in which such difficulties arise is given in Problem 2.4.13. Many examples and important issues and methods are discussed, for instance, in Chapter 6 of Dahlquist, Björk, and Anderson (1974).

2.4.4 The EM (Expectation/Maximization) Algorithm

There are many models that have the following structure. There are ideal observations, $X \sim P_\theta$ with density $p(x,\theta)$, $\theta \in \Theta \subset R^d$. Their log likelihood $l_{p,x}(\theta)$ is "easy" to maximize. Say there is a closed-form MLE or at least $l_{p,x}(\theta)$ is concave in θ. Unfortunately, we observe $S \equiv S(X) \sim Q_\theta$ with density $q(s,\theta)$ where $l_{q,s}(\theta) = \log q(s,\theta)$ is difficult to maximize; the function is not concave, difficult to compute, and so on. A fruitful way of thinking of such problems is in terms of S as representing part of X, the rest of X is "missing" and its "reconstruction" is part of the process of estimating θ by maximum likelihood. The algorithm was formalized with many examples in Dempster, Laird, and Rubin (1977), though an earlier general form goes back to Baum, Petrie, Soules, and Weiss (1970). We give a few examples of situations of the foregoing type in which it is used, and its main properties. For detailed discussion we refer to Little and Rubin (1987) and MacLachlan and Krishnan (1997). A prototypical example follows.

Example 2.4.4. *Lumped Hardy–Weinberg Data.* As in Example 2.2.6, let X_i, $i = 1, \dots, n$, be a sample from a population in Hardy–Weinberg equilibrium for a two-allele locus, $X_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})$, where $P_\theta[X = (1,0,0)] = \theta^2$, $P_\theta[X = (0,1,0)] = 2\theta(1-\theta)$, $P_\theta[X = (0,0,1)] = (1-\theta)^2$, $0 < \theta < 1$. What is observed, however, is not $\mathbf{X}$ but $\mathbf{S}$ where

$$\begin{aligned} S_i &= X_i,\ 1 \le i \le m \\ S_i &= (\epsilon_{i1} + \epsilon_{i2}, \epsilon_{i3}),\ m+1 \le i \le n. \end{aligned} \tag{2.4.4}$$

Evidently, $\mathbf{S} = \mathbf{S}(\mathbf{X})$ where $\mathbf{S}(\mathbf{X})$ is given by (2.4.4). This could happen if, for some individuals, the homozygotes of one type ($\epsilon_{i1} = 1$) could not be distinguished from the heterozygotes ($\epsilon_{i2} = 1$). The log likelihood of $\mathbf{S}$ now is

$$\begin{aligned} l_{q,s}(\theta) &= \sum_{i=1}^{m} [2\epsilon_{i1} \log\theta + \epsilon_{i2} \log 2\theta(1-\theta) + 2\epsilon_{i3} \log(1-\theta)] \\ &\quad + \sum_{i=m+1}^{n} [(\epsilon_{i1} + \epsilon_{i2}) \log(1-(1-\theta)^2) + 2\epsilon_{i3} \log(1-\theta)] \end{aligned} \tag{2.4.5}$$

a function that is of curved exponential family form. It does turn out that in this simplest case an explicit maximum likelihood solution is still possible, but the computation is clearly not as simple as in the original Hardy–Weinberg canonical exponential family example. If we suppose (say) that observations $S_1, \dots, S_m$ are not X_i but $(\epsilon_{i1}, \epsilon_{i2} + \epsilon_{i3})$, then explicit solution is in general not possible. Yet the EM algorithm, with an appropriate starting point, leads us to an MLE if it exists in both cases. □

Here is another important example.

Example 2.4.5. *Mixture of Gaussians.* Suppose $S_1, \dots, S_n$ is a sample from a population P whose density is modeled as a mixture of two Gaussian densities, $p(s, \theta) = (1-\lambda)\varphi_{\sigma_1}(s-\mu_1) + \lambda\varphi_{\sigma_2}(s-\mu_2)$ where $\boldsymbol{\theta} = (\lambda, (\mu_i, \sigma_i), i = 1, 2)$ and $0 < \lambda < 1$, $\sigma_1, \sigma_2 > 0$, $\mu_1, \mu_2 \in R$ and $\varphi_\sigma(s) = \frac{1}{\sigma}\varphi\left(\frac{s}{\sigma}\right)$. It is not obvious that this falls under our scheme but let

$$X_i = (\Delta_i, S_i),\ 1 \le i \le n \tag{2.4.6}$$

where Δ_i are independent identically distributed with $P_{\boldsymbol{\theta}}[\Delta_i = 1] = \lambda = 1 - P_{\boldsymbol{\theta}}[\Delta_i = 0]$. Suppose that given $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_n)$, the S_i are independent with

$$\mathcal{L}_{\boldsymbol{\theta}}(S_i \mid \boldsymbol{\Delta}) = \mathcal{L}_{\boldsymbol{\theta}}(S_i \mid \Delta_i) = \mathcal{N}(\Delta_i\mu_1 + (1-\Delta_i)\mu_2, \Delta_i\sigma_1^2 + (1-\Delta_i)\sigma_2^2).$$

That is, Δ_i tells us whether to sample from $\mathcal{N}(\mu_1, \sigma_1^2)$ or $\mathcal{N}(\mu_2, \sigma_2^2)$. It is easy to see (Problem 2.4.11), that under $\boldsymbol{\theta}$, $\mathbf{S}$ has the marginal distribution given previously. Thus, we can think of $\mathbf{S}$ as $S(\mathbf{X})$ where $\mathbf{X}$ is given by (2.4.6).

This five-parameter model is very rich permitting up to two modes and scales. The log likelihood similarly can have a number of local maxima and can tend to ∞ as $\boldsymbol{\theta}$ tends to the boundary of the parameter space (Problem 2.4.12). Although MLEs do not exist in these models, a local maximum close to the true $\boldsymbol{\theta}_0$ turns out to be a good "proxy" for the nonexistent MLE. The EM algorithm can lead to such a local maximum. □

The EM Algorithm. Here is the algorithm. Let

$$J(\theta \mid \theta_0) \equiv E_{\theta_0}\left(\log \frac{p(X, \theta)}{p(X, \theta_0)} \mid S(X) = s\right) \tag{2.4.7}$$

where we suppress dependence on s.

Initialize with $\theta_{\text{old}} = \theta_0$.

The first (E) step of the algorithm is to compute $J(\theta \mid \theta_{\text{old}})$ for as many values of θ as needed. If this is difficult, the EM algorithm is probably not suitable.

The second (M) step is to maximize $J(\theta \mid \theta_{\text{old}})$ as a function of θ. Again, if this step is difficult, EM is not particularly appropriate.

Then we set $\theta_{\text{new}} = \arg\max J(\theta \mid \theta_{\text{old}})$, reset $\theta_{\text{old}} = \theta_{\text{new}}$ and repeat the process.

As we shall see in important situations, including the examples, we have given, the M step is easy and the E step doable.

The rationale behind the algorithm lies in the following formulas, which we give for θ real and which can be justified easily in the case that $\mathcal{X}$ is finite (Problem 2.4.12)

$$\frac{q(s, \theta)}{q(s, \theta_0)} = E_{\theta_0}\left(\frac{p(X, \theta)}{p(X, \theta_0)} \mid S(X) = s\right) \tag{2.4.8}$$

and

$$\frac{\partial}{\partial\theta}\log q(s, \theta)\bigg|_{\theta=\theta_0} = E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log p(X, \theta) \mid S(X) = s\right)\bigg|_{\theta=\theta_0} \tag{2.4.9}$$

for all θ_0 (under suitable regularity conditions). Note that (2.4.9) follows from (2.4.8) by taking logs in (2.4.8), differentiating and exchanging E_{θ_0} and differentiation with respect to θ at θ_0. Because, formally,

$$\frac{\partial J(\theta \mid \theta_0)}{\partial \theta} = E_{\theta_0}\left(\frac{\partial}{\partial \theta}\log p(X,\theta) \mid S(X) = s\right) \tag{2.4.10}$$

and, hence,

$$\left.\frac{\partial J(\theta \mid \theta_0)}{\partial \theta}\right|_{\theta_0} = \frac{\partial}{\partial \theta}\log q(s,\theta_0) \tag{2.4.11}$$

it follows that a fixed point $\widetilde{\theta}$ of the algorithm satisfies the likelihood equation,

$$\frac{\partial}{\partial \theta}\log q(s,\widetilde{\theta}) = 0. \tag{2.4.12}$$

The main reason the algorithm behaves well follows.

Lemma 2.4.1. *If $\theta_{\text{new}}, \theta_{\text{old}}$ are as defined earlier and $S(X) = s$,*

$$q(s,\theta_{\text{new}}) \geq q(s,\theta_{\text{old}}). \tag{2.4.13}$$

Equality holds in (2.4.13) *iff the conditional distribution of X given $S(X) = s$ is the same for θ_{new} as for θ_{old} and θ_{new} maximizes $J(\theta \mid \theta_{\text{old}})$.*

Proof. We give the proof in the discrete case. However, the result holds whenever the quantities in $J(\theta \mid \theta_0)$ can be defined in a reasonable fashion. In the discrete case we appeal to the product rule. For $x \in \mathcal{X}$, $S(x) = s$

$$p(x,\theta) = q(s,\theta)r(x \mid s,\theta) \tag{2.4.14}$$

where $r(\cdot \mid \cdot,\theta)$ is the conditional frequency function of X given $S(X) = s$. Then

$$J(\theta \mid \theta_0) = \log\frac{q(s,\theta)}{q(s,\theta_0)} + E_{\theta_0}\left\{\log\frac{r(X \mid s,\theta)}{r(X \mid s,\theta_0)} \mid S(X) = s\right\}. \tag{2.4.15}$$

If $\theta_0 = \theta_{\text{old}}, \theta = \theta_{\text{new}}$,

$$\log\frac{q(s,\theta_{\text{new}})}{q(s,\theta_{\text{old}})} = J(\theta_{\text{new}} \mid \theta_{\text{old}}) - E_{\theta_{\text{old}}}\left\{\log\frac{r(X \mid s,\theta_{\text{new}})}{r(X \mid s,\theta_{\text{old}})} \mid S(X) = s\right\}. \tag{2.4.16}$$

Now, $J(\theta_{\text{new}} \mid \theta_{\text{old}}) \geq J(\theta_{\text{old}} \mid \theta_{\text{old}}) = 0$ by definition of θ_{new}. On the other hand,

$$-E_{\theta_{\text{old}}}\left\{\log\frac{r(X \mid s,\theta_{\text{new}})}{r(X \mid s,\theta_{\text{old}})} \mid S(X) = s\right\} \geq 0 \tag{2.4.17}$$

by Shannon's inequality, Lemma 2.2.1. □

The most important and revealing special case of this lemma follows.

Theorem 2.4.3. *Suppose $\{P_\theta : \theta \in \Theta\}$ is a canonical exponential family generated by (T,h) satisfying the conditions of Theorem 2.3.1. Let $S(X)$ be any statistic, then*

(a) *The EM algorithm consists of the alternation*

$$\dot{A}(\theta_{\text{new}}) = E_{\theta_{\text{old}}}(T(X) \mid S(X) = s) \tag{2.4.18}$$

$$\theta_{\text{old}} = \theta_{\text{new}}. \tag{2.4.19}$$

If a solution of (2.4.18) *exists it is necessarily unique.*

(b) *If the sequence of iterates* $\{\widehat{\theta}_m\}$ *so obtained is bounded and the equation*

$$\dot{A}(\theta) = E_\theta(T(X) \mid S(X) = s) \tag{2.4.20}$$

has a unique solution, then it converges to a limit $\widehat{\theta}^*$*, which is necessarily a local maximum of* $q(s, \theta)$.

Proof. In this case,

$$\begin{aligned} J(\theta \mid \theta_0) &= E_{\theta_0}\{(\theta - \theta_0)^T T(X) - (A(\theta) - A(\theta_0)) \mid S(X) = s\} \\ &= (\theta - \theta_0)^T E_{\theta_0}(T(X) \mid S(X) = y) - (A(\theta) - A(\theta_0)) \end{aligned} \tag{2.4.21}$$

Part (a) follows.

Part (b) is more difficult. A proof due to Wu (1983) is sketched in Problem 2.4.16. □

Example 2.4.4 (continued). **X** is distributed according to the exponential family

$$p(\mathbf{x}, \theta) = \exp\{\eta(2N_{1n}(\mathbf{x}) + N_{2n}(\mathbf{x})) - A(\boldsymbol{\eta})\}h(\mathbf{x}) \tag{2.4.22}$$

where

$$\eta = \log\left(\frac{\theta}{1-\theta}\right),\ h(\mathbf{x}) = 2^{N_{2n}(\mathbf{x})},\ A(\eta) = 2n\log(1 + e^\eta)$$

and $N_{jn} = \sum_{i=1}^n \epsilon_{ij}(x_i)$, $1 \le j \le 3$. Now,

$$A'(\eta) = 2n\theta \tag{2.4.23}$$

$$\begin{aligned} E_\theta(2N_{1n} + N_{2n} \mid \mathbf{S}) &= 2N_{1m} + N_{2m} \\ &\quad + E_\theta\left(\sum_{i=m+1}^n (2\epsilon_{i1} + \epsilon_{i2}) \mid \epsilon_{i1} + \epsilon_{i2},\ m+1 \le i \le n\right). \end{aligned} \tag{2.4.24}$$

Under the assumption that the process that causes lumping is independent of the values of the ϵ_{ij},

$$\begin{aligned} P_\theta[\epsilon_{ij} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 0] &= 0,\ 1 \le j \le 2 \\ P_\theta[\epsilon_{i1} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 1] &= \frac{\theta^2}{\theta^2 + 2\theta(1-\theta)} = \frac{\theta^2}{1 - (1-\theta)^2} \\ &= 1 - P_\theta[\epsilon_{i2} = 1 \mid \epsilon_{i1} + \epsilon_{i2} = 1]. \end{aligned}$$

Thus, we see, after some simplification, that,

$$E_\theta(2N_{1n} + N_{2n} \mid \mathbf{S}) = 2N_{1m} + N_{2m} + \frac{2}{2 - \widehat{\theta}_{\text{old}}} M_n \tag{2.4.25}$$

where

$$M_n = \sum_{i=m+1}^{n} (\epsilon_{i1} + \epsilon_{i2}).$$

Thus, the EM iteration is

$$\widehat{\theta}_{\text{new}} = \frac{2N_{1m} + N_{2m}}{n} + \frac{2}{2 - \widehat{\theta}_{\text{old}}} \frac{M_n}{n}. \tag{2.4.26}$$

It may be shown directly (Problem 2.4.12) that if $2N_{1m} + N_{2m} > 0$ and $M_n > 0$, then $\widehat{\theta}_m$ converges to the unique root of

$$\theta^2 - \frac{(2N_{3m} + N_{2m})\theta}{n} + \frac{2}{n}(N_{1m} + (1 - N_{3m})) = 0$$

in $(0, 1)$, which is indeed the MLE when $\mathbf{S}$ is observed. □

Example 2.4.6. Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be i.i.d. as (Z, Y), where $(Z, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Suppose that some of the Z_i and some of the Y_i are missing as follows: For $1 \le i \le n_1$ we observe both Z_i and Y_i, for $n_1 + 1 \le i \le n_2$, we oberve only Z_i, and for $n_2 + 1 \le i \le n$, we observe only Y_i. In this case a set of sufficient statistics is

$$T_1 = \bar{Z},\ T_2 = \bar{Y},\ T_3 = n^{-1}\sum_{i=1}^{n} Z_i^2,\ T_4 = n^{-1}\sum_{i=1}^{n} Y_i^2,\ T_5 = n^{-1}\sum_{i=1}^{n} Z_i Y_i.$$

The observed data are

$$S = \{(Z_i, Y_i) : 1 \le i \le n_1\} \cup \{Z_i : n_1 + 1 \le i \le n_2\} \cup \{Y_i : n_2 + 1 \le i \le n\}.$$

To compute $E_\theta(\mathbf{T} \mid S = s)$, where $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \theta)$, we note that for the cases with Z_i and/or Y_i observed, the conditional expected values equal their observed values. For other cases we use the properties of the bivariate normal distribution (Appendix B.4 and Section 1.4), to conclude

$$\begin{aligned} E_\theta(Y_i \mid Z_i) &= \mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1 \\ E_\theta(Y_i^2 \mid Z_i) &= [\mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1]^2 + (1 - \rho^2)\sigma_2^2 \\ E_\theta(Z_i Y_i \mid Z_i) &= [\mu_2 + \rho\sigma_2(Z_i - \mu_1)/\sigma_1] Z_i \end{aligned}$$

with the corresponding Z on Y regression equations when conditioning on Y_i (Problem 2.4.1). This completes the E-step. For the M-step, compute (Problem 2.4.1)

$$\dot{A}(\theta) = E_\theta \mathbf{T} = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \sigma_1\sigma_2\rho + \mu_1\mu_2).$$

We take $\widehat{\theta}_{\text{old}} = \widehat{\theta}_{\text{MOM}}$, where $\widehat{\theta}_{\text{MOM}}$ is the method of moment estimates $(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, r)$ (Problem 2.1.8) of θ based on the observed data, and find (Problem 2.4.1) that the M-step produces

$$\begin{aligned} \widehat{\mu}_{1,\text{new}} &= T_1(\widehat{\theta}_{\text{old}}),\ \widehat{\mu}_{2,\text{new}} = T_2(\widehat{\theta}_{\text{old}}),\ \widehat{\sigma}_{1,\text{new}}^2 = T_3(\widehat{\theta}_{\text{old}}) - \widehat{T}_1^2 \\ \widehat{\sigma}_{2,\text{new}}^2 &= T_4(\widehat{\theta}_{\text{old}}) - \widehat{T}_2^2, \\ \widehat{\rho}_{\text{new}} &= [T_5(\widehat{\theta}_{\text{old}}) - \widehat{T}_1\widehat{T}_2]/\{[T_3(\widehat{\theta}_{\text{old}}) - \widehat{T}_1][\widehat{T}_4(\widehat{\theta}_{\text{old}}) - \widehat{T}_2]\}^{\frac{1}{2}} \end{aligned} \tag{2.4.27}$$

where $T_j(\theta)$ denotes T_j with missing values replaced by the values computed in the E-step and $\widehat{T}_j = T_j(\widehat{\theta}_{\text{old}})$, $j = 1, 2$. Now the process is repeated with $\widehat{\theta}_{\text{MOM}}$ replaced by $\widehat{\theta}_{\text{new}}$. □

Because the E-step, in the context of Example 2.4.6, involves imputing missing values, the EM algorithm is often called *multiple imputation*.

Remark 2.4.1. Note that if $S(X) = X$, then $J(\theta \mid \theta_0)$ is $\log[p(X, \theta)/p(X, \theta_0)]$, which as a function of θ is maximized where the contrast $-\log p(X, \theta)$ is minimized. Also note that, in general, $-E_{\theta_0}[J(\theta \mid \theta_0)]$ is the Kullback–Leibler divergence (2.2.23).

Summary. The basic bisection algorithm for finding roots of monotone functions is developed and shown to yield a rapid way of computing the MLE in all one-parameter canonical exponential families with $\mathcal{E}$ open (when it exists). We then, in Section 2.4.2, use this algorithm as a building block for the general coordinate ascent algorithm, which yields with certainty the MLEs in k-parameter canonical exponential families with $\mathcal{E}$ open when it exists. Important variants of and alternatives to this algorithm, including the Newton–Raphson method, are discussed and introduced in Section 2.4.3 and the problems. Finally in Section 2.4.4 we derive and discuss the important EM algorithm and its basic properties.

2.5 PROBLEMS AND COMPLEMENTS

Problems for Section 2.1

1. Consider a population made up of three different types of individuals occurring in the Hardy–Weinberg proportions θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$, respectively, where $0 < \theta < 1$.

(a) Show that $T_3 = N_1/n + N_2/2n$ is a frequency substitution estimate of θ.

(b) Using the estimate of (a), what is a frequency substitution estimate of the odds ratio $\theta/(1-\theta)$?

(c) Suppose X takes the values $-1, 0, 1$ with respective probabilities p_1, p_2, p_3 given by the Hardy–Weinberg proportions. By considering the first moment of X, show that T_3 is a method of moment estimate of θ.

2. Consider n systems with failure times $X_1, \dots, X_n$ assumed to be independent and identically distributed with exponential, $\mathcal{E}(\lambda)$, distributions.

(a) Find the method of moments estimate of λ based on the first moment.

(b) Find the method of moments estimate of λ based on the second moment.

(c) Combine your answers to (a) and (b) to get a method of moment estimate of λ based on the first two moments.

(d) Find the method of moments estimate of the probability $P(X_1 \geq 1)$ that one system will last at least a month.

3. Suppose that i.i.d. $X_1, \dots, X_n$ have a beta, $\beta(\alpha_1, \alpha_2)$ distribution. Find the method of moments estimates of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ based on the first two moments.