hat{i}_X\left(\hat{\theta}\right) = \left[I - \hat{i}_{Z|X}\left(\hat{\theta}\right)\hat{i}_Y\left(\hat{\theta}\right)^{-1}\right]\hat{i}_Y\left(\hat{\theta}\right),$$

where $I$ is an identity matrix, then inverting $\hat{i}_X\left(\hat{\theta}\right)$ provides the estimate

$$\hat{var}\left\{\hat{\theta}\right\} = \hat{i}_Y\left(\hat{\theta}\right)^{-1}\left(I + \Psi'\left(\hat{\theta}\right)^T\left[I - \Psi'\left(\hat{\theta}\right)^T\right]^{-1}\right).$$

# SEM Algorithm

- This result is appealing in that it expresses the desired covariance matrix as the complete-data covariance matrix plus an incremental matrix that takes account of the uncertainty attributable to the missing data.
- When coupled with the following numerical differentiation strategy to estimate the increment, Meng and Rubin have termed this approach the supplemented EM (SEM) algorithm.
- Since numerical imprecisions in the differentiation approach affect only the estimated increment, estimation of the covariance matrix is typically more stable than the generic numerical differentiation approach.

# Monte Carlo EM

In the Monte Carl EM algorithm, the $t$-th E-step can be replaced with the following two steps:

1. Draw missing datasets $Z_1^{(t)}, \cdots, Z_{m^{(t)}}^{(t)}$ iid from $f_{Z|X}(z|x, \theta^{(t)})$. Each $Z_j^{(t)}$ is a vector of all the missing values needed to compete the observed dataset, so $Y_j = (x, Z_j)$ denotes a complete dataset where the missing values have been replaced by $Z_j$.
2. Calculate

$$\hat{Q}^{(t+1)}(\theta|\theta^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f_Y(Y_j^{(t)}|\theta).$$

Then, $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$ is Monte Carlo estimate of $Q^{(t+1)}(\theta|\theta^{(t)})$. The M-step is modified to maximize $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$.

# ECM Algorithm

- Meng and Rubin's ECM algorithm replaces the M step with a series of computationally simpler conditional maximization (CM) steps.
- Each conditional maximization is designed to be a simple optimization problem that constrains $\theta$ to a particular subspace and permits either an analytical solution or a very elementary numerical solution.
- We call the collection of simpler CM steps after the $t$-th E step a CM cycle. Thus, the $t$-th iteration of ECM is composed of the $t$-th E step and the $t$-th CM cycle.

# ECM Algorithm

- ECM Algorithm
  - Let $S$ denote the total number of CM steps in each CM cycle. For $s = 1, \cdots, S$, the $s$-th CM step in the $t$-th cycle requires the maximization of $Q(\theta|\theta^{(t)})$ subject to (or conditional on) a constraint, say

$$g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

    where $\theta^{(t+(s-1)/S)}$ is the maximizer found in the $(s-1)$th CM step of the current cycle.
  - When the entire cycle of $S$ steps of CM has been completed, we set $\theta^{(t+1)} = \theta^{(t+S/S)}$ and proceed to the E step for the $(t+1)$-th iteration.
- Clearly any ECM is a GEM algorithm, since each CM step increases $Q$.

# ECM Algorithm

- In order for ECM to be convergent, we need to ensure that each CM cycle permits search in any direction for a maximizer of $Q(\theta|\theta^{(t)})$, so that ECM effectively maximizes over the original parameter space for $\theta$ and not over some subspace.
- The art of constructing an effective ECM algorithm lies in choosing the constraints cleverly. Usually, it is natural to partition $\theta$ into $S$ sub-vectors, $\theta = (\theta_1, \cdots, \theta_S)$. Then in the sth CM step, one might seek to maximize $Q$ with respect to $\theta_s$ while holding all other components of $\theta$ fixed.
- This amounts to the constraint induced by the function $g_s(\theta) = (\theta_1, \cdots, \theta_{s-1}, \theta_{s+1}, \cdots, \theta_S)$. A maximization strategy of this type has previously been termed *iterated conditional modes*.

# EM Gradient Algorithm

- If maximization cannot be accomplished analytically, then one might consider carrying out each M step using an iterative numerical optimization approach.
- This would yield an algorithm that had nested iterative loops.
- The ECM algorithm inserts S conditional maximization steps within each iteration of the EM algorithm, also yielding nested iteration.
- To avoid the computational burden of nested looping, Lange proposed replacing the M step with a single step of Newton's method, thereby approximating the maximum without actually solving for it exactly.

# EM Gradient Algorithm

- The M step is replaced with the update given by

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} - Q''\left(\theta \mid \theta^{(t)}\right)^{-1}\Big|_{\theta=\theta^{(t)}} Q'\left(\theta \mid \theta^{(t)}\right)\Big|_{\theta=\theta^{(t)}} \\ &= \theta^{(t)} - Q''\left(\theta \mid \theta^{(t)}\right)^{-1}\Big|_{\theta=\theta^{(t)}} l'\left(\theta^{(t)} \mid \mathbf{x}\right)\end{aligned}$$

  where $l'(\theta^{(t)} \mid x)$ is the evaluation of the score function at the current iterate.

- This EM gradient algorithm has the same rate of convergence to $\hat{\theta}$ as the full EM algorithm.
- In particular, when $Y$ has an exponential family distribution with canonical parameter $\theta$, ascent is ensured.