Appealing Properties of the EM Algorithm

↳ Difficult in derivation but easily estimate

- 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.

One-dimensional Illustration of EM Algorithm

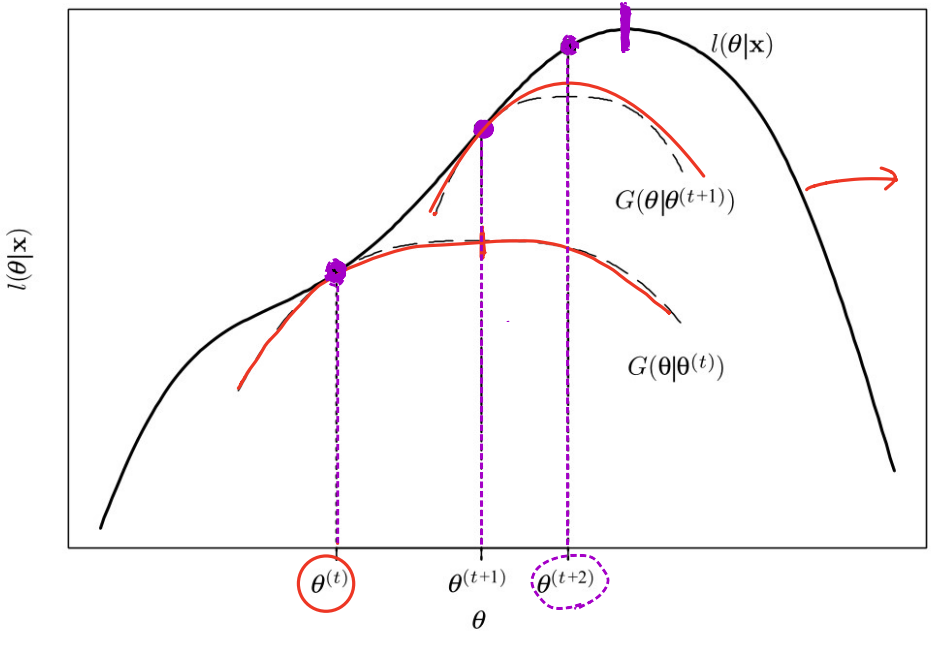

FIGURE 4.1 One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy.

Convergence

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

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

observed complete cond z|x.

Therefore,

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

$Q(\theta|\theta^{(t)})$ $H(\theta|\theta^{(t)})$

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 θ when $\theta = \theta^{(t)}$. Write

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

$$= \int -\log\left[\frac{f_{Z|X}(z \mid x, \theta)}{f_{Z|X}(z \mid x, \theta^{(t)})}\right] f_{Z|X}(z \mid x, \theta^{(t)}) dz$$

$$\geq -\log \int f_{Z|X}(z \mid x, \theta) dz = 0$$

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 θ, 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

portion of missing = 1 − portion of observed

- 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.

Mixed-effect Model

↳ coefficient : fixed effect / random effect

$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$ $\quad b_i, \varepsilon_i$ indep.

β : fixed effect (Object : estimate β)

b_i : random effect → Object : estimate variance of b_i = D

$b_i \sim N_q(0, D)$ $\quad \varepsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$

Observed-data log-likelihood

$$\ell(\beta, D, \sigma^2 \mid Y_1, \cdots, Y_n) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}(Y_i - X_i\beta)^T \Sigma_i^{-1} (Y_i - X_i\beta) - \frac{1}{2}\log|\Sigma_i| \right\}$$

where $\Sigma_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$

likelihood can be directly maximized for (β, D, σ^2) by using Newton-Raphson or Fisher scoring algorithm.

Given (D, σ^2) and hence Σ_i, we obtain β that maximizes the likelihood by solving

$$\frac{\partial \ell(\beta, D, \sigma^2 \mid Y_1, \cdots, Y_n)}{\partial \beta} = \sum_{i=1}^{n} X_i^T \Sigma_i^{-1} (Y_i - X_i\beta) = 0$$

$$\Rightarrow \hat{\beta} = \left(\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^T \Sigma_i^{-1} Y_i$$

↳ difficult to maximize. D. σ^2.

Complete-data log likelihood ← Need to use EM for D. σ^2.

ε_i. b_i indep. ⇝ Make multivariate normal distribution.

$$- \begin{pmatrix} b_i \\ \varepsilon_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D & 0 \\ 0 & \sigma^2 I_{n_i} \end{pmatrix} \right)$$

$$l_c(\beta, D, \sigma^2 \mid \varepsilon_1, \cdots, \varepsilon_N, b_1, \cdots, b_N)$$

$$= \sum_{i=1}^{n} \left(\underbrace{-\frac{1}{2} b_i^T D b_i - \frac{1}{2} \log |D|}_{\text{Normal for } b_i} \underbrace{- \frac{1}{2\sigma^2} \varepsilon_i \varepsilon_i^T - \frac{n_i}{2} \log \sigma^2}_{\text{Normal for } \varepsilon_i} \right)$$

The parameter that maximize the complete-data log-likelihood is obtained as, conditional on other parameters,

$$D = \frac{1}{N} \sum_{i=1}^{n} b_i b_i^T$$

$$\sigma^2 = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{n_i} \varepsilon_i^T \varepsilon_i$$

$$\beta = \left(\sum_{i=1}^{n_i} X_i^T X_i \right)^{-1} \sum_{i=1}^{N_i} X_i^T (Y_i - Z_i b_i)$$

$\downarrow$

Y_i - random effect

Form of EX^2 : $\quad \text{Var } X = EX^2 - (EX)^2$

$EX^2 = \text{Var } X + (EX)^2$

E-Step

- $E(b_i b_i^T \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$
- $E(\varepsilon_i^T \varepsilon_i \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$
- $E(b_i \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$

$E(b_i b_i^T \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$

$$E(b_i b_i^T \mid Y_i) = \text{Var}(b_i \mid Y_i) + E(b_i \mid Y_i) E(b_i^T \mid Y_i)$$

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} \sim \text{MVN} \left\{ \begin{pmatrix} X_i \beta \\ 0 \end{pmatrix}, \begin{pmatrix} \underbrace{Z_i D Z_i^T + \sigma^2 I_{n_i}}_{=\Sigma_i} & Z_i D \\ D Z_i^T & D \end{pmatrix} \right)$$

Let $\Sigma_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$

$$E(b_i \mid Y_i) = 0 + D Z_i^T \Sigma_i^{-1} (Y_i - X_i \beta) = D Z_i^T \Sigma_i^{-1} (Y_i - X_i \beta)$$

$$\text{Var}(b_i \mid Y_i) = D - D Z_i^T \Sigma_i^{-1} Z_i D$$

$$E(b_i b_i^T \mid Y) = \left(D - D Z_i^T \Sigma_i^{-1} Z_i D \right) + \left(D Z_i^T \Sigma_i^{-1} (Y_i - X_i \beta) \right) \left(D Z_i^T \Sigma_i^{-1} (Y_i - X_i \beta) \right)^T$$

Similarly, we use the relationship

$$E(\varepsilon_i^T \varepsilon_i \mid Y_i) = E(\varepsilon_i^T \mid Y_i) E(\varepsilon_i \mid Y_i) + \text{Var}(\varepsilon_i \mid Y_i)$$

$$\begin{pmatrix} Y_i \\ \varepsilon_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_i & \sigma I_{n_i} \\ \sigma I_{n_i} & \sigma^2 I_{n_i} \end{pmatrix} \right)$$

$$E(\varepsilon_i \mid Y_i) = 0 + \sigma \Sigma_i^{-1} (Y_i - X_i\beta).$$

$$Var(\varepsilon_i \mid Y_i) = \sigma^2 I_{n_i} - \sigma^2 \Sigma_i^{-1}.$$

M-Step $D^{(t+1)} = \frac{1}{N} \sum_i^n E(b_i b_i^T \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$

$$\sigma^{2(t+1)} = \Big/ \sum_{i=1}^n n_i \sum_{i=1}^n E(\varepsilon_i^T \varepsilon_i \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$$

$$\beta^{(t+1)} = \left(\sum_{i=1}^n X_i^T X_i \right)^{-1} \sum_{i=1}^n X_i^T E(Y_i - Z_i b_i \mid Y_i, \beta^{(t)}, D^{(t)}, \sigma^{2(t)})$$

(EX) Measurement Error Model

Consider the problem of estimating parameter.

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$X_i \sim N(\mu_x, \sigma_x^2)$ $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

↳ ε_i is independent of x_i

In simple regression, we observe (x_i, y_i)

Instead of observing (x_i, y_i), suppose we observe (z_i, y_i) → include measure error.

in the sample such that $f(y \mid x, z) = f(y \mid x)$

Assume that the cond. dist $g(z \mid x) \sim N(x, \sigma_u^2)$

given (known)

$$f(x \mid z, y) \propto f(x \mid z) f(y \mid x, z)$$

$$\propto f(x) \cdot f(z \mid x) f(y \mid x, z)$$

$N(\mu_x, \sigma_x^2)$ $N(x, \sigma_u^2)$ → $N(\beta_0 + \beta_1 x, \sigma_\varepsilon^2)$

$$\propto \exp\left(-\frac{1}{2\sigma_x^2} (x - \mu_x)^2 \right) \exp\left(-\frac{1}{2\sigma_u^2} (z - x)^2 \right) \exp\left(-\frac{\beta_1^2}{2\sigma_u^2} (x - \beta_1^{-1} y + \beta_0 \beta_1^{-1})^2 \right)$$

$$E(x \mid z, y) = \frac{\mu_x/\sigma_x^2 + z/\sigma_u^2 + \beta_1(y-\beta_0)/\sigma_e^2}{1/\sigma_x^2 + 1/\sigma_u^2 + \beta_1^2/\sigma_e^2}$$

$$Var(x \mid z, y) = \left(1/\sigma_x^2 + 1/\sigma_u^2 + \beta_1^2/\sigma_e^2\right)^{-1}$$

E - Step

$$E(x \mid z, y; \theta^{(t)}) = \frac{\mu_x^{(t)}/\sigma_x^{2(t)} + z/\sigma_u^2 + \beta_1^{(t)}(y-\beta_0^{(t)})/\sigma_e^{2(t)}}{1/\sigma_x^{2(t)} + 1/\sigma_u^2 + \beta_1^{(t)2}/\sigma_e^{2(t)}} = \hat{x}^{(t)}$$

$$E(x^2 \mid z, y; \theta^{(t)}) = \left(\hat{x}^{(t)}\right)^2 + \left(1/\sigma_x^{2(t)} + 1/\sigma_u^2 + \beta_1^{(t)2}/\sigma_e^{2(t)}\right)^{-1}$$

M - Step

$$\mu_x^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{x}^{(t)}$$

$$\sigma_x^{2(t+1)} = \frac{1}{n}\sum_{i=1}^{n} E(x^2 \mid z, y; \theta^{(t)}) - \left(\mu_x^{(t+1)}\right)^2$$

$$\beta_1^{(t+1)} = 1/\sigma_x^{2(t+1)}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^{(t)} y_i - \mu_x^{(t+1)} \underbrace{\hat{\mu}_y}_{=\frac{1}{n}\sum_{i=1}^{n} y_i}\right)$$

$$\beta_0^{(t+1)} = \hat{\mu}_y - \beta_1^{(t+1)} \mu_x^{(t+1)}$$

$$\sigma_e^{2(t+1)} = \hat{\sigma}_y^2 - \left(\beta_1^{(t+1)}\right)^2 \sigma_x^{2(t+1)}$$

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

likelihood x prior x marginal density of y → normalizing constant.

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

prior. ignore. complete-data likelihood

  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 θ 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

EM algorithm. ⊢ Improving E-Step MCEM.
⊢ Improving M-Step. ECM / EM Gradient.

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

$Y = (X, Z)$ ↳ missing → impute. (multiple imputation)

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

Improving M-Step

maximization.
multiple parameter (high-dimension)
↳ can make cond. density

- 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 θ 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 θ 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 θ 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 θ_s while holding all other components of θ 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 θ, ascent is ensured.