CHAPTER 4

EM OPTIMIZATION METHODS

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. The strategy's statistical foundations and effectiveness in a variety of statistical problems were shown in a seminal paper by Dempster, Laird, and Rubin [150]. Other references on EM and related methods include [409, 413, 449, 456, 625]. The popularity of the EM algorithm stems from how simple it can be to implement and how reliably it can find the global optimum through stable, uphill steps.

In a frequentist setting, we may conceive of observed data generated from random variables $\mathbf{X}$ along with missing or unobserved data from random variables $\mathbf{Z}$. We envision complete data generated from $\mathbf{Y} = (\mathbf{X}, \mathbf{Z})$. Given observed data $\mathbf{x}$, we wish to maximize a likelihood $L(\boldsymbol{\theta}|\mathbf{x})$. Often it will be difficult to work with this likelihood and easier to work with the densities of $\mathbf{Y}|\boldsymbol{\theta}$ and $\mathbf{Z}|(\mathbf{x}, \boldsymbol{\theta})$. The EM algorithm sidesteps direct consideration of $L(\boldsymbol{\theta}|\mathbf{x})$ by working with these easier densities.

In a Bayesian application, interest often focuses on estimating the mode of a posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$. Again, optimization can sometimes be simplified by consideration of unobserved random variables $\boldsymbol{\psi}$ in addition to the parameters of interest, $\boldsymbol{\theta}$.

The missing data may not truly be missing: They may be only a conceptual ploy that simplifies the problem. In this case, $\mathbf{Z}$ is often referred to as *latent*. It may seem counterintuitive that optimization sometimes can be simplified by introducing this new element into the problem. However, examples in this chapter and its references illustrate the potential benefit of this approach. In some cases, the analyst must draw upon his or her creativity and cleverness to invent effective latent variables; in other cases, there is a natural choice.

4.1 MISSING DATA, MARGINALIZATION, AND NOTATION

Whether $\mathbf{Z}$ is considered latent or missing, it may be viewed as having been removed from the complete $\mathbf{Y}$ through the application of some many-to-fewer mapping, $\mathbf{X} = M(\mathbf{Y})$. Let $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ and $f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})$ denote the densities of the observed data and the complete data, respectively. The latent- or missing-data assumption

Computational Statistics, Second Edition. Geof H. Givens and Jennifer A. Hoeting.

amounts to a marginalization model in which we observe $\mathbf{X}$ having density $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = \int_{\{\mathbf{y}:M(\mathbf{y})=\mathbf{x}\}} f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y}$. Note that the conditional density of the missing data given the observed data is $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})/f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$.

In Bayesian applications focusing on the posterior density for parameters of interest, $\boldsymbol{\theta}$, there are two manners in which we may consider the posterior to represent a marginalization of a broader problem. First, it may be sensible to view the likelihood $L(\boldsymbol{\theta}|\mathbf{x})$ as a marginalization of the complete-data likelihood $L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$. In this case the missing data are $\mathbf{z}$, and we use the same sort of notation as above. Second, we may consider there to be missing parameters $\boldsymbol{\psi}$, whose inclusion simplifies Bayesian calculations even though $\boldsymbol{\psi}$ is of no interest itself. Fortunately, under the Bayesian paradigm there is no practical distinction between these two cases. Since $\mathbf{Z}$ and $\boldsymbol{\psi}$ are both missing random quantities, it matters little whether we use notation that suggests the missing variables to be unobserved data or parameters. In cases where we adopt the frequentist notation, the reader may replace the likelihood and $\mathbf{Z}$ by the posterior and $\boldsymbol{\psi}$, respectively, to consider the Bayesian point of view.

In the literature about EM, it is traditional to adopt notation that reverses the roles of $\mathbf{X}$ and $\mathbf{Y}$ compared to our usage. We diverge from tradition, using $\mathbf{X} = \mathbf{x}$ to represent observed data as everywhere else in this book.

4.2 THE EM ALGORITHM

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

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left\{\log L(\boldsymbol{\theta}|\mathbf{Y}) \;\middle|\; \mathbf{x}, \boldsymbol{\theta}^{(t)}\right\} \tag{4.1}$$

$$= E\left\{\log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) \;\middle|\; \mathbf{x}, \boldsymbol{\theta}^{(t)}\right\} \tag{4.2}$$

$$= \int \left[\log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})\right] f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z}, \tag{4.3}$$

where (4.3) emphasizes that $\mathbf{Z}$ is the only random part of $\mathbf{Y}$ once we are given $\mathbf{X} = \mathbf{x}$.

EM is initiated from $\boldsymbol{\theta}^{(0)}$ then alternates between two steps: E for expectation and M for maximization. The algorithm is summarized as:

1. **E step**: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.
2. **M step**: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(t+1)}$ equal to the maximizer of Q.
3. Return to the E step unless a stopping criterion has been met.

Stopping criteria for optimization problems are discussed in Chapter 2. In the present case, such criteria are usually built upon $(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})$ or $\left|Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\right|$.

Example 4.1 (Simple Exponential Density) To understand the EM notation, consider a trivial example where $Y_1, Y_2 \sim$ i.i.d. Exp(θ). Suppose $y_1 = 5$ is observed but the value y_2 is missing. The complete-data log likelihood function is $\log L(\theta|\mathbf{y}) = \log f_{\mathbf{Y}}(\mathbf{y}|\theta) = 2\log\{\theta\} - \theta y_1 - \theta y_2$. Taking the conditional expectation of $\log L(\theta|\mathbf{Y})$ yields $Q(\theta|\theta^{(t)}) = 2\log\{\theta\} - 5\theta - \theta/\theta^{(t)}$, since $E\{Y_2|y_1, \theta^{(t)}\} = E\{Y_2|\theta^{(t)}\} = 1/\theta^{(t)}$ follows from independence. The maximizer of $Q(\theta|\theta^{(t)})$ with respect to θ is easily found to be the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$. Solving for θ provides the updating equation $\theta^{(t+1)} = 2\theta^{(t)}/(5\theta^{(t)} + 1)$. Note here that the E step and M step do not need to be rederived at each iteration: Iterative application of the updating formula starting from some initial value provides estimates that converge to $\hat{\theta} = 0.2$.

This example is not realistic. The maximum likelihood estimate of θ from the observed data can be determined from elementary analytic methods without reliance on any fancy numerical optimization strategy like EM. More importantly, we will learn that taking the required expectation is tricker in real applications, because one needs to know the conditional distribution of the complete data given the missing data. □

Example 4.2 (Peppered Moths) The peppered moth, *Biston betularia*, presents a fascinating story of evolution and industrial pollution [276]. The coloring of these moths is believed to be determined by a single gene with three possible alleles, which we denote C, I, and T. Of these, C is dominant to I, and T is recessive to I. Thus the genotypes CC, CI, and CT result in the *carbonaria* phenotype, which exhibits solid black coloring. The genotype TT results in the *typica* phenotype, which exhibits light-colored patterned wings. The genotypes II and IT produce an intermediate phenotype called *insularia*, which varies widely in appearance but is generally mottled with intermediate color. Thus, there are six possible genotypes, but only three phenotypes are measurable in field work.

In the United Kingdom and North America, the *carbonaria* phenotype nearly replaced the paler phenotypes in areas affected by coal-fired industries. This change in allele frequencies in the population is cited as an instance where we may observe microevolution occurring on a human time scale. The theory (supported by experiments) is that "differential predation by birds on moths that are variously conspicuous against backgrounds of different reflectance" (p. 88) induces selectivity that favors the *carbonaria* phenotype in times and regions where sooty, polluted conditions reduce the reflectance of the surface of tree bark on which the moths rest [276]. Not surprisingly, when improved environmental standards reduced pollution, the prevalence of the lighter-colored phenotypes increased and that of *carbonaria* plummeted.

Thus, it is of interest to monitor the allele frequencies of C, I, and T over time to provide insight on microevolutionary processes. Further, trends in these frequencies also provide an interesting biological marker to monitor air quality. Within a sufficiently short time period, an approximate model for allele frequencies can be built from the Hardy–Weinberg principle that each genotype frequency in a population in Hardy–Weinberg equilibrium should equal the product of the corresponding allele frequencies, or double that amount when the two alleles differ (to account for uncertainty in the parental source) [15, 316]. Thus, if the allele frequencies in the population are p_{C}, p_{I}, and p_{T}, then the genotype frequencies should be p_{C}^2, $2p_{\mathrm{C}}p_{\mathrm{I}}$,

$2p_C p_T$, p_I^2, $2p_I p_T$, and p_T^2 for genotypes CC, CI, CT, II, IT, and TT, respectively. Note that $p_C + p_I + p_T = 1$.

Suppose we capture n moths, of which there are n_C, n_I, and n_T of the *carbonaria*, *insularia*, and *typica* phenotypes, respectively. Thus, $n = n_C + n_I + n_T$. Since each moth has two alleles in the gene in question, there are $2n$ total alleles in the sample. If we knew the genotype of each moth rather than merely its phenotype, we could generate genotype counts n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, and n_{TT}, from which allele frequencies could easily be tabulated. For example, each moth with genotype CI contributes one C allele and one I allele, whereas a II moth contributes two I alleles. Such allele counts would immediately provide estimates of p_C, p_I, and p_T. It is far less clear how to estimate the allele frequencies from the phenotype counts alone.

In the EM notation, the observed data are $\mathbf{x} = (n_C, n_I, n_T)$ and the complete data are $\mathbf{y} = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$. The mapping from the complete data to the observed data is $\mathbf{x} = M(\mathbf{y}) = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$. We wish to estimate the allele probabilities, p_C, p_I, and p_T. Since $p_T = 1 - p_C - p_I$, the parameter vector for this problem is $\mathbf{p} = (p_C, p_I)$, but for notational brevity we often refer to p_T in what follows.

The complete data log likelihood function is multinomial:

$$\begin{aligned}\log f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p}) = {} & n_{CC}\log\{p_C^2\} + n_{CI}\log\{2p_C p_I\} + n_{CT}\log\{2p_C p_T\} \\ & + n_{II}\log\{p_I^2\} + n_{IT}\log\{2p_I p_T\} + n_{TT}\log\{p_T^2\} \\ & + \log\binom{n}{n_{CC}\quad n_{CI}\quad n_{CT}\quad n_{II}\quad n_{IT}\quad n_{TT}}. \end{aligned} \tag{4.4}$$

The complete data are not all observed. Let $\mathbf{Y} = (N_{CC}, N_{CI}, N_{CT}, N_{II}, N_{IT}, n_{TT})$, since we know $N_{TT} = n_{TT}$ but the other frequencies are not directly observed. To calculate $Q(\mathbf{p}|\mathbf{p}^{(t)})$, notice that conditional on n_C and a parameter vector $\mathbf{p}^{(t)} = (p_C^{(t)}, p_I^{(t)})$, the latent counts for the three *carbonaria* genotypes have a three-cell multinomial distribution with count parameter n_C and cell probabilities proportional to $(p_C^{(t)})^2$, $2p_C^{(t)}p_I^{(t)}$, and $2p_C^{(t)}p_T^{(t)}$. A similar result holds for the two insularia cells. Thus the expected values of the first five random parts of (4.4) are

$$E\{N_{CC}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{CC}^{(t)} = \frac{n_C(p_C^{(t)})^2}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}, \tag{4.5}$$

$$E\{N_{CI}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{CI}^{(t)} = \frac{2n_C p_C^{(t)} p_I^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}, \tag{4.6}$$

$$E\{N_{CT}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{CT}^{(t)} = \frac{2n_C p_C^{(t)} p_T^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}, \tag{4.7}$$

$$E\{N_{II}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{II}^{(t)} = \frac{n_I(p_I^{(t)})^2}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}}, \tag{4.8}$$

$$E\{N_{IT}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{IT}^{(t)} = \frac{2n_I p_I^{(t)} p_T^{(t)}}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}}. \tag{4.9}$$

Finally, we know $n_{TT} = n_T$, where n_T is observed. The multinomial coefficient in the likelihood has a conditional expectation, say $k(n_C, n_I, n_T, \mathbf{p}^{(t)})$, that does not depend on $\mathbf{p}$. Thus, we have found

$$\begin{aligned} Q(\mathbf{p}|\mathbf{p}^{(t)}) = {} & n_{CC}^{(t)} \log\{p_C^2\} + n_{CI}^{(t)} \log\{2p_C p_I\} \\ & + n_{CT}^{(t)} \log\{2p_C p_T\} + n_{II}^{(t)} \log\{p_I^2\} \\ & + n_{IT}^{(t)} \log\{2p_I p_T\} + n_{TT} \log\{p_T^2\} + k(n_C, n_I, n_T, \mathbf{p}^{(t)}). \end{aligned} \tag{4.10}$$

Recalling $p_T = 1 - p_C - p_I$ and differentiating with respect to p_C and p_I yields

$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}, \tag{4.11}$$

$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}. \tag{4.12}$$

Setting these derivatives equal to zero and solving for p_C and p_I completes the M step, yielding

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}, \tag{4.13}$$

$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}, \tag{4.14}$$

$$p_T^{(t+1)} = \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{2n}, \tag{4.15}$$

where the final expression is derived from the constraint that the probabilities sum to one. If the tth latent counts were true, the number of *carbonaria* alleles in the sample would be $2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}$. There are $2n$ total alleles in the sample. Thus, the EM update consists of setting the elements of $\mathbf{p}^{(t+1)}$ equal to the phenotypic frequencies that would result from the tth latent genotype counts.

Suppose the observed phenotype counts are $n_C = 85$, $n_I = 196$, and $n_T = 341$. Table 4.1 shows how the EM algorithm converges to the MLEs, roughly $\hat{p}_C = 0.07084$, $\hat{p}_I = 0.18874$, and $\hat{p}_T = 0.74043$. Finding a precise estimate of $\hat{p}_I$ is slower than for $\hat{p}_C$, since the likelihood is flatter along the p_I coordinate.

The last three columns of Table 4.1 show convergence diagnostics. A relative convergence criterion,

$$R^{(t)} = \frac{\|\mathbf{p}^{(t)} - \mathbf{p}^{(t-1)}\|}{\|\mathbf{p}^{(t-1)}\|}, \tag{4.16}$$

summarizes the total amount of relative change in $\mathbf{p}^{(t)}$ from one iteration to the next, where $\|\mathbf{z}\| = (\mathbf{z}^T\mathbf{z})^{1/2}$. For illustrative purposes, we also include $D_C^{(t)} = (p_C^{(t)} - \hat{p}_C)/(p_C^{(t-1)} - \hat{p}_C)$ and the analogous quantity $D_I^{(t)}$. These ratios quickly converge to constants, confirming that the EM rate of convergence is linear as defined by (2.19). □

TABLE 4.1 EM results for peppered moth example. The diagnostic quantities $R^{(t)}$, $D_{\mathsf{C}}^{(t)}$, and $D_{\mathsf{I}}^{(t)}$ are defined in the text.

t	$p_{\mathsf{C}}^{(t)}$	$p_{\mathsf{I}}^{(t)}$	$R^{(t)}$	$D_{\mathsf{C}}^{(t)}$	$D_{\mathsf{I}}^{(t)}$
0	0.333333	0.333333			
1	0.081994	0.237406	5.7×10^{-1}	0.0425	0.337
2	0.071249	0.197870	1.6×10^{-1}	0.0369	0.188
3	0.070852	0.190360	3.6×10^{-2}	0.0367	0.178
4	0.070837	0.189023	6.6×10^{-3}	0.0367	0.176
5	0.070837	0.188787	1.2×10^{-3}	0.0367	0.176
6	0.070837	0.188745	2.1×10^{-4}	0.0367	0.176
7	0.070837	0.188738	3.6×10^{-5}	0.0367	0.176
8	0.070837	0.188737	6.4×10^{-6}	0.0367	0.176

Example 4.3 (Bayesian Posterior Mode) Consider a Bayesian problem with likelihood $L(\boldsymbol{\theta}|\mathbf{x})$, prior $f(\boldsymbol{\theta})$, and missing data or parameters $\mathbf{Z}$. To find the posterior mode, the E step requires

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\{\log\{L(\boldsymbol{\theta}|\mathbf{Y})f(\boldsymbol{\theta})k(\mathbf{Y})\}|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} \\ &= E\{\log L(\boldsymbol{\theta}|\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} + \log f(\boldsymbol{\theta}) + E\{\log k(\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\}, \end{aligned} \tag{4.17}$$

where the final term in (4.17) is a normalizing constant that can be ignored because Q is to be maximized with respect to $\boldsymbol{\theta}$. 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. Section 4.3.2 describes a variety of methods for facilitating the M step in difficult situations. □

4.2.1 Convergence

To investigate the convergence properties of the EM algorithm, we begin by showing that each maximization step increases the observed-data log likelihood, $l(\boldsymbol{\theta}|\mathbf{x})$. First, note that the log of the observed-data density can be reexpressed as

$$\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) - \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}). \tag{4.18}$$

Therefore,

$$E\{\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} = E\{\log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} - E\{\log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\},$$

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

$$\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{4.19}$$

where

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left\{\log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x}, \boldsymbol{\theta}) \middle| \mathbf{x}, \boldsymbol{\theta}^{(t)}\right\}. \tag{4.20}$$

The importance of (4.19) becomes apparent after we show that $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is maximized with respect to $\boldsymbol{\theta}$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. To see this, write

$$\begin{aligned} H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) & \\ &= E\left\{ \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) - \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x}, \boldsymbol{\theta}) \,\middle|\, \mathbf{x}, \boldsymbol{\theta}^{(t)} \right\} \\ &= \int -\log \left[\frac{f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})} \right] f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z} \\ &\geq -\log \int f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})\, d\mathbf{z} \\ &= 0. \end{aligned} \tag{4.21}$$

Equation (4.21) follows from an application of Jensen's inequality, since $-\log u$ is strictly convex in u.

Thus, any $\boldsymbol{\theta} \neq \boldsymbol{\theta}^{(t)}$ makes $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ smaller than $H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$. In particular, if we choose $\boldsymbol{\theta}^{(t+1)}$ to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$, then

$$\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}^{(t+1)}) - \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}^{(t)}) \geq 0, \tag{4.22}$$

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

Choosing $\boldsymbol{\theta}^{(t+1)}$ at each iteration to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ constitutes the standard EM algorithm. If instead we simply select any $\boldsymbol{\theta}^{(t+1)}$ for which $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, then the resulting algorithm is called generalized EM, or GEM. In either case, a step that increases Q increases the log likelihood. Conditions under which this guaranteed ascent ensures convergence to an MLE are explored in [60, 676].

Having established this result, we next consider the order of convergence for the method. The EM algorithm defines a mapping $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})$ where the function $\boldsymbol{\Psi}(\boldsymbol{\theta}) = (\Psi_1(\boldsymbol{\theta}), \ldots, \Psi_p(\boldsymbol{\theta}))$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. When EM converges, it converges to a fixed point of this mapping, so $\hat{\boldsymbol{\theta}} = \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})$. Let $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ denote the Jacobian matrix whose (i, j)th element is $d\Psi_i(\boldsymbol{\theta})/d\theta_j$. Taylor series expansion of $\boldsymbol{\Psi}$ yields

$$\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}} \approx \boldsymbol{\Psi}'(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}), \tag{4.23}$$

since $\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}} = \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)}) - \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})$. Comparing this result with (2.19), we see that the EM algorithm has linear convergence when $p = 1$. For $p > 1$, convergence is still linear provided that the observed information, $-\mathbf{l}''(\hat{\boldsymbol{\theta}}|\mathbf{x})$, is positive definite. More precise details regarding convergence are given in [150, 449, 452, 455].

The global rate of EM convergence is defined as

$$\rho = \lim_{t \to \infty} \frac{\|\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}}\|}{\|\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}\|}. \tag{4.24}$$

It can be shown that ρ equals the largest eigenvalue of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ when $-\mathbf{l}''(\hat{\boldsymbol{\theta}}|\mathbf{x})$ is positive definite. In Sections 4.2.3.1 and 4.2.3.2 we will examine how $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ is a matrix of the fractions of missing information. Therefore, ρ effectively serves as a scalar

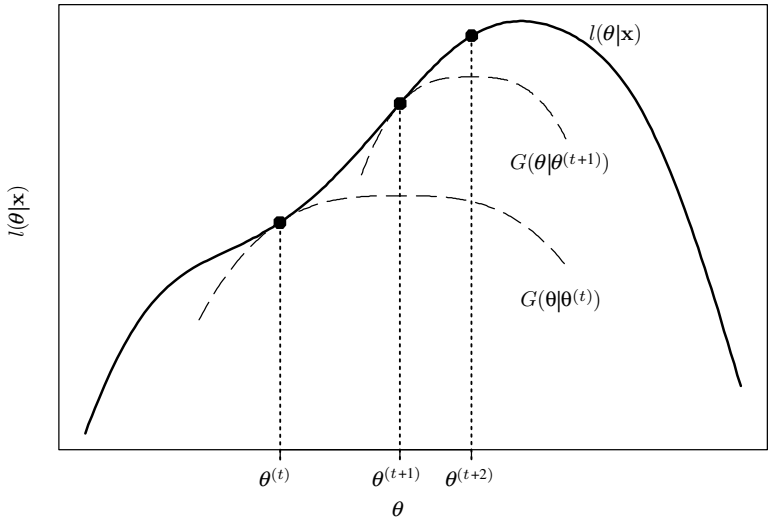

FIGURE 4.1 One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy.

summary of the overall proportion of missing information. 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. Thus, 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. Section 4.3.3 discusses methods for accelerating EM convergence.

To further understand how EM works, note from (4.21) that

$$l(\boldsymbol{\theta}|\mathbf{x}) \geq Q(\boldsymbol{\theta}|\theta^{(t)}) + l(\boldsymbol{\theta}^{(t)}|\mathbf{x}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{4.25}$$

Since the last two terms in $G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ are independent of $\boldsymbol{\theta}$, the functions Q and G are maximized at the same $\boldsymbol{\theta}$. Further, G is tangent to l at $\boldsymbol{\theta}^{(t)}$ 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. This idea is illustrated in Figure 4.1. Each E step amounts to forming the minorizing function G, and each M step amounts to maximizing it to provide an uphill step.

Temporarily replacing l by a minorizing function is an example of a more general strategy known as *optimization transfer*. Links to the EM algorithm and other statistical applications of optimization transfer are surveyed in [410]. In mathematical applications where it is standard to pose optimizations as minimizations, one typically refers to *majorization*, as one could achieve by majorizing the negative log likelihood using $-G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

4.2.2 Usage in Exponential Families

When the complete data are modeled to have a distribution in the exponential family, the density of the data can be written as $f(\mathbf{y}|\boldsymbol{\theta}) = c_1(\mathbf{y})c_2(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\}$, where $\boldsymbol{\theta}$ is a vector of natural parameters and $\mathbf{s}(\mathbf{y})$ is a vector of sufficient statistics. In this case, the E step finds

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = k + \log c_2(\boldsymbol{\theta}) + \int \boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z}, \tag{4.26}$$

where k is a quantity that does not depend on $\boldsymbol{\theta}$. To carry out the M step, set the gradient of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ equal to zero. This yields

$$\frac{-\mathbf{c}_2'(\boldsymbol{\theta})}{c_2(\boldsymbol{\theta})} = \int \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z} \tag{4.27}$$

after rearranging terms and adopting the obvious notational shortcut to vectorize the integral of a vector. It is straightforward to show that $\mathbf{c}_2'(\boldsymbol{\theta}) = -c_2(\boldsymbol{\theta}) E\{\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta}\}$. Therefore, (4.27) implies that the M step is completed by setting $\boldsymbol{\theta}^{(t+1)}$ equal to the $\boldsymbol{\theta}$ that solves

$$E\{\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta}\} = \int \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z}. \tag{4.28}$$

Aside from replacing $\boldsymbol{\theta}^{(t)}$ with $\boldsymbol{\theta}^{(t+1)}$, the form of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is unchanged for the next E step, and the next M step solves the same optimization problem. Therefore, the EM algorithm for exponential families consists of:

1. **E step**: Compute the expected values of the sufficient statistics for the complete data, given the observed data and using the current parameter guesses, $\boldsymbol{\theta}^{(t)}$. Let $\mathbf{s}^{(t)} = E\{\mathbf{s}(\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} = \int \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z}$.
2. **M step**: Set $\boldsymbol{\theta}^{(t+1)}$ to the value that makes the unconditional expectation of the sufficient statistics for the complete data equal to $\mathbf{s}^{(t)}$. In other words, $\boldsymbol{\theta}^{(t+1)}$ solves $E\{\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta})\} = \mathbf{s}^{(t)}$.
3. Return to the E step unless a convergence criterion has been met.

Example 4.4 (Peppered Moths, Continued) The complete data in Example 4.2 arise from a multinomial distribution, which is in the exponential family. The sufficient statistics are, say, the first five genotype counts (with the sixth derived from the constraint that the counts total n), and the natural parameters are the corresponding log probabilities seen in (4.4). The first three conditional expectations for the E step are $s_{\mathrm{CC}}^{(t)} = n_{\mathrm{CC}}^{(t)}$, $s_{\mathrm{CI}}^{(t)} = n_{\mathrm{CI}}^{(t)}$, and $s_{\mathrm{CT}}^{(t)} = n_{\mathrm{CT}}^{(t)}$, borrowing notation from (4.5)–(4.9) and indexing the components of $\mathbf{s}^{(t)}$ in the obvious way. The unconditional expectations of the first three sufficient statistics are np_{C}^2, $2np_{\mathrm{C}}p_{\mathrm{I}}$, and $2np_{\mathrm{C}}p_{\mathrm{T}}$. Equating these three expressions with the conditional expectations given above and solving for p_{C} constitutes the M step for p_{C}. Summing the three equations gives $np_{\mathrm{C}}^2 + 2np_{\mathrm{C}}p_{\mathrm{I}} + 2np_{\mathrm{C}}p_{\mathrm{T}} = n_{\mathrm{CC}}^{(t)} + n_{\mathrm{CI}}^{(t)} + n_{\mathrm{CT}}^{(t)}$, which reduces to the update given in (4.13). EM updates

for p_{I} and p_{T} are found analogously, on noting the constraint that the three probabilities sum to 1. □

4.2.3 Variance Estimation

In a maximum likelihood setting, the EM algorithm is used to find an MLE but does not automatically produce an estimate of the covariance matrix of the MLEs. Typically, we would use the asymptotic normality of the MLEs to justify seeking an estimate of the Fisher information matrix. One way to estimate the covariance matrix, therefore, is to compute the observed information, $-\mathbf{l}''(\hat{\boldsymbol{\theta}}|\mathbf{x})$, where $\mathbf{l}''$ is the Hessian matrix of second derivatives of $\log L(\boldsymbol{\theta}|\mathbf{x})$.

In a Bayesian setting, an estimate of the posterior covariance matrix for $\boldsymbol{\theta}$ can be motivated by noting the asymptotic normality of the posterior [221]. This requires the Hessian of the log posterior density.

In some cases, the Hessian may be computed analytically. In other cases, the Hessian may be difficult to derive or code. In these instances, a variety of other methods are available to simplify the estimation of the covariance matrix.

Of the options described below, the SEM (supplemented EM) algorithm is easy to implement while generally providing fast, reliable results. Even easier is bootstrapping, although for very complex problems the computational burden of nested looping may be prohibitive. These two approaches are recommended, yet the other alternatives can also be useful in some settings.

4.2.3.1 Louis's Method Taking second partial derivatives of (4.19) and negating both sides yields

$$-\mathbf{l}''(\boldsymbol{\theta}|\mathbf{x}) = -\mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})\big|_{\boldsymbol{\omega}=\boldsymbol{\theta}} + \mathbf{H}''(\boldsymbol{\theta}|\boldsymbol{\omega})\big|_{\boldsymbol{\omega}=\boldsymbol{\theta}}, \tag{4.29}$$

where the primes on $\mathbf{Q}''$ and $\mathbf{H}''$ denote derivatives with respect to the first argument, namely $\boldsymbol{\theta}$.

Equation (4.29) can be rewritten as

$$\hat{\mathbf{I}}_{\mathbf{X}}(\boldsymbol{\theta}) = \hat{\mathbf{I}}_{\mathbf{Y}}(\boldsymbol{\theta}) - \hat{\mathbf{I}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta}), \tag{4.30}$$

where $\hat{\mathbf{I}}_{\mathbf{X}}(\boldsymbol{\theta}) = -\mathbf{l}''(\boldsymbol{\theta}|\mathbf{x})$ is the observed information, and $\hat{\mathbf{I}}_{\mathbf{Y}}(\boldsymbol{\theta})$ and $\hat{\mathbf{I}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta})$ will be called the complete information and the missing information, respectively. Interchanging integration and differentiation (when possible), we have

$$\hat{\mathbf{I}}_{\mathbf{Y}}(\boldsymbol{\theta}) = -\mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})\big|_{\boldsymbol{\omega}=\boldsymbol{\theta}} = -E\{\mathbf{l}''(\boldsymbol{\theta}|\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}\}, \tag{4.31}$$

which is reminiscent of the Fisher information defined in (1.28). This motivates calling $\hat{\mathbf{I}}_{\mathbf{Y}}(\boldsymbol{\theta})$ the complete information. A similar argument holds for $-\mathbf{H}''$. Equation (4.30), stating that the observed information equals the complete information minus the missing information, is a result termed the *missing-information principle* [424, 673].

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

$$\hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta}) = \text{var}\left\{\frac{d \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x}, \boldsymbol{\theta})}{d\boldsymbol{\theta}}\right\} \tag{4.32}$$

where the variance is taken with respect to $f_{\mathbf{Z}|\mathbf{X}}$. Further, since the expected score is zero at $\hat{\boldsymbol{\theta}}$,

$$\hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}}) = \int \mathbf{S}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}})\mathbf{S}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}})^{\mathsf{T}} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{X}, \hat{\boldsymbol{\theta}})\, d\mathbf{z}, \tag{4.33}$$

where

$$\mathbf{S}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta}) = \frac{d \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{d\boldsymbol{\theta}}.$$

The missing-information principle enables us to express $\hat{\mathbf{i}}_{\mathbf{X}}(\boldsymbol{\theta})$ in terms of the complete-data likelihood and the conditional density of the missing data given the observed data, while avoiding calculations involving the presumably complicated marginal likelihood of the observed data. This approach can be easier to derive and code in some instances, but it is not always significantly simpler than direct calculation of $-\mathbf{l}''(\hat{\boldsymbol{\theta}}|\mathbf{x})$.

If $\hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta})$ or $\hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta})$ is difficult to compute analytically, it may be estimated via the Monte Carlo method (see Chapter 6). For example, the simplest Monte Carlo estimate of $\hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta})$ is

$$\frac{1}{m}\sum_{i=1}^{m} -\frac{d^2 \log f_{\mathbf{Y}}(\mathbf{y}_i|\boldsymbol{\theta})}{d\boldsymbol{\theta} \cdot d\boldsymbol{\theta}}, \tag{4.34}$$

where for $i = 1, \ldots, m$, the $\mathbf{y}_i = (\mathbf{x}, \mathbf{z}_i)$ are simulated complete datasets consisting of the observed data and i.i.d. imputed missing-data values $\mathbf{z}_i$ drawn from $f_{\mathbf{Z}|\mathbf{X}}$. Similarly, a simple Monte Carlo estimate of $\hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta})$ is the sample variance of the values of

$$-\left[d \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_i|\mathbf{x}, \boldsymbol{\theta})\right]/d\boldsymbol{\theta}$$

obtained from such a collection of $\mathbf{z}_i$.

Example 4.5 (Censored Exponential Data) Suppose we attempt to observed complete data under the model $Y_1, \ldots, Y_n \sim$ i.i.d. Exp(λ), but some cases are right-censored. Thus, the observed data are $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_i = (\min(y_i, c_i), \delta_i)$, the c_i are the censoring levels, and $\delta_i = 1$ if $y_i \leq c_i$ and $\delta_i = 0$ otherwise.

The complete-data log likelihood is $l(\lambda|y_1, \ldots, y_n) = n \log \lambda - \lambda \sum_{i=1}^n y_i$. Thus,

$$Q(\lambda|\lambda^{(t)}) = E\{l(\lambda|Y_1, \ldots, Y_n)|\mathbf{x}, \lambda^{(t)}\} \tag{4.35}$$

$$= n \log \lambda - \lambda \sum_{i=1}^n E\{Y_i|x_i, \lambda^{(t)}\}$$

$$= n \log \lambda - \lambda \sum_{i=1}^n \left[y_i \delta_i + \left(c_i + \frac{1}{\lambda^{(t)}} \right) (1 - \delta_i) \right] \tag{4.36}$$

$$= n \log \lambda - \lambda \sum_{i=1}^n \left[y_i \delta_i + c_i (1 - \delta_i) \right] - \frac{C\lambda}{\lambda^{(t)}}, \tag{4.37}$$

where $C = \sum_{i=1}^n (1 - \delta_i)$ denotes the number of censored cases. Note that (4.36) follows from the memoryless property of the exponential distribution. Therefore, $-Q''(\lambda|\lambda^{(t)}) = n/\lambda^2$.

The unobserved outcome for a censored case, Z_i, has density $f_{Z_i|X}(z_i|x, \lambda) = \lambda \exp\{-\lambda(z_i - c_i)\} 1_{\{z_i > c_i\}}$. Calculating $\hat{i}_{\mathbf{Z}|\mathbf{X}}(\lambda)$ as in (4.32), we find

$$\frac{d \log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x}, \lambda)}{d\lambda} = C/\lambda - \sum_{\{i:\ \delta_i = 0\}} (Z_i - c_i). \tag{4.38}$$

The variance of this expression with respect to $f_{Z_i|X}$ is

$$\hat{i}_{\mathbf{Z}|\mathbf{X}}(\lambda) = \sum_{\{i:\ \delta_i = 0\}} \text{var}\{Z_i - c_i\} = \frac{C}{\lambda^2}, \tag{4.39}$$

since $Z_i - c_i$ has an Exp(λ) distribution.

Thus, applying Louis's method,

$$\hat{i}_{\mathbf{X}}(\lambda) = \frac{n}{\lambda^2} - \frac{C}{\lambda^2} = \frac{U}{\lambda^2}, \tag{4.40}$$

where $U = \sum_{i=1}^n \delta_i$ denotes the number of uncensored cases. For this elementary example, it is easy to confirm by direct analysis that $-l''(\lambda|\mathbf{x}) = U/\lambda^2$. □

4.2.3.2 SEM Algorithm Recall that Ψ denotes the EM mapping, having fixed point $\hat{\boldsymbol{\theta}}$ and Jacobian matrix $\Psi'(\boldsymbol{\theta})$ with (i, j)th element equaling $d\Psi_i(\boldsymbol{\theta})/d\theta_j$. Dempster et al. [150] show that

$$\Psi'(\hat{\boldsymbol{\theta}})^{\mathsf{T}} = \hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1} \tag{4.41}$$

in the terminology of (4.30).

If we reexpress the missing information principle in (4.30) as

$$\hat{\mathbf{i}}_{\mathbf{X}}(\hat{\boldsymbol{\theta}}) = \left[\mathbf{I} - \hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1} \right] \hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}}), \tag{4.42}$$

where $\mathbf{I}$ is an identity matrix, and substitute (4.41) into (4.42), then inverting $\hat{\mathbf{i}}_{\mathbf{X}}(\hat{\boldsymbol{\theta}})$ provides the estimate

$$\widehat{\text{var}}\{\hat{\boldsymbol{\theta}}\} = \hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1}\left(\mathbf{I} + \boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^{\mathsf{T}}[\mathbf{I} - \boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^{\mathsf{T}}]^{-1}\right). \tag{4.43}$$

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* [453]. 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 described in Section 4.2.3.5.

Estimation of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ proceeds as follows. The first step of SEM is to run the EM algorithm to convergence, finding the maximizer $\hat{\boldsymbol{\theta}}$. The second step is to restart the algorithm from, say, $\boldsymbol{\theta}^{(0)}$. Although one may restart from the original starting point, it is preferable to choose $\boldsymbol{\theta}^{(0)}$ to be closer to $\hat{\boldsymbol{\theta}}$.

Having thus initialized SEM, we begin SEM iterations for $t = 0, 1, 2, \ldots$. The $(t+1)$th SEM iteration begins by taking a standard E step and M step to produce $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$. Next, for $j = 1, \ldots, p$, define $\boldsymbol{\theta}^{(t)}(j) = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p)$ and

$$r_{ij}^{(t)} = \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j} \tag{4.44}$$

for $i = 1, \ldots, p$, recalling that $\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$. This ends one SEM iteration. The $\Psi_i(\boldsymbol{\theta}^{(t)}(j))$ values are the estimates produced by applying one EM cycle to $\boldsymbol{\theta}^{(t)}(j)$ for $j = 1, \ldots, p$.

Notice that the (i, j)th element of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ equals $\lim_{t\to\infty} r_{ij}^{(t)}$. We may consider each element of this matrix to be precisely estimated when the sequence of $r_{ij}^{(t)}$ values stabilizes for $t \geq t_{ij}^*$. Note that different numbers of iterations may be needed for precise estimation of different elements of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$. When all elements have stabilized, SEM iterations stop and the resulting estimate of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ is used to determine $\widehat{\text{var}}\{\hat{\boldsymbol{\theta}}\}$ as given in (4.43).

Numerical imprecision can cause the resulting covariance matrix to be slightly asymmetric. Such asymmetry can be used to diagnose whether the original EM procedure was run to sufficient precision and to assess how many digits are trustworthy in entries of the estimated covariance matrix. Difficulties also arise if $\mathbf{I} - \boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^{\mathsf{T}}$ is not positive semidefinite or cannot be inverted numerically; see [453]. It has been suggested that transforming $\boldsymbol{\theta}$ to achieve an approximately normal likelihood can lead to faster convergence and increased accuracy of the final solution.

Example 4.6 (Peppered Moths, Continued) The results from Example 4.2 can be supplemented using the approach of Meng and Rubin [453]. Stable, precise results are obtained within a few SEM iterations, starting from $p_{\text{C}}^{(0)} = 0.07$ and

$p_{\mathrm{I}}^{(t)} = 0.19$. Standard errors for $\hat{p}_{\mathrm{C}}$, $\hat{p}_{\mathrm{I}}$, and $\hat{p}_{\mathrm{T}}$ are 0.0074, 0.0119, and 0.0132, respectively. Pairwise correlations are $\mathrm{cor}\{\hat{p}_{\mathrm{C}}, \hat{p}_{\mathrm{I}}\} = -0.14$, $\mathrm{cor}\{\hat{p}_{\mathrm{C}}, \hat{p}_{\mathrm{T}}\} = -0.44$, and $\mathrm{cor}\{\hat{p}_{\mathrm{I}}, \hat{p}_{\mathrm{T}}\} = -0.83$. Here, SEM was used to obtain results for $\hat{p}_{\mathrm{C}}$ and $\hat{p}_{\mathrm{I}}$, and elementary relationships among variances, covariances, and correlations were used to extend these results for $\hat{p}_{\mathrm{T}}$ since the estimated probabilities sum to 1. □

It may seem inefficient not to begin SEM iterations until EM iterations have ceased. An alternative would be to attempt to estimate the components of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ as EM iterations progress, using

$$\tilde{r}_{ij}^{(t)} = \frac{\Psi_i(\theta_1^{(t-1)}, \ldots, \theta_{j-1}^{(t-1)}, \theta_j^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}) - \Psi_i(\boldsymbol{\theta}^{(t-1)})}{\theta_j^{(t)} - \theta_j^{(t-1)}}. \tag{4.45}$$

However, Meng and Rubin [453] argue that this approach will not require fewer iterations overall, that the extra steps required to find $\hat{\boldsymbol{\theta}}$ first can be offset by starting SEM closer to $\hat{\boldsymbol{\theta}}$, and that the alternative is numerically less stable. Jamshidian and Jennrich survey a variety of methods for numerically differentiating $\boldsymbol{\Psi}$ or $\mathbf{l}'$ itself, including some they consider superior to SEM [345].

4.2.3.3 Bootstrapping Thorough discussion of bootstrapping is given in Chapter 9. In its simplest implementation, bootstrapping to obtain an estimated covariance matrix for EM would proceed as follows for i.i.d. observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$:

1. Calculate $\hat{\boldsymbol{\theta}}_{\mathrm{EM}}$ using a suitable EM approach applied to $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $j = 1$ and set $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}_{\mathrm{EM}}$.
2. Increment j. Sample *pseudo-data* $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$ completely at random from $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with replacement.
3. Calculate $\hat{\boldsymbol{\theta}}_j$ by applying the same EM approach to the pseudo-data $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$.
4. Stop if j is large enough; otherwise return to step 2.

For most problems, a few thousand iterations will suffice. At the end of the process, we have generated a collection of parameter estimates, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B$, where B denotes the total number of iterations used. Then the sample variance of these B estimates is the estimated variance of $\hat{\boldsymbol{\theta}}$. Conveniently, other aspects of the sampling distribution of $\hat{\boldsymbol{\theta}}$, such as correlations and quantiles, can be estimated using the corresponding sample estimates based on $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B$. Note that bootstrapping embeds the EM loop in a second loop of B iterations. This nested looping can be computationally burdensome when the solution of each EM problem is slow because of a high proportion of missing data or high dimensionality.

4.2.3.4 Empirical Information When the data are i.i.d., note that the score function is the sum of individual scores for each observation:

$$\frac{d \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbf{l}'(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \mathbf{l}'(\boldsymbol{\theta}|\mathbf{x}_i), \tag{4.46}$$

where we write the observed dataset as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Since the Fisher information matrix is defined to be the variance of the score function, this suggests estimating the information using the sample variance of the individual scores. The *empirical information* is defined as

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{l}'(\boldsymbol{\theta}|\mathbf{x}_i)\mathbf{l}'(\boldsymbol{\theta}|\mathbf{x}_i)^{\mathsf{T}} - \frac{1}{n^2}\mathbf{l}'(\boldsymbol{\theta}|\mathbf{x})\mathbf{l}'(\boldsymbol{\theta}|\mathbf{x})^{\mathsf{T}}. \tag{4.47}$$

This estimate has been discussed in the EM context in [450, 530]. The appeal of this approach is that all the terms in (4.47) are by-products of the M step: No additional analysis is required. To see this, note that $\boldsymbol{\theta}^{(t)}$ maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - l(\boldsymbol{\theta}|\mathbf{x})$ with respect to $\boldsymbol{\theta}$. Therefore, taking derivatives with respect to $\boldsymbol{\theta}$,

$$Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \mathbf{l}'(\boldsymbol{\theta}|\mathbf{x})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \tag{4.48}$$

Since $\mathbf{Q}'$ is ordinarily calculated at each M step, the individual terms in (4.47) are available.

4.2.3.5 Numerical Differentiation To estimate the Hessian, consider computing the numerical derivative of $\mathbf{l}'$ at $\hat{\boldsymbol{\theta}}$, one coordinate at a time, using (1.10). The first row of the estimated Hessian can be obtained by adding a small perturbation to the first coordinate of $\hat{\boldsymbol{\theta}}$, then computing the ratio of the difference between $\mathbf{l}'(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and at the perturbed value, relative to the magnitude of the perturbation. The remaining rows of the Hessian are approximated similarly. If a perturbation is too small, estimated partial derivatives may be inaccurate due to roundoff error; if a perturbation is too big, the estimates may also be inaccurate. Such numerical differentiation can be tricky to automate, especially when the components of $\hat{\boldsymbol{\theta}}$ have different scales. More sophisticated numerical differentiation strategies are surveyed in [345].

4.3 EM VARIANTS

4.3.1 Improving the E Step

The E step requires finding the expected log likelihood of the complete data conditional on the observed data. We have denoted this expectation as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. When this expectation is difficult to compute analytically, it can be approximated via Monte Carlo (see Chapter 6).

4.3.1.1 Monte Carlo EM Wei and Tanner [656] propose that the tth E step can be replaced with the following two steps:

1. Draw missing datasets $\mathbf{Z}_1^{(t)}, \ldots, \mathbf{Z}_{m^{(t)}}^{(t)}$ i.i.d. from $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})$. Each $\mathbf{Z}_j^{(t)}$ is a vector of all the missing values needed to complete the observed dataset, so $\mathbf{Y}_j = (\mathbf{x}, \mathbf{Z}_j)$ denotes a completed dataset where the missing values have been replaced by $\mathbf{Z}_j$.

2. Calculate $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \left(1/m^{(t)}\right) \sum_{j=1}^{m^{(t)}} \log f_{\mathbf{Y}}(\mathbf{Y}_j^{(t)}|\boldsymbol{\theta})$.

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

The recommended strategy is to let $m^{(t)}$ be small during early EM iterations and to increase $m^{(t)}$ as iterations progress to reduce the Monte Carlo variability introduced in $\hat{Q}$. Nevertheless, this *Monte Carlo EM algorithm (MCEM)* will not converge in the same sense as ordinary EM. As iterations proceed, values of $\boldsymbol{\theta}^{(t)}$ will eventually bounce around the true maximum, with a precision that depends on $m^{(t)}$. Discussion of the asymptotic convergence properties of MCEM is provided in [102]. A stochastic alternative to MCEM is discussed in [149].

Example 4.7 (Censored Exponential Data, Continued) In Example 4.5, it was easy to compute the conditional expectation of $l(\lambda|\mathbf{Y}) = n \log \lambda - \lambda \sum_{i=1}^{n} Y_i$ given the observed data. The result, given in (4.37), can be maximized to provide the ordinary EM update,

$$\lambda^{(t+1)} = \frac{n}{\sum_{i=1}^{n} x_i + C/\lambda^{(t)}}. \tag{4.49}$$

Application of MCEM is also easy. In this case,

$$\hat{Q}^{(t+1)}(\lambda|\lambda^{(t)}) = n \log \lambda - \frac{\lambda}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \mathbf{Y}_j^{\top} \mathbf{1}, \tag{4.50}$$

where $\mathbf{1}$ is a vector of ones and $\mathbf{Y}_j$ is the jth completed dataset comprising the uncensored data and simulated data $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jC})$ with $Z_{jk} - c_k \sim$ i.i.d. Exp$(\lambda^{(t)})$ for $k = 1, \ldots, C$ to replace the censored values. Setting $\hat{Q}'(\lambda|\lambda^{(t)}) = 0$ and solving for λ yields

$$\lambda^{(t+1)} = \frac{n}{\sum_{j=1}^{m^{(t)}} \mathbf{Y}_j^{\top} \mathbf{1}/m^{(t)}} \tag{4.51}$$

as the MCEM update.

The website for this book provides $n = 30$ observations, including $C = 17$ censored observations. Figure 4.2 compares the performance of MCEM and ordinary EM for estimating λ with these data. Both methods easily find the MLE $\hat{\lambda} = 0.2185$. For MCEM, we used $m^{(t)} = 5^{1+\lfloor t/10 \rfloor}$, where $\lfloor z \rfloor$ denotes the integer part of z. Fifty iterations were used altogether. Both algorithms were initiated from $\lambda^{(0)} = 0.5042$, which is the mean of all 30 data values disregarding censoring. □

4.3.2 Improving the M Step

One of the appeals of the EM algorithm is that the derivation and maximization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is often simpler than incomplete-data maximum likelihood calculations, since $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ relates to the complete-data likelihood. In some cases, however, the M step cannot be carried out easily even though the E step yielding $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is

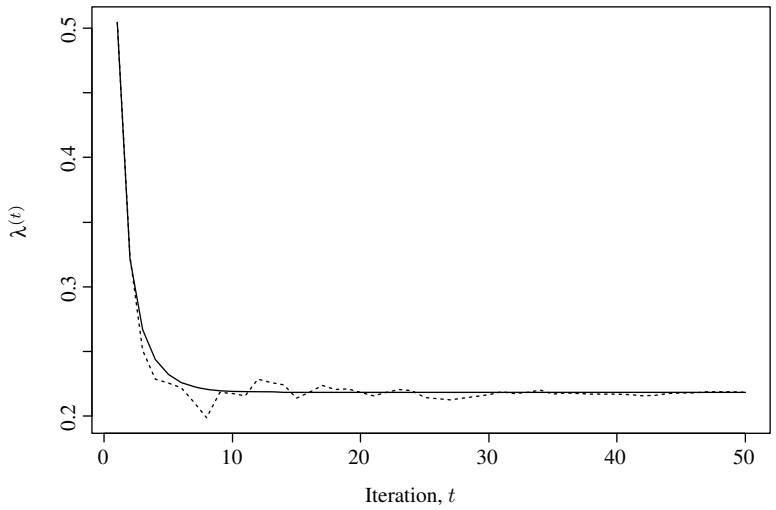

FIGURE 4.2 Comparison of iterations for EM (solid) and MCEM (dotted) for the censored exponential data discussed in Example 4.7.

straightforward. Several strategies have been proposed to facilitate the M step in such cases.

4.3.2.1 ECM Algorithm Meng and Rubin's *ECM algorithm* replaces the M step with a series of computationally simpler conditional maximization (CM) steps [454]. Each conditional maximization is designed to be a simple optimization problem that constrains $\boldsymbol{\theta}$ 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 tth E step a CM *cycle*. Thus, the tth iteration of ECM is composed of the tth E step and the tth CM cycle. Let S denote the total number of CM steps in each CM cycle. For $s = 1, \ldots, S$, the sth CM step in the tth cycle requires the maximization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ subject to (or conditional on) a constraint, say

$$\mathbf{g}_s(\boldsymbol{\theta}) = \mathbf{g}_s(\boldsymbol{\theta}^{(t+(s-1)/S)}) \tag{4.52}$$

where $\boldsymbol{\theta}^{(t+(s-1)/S)}$ 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 $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+S/S)}$ and proceed to the E step for the $(t+1)$th iteration.

Clearly any ECM is a GEM algorithm (Section 4.2.1), since each CM step increases Q. In order for ECM to be convergent, we need to ensure that each CM cycle permits search in any direction for a maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, so that ECM effectively maximizes over the original parameter space for $\boldsymbol{\theta}$ and not over some subspace. Precise conditions are discussed in [452, 454]; extensions of this method include [415, 456].

The art of constructing an effective ECM algorithm lies in choosing the constraints cleverly. Usually, it is natural to partition $\boldsymbol{\theta}$ into S subvectors, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$.

Then in the sth CM step, one might seek to maximize Q with respect to $\boldsymbol{\theta}_s$ while holding all other components of $\boldsymbol{\theta}$ fixed. This amounts to the constraint induced by the function $g_s(\boldsymbol{\theta}) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_S)$. A maximization strategy of this type has previously been termed *iterated conditional modes* [36]. If the conditional maximizations are obtained by finding the roots of score functions, the CM cycle can also be viewed as a Gauss–Seidel iteration (see Section 2.2.5).

Alternatively, the sth CM step might seek to maximize Q with respect to all other components of $\boldsymbol{\theta}$ while holding $\boldsymbol{\theta}_s$ fixed. In this case, $g_s(\boldsymbol{\theta}) = \boldsymbol{\theta}_s$. Additional systems of constraints can be imagined, depending on the particular problem context. A variant of ECM inserts an E step between each pair of CM steps, thereby updating Q at every stage of the CM cycle.

Example 4.8 (Multivariate Regression with Missing Values) A particularly illuminating example given by Meng and Rubin [454] involves multivariate regression with missing values. Let $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be n independent d-dimensional vectors observed from the d-variate normal model given by

$$\mathbf{U}_i \sim N_d\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \tag{4.53}$$

for $\mathbf{U}_i = (U_{i1}, \ldots, U_{id})$ and $\boldsymbol{\mu}_i = \mathbf{V}_i\boldsymbol{\beta}$, where the $\mathbf{V}_i$ are known $d \times p$ design matrices, $\boldsymbol{\beta}$ is a vector of p unknown parameters, and $\boldsymbol{\Sigma}$ is a $d \times d$ unknown variance–covariance matrix. There are many cases where $\boldsymbol{\Sigma}$ has some meaningful structure, but we consider $\boldsymbol{\Sigma}$ to be unstructured for simplicity. Suppose that some elements of some $\mathbf{U}_i$ are missing.

Begin by reordering the elements of $\mathbf{U}_i$, $\boldsymbol{\mu}_i$, and the rows of $\mathbf{V}_i$ so that for each i, the observed components of $\mathbf{U}_i$ are first and any missing components are last. For each $\mathbf{U}_i$, denote by $\boldsymbol{\beta}_i$ and $\boldsymbol{\Sigma}_i$ the corresponding reorganizations of the parameters. Thus, $\boldsymbol{\beta}_i$ and $\boldsymbol{\Sigma}_i$ are completely determined by $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and the pattern of missing data: They do not represent an expansion of the parameter space.

This notational reorganization allows us to write $\mathbf{U}_i = (\mathbf{U}_{\text{obs},i}, \mathbf{U}_{\text{miss},i})$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{\text{obs},i}, \boldsymbol{\mu}_{\text{miss},i})$, and

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{\text{obs},i} & \boldsymbol{\Sigma}_{\text{cross},i} \\ \boldsymbol{\Sigma}_{\text{cross},i}^{\mathsf{T}} & \boldsymbol{\Sigma}_{\text{miss},i} \end{pmatrix}. \tag{4.54}$$

The full set of observed data can be denoted $\mathbf{U}_{\text{obs}} = (\mathbf{U}_{\text{obs},1}, \ldots, \mathbf{U}_{\text{obs},n})$.

The observed-data log likelihood function is

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{u}_{\text{obs}}) = -\frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{\Sigma}_{\text{obs},i}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{u}_{\text{obs},i} - \boldsymbol{\mu}_{\text{obs},i})^{\mathsf{T}}\boldsymbol{\Sigma}_{\text{obs},i}^{-1}(\mathbf{u}_{\text{obs},i} - \boldsymbol{\mu}_{\text{obs},i})$$

up to an additive constant. This likelihood is quite tedious to work with or to maximize. Note, however, that the complete-data sufficient statistics are given by $\sum_{i=1}^{n} U_{ij}$ for $j = 1, \ldots, d$ and $\sum_{i=1}^{n} U_{ij}U_{ik}$ for $j, k = 1, \ldots, d$. Thus, the E step amounts to finding the expected values of these sufficient statistics conditional on the observed data and current parameter values $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$.

Now for $j = 1, \ldots, d$

$$E\left\{ \sum_{i=1}^{n} U_{ij} \middle| \mathbf{u}_{\text{obs}}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)} \right\} = \sum_{i=1}^{n} a_{ij}^{(t)}, \tag{4.55}$$

where

$$a_{ij}^{(t)} = \begin{cases} \alpha_{ij}^{(t)} & \text{if } U_{ij} \text{ is missing,} \\ u_{ij} & \text{if } U_{ij} = u_{ij} \text{ is observed,} \end{cases} \tag{4.56}$$

and $\alpha_{ij}^{(t)} = E\{U_{ij} | \mathbf{u}_{\text{obs},i}, \boldsymbol{\beta}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}\}$. Similarly, for $j, k = 1, \ldots, d$,

$$E\left\{ \sum_{i=1}^{n} U_{ij} U_{ik} \middle| \mathbf{u}_{\text{obs}}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)} \right\} = \sum_{i=1}^{n} (a_{ij}^{(t)} a_{ik}^{(t)} + b_{ijk}^{(t)}), \tag{4.57}$$

where

$$b_{ijk}^{(t)} = \begin{cases} \gamma_{ijk}^{(t)} & \text{if } U_{ij} \text{ and } U_{ik} \text{ are both missing,} \\ 0 & \text{otherwise,} \end{cases} \tag{4.58}$$

and $\gamma_{ijk}^{(t)} = \text{cov}\{U_{ij}, U_{ik} | \mathbf{u}_{\text{obs},i}, \boldsymbol{\beta}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}\}$.

Fortunately, the derivation of the $\alpha_{ij}^{(t)}$ and $\gamma_{ijk}^{(t)}$ is fairly straightforward. The conditional distribution of $\mathbf{U}_{\text{miss},i} | (\mathbf{u}_{\text{obs},i}, \boldsymbol{\beta}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)})$ is

$$N\left(\boldsymbol{\mu}_{\text{miss},i}^{(t)} + \boldsymbol{\Sigma}_{\text{cross},i} \boldsymbol{\Sigma}_{\text{miss},i}^{-1} (\mathbf{u}_{\text{obs},i} - \boldsymbol{\mu}_{\text{obs},i}^{(t)}), \boldsymbol{\Sigma}_{\text{obs},i} - \boldsymbol{\Sigma}_{\text{cross},i} \boldsymbol{\Sigma}_{\text{miss},i}^{-1} \boldsymbol{\Sigma}_{\text{cross},i}^{T} \right).$$

The values for $\alpha_{ij}^{(t)}$ and $\boldsymbol{\gamma}_{ijk}^{(t)}$ can be read from the mean vector and variance–covariance matrix of this distribution, respectively. Knowing these, $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ can be formed following (4.26).

Having thus achieved the E step, we turn now to the M step. The high dimensionality of the parameter space and the complexity of the observed-data likelihood renders difficult any direct implementation of the M step, whether by direct maximization or by reference to the exponential family setup. However, implementing an ECM strategy is straightforward using $S = 2$ conditional maximization steps in each CM cycle.

Treating $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ separately allows easy constrained optimizations of Q. First, if we impose the constraint that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(t)}$, then we can maximize the constrained version of $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ with respect to $\boldsymbol{\beta}$ by using the weighted least squares estimate

$$\boldsymbol{\beta}^{(t+1/2)} = \left(\sum_{i=1}^{n} \mathbf{V}_i^{\mathsf{T}} (\boldsymbol{\Sigma}_i^{(t)})^{-1} \mathbf{V}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{V}_i^{\mathsf{T}} (\boldsymbol{\Sigma}_i^{(t)})^{-1} \mathbf{a}_i^{(t)} \right), \tag{4.59}$$

where $\mathbf{a}_i^{(t)} = (a_{i1}^{(t)}, \ldots, a_{id}^{(t)})^{\mathsf{T}}$ and $\boldsymbol{\Sigma}_i^{(t)}$ is treated as a known variance–covariance matrix. This ensures that $Q(\boldsymbol{\beta}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \geq Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)})$. This constitutes the first of two CM steps.

The second CM step follows from the fact that setting $\boldsymbol{\Sigma}^{(t+2/2)}$ equal to

$$E\left\{ \frac{1}{n} \sum_{i=1}^{n} (\mathbf{U}_i - \mathbf{V}_i \boldsymbol{\beta}^{(t+1/2)})(\mathbf{U}_i - \mathbf{V}_i \boldsymbol{\beta}^{(t+1/2)})^{\mathsf{T}} \;\middle|\; \mathbf{u}_{\text{obs}}, \boldsymbol{\beta}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t)} \right\} \tag{4.60}$$

maximizes $Q(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ with respect to $\boldsymbol{\Sigma}$ subject to the constraint that $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t+1/2)}$, because this amounts to plugging in $\alpha_{ij}^{(t)}$ and $\gamma_{ijk}^{(t)}$ values where necessary and computing the sample covariance matrix of the completed data. This update guarantees

$$\begin{aligned} Q(\boldsymbol{\beta}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t+2/2)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)}) &\geq Q(\boldsymbol{\beta}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \\ &\geq Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)}). \end{aligned} \tag{4.61}$$

Together, the two CM steps yield $(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)}) = (\boldsymbol{\beta}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t+2/2)})$ and ensure an increase in the Q function.

The E step and the CM cycle described here can each be implemented using familiar closed-form analytic results; no numerical integration or maximization is required. After updating the parameters with the CM cycle described above, we return to another E step, and so forth. In summary, ECM alternates between (i) creating updated complete datasets and (ii) sequentially estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in turn by fixing the other at its current value and using the current completed-data component. □

4.3.2.2 EM Gradient Algorithm If maximization cannot be accomplished analytically, then one might consider carrying out each M step using an iterative numerical optimization approach like those discussed in Chapter 2. 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 [407]. The M step is replaced with the update given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})^{-1} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathbf{Q}'(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \tag{4.62}$$

$$= \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})^{-1} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathbf{l}'(\boldsymbol{\theta}^{(t)} | \mathbf{x}), \tag{4.63}$$

where $\mathbf{l}'(\boldsymbol{\theta}^{(t)} | \mathbf{x})$ is the evaluation of the score function at the current iterate. Note that (4.63) follows from the observation in Section 4.2.3.4 that $\boldsymbol{\theta}^{(t)}$ maximizes $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) - l(\boldsymbol{\theta} | \mathbf{x})$. This *EM gradient algorithm* has the same rate of convergence to $\hat{\boldsymbol{\theta}}$ as the full EM algorithm. Lange discusses conditions under which ascent can be ensured, and

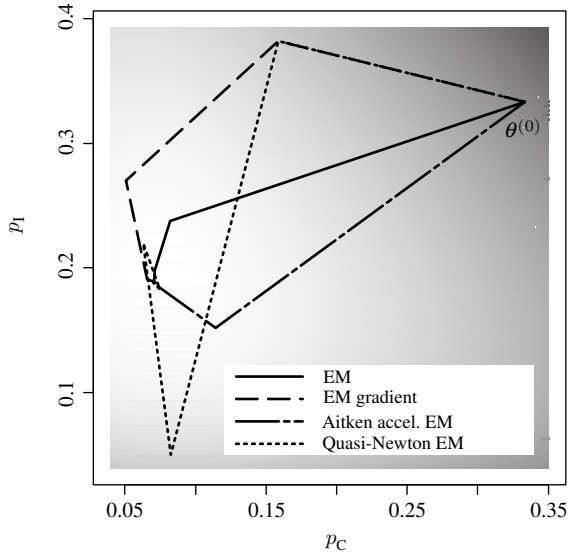

FIGURE 4.3 Steps taken by the EM gradient algorithm (long dashes). Ordinary EM steps are shown with the solid line. Steps from two methods from later sections (Aitken and quasi-Newton acceleration) are also shown, as indicated in the key. The observed-data log likelihood is shown with the gray scale, with light shading corresponding to high likelihood. All algorithms were started from $p_C = p_I = \frac{1}{3}$.

scalings of the update increment to speed convergence [407]. In particular, when $\mathbf{Y}$ has an exponential family distribution with canonical parameter $\boldsymbol{\theta}$, ascent is ensured and the method matches that of Titterington [634]. In other cases, the step can be scaled down to ensure ascent (as discussed in Section 2.2.2.1), but inflating steps speeds convergence. For problems with a high proportion of missing information, Lange suggests considering doubling the step length [407].

Example 4.9 (Peppered Moths, Continued) Continuing Example 4.2, we apply the EM gradient algorithm to these data. It is straightforward to show

$$\frac{d^2 Q(\mathbf{p}|\mathbf{p}^{(t)})}{dp_C^2} = -\frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C^2} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1 - p_C - p_I)^2}, \tag{4.64}$$

$$\frac{d^2 Q(\mathbf{p}|\mathbf{p}^{(t)})}{dp_I^2} = -\frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I^2} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1 - p_C - p_I)^2}, \tag{4.65}$$

and

$$\frac{d^2 Q(\mathbf{p}|\mathbf{p}^{(t)})}{dp_C dp_I} = -\frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1 - p_C - p_I)^2}. \tag{4.66}$$

Figure 4.3 shows the steps taken by the resulting EM gradient algorithm, starting from $p_C = p_I = p_T = \frac{1}{3}$. Step halving was implemented to ensure ascent. The first

step heads somewhat in the wrong direction, but in subsequent iterations the gradient steps progress quite directly uphill. The ordinary EM steps are shown for comparison in this figure. □

4.3.3 Acceleration Methods

The slow convergence of the EM algorithm is a notable drawback. Several techniques have been suggested for using the relatively simple analytic setup from EM to motivate particular forms for Newton-like steps. In addition to the two approaches described below, approaches that cleverly expand the parameter space in manners that speed convergence without affecting the marginal inference about $\boldsymbol{\theta}$ are topics of recent interest [421, 456].

4.3.3.1 Aitken Acceleration Let $\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)}$ be the next iterate obtained by the standard EM algorithm from $\boldsymbol{\theta}^{(t)}$. Recall that the Newton update to maximize the log likelihood would be

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})^{-1}\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x}). \tag{4.67}$$

The EM framework suggests a replacement for $\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x})$. In Section 4.2.3.4 we noted that $\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x}) = \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$. Expanding $\mathbf{Q}'$ around $\boldsymbol{\theta}^{(t)}$, evaluated at $\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)}$, yields

$$\mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)}} \approx \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)} - \boldsymbol{\theta}^{(t)}), \tag{4.68}$$

where $\hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta}^{(t)})$ is defined in (4.31). Since $\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)}$ maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$, the left-hand side of (4.68) equals zero. Therefore

$$\mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \approx \hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)} - \boldsymbol{\theta}^{(t)}). \tag{4.69}$$

Thus, from (4.67) we arrive at

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})^{-1}\hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}_{\mathrm{EM}}^{(t+1)} - \boldsymbol{\theta}^{(t)}). \tag{4.70}$$

This update—relying on the approximation in (4.69)—is an example of a general strategy known as *Aitken acceleration* and was proposed for EM by Louis [424]. Aitken acceleration of EM is precisely the same as applying the Newton–Raphson method to find a zero of $\boldsymbol{\Psi}(\boldsymbol{\theta}) - \boldsymbol{\theta}$, where $\boldsymbol{\Psi}$ is the mapping defined by the ordinary EM algorithm producing $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})$ [343].

Example 4.10 (Peppered Moths, Continued) This acceleration approach can be applied to Example 4.2. Obtaining $\mathbf{l}''$ is analytically more tedious than the simpler derivations employed for other EM approaches to this problem. Figure 4.3 shows the Aitken accelerated steps, which converge quickly to the solution. The procedure was started from $p_{\mathrm{C}} = p_{\mathrm{I}} = p_{\mathrm{T}} = \frac{1}{3}$, and step halving was used to ensure ascent. □

Aitken acceleration is sometimes criticized for potential numerical instabilities and convergence failures [153, 344]. Further, when $\mathbf{l}''(\boldsymbol{\theta}|\mathbf{x})$ is difficult to compute, this approach cannot be applied without overcoming the difficulty [20, 345, 450].

Section 4.2.1 noted that the EM algorithm converges at a linear rate that depends on the fraction of missing information. The updating increment given in (4.70) is, loosely speaking, scaled by the ratio of the complete information to the observed information. Thus, when a greater proportion of the information is missing, the nominal EM steps are inflated more.

Newton's method converges quadratically, but (4.69) only becomes a precise approximation as $\boldsymbol{\theta}^{(t)}$ nears $\hat{\boldsymbol{\theta}}$. Therefore, we should only expect this acceleration approach to enhance convergence only as preliminary iterations hone $\boldsymbol{\theta}$ sufficiently. The acceleration should not be employed without having taken some initial iterations of ordinary EM so that (4.69) holds.

4.3.3.2 Quasi-Newton Acceleration The quasi-Newton optimization method discussed in Section 2.2.2.3 produces updates according to

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - (\mathbf{M}^{(t)})^{-1}\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x}) \tag{4.71}$$

for maximizing $\mathbf{l}(\boldsymbol{\theta}|\mathbf{x})$ with respect to $\boldsymbol{\theta}$, where $\mathbf{M}^{(t)}$ is an approximation to $\mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})$. Within the EM framework, one can decompose $\mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})$ into a part computed during EM and a remainder. By taking two derivatives of (4.19), we obtain

$$\mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x}) = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{H}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \tag{4.72}$$

at iteration t. The remainder is the last term in (4.72); suppose we approximate it by $\mathbf{B}^{(t)}$. Then by using

$$\mathbf{M}^{(t)} = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{B}^{(t)} \tag{4.73}$$

in (4.71) we obtain a *quasi-Newton EM acceleration*.

A key feature of the approach is how $\mathbf{B}^{(t)}$ approximates $\mathbf{H}''(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$. The idea is to start with $\mathbf{B}^{(0)} = \mathbf{0}$ and gradually accumulate information about $\mathbf{H}''$ as iterations progress. The information is accumulated using a sequence of secant conditions, as is done in ordinary quasi-Newton approaches (Section 2.2.2.3).

Specifically, we can require that $\mathbf{B}^{(t)}$ satisfy the secant condition

$$\mathbf{B}^{(t+1)}\mathbf{a}^{(t)} = \mathbf{b}^{(t)}, \tag{4.74}$$

where

$$\mathbf{a}^{(t)} = \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} \tag{4.75}$$

and

$$\mathbf{b}^{(t)} = \mathbf{H}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t+1)}} - \mathbf{H}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \tag{4.76}$$

Recalling the update (2.49), we can satisfy the secant condition by setting

$$\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} + c^{(t)}\mathbf{v}^{(t)}(\mathbf{v}^{(t)})^{\mathsf{T}}, \tag{4.77}$$

where $\mathbf{v}^{(t)} = \mathbf{b}^{(t)} - \mathbf{B}^{(t)}\mathbf{a}^{(t)}$ and $c^{(t)} = 1/[(\mathbf{v}^{(t)})^T\mathbf{a}^{(t)}]$.

Lange proposed this *quasi-Newton* EM algorithm, along with several suggested strategies for improving its performance [408]. First, he suggested starting with $\mathbf{B}^{(0)} = \mathbf{0}$. Note that this implies that the first increment will equal the EM gradient increment. Indeed, the EM gradient approach is exact Newton–Raphson for maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, whereas the approach described here evolves into approximate Newton–Raphson for maximizing $l(\boldsymbol{\theta}|\mathbf{x})$.

Second, Davidon's [134] update is troublesome if $(\mathbf{v}^{(t)})^{\mathsf{T}}\mathbf{a}^{(t)} = 0$ or is small compared to $\|\mathbf{v}^{(t)}\| \cdot \|\mathbf{a}^{(t)}\|$. In such cases, we may simply set $\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)}$.

Third, there is no guarantee that $\mathbf{M}^{(t)} = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{B}^{(t)}$ will be negative definite, which would ensure ascent at the tth step. Therefore, we may scale $\mathbf{B}^{(t)}$ and use $\mathbf{M}^{(t)} = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \alpha^{(t)}\mathbf{B}^{(t)}$ where, for example, $\alpha^{(t)} = 2^{-m}$ for the smallest positive integer that makes $\mathbf{M}^{(t)}$ negative definite.

Finally, note that $\mathbf{b}^{(t)}$ may be expressed entirely in terms of $\mathbf{Q}'$ functions since

$$\begin{aligned}
\mathbf{b}^{(t)} &= \mathbf{H}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t+1)}} - \mathbf{H}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} && (4.78)\\
&= 0 - \mathbf{H}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} && (4.79)\\
&= \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. && (4.80)
\end{aligned}$$

Equation (4.79) follows from (4.19) and the fact that $l(\boldsymbol{\theta}|\mathbf{x}) - \mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ has its minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. The derivative at this minimum must be zero, forcing $\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x}) = \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$, which allows (4.80).

Example 4.11 (Peppered Moths, Continued) We can apply quasi-Newton acceleration to Example 4.2, using the expressions for $\mathbf{Q}''$ given in (4.64)–(4.66) and obtaining $\mathbf{b}^{(t)}$ from (4.80). The procedure was started from $p_{\mathsf{C}} = p_{\mathsf{I}} = p_{\mathsf{T}} = \frac{1}{3}$ and $\mathbf{B}^{(0)} = \mathbf{0}$, with step halving to ensure ascent.

The results are shown in Figure 4.3. Note that $\mathbf{B}^{(0)} = \mathbf{0}$ means that the first quasi-Newton EM step will match the first EM gradient step. The second quasi-Newton EM step completely overshoots the ridge of highest likelihood, resulting in a step that is just barely uphill. In general, the quasi-Newton EM procedure behaves like other quasi-Newton methods: There can be a tendency to step beyond the solution or to converge to a local maximum rather than a local minimum. With suitable safeguards, the procedure is fast and effective in this example. □

The quasi-Newton EM requires the inversion of $\mathbf{M}^{(t)}$ at step t. Lange et al. describe a quasi-Newton approach based on the approximation of $-\mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})$ by some $\mathbf{M}^{(t)}$ that relies on an inverse-secant update [409, 410]. In addition to avoiding computationally burdensome matrix inversions, such updates to $\boldsymbol{\theta}^{(t)}$ and $\mathbf{B}^{(t)}$ can be expressed entirely in terms of $\mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x})$ and ordinary EM increments when the M step is solvable.

TABLE 4.2 Frequencies of respondents reporting numbers of risky sexual encounters; see Problem 4.2.

Encounters, i	0	1	2	3	4	5	6	7	8
Frequency, n_i	379	299	222	145	109	95	73	59	45
Encounters, i	9	10	11	12	13	14	15	16	
Frequency, n_i	30	24	12	4	2	0	1	1	

Jamshidian and Jennrich elaborate on inverse-secant updating and discuss the more complex BFGS approach [344]. These authors also provide a useful survey of a variety of EM acceleration approaches and a comparison of effectiveness. Some of their approaches converge faster on examples than does the approach described above. In a related paper, they present a conjugate gradient acceleration of EM [343].

PROBLEMS

4.1. Recall the peppered moth analysis introduced in Example 4.2. In the field, it is quite difficult to distinguish the *insularia* and *typica* phenotypes due to variations in wing color and mottle. In addition to the 622 moths mentioned in the example, suppose the sample collected by the researchers actually included $n_U = 578$ more moths that were known to be *insularia* or *typical* but whose exact phenotypes could not be determined.

- **a.** Derive the EM algorithm for maximum likelihood estimation of p_C, p_I, and p_I for this modified problem having observed data n_C, n_I, n_T, and n_U as given above.
- **b.** Apply the algorithm to find the MLEs.
- **c.** Estimate the standard errors and pairwise correlations for $\hat{p}_C$, $\hat{p}_I$, and $\hat{p}_I$ using the SEM algorithm.
- **d.** Estimate the standard errors and pairwise correlations for $\hat{p}_C$, $\hat{p}_I$, and $\hat{p}_I$ by bootstrapping.
- **e.** Implement the EM gradient algorithm for these data. Experiment with step halving to ensure ascent and with other step scalings that may speed convergence.
- **f.** Implement Aitken accelerated EM for these data. Use step halving.
- **g.** Implement quasi-Newton EM for these data. Compare performance with and without step halving.
- **h.** Compare the effectiveness and efficiency of the standard EM algorithm and the three variants in (e), (f), and (g). Use step halving to ensure ascent with the three variants. Base your comparison on a variety of starting points. Create a graph analogous to Figure 4.3.

4.2. Epidemiologists are interested in studying the sexual behavior of individuals at risk for HIV infection. Suppose 1500 gay men were surveyed and each was asked how many risky sexual encounters he had in the previous 30 days. Let n_i denote the number of respondents reporting i encounters, for $i = 1, \ldots, 16$. Table 4.2 summarizes the responses.

These data are poorly fitted by a Poisson model. It is more realistic to assume that the respondents comprise three groups. First, there is a group of people who, for whatever reason, report zero risky encounters even if this is not true. Suppose a respondent has probability α of belonging to this group.

With probability β, a respondent belongs to a second group representing typical behavior. Such people respond truthfully, and their numbers of risky encounters are assumed to follow a Poisson(μ) distribution.

Finally, with probability $1 - \alpha - \beta$, a respondent belongs to a high-risk group. Such people respond truthfully, and their numbers of risky encounters are assumed to follow a Poisson(λ) distribution.

The parameters in the model are α, β, μ, and λ. At the tth iteration of EM, we use $\boldsymbol{\theta}^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \mu^{(t)}, \lambda^{(t)})$ to denote the current parameter values. The likelihood of the observed data is given by

$$L(\boldsymbol{\theta}|n_0, \ldots, n_{16}) \propto \prod_{i=0}^{16} \left[\frac{\pi_i(\boldsymbol{\theta})}{i!}\right]^{n_i}, \tag{4.81}$$

where

$$\pi_i(\boldsymbol{\theta}) = \alpha 1_{\{i=0\}} + \beta\mu^i \exp\{-\mu\} + (1 - \alpha - \beta)\lambda^i \exp\{-\lambda\} \tag{4.82}$$

for $i = 1, \ldots, 16$.

The observed data are $n_0, \ldots, n_{16}$. The complete data may be construed to be $n_{z,0}, n_{t,0}, \ldots, n_{t,16}$, and $n_{p,0}, \ldots, n_{p,16}$, where $n_{k,i}$ denotes the number of respondents in group k reporting i risky encounters and $k = z$, t, and p correspond to the zero, typical, and promiscuous groups, respectively. Thus, $n_0 = n_{z,0} + n_{t,0} + n_{p,0}$ and $n_i = n_{t,i} + n_{p,i}$ for $i = 1, \ldots, 16$. Let $N = \sum_{i=0}^{16} n_i = 1500$.

Define

$$z_0(\boldsymbol{\theta}) = \frac{\alpha}{\pi_0(\boldsymbol{\theta})}, \tag{4.83}$$

$$t_i(\boldsymbol{\theta}) = \frac{\beta\mu^i \exp\{-\mu\}}{\pi_i(\boldsymbol{\theta})}, \tag{4.84}$$

$$p_i(\boldsymbol{\theta}) = \frac{(1 - \alpha - \beta)\lambda^i \exp\{-\lambda\}}{\pi_i(\boldsymbol{\theta})} \tag{4.85}$$

for $i = 0, \ldots, 16$. These correspond to probabilities that respondents with i risky encounters belong to the various groups.

a. Show that the EM algorithm provides the following updates:

$$\alpha^{(t+1)} = \frac{n_0 z_0(\boldsymbol{\theta}^{(t)})}{N}, \tag{4.86}$$

$$\beta^{(t+1)} = \sum_{i=0}^{16} \frac{n_i t_i(\boldsymbol{\theta}^{(t)})}{N}, \tag{4.87}$$

$$\mu^{(t+1)} = \frac{\sum_{i=0}^{16} i\, n_i t_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(t)})}, \tag{4.88}$$

$$\lambda^{(t+1)} = \frac{\sum_{i=0}^{16} i\, n_i p_i(\boldsymbol{\theta}^{(t)})}{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(t)})}. \tag{4.89}$$

b. Estimate the parameters of the model, using the observed data.

c. Estimate the standard errors and pairwise correlations of your parameter estimates, using any available method.

4.3. The website for this book contains 50 trivariate data points drawn from the $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Some data points have missing values in one or more coordinates. Only 27 of the 50 observations are complete.

a. Derive the EM algorithm for joint maximum likelihood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It is easiest to recall that the multivariate normal density is in the exponential family.

b. Determine the MLEs from a suitable starting point. Investigate the performance of the algorithm, and comment on your results.

c. Consider Bayesian inference for $\boldsymbol{\mu}$ when

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 & 1.2 \\ 0.6 & 0.5 & 0.5 \\ 1.2 & 0.5 & 3.0 \end{pmatrix}$$

is known. Assume independent priors for the three elements of $\boldsymbol{\mu}$. Specifically, let the jth prior be

$$f(\mu_j) = \frac{\exp\{-(\mu_j - \alpha_j)/\beta_j\}}{\beta_j\left[1 + \exp\{-(\mu_j - \alpha_j)/\beta_j\}\right]^2},$$

where $(\alpha_1, \alpha_2, \alpha_3) = (2, 4, 6)$ and $\beta_j = 2$ for $j = 1, 2, 3$. Comment on difficulties that would be faced in implementing a standard EM algorithm for estimating the posterior mode for $\boldsymbol{\mu}$. Implement a gradient EM algorithm, and evaluate its performance.

d. Suppose that $\boldsymbol{\Sigma}$ is unknown in part (c) and that an improper uniform prior is adopted, that is, $f(\Sigma) \propto 1$ for all positive definite $\boldsymbol{\Sigma}$. Discuss ideas for how to estimate the posterior mode for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

4.4. Suppose we observe lifetimes for 14 gear couplings in certain mining equipment, as given in Table 4.3 (in years). Some of these data are right censored because the equipment was replaced before the gear coupling failed. The censored data are in parentheses; the actual lifetimes for these components may be viewed as missing.

Model these data with the Weibull distribution, having density function $f(x) = abx^{b-1}\exp\{-ax^b\}$ for $x > 0$ and parameters a and b. Recall that Problem 2.3 in Chapter 2 provides more details about such models. Construct an EM algorithm to estimate a and b. Since the Q function involves expectations that are analytically unavailable, adopt the MCEM strategy where necessary. Also, optimization of Q cannot be completed analytically. Therefore, incorporate the ECM strategy of conditionally maximizing with respect to each parameter separately, applying a one-dimensional

TABLE 4.3 Fourteen lifetimes for mining equipment gear couplings, in years. Right-censored values are in parenthesis. In these cases, we know only that the lifetime was at least as long as the given value.

(6.94)	5.50	4.54	2.14	(3.65)	(3.40)	(4.38)
10.24	4.56	9.42	(4.55)	(4.15)	5.64	(10.23)

Newton-like optimizer where needed. Past observations suggest $(a, b) = (0.003, 2.5)$ may be a suitable starting point. Discuss the convergence properties of the procedure you develop, and the results you obtain. What are the advantages and disadvantages of your technique compared to direct maximization of the observed-data likelihood using, say, a two-dimensional quasi-Newton approach?

4.5. A *hidden Markov model (HMM)* can be used to describe the joint probability of a sequence of unobserved (hidden) discrete-state variables, $\mathbf{H} = (H_0, \ldots, H_n)$, and a sequence of corresponding observed variables $\mathbf{O} = (O_0, \ldots, O_n)$ for which O_i is dependent on H_i for each i. We say that H_i emits O_i; consideration here is limited to discrete emission variables. Let the state spaces for elements of $\mathbf{H}$ and $\mathbf{O}$ be $\mathcal{H}$ and $\mathcal{E}$, respectively.

Let $\mathbf{O}_{\leq j}$ and $\mathbf{O}_{>j}$ denote the portions of $\mathbf{O}$ with indices not exceeding j and exceeding j, respectively, and define the analogous partial sequences for $\mathbf{H}$. Under an HMM, the H_i have the Markov property

$$P[H_i|\mathbf{H}_{\leq i-1}, O_0] = P[H_i|H_{i-1}] \tag{4.90}$$

and the emissions are conditionally independent, so

$$P[O_i|\mathbf{H}, \mathbf{O}_{\leq i-1}, \mathbf{O}_{>i}] = P[O_i|H_i]. \tag{4.91}$$

Time-homogeneous transitions of the hidden states are governed by transition probabilities $p(h, h^*) = P[H_{i+1} = h^*|H_i = h]$ for $h, h^* \in \mathcal{H}$. The distribution for H_0 is parameterized by $\pi(h) = P[H_0 = h]$ for $h \in \mathcal{H}$. Finally, define emission probabilities $e(h, o) = P[O_i = o|H_i = h]$ for $h \in \mathcal{H}$ and $o \in \mathcal{E}$. Then the parameter set $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{P}, \mathbf{E})$ completely parameterizes the model, where $\boldsymbol{\pi}$ is a vector of initial-state probabilities, $\mathbf{P}$ is a matrix of transition probabilities, and $\mathbf{E}$ is a matrix of emission probabilities.

For an observed sequence $\mathbf{o}$, define the *forward variables* to be

$$\alpha(i, h) = P[\mathbf{O}_{\leq i} = \mathbf{o}_{\leq i},\ H_i = h] \tag{4.92}$$

and the *backward variables* to be

$$\beta(i, h) = P[\mathbf{O}_{>i} = \mathbf{o}_{>i}|H_i = h] \tag{4.93}$$

for $i = 1, \ldots, n$ and each $h \in \mathcal{H}$. Our notation suppresses the dependence of the forward and backward variables on $\boldsymbol{\theta}$. Note that

$$P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}] = \sum_{h\in\mathcal{H}} \alpha(n, h) = \sum_{h\in\mathcal{H}} \pi(h)e(h, o_0)\beta(0, h). \tag{4.94}$$

The forward and backward variables are also useful for computing the probability that state h occurred at the ith position of the sequence given $\mathbf{O} = \mathbf{o}$ according to $P[H_i = h|\mathbf{O} = \mathbf{o}, \boldsymbol{\theta}] = \sum_{h\in H} \alpha(i, h)\beta(i, h)/P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]$, and expectations of functions of the states with respect to these probabilities.

a. Show that the following algorithms can be used to calculate $\alpha(i, h)$ and $\beta(i, h)$. The *forward algorithm* is

- Initialize $\alpha(0, h) = \pi(h)e(h, o_0)$.
- For $i = 0, \ldots, n-1$, let $\alpha(i+1, h) = \sum_{h^* \in \mathcal{H}} \alpha(i, h^*)p(h^*, h)e(h, o_{i+1})$.

The *backward algorithm* is

- Initialize $\beta(n, h) = 1$.
- For $i = n, \ldots, 1$, let $\beta(i-1, h) = \sum_{h^* \in \mathcal{H}} p(h, h^*)e(h^*, o_i)\beta(h, i)$.

These algorithms provide very efficient methods for finding $P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]$ and other useful probabilities, compared to naively summing over all possible sequences of states.

b. Let $N(h)$ denote the number of times $H_0 = h$, let $N(h, h^*)$ denote the number of transitions from h to h^*, and let $N(h, o)$ denote the number of emissions of o when the underlying state is h. Prove that these random variables have the following expectations:

$$E\{N(h)\} = \frac{\alpha(0, h)\beta(0, h)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]}, \tag{4.95}$$

$$E\{N(h, h^*)\} = \sum_{i=0}^{n-1} \frac{\alpha(i, h)p(h, h^*)e(h^*, o_{i+1})\beta(i+1, h^*)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]}, \tag{4.96}$$

$$E\{N(h, o)\} = \sum_{i:\ O_i = o} \frac{\alpha(i, h)\beta(i, h)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]}. \tag{4.97}$$

c. The *Baum–Welch algorithm* efficiently estimates the parameters of an HMM [25]. Fitting these models has proven extremely useful in diverse applications including statistical genetics, signal processing and speech recognition, problems involving environmental time series, and Bayesian graphical networks [172, 236, 361, 392, 523]. Starting from some initial values $\boldsymbol{\theta}^{(0)}$, the Baum–Welch algorithm proceeds via iterative application of the following update formulas:

$$\pi(h)^{(t+1)} = \frac{E\{N(h)|\boldsymbol{\theta}^{(t)}\}}{\sum_{h^* \in \mathcal{H}} E\{N(h^*)|\boldsymbol{\theta}^{(t)}\}}, \tag{4.98}$$

$$p(h, h^*)^{(t+1)} = \frac{E\{N(h, h^*)|\boldsymbol{\theta}^{(t)}\}}{\sum_{h^{**} \in \mathcal{H}} E\{N(h, h^{**})|\boldsymbol{\theta}^{(t)}\}}, \tag{4.99}$$

$$e(h, o)^{(t+1)} = \frac{E\{N(h, o)|\boldsymbol{\theta}^{(t)}\}}{\sum_{o^* \in \mathcal{E}} E\{N(h, o^*)|\boldsymbol{\theta}^{(t)}\}}. \tag{4.100}$$

Prove that the Baum–Welch algorithm is an EM algorithm. It is useful to begin by noting that the complete data likelihood is given by

$$\prod_{h \in \mathcal{H}} \pi(h)^{N(h)} \prod_{h \in \mathcal{H}} \prod_{o \in \mathcal{E}} e(h, o)^{N(h,o)} \prod_{h \in \mathcal{H}} \prod_{h^* \in \mathcal{H}} p(h, h^*)^{N(h,h^*)}. \tag{4.101}$$

d. Consider the following scenario. In Flip's left pocket is a penny; in his right pocket is a dime. On a fair toss, the probability of showing a head is p for the penny

and d for the dime. Flip randomly chooses a coin to begin, tosses it, and reports the outcome (heads or tails) without revealing which coin was tossed. Then, Flip decides whether to use the same coin for the next toss, or to switch to the other coin. He switches coins with probability s, and retains the same coin with probability $1 - s$. The outcome of the second toss is reported, again not revealing the coin used. This process is continued for a total of 200 coin tosses. The resulting sequence of heads and tails is available from the website for this book. Use the Baum–Welch algorithm to estimate p, d, and s.

e. Only for students seeking extra challenge: Derive the Baum–Welch algorithm for the case when the dataset consists of M independent observation sequences arising from a HMM. Simulate such data, following the coin example above. (You may wish to mimic the single-sequence data, which were simulated using $p = 0.25$, $d = 0.85$, and $s = 0.1$.) Code the Baum–Welch algorithm, and test it on your simulated data.

In addition to considering multiple sequences, HMMs and the Baum–Welch algorithm can be generalized for estimation based on more general emission variables and emission and transition probabilities that have more complex parameterizations, including time inhomogeneity.