Missing Data: Theory & Methods

Chapter 4

Advanced Topics of the EM

Contents

Nonparametric EM

- So far we have considered using the EM algorithm to compute the MLE in parametric models with incomplete data.
- The EM can be extended to compute the MLE in nonparametric (and semiparametric) models.
- First we need to think about how MLE works in a nonparametric setting.
- Let $Y_1, \cdots, Y_n$ be an iid sample and $Y_i \sim F$, where F is an unknown cumulative distribution function.
- In the parametric setting, $F'(y) = f(y;\theta)$, where θ is a finite-dimensional parameter indexing the density f, say $f(y;\mu;\sigma^2) = \sigma^{-1}\phi\left(\frac{y-\mu}{\sigma}\right)$, where ϕ is the standard normal density.

Nonparametric EM

- Then the inference about F is essentially the same as the inference about θ, which can be carried out by maximizing the (parametric) likelihood

$$L(D;\theta) = \prod_{i=1}^{n} f(Y_i;\theta).$$

- Now, suppose we do not make any parametric assumptions on F and let it range over all possible (continuous) distribution functions. Denote its density by $f = F'$. Then, the likelihood becomes

$$L(D;f) = \prod_{i=1}^{n} f(Y_i).$$

- It might be tempting to define the MLE for f as

$$\widehat{f} := \arg\max_{f} L(D;f), \quad \text{subject to } \int f(y)dy = 1. \tag{4.1}$$

Nonparametric EM

- However, the optimization problem in (4.1) is an ill posed one, even if f is assumed to be continuous.
- Because $L(D; f)$ depends only on the values of f at the observed points $Y_1, \cdots, Y_n$, it can take arbitrarily high spikes at the Y_i, as long as the widths of the spikes are small enough so that the constraint $\int f(y)dy = 1$ is satisfied.
- One way to circumvent this problem is to add to $-\log L(D; f)$ a penalty function, e.g., $\int f'^2(y)dy$ that penalizes unsmoothness of f and then find the minimizer of the penalized negative log-likelihood (Green and Silverman, 1993). Another way is to use Kernel smoothed methods. The drawback of such approaches is that the amount of smoothness is usually controlled some arbitrarily chosen "tuning parameter".

Nonparametric EM

- Alternatively, one can focus on the distribution function F rather than its density, and restrict the space of F to all discrete distribution functions taking jumps at the Y_i. So that we have a nonparametric likelihood (or empirical likelihood):

$$L(D; F) = \prod_{i=1}^{n} F\{Y_i\},$$

  where $F\{y\} = \Pr(Y = y)$.

- Suppose $y_{(1)} < \cdots < y_{(m)}$ are the ordered distinct values of the Y_i. Write $d_k = F\{y_{(k)}\}$, $k = 1, \cdots, m$. Then the log-likelihood can be written as

$$\sum_{k=1}^{m} n_k \log d_k, \quad \text{subject to } \sum_{k=1}^{m} d_k = 1, \tag{4.2}$$

  where n_k is the number of the Y_i that are equal to $y_{(k)}$, i.e., $n_k = \sum_{i=1}^{n} I(Y_i = y_{(k)})$.

Nonparametric EM

- Thus, the discretized log-likelihood take the form of a multinomial log-likelihood, and it is easy to obtain

$$\widehat{d}_k = \frac{n_k}{n}.$$

- Therefore, the nonparametric MLE (NPMLE) for F is

$$\widehat{F}(y) = \sum_{y_{(k)} \le y} d_k = n^{-1} \sum_{y_{(k)} \le y} \sum_{i=1}^{n} I(Y_i = y_{(k)}) = n^{-1} \sum_{i=1}^{n} I(Y_i \le y),$$

  which is the same as the familiar "empirical distribution" function.

- Note back in (4.2), the n_k, and thus the functions $I(Y_i \le y)$, are sufficient statistics.

Nonparametric EM

Right-censored data

- Univariate coarsened data usually takes the form of censoring (Cox, 1972), which often occurs in medical studies where subjects are followed to register the time to certain event, e.g., death, onset of disease, and etc.
- Because the patient cannot be followed indefinitely till the event occurs, at some point in time the study will be terminated, and those without having the event during follow-up are said to be (right) censored.
- Let Y denote the event time and C denote the study end time, then the observed data consists of $(\delta, Y \wedge C)$, where $\delta = I(Y \leq C)$ and $a \wedge b = \min(a, b)$.
- In missing data terminology, the coarsening rule is

$$Y \to \begin{cases} Y, & \text{if } \delta = 1 \\ (C, \{Y > C\}), & \text{if } \delta = 0. \end{cases}$$

Nonparametric EM

Right-censored data

- One can show that when $C \perp\!\!\!\perp Y$, this coarsening rule is at random (MAR). So that the MLE based on the observed data leads to valid inference on the distribution of Y.
- Because of the independent censoring assumption, all inference can be conditioned upon the censoring times. So, for simplicity, we treat the censoring times as fixed, and use small letter c to denote their values.
- The observed data likelihood based on $(\delta_i, Y_i \wedge c_i)$, $i = 1, \cdots, n$, is

$$\prod_{i=1}^{n} F\{Y_i\}^{\delta_i} \{1 - F(c_i)\}^{1-\delta_i}.$$

- The best known estimator for F in the right censored data is the Kaplan Meier estimator (Kaplan and Meier, 1958), which can shown to be equivalent to the NPMLE.

Nonparametric EM

Right-censored data

- Alternatively, we can use the EM algorithm to compute the NPMLE based on the uncensored full data Y_i.
- Specifically, by inspecting the observed-data likelihood, we see that we need to discretize F at the those Y_i with $\delta_i = 1$.
- Suppose $y_{(1)} < \cdots < y_{(m)}$ are the ordered distinct values of the Y_i with $\delta_i = 1$. Write $d_k = F\{y_{(k)}\}$, $k = 1, \cdots, m$. Now our aim is to estimate the d_k.
- Recall that the full data MLE gives that $\widehat{d_k} = n^{-1} \sum_{i=1}^{n} I(Y_i = y_{(k)})$ and that the $I(Y_i = y_{(k)})$ is the sufficient statistic. Hence the M step at the $(j+1)$th iteration is given by

$$d_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} E[I(Y_i = y_{(k)}) | \delta_i, Y_i \wedge c_i; F^{(j)}].$$

Right-censored data

- Now, the E step. We have that

$$E[I(Y_i = y_{(k)})|\delta_i = 1, Y_i \wedge c_i] = I(Y_i = y_{(k)}),$$

and

$$\begin{aligned} E[I(Y_i = y_{(k)})|\delta_i = 0, Y_i \wedge c_i] &= E[I(Y_i = y_{(k)})|Y_i > c_i] \\ &= I(c_i < y_{(k)}) \frac{d_k^{(j)}}{1 - F^{(j)}(c_i)}, \end{aligned}$$

where $F^{(j)}(y) = \sum_{y_{(k)} \leq y} d_k$.

- Hence,

$$d_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} \left\{ \delta_i I(Y_i = y_{(k)}) + (1 - \delta_i) I(c_i < y_{(k)}) \frac{d_k^{(j)}}{\sum_{y_{(l)} > c_i} d_l^{(j)}} \right\}. \quad (4.3)$$

Right-censored data

- An alternative "functional" form of the iteration can be derived as follows.
- Note that by the form of full-data NPMLE, we have

$$F^{(j+1)}(y) = n^{-1} \sum_{i=1}^{n} E[I(Y_i \le y) | \delta_i, Y_i \wedge c_i; F^{(j)}]$$
$$= n^{-1} \sum_{i=1}^{n} \left\{ \delta_i I(Y_i \le y) + (1-\delta_i) I(c_i < y) \frac{F^{(j)}(y) - F^{(j)}(c_i)}{1 - F^{(j)}(c_i)} \right\}.$$

- Clearly, this is equivalent to (4.3), which is more useful in actual implementation of the algorithm.

Nonparametric EM

Interval-censored data

- Interval-censored data usually arise if the event of interest is asymptomatic so that the investigators conduct periodic tests to find out whether it's occurred or not.
- When each subject has only one test, the resulting data are called **current status** data.
- Again, the full data of interest are the underlying event times $Y_1, \cdots, Y_n$. For the ith subject, denote the test time as u_i, and then the test result is indicated by $\delta_i \equiv I(Y_i \leq u_i)$. So the observed data consist of

$$(\delta_i, u_i), \;\; i = 1, \cdots, n.$$

- Unlike right-censored data, we never observe Y_i exactly (even if $Y_i \leq u_i$).

Nonparametric EM

Interval-censored data

- The observed data likelihood is
$$\prod_{i=1}^{n} F(u_i)^{\delta_i} \{1 - F(u_i)\}^{1-\delta_i}.$$
- Note that the likelihood needs no "discretization" as it contains no density function. This is the result of the Y_i being never exactly observed.
- Also note that the likelihood depends on F only through the $F(u_i)$. Write $u_{(1)} < \cdots < u_{(m)}$ as the distinct ordered values of the u_i. Then, the function F should satisfy the monotonicity constraint $0 \leq F(u_{(1)}) \leq \cdots \leq F(u_{(m)}) \leq 1$, or, $F(u_{(k)}) - F(u_{(k-1)}) \geq 0$, $k = 1, \cdots, m, m+1$, where $u_{(0)} = 0$ and $u_{(m+1)} = \infty$.
- At those $u_{(k)}$ that correspond to u_i with $\delta_i = 0$, increasing $F(u_{(k)})$ serves only to decrease to likelihood. So from the MLE perspective, we must have $F(u_{(k)}) = F(u_{(k-1)})$.

Interval-censored data

- So we might as well re-define $u_{(1)} < \cdots < u_{(m)}$ as the distinct ordered values of the u_i with $\delta_i = 1$.
- A direct approach to computing the NPMLE based on the observed-data likelihood is to treat it as an isotonotic regression problem (Greeneboom and Wellner, 1992).
- Here we use the EM based on the full data. Because we can only "identify" F at $u_{(1)}, \cdots, u_{(m)}$, the distribution of the probability mass $d_k := F(u_{(k)}) - F(u_{(k-1)})$ on $(u_{(k-1)}, u_{(k)}]$ is irrelevant. For simplicity, we put it on the right end. That is, we let $d_k = \Pr(Y = u_{(k)})$.
- Similar to the right-censored data, the M step at the $(j+1)$th iteration is given by

$$d_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} E[I(Y_i = y_{(k)}) | \delta_i, u_i; F^{(j)}].$$

Interval-censored data

- For the E step, we have that

$$E[I(Y_i = u_{(k)})|\delta_i = 1, u_i; F^{(j)}] = I(u_i \geq u_{(k)}) \frac{d_k^{(j)}}{F^{(j)}(u_i)},$$

and

$$E[I(Y_i = u_{(k)})|\delta_i = 0, u_i; F^{(j)}] = I(u_i < u_{(k)}) \frac{d_k^{(j)}}{1 - F^{(j)}(u_i)}.$$

- Hence,

$$d_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} \left\{ \delta_i I(u_i \geq u_{(k)}) \frac{d_k^{(j)}}{F^{(j)}(u_i)} + (1 - \delta_i) I(u_i < u_{(k)}) \frac{d_k^{(j)}}{1 - F^{(j)}(u_i)} \right\}.$$

Interval-censored data

- Similarly, for a functional form of the EM, we have that

$$F^{(j+1)}(y) = n^{-1} \sum_{i=1}^{n} \Big\{ \delta_i I(u_i \geq y) \frac{F^{(j)}(u_i) - F^{(j)}(y)}{F^{(j)}(u_i)} + (1-\delta_i) I(u_i < y) \frac{F^{(j)}(y) - F^{(j)}(u_i)}{1 - F^{(j)}(u_i)} \Big\}.$$

- The algorithm can be generalized to arbitrary interval-censored data arising from multiple periodic tests .
- Suppose Y_i is interval censored such that we only know that it lies between an interval (l_i, r_i). (In the current status case, if $\delta_1 = 1$, then $l_i = 0$ and $r_i = u_i$; if $\delta_i = 0$, then $l_i = u_i$ and $r_i = \infty$).

Nonparametric EM

Interval-censored data

- Then, a similar derivation shows that the EM step for the NPMLE is given by

$$F^{(j+1)}(y) = n^{-1} \sum_{i=1}^{n} \frac{F^{(j)}(y \wedge r_i) - F^{(j)}(y \wedge l_i)}{F^{(j)}(r_i) - F^{(j)}(l_i)}.$$

- One can similarly derive a corresponding "algebraic form".

Nonparametric EM

A mouse tumorigenicity study

- Consider an example from a tumorigenicity study described in Hoel and Walburg (1972).
- One hundred and forty-four RFM mice were assigned to either a germ-free or a conventional environment, and were sacrificed after several years to detect the presence of lung tumor.
- The purpose of this study is to compare the time from beginning of the study until the time to observe a tumor.
- Since lung tumors are nonlethal and cannot be observed before death in RFM mice, it is appropriate to treat this data as current status data.
- We use the EM algorithm to estimate the survival functions for the two groups of mice separately.

A mouse tumorigenicity study

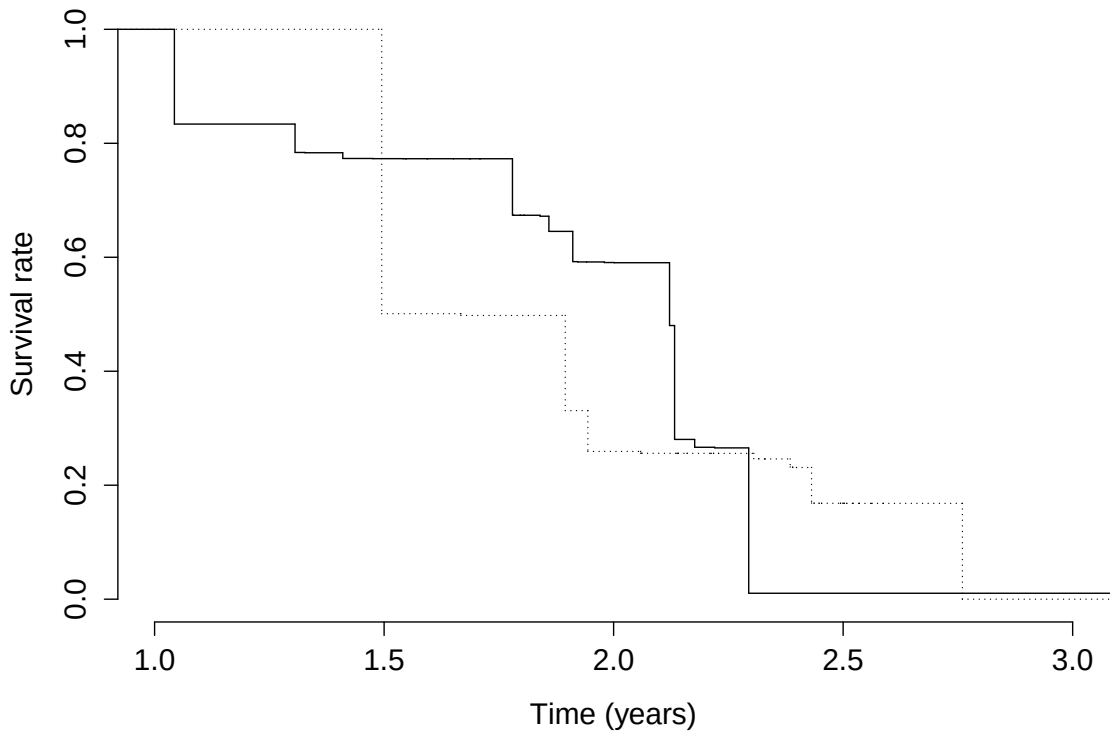

Figure 4.1: Nonparametric maximum likelihood estimates for the survival functions for tumor occurence in mice. Solid: germ-free; dotted: conventional.

Contents

Supplemented EM

- For variance calculation, the Louis formula is a convenient method for models with fair complexity.
- However, the method involves extra analytical calculations that may not be feasible for more complex problems.
- Meng and Rubin (1991) proposed a supplemented EM algorithm (SEM) that provides a numerical approximation to the information as a bi-product of the EM calculations themselves.
- This approach is based on the derivative of the mapping defined by the iterations of the EM algorithm.

Supplemented EM

- Define $\mathcal{M}(\theta)$ to be the value of θ^* that maximizes $Q(\theta^*|\theta)$, i.e.,

$$\mathcal{M}(\theta) = \arg\max_{\theta^*} Q(\theta^*|\theta).$$

- Dempster, Laird and Rubin (1977) observed that this the derivative of $\mathcal{M}(\theta)$ satisfies

$$\frac{\partial}{\partial \theta^{\mathrm{T}}}\mathcal{M}(\widehat{\theta}) = -\frac{\partial^2 H(\theta|\widehat{\theta})}{\partial \theta^{\otimes 2}}\bigg|_{\theta=\widehat{\theta}} \left\{ -\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial \theta^{\otimes 2}}\bigg|_{\theta=\widehat{\theta}} \right\}^{-1}, \qquad (4.4)$$

  where

$$\frac{\partial^2 H(\theta|\widehat{\theta})}{\partial \theta^{\otimes 2}} = E\left[\frac{\partial^2}{\partial \theta^{\otimes 2}} \log L(D_{\mathsf{mis}}|D_{\mathsf{obs}};\theta) \Big| D_{\mathsf{obs}};\widehat{\theta} \right].$$

- Justification of (4.4) is relegated to the Addendum.

Supplemented EM

- Recall from §2 that

$$-\left.\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial\theta^{\otimes 2}}\right|_{\theta=\widehat{\theta}} = I(D;\widehat{\theta})$$

  represents the full data information and $-\frac{\partial^2 H(\theta|\widehat{\theta})}{\partial\theta^{\otimes 2}}|_{\theta=\widehat{\theta}}$ is the missing data information.

- By (2.2), we have that

$$\begin{aligned} I(D_{\mathsf{obs}};\widehat{\theta}) &= -\left.\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial\theta^{\otimes 2}}\right|_{\theta=\widehat{\theta}} + \left.\frac{\partial^2 H(\theta|\widehat{\theta})}{\partial\theta^{\otimes 2}}\right|_{\theta=\widehat{\theta}} \\ &= \left\{ I - \frac{\partial}{\partial\theta^{\mathrm{T}}}\mathcal{M}(\widehat{\theta}) \right\} \left\{ -\left.\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial\theta^{\otimes 2}}\right|_{\theta=\widehat{\theta}} \right\}, \end{aligned}$$

  where I is the identity matrix.

Supplemented EM

- Thus if $\frac{\partial}{\partial \theta^{\mathrm{T}}}\mathcal{M}(\widehat{\theta})$ can be estimated and $-\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial \theta^{\otimes 2}}|_{\theta=\widehat{\theta}}$ computed, then the observed information can be estimated without further computation.
- In the SEM algorithm, the derivatives in $\frac{\partial}{\partial \theta^{\mathrm{T}}}\mathcal{M}(\widehat{\theta})$ are approximated with numerical differences.
- Specifically, set $\theta(k) = (\widehat{\theta}_1, \cdots, \widehat{\theta}_{k-1}, \theta_k, \widehat{\theta}_{k+1}, \cdots, \widehat{\theta}_p)^{\mathrm{T}}$, where θ_k is some value close to $\widehat{\theta}_k$.
- Use the M step of the EM algorithm to compute $\mathcal{M}(\theta(k))$, that is,

$$\mathcal{M}(\theta(k)) = \arg\max_{\theta^*} Q(\theta^*|\theta(k)).$$

- Then, approximate the kth component of $\frac{\partial}{\partial \theta^{\mathrm{T}}}\mathcal{M}(\widehat{\theta})$ by

$$\frac{\mathcal{M}(\theta(k))^{\mathrm{T}} - \mathcal{M}(\widehat{\theta})^{\mathrm{T}}}{\theta_k - \widehat{\theta}_k}.$$

Supplemented EM

- Since at the exact solution,$\mathcal{M}(\widehat{\theta}) = \widehat{\theta}$, this substitution could also be made in the previous expression.
- However, as Gray (2002) noted, usually the EM algorithm is terminated a little short of the exact maximizer, so $\mathcal{M}(\widehat{\theta})$ will not be exactly equal to $\widehat{\theta}$, and a small difference between $\mathcal{M}(\widehat{\theta})$ and $\widehat{\theta}$ can affect the accuracy of the results.

Supplemented EM

- The remaining question is how to choose $\theta \equiv (\theta_1, \cdots, \theta_p)^{\mathrm{T}}$ for the differences.
- Numerical differences as approximations to derivatives can be quite sensitive to the size of the difference used.
- Meng and Rubin (1991) proposed using some $\theta^{(j_0)}$ along the way of the EM iterations.
- That is, the kth component of $\frac{\partial}{\partial \theta^{\mathrm{T}}} \mathcal{M}(\widehat{\theta})$ is approximated by

$$\frac{\mathcal{M}(\theta^{(j_0)}(k))^{\mathrm{T}} - \mathcal{M}(\widehat{\theta})^{\mathrm{T}}}{\theta_k^{(j_0)} - \widehat{\theta}_k},$$

where $\theta^{(j_0)}(k) = (\widehat{\theta}_1, \cdots, \widehat{\theta}_{k-1}, \theta_k^{(j_0)}, \widehat{\theta}_{k+1}, \cdots, \widehat{\theta}_p)^{\mathrm{T}}$.

Contents

Expectation/Conditional Maximization (ECM)

- In many applications, the M step is difficult and would require iterative search methods for finding the maximum.
- Recall from §2.1 though that generally it is not necessary to find the exact maximum in the M step for the EM algorithm to result in a monotonic increase of observed data log-likelihood.
- In applications where iterative search methods are needed, sometimes subsets of the parameters can be maximized much more easily than the full parameter vector.
- Suppose that $\theta = (\theta_1, \theta_2)$, and that $Q(\theta_1, \theta_2 | \theta_1^{(j)}, \theta_2^{(j)})$ is easy to maximize over either θ_1 or θ_2 with the other held fixed, but that jointly maximizing over both is more difficult.

Expectation/Conditional Maximization (ECM)

- In these settings Meng and Rubin (1993) proposed using a generalization of the EM algorithm called the Expectation/Conditional Maximization (ECM) algorithm.
- Given the parameter values from the previous iteration (or the initial values), the algorithm proceeds by computing $Q(\theta_1, \theta_2 | \theta_1^{(j)}, \theta_2^{(j)})$ in the E-step as before.
- Next, $Q(\theta_1, \theta_2^{(j)} | \theta_1^{(j)}, \theta_2^{(j)})$ is maximized over θ_1 to obtain $\theta_1^{(j+1)}$. Then $Q(\theta_1^{(j+1)}, \theta_2 | \theta_1^{(j)}, \theta_2^{(j)})$ is maximized over θ_2 to obtain $\theta_2^{(j+1)}$.
- Then the algorithm returns to the E-step, computing the expectation at the new parameter values $(\theta_1^{(j+1)}, \theta_2^{(j+1)})$, with the steps repeated until convergence.

Expectation/Conditional Maximization (ECM)

- By definition,

$$Q(\theta_1^{(j+1)}, \theta_2^{(j+1)} | \theta_1^{(j)}, \theta_2^{(j)}) \geq Q(\theta_1^{(j+1)}, \theta_2^{(j)} | \theta_1^{(j)}, \theta_2^{(j)})$$
$$\geq Q(\theta_1^{(j)}, \theta_2^{(j)} | \theta_1^{(j)}, \theta_2^{(j)}).$$

  So, by the proof of Proposition 2.2, the observed-data log-likelihood is always increasing.
- The two maximizations within the M step could be iterated, to get improved M-step estimates before returning to the next E-step (often such iteration would converge to the joint maximizers of $Q(\theta_1, \theta_2 | \theta_1^{(j)}, \theta_2^{(j)})$, in which case this algorithm would be an EM algorithm with a special type of computations in the M step).
- However, since the EM algorithm often requires many iterations through both the E and M steps, there is often little advantage to further iteration within each M step.

Expectation/Conditional Maximization (ECM)

- The phrase "conditional maximization" in ECM is used to denote the process of maximizing over a subset of the parameters with the other parameters held fixed.
- Note that it should not be misunderstood as maximizing the likelihood for a conditional distribution. All that is meant by "conditional" is fixing the values of a subset of the parameters.
- The algorithm as described above has an obvious generalization to settings where there are three or more different subsets of parameters, each of which is easy to maximize when the others are held fixed.

Expectation/Conditional Maximization (ECM)

Example: Multivariate normal regression with incomplete response

- Suppose the complete data response vectors Y_i are independent with

$$Y_i|Z_i \sim N(Z_i\beta, V), \qquad i = 1, \cdots, n,$$

  where $Y_{i1}, \cdots, Y_{ik}$, Z_i is a $k \times p$ matrix of covariates (usually including a constant term), β is a p-vector of unknown parameters, and the covariance matrix V is only required to be positive definite (which means it has $k(k+1)/2$ unknown parameters).
- Suppose that Y_i is incomplete for some cases. For example, the components of Y_i could be longitudinal measurements, and some subjects may drop out before completing all measurements. But missinness could also be nonmonotone (see §3.4).
- The missingness mechanism is assumed to be MAR.

Expectation/Conditional Maximization (ECM)

Example: Multivariate normal regression with incomplete response

- Let $\theta = (\beta, V)$. The full-data log-likelihood is

$$\begin{aligned} l_n(D;\theta) &= -\frac{n}{2}\log\det V - \frac{1}{2}\sum_{i=1}^{n}(Y_i - Z_i\beta)^{\mathrm{T}}V^{-1}(Y_i - Z_i\beta) \\ &= -\frac{n}{2}\log\det V - \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}(Y_i - Z_i\beta)^{\otimes 2}V^{-1}\right). \end{aligned}$$

- Denote the observed response vector for the ith subject as $M_i(Y_i)$.
- The E step at the $(j+1)$th iteration computes

$$Q(\beta, V|\theta^{(j)}) = -\frac{n}{2}\log\det V - \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}\left(\widehat{C}_i^{(j)} + (\widehat{Y}_i^{(j)} - Z_i\beta)^{\otimes 2}\right)V^{-1}\right), \tag{4.5}$$

Expectation/Conditional Maximization (ECM)

Example: Multivariate normal regression with incomplete response

where $\widehat{Y}_i^{(j)} = E[Y_i|M_i(Y_i), Z_i; \theta^{(j)}]$, and $\widehat{C}_i^{(j)} = \mathrm{Var}[Y_i|M_i(Y_i), Z_i; \theta^{(j)}]$.

- Maximization of $Q(\beta, V|\theta^{(j)})$ jointly over β and V is hard, generally requiring iterative methods.
- However, it is easy to maximize $Q(\beta, V|\theta^{(j)})$ over β for fixed V, and *vice versa*, using standard techniques.
- So, the conditional maximization proceeds as follows.
- First

$$\begin{aligned}\beta^{(j+1)} &= \arg\max_{\beta} Q(\beta, V^{(j)}|\theta^{(j)}) \\ &= \left(\sum_{i=1}^{n} Z_i^{\mathrm{T}} {V^{(j)}}^{-1} Z_i\right)^{-1} \sum_{i=1}^{n} Z_i^{\mathrm{T}} {V^{(j)}}^{-1} \widehat{Y}_i^{(j)}.\end{aligned}$$

Expectation/Conditional Maximization (ECM)

Example: Multivariate normal regression with incomplete response

- Then,

$$V^{(j+1)} = \arg\max_{V} Q(\beta^{(j+1)}, V|\theta^{(j)})$$
$$= n^{-1} \sum_{i=1}^{n} \left(\widehat{C}_i^{(j)} + (\widehat{Y}_i^{(j)} - Z_i \beta^{(j+1)})^{\otimes 2} \right).$$

- Although joint maximizers are not computed at each iteration, generally this algorithm will still converge to the maximizers of the observed-data likelihood.
- On the other hand, direct maximization of the observed-data likelihood would require an iterative search with possibly lots of determinant and matrix inversion calculations within each step.

Concluding Remarks about EM Algorithm

- The EM algorithm provides a convenient framework to approach estimation in incomplete data problems.
- It is widely used in the statistical literature, and there are many other extensions and variations than those discussed above.
- For example, Liu and Rubin (1994) extend the ECM algorithm to an ECME algorithm, where in some of the sub-steps the observed-data likelihood is directly maximized over a subset of the parameters. Other extensions are discussed in Meng and van Dyk (1997).

Concluding Remarks about EM Algorithm

- The EM algorithm is most useful when
  1. Computing the conditional expectation of the full-data log-likelihood is easier than directly computing the observed-data likelihood;
  2. Maximizing $Q(\theta|\theta^{(j)})$, or at least finding sufficiently improved values of θ through the ECM algorithm or other means, is sufficiently fast.
- Since the EM algorithm usually converges at a fairly slow rate (linear as opposed to the quadratic rate of Newton-Raphson), individual iterations have to be fast or it will not be competitive with directly maximizing the observed data likelihood.
- Another important reason for using the EM algorithm is to compute MLEs from incomplete data using available software for fitting full data.

Concluding Remarks about EM Algorithm

- As has been seen in the foregoing chapters, in many problems the M step involves full-data estimates computed based on modified weights.
- So, provided the E step calculations can be programmed, it is then often straightforward to compute the steps in the EM algorithm without much additional work.
- Admittedly, in some (probably many) incomplete data problems, the EM algorithm is less advantageous as compared to the competing methods such as multiple imputation and Bayesian inference, particularly when the E step is hard to program and requires Monte-Carlo integration.
- However, the framework of EM is still attractive in that it is built under a strictly frequentist paradigm.

References

- Cox, D. R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society: Series B, 34, 187-220.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1-38.
- Green, P. J. and Silverman, B. W. (1993). Nonparametric regression and generalized linear models: a roughness penalty approach. CRC Press.
- Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. New York: Springer.
- Hoel, D. G. and Walburg, H. E. (1972). Statistical analysis of survival experiments. Journal of the National Cancer Institute, 49, 361-372.
- Kaplan, E. L. andMeier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.
- Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika, 633-648.
- Meng, X. L. andRubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association, 86, 899-909.
- Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80, 267-278.
- Meng, X. L. and Van Dyk, D. (1997). The EM Algorithm: an Old FolkâĂŘsong Sung to a Fast New Tune. Journal of the Royal Statistical Society: Series B, 59, 511-567.