5.4 Problems

5.1 Use a numerical maximizer to find the maximum of

$$f(x) = [\cos(50x) + \sin(20x)]^2 .$$

Compare to the results of a stochastic exploration. (*Note:* If you use `R` the functions are `optim` and `optimize`.)

5.2 For each of the likelihood functions in Exercise 1.1, find the maximum with both `R` function `optimize` and stochastic exploration.

5.3 (Ó Ruanaidh and Fitzgerald 1996) Consider the setup of Example 5.1.

(a) Show that (5.2) holds.

(b) Discuss the validity of the approximation

$$\begin{aligned}\mathbf{x}^t\mathbf{x} - \mathbf{x}^t G(G^tG)^{-1}G^t\mathbf{x} &\simeq \sum_{i=1}^N x_i^2 - \frac{2}{N}\left(\sum_{i=1}^N x_i\cos(\omega t_i) + \sum_{i=1}^N x_i \sin(\omega t_i)\right)^2 \\ &= \sum_{i=1}^N x_i^2 - 2S_N .\end{aligned}$$

(c) Show that $\pi(\omega|\mathbf{x})$ can be approximated by

$$\pi(\omega|\mathbf{x}) \propto \left[1 - \frac{2S_N}{\sum_{i=1}^N x_i^2}\right]^{(2-N)/2} .$$

5.4 For the situation of Example 5.5:

(a) Reproduce Figure 5.4.

(b) For a range of values of r, and a range of values of c, where $T_t = c/\log(t)$, examine the trajectories of $(x^{(t)}, h(x^{(t)}))$. Comment on the behaviors and recommend an optimal choice of r and c.

5.5 Given a simple Ising model with energy function

$$h(s) = \beta \sum_{(u,v)\in\mathcal{N}} s_u s_v ,$$

where $\mathcal{N}$ is the neighborhood relation that u and v are neighbors either horizontally or vertically, apply the algorithm $[A.19]$ to the cases $\beta = 0.4$ and $\beta = 4.5$.

5.6 Here we will outline a proof of Theorem 5.10. It is based on *the Laplace approximation* (see Section 3.4), and is rather painful. The treatment of the error terms here is a little cavalier; see Tierney and Kadane (1986) or Schervish (1995) for complete details.

(a) Expand $h(\theta)$ around θ^* up to third order terms to establish

$$\int \theta e^{\lambda h(\theta)} d(\theta) = e^{\lambda h(\theta^*)} \int \theta e^{\lambda h''(\theta^*)(\theta-\theta^*)^2/2} e^{\lambda h'''(\theta^*)(\theta-\theta^*)^3/6} e^{\lambda \mathcal{O}(|\theta-\theta^*|^4)},$$

where $\mathcal{O}(|\theta-\theta^*|^4) \leq C(|\theta-\theta^*|^4)$ for some constant C and $|\theta-\theta^*|$ small enough.

(b) Use Taylor series expansions to show that

$$e^{\lambda\mathcal{O}(|\theta-\theta^*|^4)} = 1 + \lambda\mathcal{O}(|\theta-\theta^*|^4)$$

and

$$e^{\lambda h'''(\theta^*)(\theta-\theta^*)^3/6} = 1 + \lambda h'''(\theta^*)(\theta-\theta^*)^3/6 + \lambda\mathcal{O}(|\theta-\theta^*|^6).$$

(c) Substitute the part (b) expressions into the integral of part (a), and write $\theta = t + \theta^*$, where $t = \theta - \theta^*$, to obtain

$$\begin{aligned}\int \theta e^{\lambda h(\theta)} d(\theta) = e^{\lambda h(\theta^*)}\theta^* \int e^{\lambda h''(\theta^*)t^2/2}[1 + \lambda h'''(\theta^*)t^3/6 + \lambda\mathcal{O}(|t|)^4]dt \\ + e^{\lambda h(\theta^*)} \int t e^{\lambda h''(\theta^*)t^2/2}[1 + \lambda h'''(\theta^*)t^3/6 + \lambda\mathcal{O}(|t|)^4]dt\,.\end{aligned}$$

(d) All of the integrals in part (*c*) can be evaluated. Show that the ones involving odd powers of t in the integrand are zero by symmetry, and

$$\left|\int \lambda|t|^4 e^{\lambda h''(\theta^*)t^2/2}dt\right| \le \frac{M}{\lambda^{3/2}},$$

which bounds the error terms.

(e) Show that we now can write

$$\int \theta e^{\lambda h(\theta)} d(\theta) = \theta^* e^{\lambda h(\theta^*)}\left[\sqrt{-2\pi}h''(\theta^*) + \frac{M}{\lambda^{3/2}}\right],$$

$$\int e^{\lambda h(\theta)} d(\theta) = e^{\lambda h(\theta^*)}\left[\sqrt{-2\pi}h''(\theta^*) + \frac{M}{\lambda^{3/2}}\right],$$

and hence

$$\frac{\int \theta e^{\lambda h(\theta)} d(\theta)}{\int e^{\lambda h(\theta)} d(\theta)} = \theta^*\left[1 + \frac{M}{\lambda^{3/2}}\right],$$

completing the proof.

(f) Show how Corollary 5.11 follows by starting with $\int b(\theta)e^{\lambda h(\theta)}d(\theta)$ instead of $\int \theta e^{\lambda h(\theta)}d(\theta)$. Use a Taylor series on $b(\theta)$.

5.7 For the Student's t distribution $\mathcal{T}_\nu$:

(a) Show that the distribution $\mathcal{T}_\nu$ can be expressed as the (continuous) mixture of a normal and of a chi squared distribution. (*Hint:* See Section 2.2.)

(b) When $X \sim \mathcal{T}_\nu$, given a function $h(x)$ derive a representation of $h(x) = \mathbb{E}[H(x,Z)|x]$, where $Z \sim \mathcal{G}a((\alpha-1)/2, \alpha/2)$.

5.8 For the normal mixture of Example 5.19, assume that $\sigma_1 = \sigma_2 = 1$ for simplicity's sake.

(a) Show that the model does belong to the missing data category by considering the vector $\mathbf{z} = (z_1, \ldots, z_n)$ of allocations of the observations x_i to the first and second components of the mixture.

(b) Show that the mth iteration of the EM algorithm consists in replacing the allocations z_i by their expectation

$$\begin{aligned}\mathbb{E}[Z_i|x_i, \mu_1^{(m-1)}, \mu_2^{(m-1)}] =&(1-p)\varphi(x_i; \mu_2^{(m-1)}, 1)\Big/ \\ &\{p\varphi(x_i; \mu_1^{(m-1)}, 1) + (1-p)\varphi(x_i; \mu_2^{(m-1)}, 1)\}\end{aligned}$$

and then setting the new values of the means as

$$\mu_1^{(m)} = \sum \mathbb{E}[1 - Z_i | x_i, \mu_1^{(m-1)}, \mu_2^{(m-1)}] x_i \Big/ \sum \mathbb{E}[1 - Z_i | x_i, \mu_1^{(m-1)}, \mu_2^{(m-1)}]$$

and

$$\mu_2^{(m)} = \sum \mathbb{E}[Z_i | x_i, \mu_1^{(m-1)}, \mu_2^{(m-1)}] x_i \Big/ \sum \mathbb{E}[Z_i | x_i, \mu_1^{(m-1)}, \mu_2^{(m-1)}] \,.$$

5.9 Suppose that the random variable X has a mixture distribution (1.3); that is, the X_i are independently distributed as

$$X_i \sim \theta g(x) + (1-\theta) h(x), \quad i = 1, \dots, n,$$

where $g(\cdot)$ and $h(\cdot)$ are known. An EM algorithm can be used to find the ML estimator of θ. Introduce $Z_1, \dots, Z_n$, where Z_i indicates from which distribution X_i has been drawn, so

$$X_i | Z_i = 1 \sim g(x),$$
$$X_i | Z_i = 0 \sim h(x).$$

(a) Show that the complete data likelihood can be written

$$L^c(\theta | \mathbf{x}, \mathbf{z}) = \prod_{i=1}^n \left[z_i g(x_i) + (1 - z_i) h(x_i) \right] \theta^{z_i} (1-\theta)^{1-z_i}.$$

(b) Show that $\mathbb{E}[Z_i|\theta, x_i] = \theta g(x_i)/[\theta g(x_i) + (1-\theta)h(x_i)]$ and, hence, that the EM sequence is given by

$$\hat\theta_{(j+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\hat\theta_{(j)} g(x_i)}{\hat\theta_{(j)} g(x_i) + (1 - \hat\theta_{(j)}) h(x_i)}.$$

(c) Show that $\hat\theta_{(j)}$ converges to $\hat\theta$, a maximum likelihood estimator of θ.

5.10 Consider the sample $\mathbf{x} = (0.12, 0.17, 0.32, 0.56, 0.98, 1.03, 1.10, 1.18, 1.23, 1.67, 1.68, 2.33)$, generated from an exponential mixture

$$p\,\mathcal{E}xp(\lambda) + (1-p)\,\mathcal{E}xp(\mu).$$

(a) Show that the likelihood $h(p, \lambda, \mu)$ can be expressed as $\mathbb{E}[H(x, Z)]$, where $z = (z_1, \dots, z_{12})$ corresponds to the vector of allocations of the observations x_i to the first and second components of the mixture; that is, for $i = 1, \dots, 12$,

$$P(z_i = 1) = 1 - P(z_i = 2) = \frac{p\lambda \exp(-\lambda x_i)}{p\lambda \exp(-\lambda x_i) + (1-p)\mu \exp(-\mu x_i)}.$$

(b) Compare the performances of $[A.22]$ with those of the EM algorithm in this setup.

5.11 This problem refers to Section 5.5.4.

(a) Show that $\int \tilde h(x|\theta) dx = \frac{1}{c(\theta)}$.

(b) Show that for any two values θ and η,

$$\log \frac{h(x|\theta)}{h(x|\eta)} = \log \frac{\tilde{h}(x|\theta)}{\tilde{h}(x|\eta)} + \frac{c(\theta)}{c(\eta)} = \log \frac{\tilde{h}(x|\theta)}{\tilde{h}(x|\eta)} - \log \mathbb{E}\left(\frac{\tilde{h}(X|\theta)}{\tilde{h}(X|\eta)}\right),$$

where $X \sim h(x|\eta)$. (*Hint:* The last equality follows from part (a) and an importance sampling argument.)

(c) Thus, establish the validity of the approximation

$$\max_x h(x|\theta) \approx \max_x \left\{ \log \frac{\tilde{h}(x|\theta)}{\tilde{h}(x|\eta)} - \log \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{h}(x_i|\theta)}{\tilde{h}(x_i|\eta)} \right\},$$

where the X_i's are generated from $h(x|\eta)$

5.12 Consider $h(\alpha)$, the likelihood of the beta $\mathcal{B}(\alpha, \alpha)$ distribution associated with the observation $x = 0.345$.

(a) Express the normalizing constant, $c(\alpha)$, in $h(\alpha)$ and show that it cannot be easily computed when α is not an integer.

(b) Examine the approximation of the ratio $c(\alpha)/c(\alpha_0)$, for $\alpha_0 = 1/2$ by the method of Geyer and Thompson (1992) (see Example 5.25).

(c) Compare this approach with the alternatives of Chen and Shao (1997), detailed in Problems 4.1 and 4.2.

5.13 Consider the function

$$h(\theta) = \frac{||\theta||^2(p + ||\theta||^2)(2p - 2 + ||\theta||^2)}{(1 + ||\theta||^2)(p + 1 + ||\theta||^2)(p + 3 + ||\theta||^2)},$$

when $\theta \in \mathbb{R}^p$ and $p = 10$.

(a) Show that the function $h(\theta)$ has a unique maximum.

(b) Show that $h(\theta)$ can be expressed as $\mathbb{E}[H(\theta, Z)]$, where $z = (z_1, z_2, z_3)$ and $Z_i \sim \mathcal{E}xp(1/2)$ $(i = 1, 2, 3)$. Deduce that $f(z|x)$ does not depend on x in (5.26).

(c) When $g(z) = \exp(-\alpha\{z_1 + z_2 + z_3\})$, show that the variance of (5.26) is infinite for some values of $t = ||\theta||^2$ when $\alpha > 1/2$. Identify A_2, the set of values of t for which the variance of (5.26) is infinite when $\alpha = 2$.

(d) Study the behavior of the estimate (5.26) when t goes from A_2 to its complement A_2^c to see if the infinite variance can be detected in the evaluation of $h(t)$.

5.14 In the exponential family, EM computations are somewhat simplified. Show that if the complete data density f is of the form

$$f(y, z|\theta) = h(y, z) \exp\{\eta(\theta)T(y, z) - B(\theta)\},$$

then we can write

$$Q(\theta|\theta^*, \mathbf{y}) = \mathbb{E}_{\theta^*}[\log h(\mathbf{y}, \mathbf{Z})] + \sum \eta_i(\theta)\mathbb{E}_{\theta^*}[T_i|\mathbf{y}] - B(\theta).$$

Deduce that calculating the complete-data MLE only involves the simpler expectation $\mathbb{E}_{\theta^*}[T_i|\mathbf{y}]$.

5.15 For density functions f and g, we define the *entropy distance* between f and g, with respect to f (also known as *Kullback–Leibler information of g at f* or *Kullback–Leibler distance between g and f*) as

$$\mathbb{E}_f[\log(f(X)/g(X))] = \int \log\left[\frac{f(x)}{g(x)}\right] f(x)\, dx.$$

(a) Use Jensen's inequality to show that $\mathbb{E}_f[\log(f(X)/g(X))] \geq 0$ and, hence, that the entropy distance is always non-negative, and equals zero if $f = g$.
(b) The inequality in part (a) implies that $\mathbb{E}_f \log[g(X)] \leq \mathbb{E}_f \log[f(X)]$. Show that this yields (5.15).

(*Note:* Entropy distance was explored by Kullback 1968; for an exposition of its properties, see, for example, Brown 1986. Entropy distance has, more recently, found many uses in Bayesian analysis see, for example, Berger 1985, Bernardo and Smith 1994, or Robert 1996b.)

5.16 Refer to Example 5.17
(a) Verify that the missing data density is given by (5.16).
(b) Show that $\mathbb{E}_{\theta'}[Z_1] = \theta' + \frac{\phi(a-\theta')}{1-\Phi(a-\theta')}$, and that the EM sequence is given by (5.17).
(c) Reproduce Figure 5.6. In particular, examine choices of M that will bring the MCEM sequence closed to the EM sequence.

5.17 The *probit model* is a model with a covariate $X \in \mathbb{R}^p$ such that

$$Y|X = x \sim \mathcal{B}(\Phi(x^T\beta)\,,$$

where $\beta \in \mathbb{R}^p$ and Φ denotes the $\mathcal{N}(0,1)$ cdf.
(a) Give the likelihood associated with a sample $((x_1, y_1), \ldots, (x_n, y_n))$.
(b) Show that, if we associate with each observation (x_i, y_i) a missing variable Z_i such that

$$Z_i|X_i = x \sim \mathcal{N}(x^T\beta, 1) \qquad Y_i = \mathbb{I}_{Z_i>0}\,,$$

iteration m of the associated EM algorithm is the expected least squares estimator

$$\beta_{(m)} = (X^T X)^{-1} X^T\, \mathbb{E}_{\beta_{(m-1)}}[\mathbf{Z}|\mathbf{x}, \mathbf{y}]\,,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, and $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$, and X is the matrix with columns made of the x_i's.
(c) Give the value of $\mathbb{E}_\beta[z_i|x_i, y_i]$.

5.18 The following are genotype data on blood type.

Genotype	Probability	Observed	Probability	Frequency
AA	p_A^2	A	$p_A^2 + 2p_Ap_O$	$n_A = 186$
AO	$2p_Ap_O$			
BB	p_B^2	B	$p_B^2 + 2p_Bp_O$	$n_B = 38$
BO	$2p_Bp_O$			
AB	$2p_Ap_B$	AB	p_Ap_B	$n_{AB} = 13$
OO	p_O^2	O	p_O^2	$n_O = 284$

Because of dominance, we can only observe the genotype in the third column, with probabilities given by the fourth column. The interest is in estimating the allele frequencies p_A, p_B, and p_O (which sum to 1).
(a) Under a multinomial model, verify that the observed data likelihood is proportional to

$$(p_A^2 + 2p_Ap_O)^{n_A}(p_B^2 + 2p_Bp_O)^{n_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}\,.$$

(b) With missing data Z_A and Z_B, verify the complete data likelihood

$$(p_A^2)^{Z_A}(2p_Ap_O)^{n_A-Z_A}(p_B^2)^{Z_B}(2p_Bp_O)^{n_B-Z_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}\,.$$

(c) Verify that the missing data distribution is

$$Z_A \sim \text{binomial}\left(n_A, \frac{p_A^2}{p_A^2 + 2p_A p_O}\right) \text{ and } Z_B \sim \text{binomial}\left(n_B, \frac{p_B^2}{p_B^2 + 2p_B p_O}\right),$$

and write an EM algorithm to estimate p_A, p_B, and p_O.

5.19 Cox and Snell (1981) report data on survival time Y in weeks and $\log_{10}$ (initial white blood count), x, for 17 patients suffering from leukemia as follows.

x	3.36	2.88	3.63	3.41	3.78	4.02	4.0	4.23	3.73	3.85	3.97	4.51	4.54	5.0	5.0	4.72	5.0
Y	65	156	100	134	16	108	121	4	39	143	56	26	22	1	1	5	65

They suggest that an appropriate model for these data is

$$Y_i = b_0 \exp b_1(x_x - \bar{x})\varepsilon_i, \quad \varepsilon_i \sim \mathcal{E}xp(1),$$

which leads to the likelihood function

$$L(b_0, b_1) = \prod_i b_0 \exp b_1(x_x - \bar{x}) \exp\{-y_i/(b_0 \exp b_1(x_x - \bar{x}))\}.$$

(a) Differentiation of the log likelihood function leads to the following equations for the MLEs:

$$\begin{aligned} \hat{b}_0 &= \frac{1}{n} \sum_i y_i \exp(-\hat{b}_1(x_i - \bar{x})) \\ 0 &= \sum_i y_i(x_i - \bar{x}) \exp(-\hat{b}_1(x_i - \bar{x})). \end{aligned}$$

Solve these equations for the MLEs ($\hat{b}_0 = 51.1$, $\hat{b}_1 = -1.1$).

(b) It is unusual for survival times to be uncensored. Suppose there is censoring in that the experiment is stopped after 90 weeks (hence any Y_i greater than 90 is replaced by a 90, resulting in 6 censored observations.) Order the data so that the first m observations are uncensored, and the last $n - m$ are censored. Show that the observed-data likelihood is

$$\begin{aligned} L(b_0, b_1) &= \prod_{i=1}^{m} b_0 \exp b_1(x_x - \bar{x}) \exp\{-y_i/(b_0 \exp(b_1(x_x - \bar{x})))\} \\ &\quad \times \prod_{i=m+1}^{n} (1 - F(a|x_i)), \end{aligned}$$

where

$$\begin{aligned} F(a_| x_i) &= \int_0^a b_0 \exp b_1(x_x - \bar{x}) \exp\{-t/(b_0 \exp(b_1(x_x - \bar{x})))\} dt \\ &= 1 - \exp\{-a/(b_0 \exp(b_1(x_x - \bar{x})))\}. \end{aligned}$$

(c) If we let Z_i denote the missing data, show that the Z_i are independent with distribution

$$Z_i \sim \frac{b_0 \exp b_1(x_x - \bar{x}) \exp\{-z/(b_0 \exp(b_1(x_x - \bar{x})))\}}{1 - F(a|x_i)}, \quad 0 < z < \infty.$$

(d) Verify that the complete data likelihood is

$$L(b_0, b_1) = \prod_{i=1}^{m} b_0 \exp b_1(x_x - \bar{x}) \exp\{-y_i/(b_0 \exp(b_1(x_x - \bar{x})))\}$$
$$\times \prod_{i=m+1}^{n} b_0 \exp b_1(x_x - \bar{x}) \exp\{-z_i/(b_0 \exp(b_1(x_x - \bar{x})))\},$$

and the expected complete-data log-likelihood is obtained by replacing the Z_i by their expected value

$$\mathbb{E}[Z_i] = (a + b_0 \exp(b_1(x_x - \bar{x}))\frac{1 - F(a|x_i)}{F(a|x_i)}.$$

(c) Implement an EM algorithm to obtain the MLEs of b_0 and b_1.

5.20 Consider the following 12 observations from $\mathcal{N}_2(0, \Sigma)$, with σ_1^2, σ_2^2, and ρ unknown:

x_1	1	1	−1	−1	2	2	−2	−2	–	–	–	–
x_2	1	−1	1	−1	–	–	–	–	2	2	−2	−2

where "−" represents a missing value.

(a) Show that the likelihood function has global maxima at $\rho = \pm 1/2$, $\sigma_1^2 = \sigma_2^2 = 8/3$, and a saddlepoint at $\rho = 0$, $\sigma_1^2 = \sigma_2^2 = 5/2$.

(b) Show that if an EM sequence starts with $\rho = 0$, then it remains at $\rho = 0$ for all subsequent iterations.

(c) Show that if an EM sequence starts with ρ bounded away from zero, it will converge to a maximum.

(d) Take into account roundoff errors; that is, the fact that $\lfloor x_i \rfloor$ is observed instead of x_i.

(*Note:* This problem is due to Murray 1977 and is discussed by Wu 1983.)

5.21 (Ó Ruanaidh and Fitzgerald 1996) Consider an AR(p) model

$$X_t = \sum_{j=1}^{p} \theta_j X_{t-j} + \epsilon_t,$$

with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, observed for $t = p+1, \ldots, m$. The future values $X_{m+1}, \ldots, X_n$ are considered to be missing data. The initial values $x_1, \ldots, x_p$ are taken to be zero.

(a) Give the expression of the observed and complete-data likelihoods.

(b) Give the conditional maximum likelihood estimators of θ, σ and $\mathbf{z} = (X_{m+1}, \ldots, X_n)$; that is, the maximum likelihood estimators when the two other parameters are fixed.

(c) Detail the E- and M-steps of the EM algorithm in this setup, when applied to the future values $\mathbf{z}$ and when σ is fixed.

5.22 We observe independent Bernoulli variables $X_1, \ldots, X_n$, which depend on unobservable variables Z_i distributed independently as $\mathcal{N}(\zeta, \sigma^2)$, where

$$X_i = \begin{cases} 0 & \text{if } Z_i \leq u \\ 1 & \text{if } Z_i > u. \end{cases}$$

Assuming that u is known, we are interested in obtaining MLEs of ζ and σ^2.

(a) Show that the likelihood function is

$$p^S(1-p)^{n-S},$$

where $S = \sum x_i$ and

$$p = P(Z_i > u) = \Phi\left(\frac{\zeta - u}{\sigma}\right) .$$

(b) If we consider $z_1, \ldots, z_n$ to be the complete data, show that the complete data likelihood is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(z_i - \zeta)^2\right\}$$

and the expected complete-data log-likelihood is

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\mathbb{E}[Z_i^2|x_i] - 2\zeta\mathbb{E}[Z_i|x_i] + \zeta^2\right).$$

(c) Show that the EM sequence is given by

$$\begin{aligned}
\hat{\zeta}_{(j+1)} &= \frac{1}{n}\sum_{i=1}^{n} t_i(\hat{\zeta}_{(j)}, \hat{\sigma}^2_{(j)}) \\
\hat{\sigma}^2_{(j+1)} &= \frac{1}{n}\left[\sum_{i=1}^{n} v_i(\hat{\zeta}_{(j)}, \hat{\sigma}^2_{(j)}) - \frac{1}{n}\left(\sum_{i=1}^{n} t_i(\hat{\zeta}_{(j)}, \hat{\sigma}^2_{(j)})\right)^2\right] ,
\end{aligned}$$

where

$$t_i(\zeta, \sigma^2) = \mathbb{E}[Z_i|x_i, \zeta, \sigma^2] \quad \text{and} \quad v_i(\zeta, \sigma^2) = \mathbb{E}[Z_i^2|x_i, \zeta, \sigma^2] .$$

(d) Show that

$$\begin{aligned}
\mathbb{E}[Z_i|x_i, \zeta, \sigma^2] &= \zeta + \sigma H_i\left(\frac{u-\zeta}{\sigma}\right) , \\
\mathbb{E}[Z_i^2|x_i, \zeta, \sigma^2] &= \zeta^2 + \sigma^2 + \sigma(u+\zeta)H_i\left(\frac{u-\zeta}{\sigma}\right) ,
\end{aligned}$$

where

$$H_i(t) = \begin{cases} \frac{\varphi(t)}{1-\Phi(t)} & \text{if } X_i = 1 \\ -\frac{\varphi(t)}{\Phi(t)} & \text{if } X_i = 0. \end{cases}$$

(e) Show that $\hat{\zeta}_{(j)}$ converges to $\hat{\zeta}$ and that $\hat{\sigma}^2_{(j)}$ converges to $\hat{\sigma}^2$, the MLEs of ζ and σ^2, respectively.

5.23 Referring to Example 5.18

(a) Show that if $Z_i \sim \mathcal{M}(m_i; p_1, \ldots, p_k)$, $i = 1, \ldots, n$, then $\sum_i Z_i \sim \mathcal{M}(\sum_i m_i; p_1, \ldots, p_k)$.

(b) Show that the incomplete data likelihood is given by

$$\sum_{\mathbf{x}} L(\mathbf{p}|\mathbf{z},\mathbf{x}) = \left[\prod_{i=1}^{m}\sum_{x_i=0}^{\infty}\binom{n_i+x_i}{T_{i1},\ldots,T_{i4},x_i}p_1^{T_{i1}}\cdots p_4^{T_{i4}}p_5^{x_i}\right] \times\left[\prod_{i=m+1}^{n}\binom{n_i}{T_{i1},\ldots,T_{i4}}\prod_{j=1}^{5}p_j^{T_{ij}}\right]$$
$$= \left[\prod_{i=1}^{m}\frac{n_i!}{T_{i1}!\cdots T_{i4}!}\frac{p_1^{T_{i1}}\cdots p_4^{T_{i4}}}{(1-p_5)^{n_i+1}}\right]\left[\prod_{i=m+1}^{n}\binom{n_i}{T_{i1},\ldots,T_{i4}}\prod_{j=1}^{4}p_j^{T_{ij}}\right].$$

(c) Verify equation (5.19).
(d) Verify the equations of the expected log-likelihood and the EM sequence.

5.24 Recall the censored Weibull model of Problems 1.11–1.13. By extending the likelihood technique of Example 5.14, use the EM algorithm to fit the Weibull model, accounting for the censoring. Use the data of Problem 1.13 and fit the three cases outlined there.

5.25 An alternate implementation of the Monte Carlo EM might be, for $Z_1,\ldots,Z_m \sim k(\mathbf{z}|\mathbf{x},\theta)$, to iteratively maximize

$$\log \hat{L}(\theta|\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\{\log L^c(\theta|\mathbf{x},\mathbf{z}_i) - \log k(\mathbf{z}_i|\theta,x)\}$$

(which might more accurately be called Monte Carlo maximum likelihood).
(a) Show that $\hat{L}(\theta|\mathbf{x}) \to L(\theta|\mathbf{x})$ as $m\to\infty$.
(b) Show how to use $\hat{L}(\theta|\mathbf{x})$ to obtain the MLE in Example 5.22. (*Warning:* This is difficult.)

5.26 For the situation of Example 5.21, data $(x_1,x_2,x_3,x_4) = (125,18,20,34)$ are collected.
(a) Use the EM algorithm to find the MLE of θ.
(b) Use the Monte Carlo EM algorithm to find the MLE of θ. Compare your results to those of part (a).

5.27 For the situation of Example 5.22:
(a) Verify the formula for the likelihood function.
(b) Show that the complete-data MLEs are given by $\hat{\theta}_k = \frac{1}{nt}\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ijk} + \hat{z}_{ijk}$.

5.28 For the model of Example 5.22, Table 5.7 contains data on the movement between 5 zones of 18 tree swallows with $m = t = 5$, where a 0 denotes that the bird was not captured.
(a) Using the MCEM algorithm of Example 5.22, calculate the MLEs for $\theta_1,\ldots,\theta_5$ and $p_1,\ldots,p_5$.
(b) Assume now that state 5 represents the death of the animal. Rewrite the MCEM algorithm to reflect this, and recalculate the MLEs. Compare them to the answer in part (a).

5.29 Referring to Example 5.23
(a) Verify the expression for the second derivative of the log-likelihood.
(b) Reproduce the EM estimator and its standard error.
(c) Estimate θ, and its standard error, using the Monte Carlo EM algorithm. Compare the results to those of the EM algorithm.

	Time						Time						Time				
	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
a	2	2	0	0	0	g	1	1	1	5	0	m	1	1	1	1	1
b	2	2	0	0	0	h	4	2	0		0	n	2	2	1	0	0
c	4	1	1	2	0	i	5	5	5	5	0	o	4	2	2	0	0
d	4	2	0	0	0	j	2	2	0	0	0	p	1	1	1	1	0
e	1	1	0	0	0	k	2	5	0	0	0	q	1	0	0	4	0
f	1	1	0	0	0	l	1	1	0	0	0	s	2	2	0	0	0

Table 5.7. Movement histories of 18 tree swallows over 5 time periods (*Source:* Scherrer 1997.)

5.30 Referring to Section 5.3.4

(a) Show how (5.22) can be derived from (5.21).

(b) The derivation in Section 5.3.4 is valid for vector parameters $\theta = (\theta_1, \ldots, \theta_p)$. For a function $h(\theta)$, define

$$\frac{\partial}{\partial \theta} h(\theta) = \left\{ \frac{\partial}{\partial \theta_i} h(\theta)) \right\}_{ij} = h^{(1)}$$

$$\frac{\partial^2}{\partial \theta^2} h(\theta) = \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} h(\theta)) \right\}_{ij} = h^{(2)} ,$$

where $h^{(1)}$ is a $p \times 1$ vector and $h^{(2)}$ is a $p \times p$ matrix. Using this notation, show that (5.22) becomes

$$\begin{aligned}\log L(\theta|\mathbf{x})^{(2)} &= \mathbb{E}\left(\log L(\theta|\mathbf{x},\mathbf{z})^{(2)}\right) + \mathbb{E}\left[\left(\log L(\theta|\mathbf{x},\mathbf{z})^{(1)}\right)\left(\log L(\theta|\mathbf{x},\mathbf{z})^{(1)}\right)'\right] \\ &\quad - \left[\mathbb{E}\left(\log L(\theta|\mathbf{x},\mathbf{z})^{(1)}\right)\right]\left[\mathbb{E}\left(\log L(\theta|\mathbf{x},\mathbf{z})^{(1)}\right)\right]' .\end{aligned}$$

(c) Use the equation in part(b) to attach standard errors to the EM estimates in Example 5.18

5.31 (For baseball fans only) It is typical for baseball announcers to report biased information, intended to overstate a player's ability. If we consider a sequence of at-bats as Bernoulli trials, we are likely to hear the report of a maximum (the player is 8-out-of-his-last-17) rather than an ordinary average. Assuming that $X_1, X_2, \ldots, X_n$ are the Bernoulli random variables representing a player's sequence of at-bats (1=hit, 0=no hit), a biased report is the observance of k^*, m^*, and r^*, where

$$r^* = \frac{k^*}{m^*} \geq \max_{m^* \leq i < n} \frac{X_n + X_{n-1} + \cdots + X_{n-i}}{i+1} .$$

If we assume that $\mathbb{E}[X_i] = \theta$, then θ is the player's true batting ability and the parameter of interest. Estimation of θ is difficult using only k^*, m^*, and r^*, but it can be accomplished with an EM algorithm. With observed data (k^*, m^*, r^*), let $\mathbf{z} = (z_1, \ldots, z_{n-m^*-1})$ be the augmented data. (This is a sequence of 0's and 1's that are commensurate with the observed data. Note that X_{n-m^*} is certain to equal 0.)

(a) Show that the EM sequence is given by

$$\hat{\theta}_{j+1} = \frac{k^* + \mathbb{E}[S_Z|\hat{\theta}_j]}{n},$$

where $\mathbb{E}[S_Z|\hat{\theta}_j]$ is the expected number of successes in the missing data, assuming that $\hat{\theta}_j$ is the true value of θ.

(b) Give an algorithm for computing the sequence $(\hat{\theta}_j)$. Use a Monte Carlo approximation to evaluate the expectation.

At-Bat	k^*	m^*	$\hat{\theta}$	EM MLE
339	12	39	0.298	0.240
340	47	155	0.297	0.273
341	13	41	0.299	0.251
342	13	42	0.298	0.245
343	14	43	0.300	0.260
344	14	44	0.299	0.254
345	4	11	0.301	0.241
346	5	11	0.303	0.321

Table 5.8. A portion of the 1992 batting record of major-league baseball player Dave Winfield.

(c) For the data given in Table 5.8, implement the Monte Carlo EM algorithm and calculate the EM estimates.

(*Note:* The "true batting average" $\hat{\theta}$ cannot be computed from the given data and is only included for comparison. The selected data EM MLEs are usually biased downward, but also show a large amount of variability. See Casella and Berger 1994 for details.)

5.32 The following dataset gives independent observations of $Z = (X, Y) \sim \mathcal{N}_2(0, \Sigma)$ with missing data $*$.

x	1.17	-0.98	0.18	0.57	0.21	*	*	*
y	0.34	-1.24	-0.13	*	*	-0.12	-0.83	1.64

(a) Show that the observed likelihood is

$$\prod_{i=1}^{3} \left\{ |\Sigma|^{-1/2} \, e^{-z_i^t \Sigma^{-1} z_i/2} \right\} \sigma_1^{-2} \, e^{-(x_4^2+x_5^2)/2\sigma_1^2} \sigma_2^{-3} \, e^{-(y_6^2+y_7^2+y_8^2)/2\sigma_2^2} .$$

(b) Examine the consequence of the choice of $\pi(\Sigma) \propto |\Sigma|^{-1}$ on the posterior distribution of Σ.

(c) Show that the missing data can be simulated from

$$X_i^\star \sim \mathcal{N}\left(\rho \frac{\sigma_1}{\sigma_2} y_i, \sigma_1^2(1-\rho^2)\right) \qquad (i = 6, 7, 8),$$
$$Y_i^\star \sim \mathcal{N}\left(\rho \frac{\sigma_2}{\sigma_1} x_i, \sigma_2^2(1-\rho^2)\right) \qquad (i = 4, 5),$$

to derive a stochastic EM algorithm.

(d) Derive an efficient simulation method to obtain the MLE of the covariance matrix Σ.

5.33 The EM algorithm can also be implemented in a Bayesian hierarchical model to find a posterior mode. Suppose that we have the hierarchical model

$$\begin{aligned} X|\theta &\sim f(x|\theta) \,, \\ \theta|\lambda &\sim \pi(\theta|\lambda) \,, \\ \lambda &\sim \gamma(\lambda) \,, \end{aligned}$$

where interest would be in estimating quantities from $\pi(\theta|x)$. Since

$$\pi(\theta|x) = \int \pi(\theta, \lambda|x) d\,\lambda,$$

where $\pi(\theta, \lambda|x) = \pi(\theta|\lambda, x)\pi(\lambda|x)$, the EM algorithm is a candidate method for finding the mode of $\pi(\theta|x)$, where λ would be used as the augmented data.

(a) Define $k(\lambda|\theta, x) = \pi(\theta, \lambda|x)/\pi(\theta|x)$ and show that

$$\log \pi(\theta|x) = \int \log \pi(\theta, \lambda|x) k(\lambda|\theta^*, x) d\,\lambda - \int \log k(\lambda|\theta, x) k(\lambda|\theta^*, x) d\,\lambda.$$

(b) If the sequence $(\hat{\theta}_{(j)})$ satisfies

$$\max_{\theta} \int \log \pi(\theta, \lambda|x) k(\lambda|\theta_{(j)}, x) d\,\lambda = \int \log \pi(\theta_{(j+1)}, \lambda|x) k(\lambda|\theta_{(j)}, x) d\,\lambda,$$

show that $\log \pi(\theta_{(j+1)}|x) \geq \log \pi(\theta_{(j)}|x)$. Under what conditions will the sequence $(\hat{\theta}_{(j)})$ converge to the mode of $\pi(\theta|x)$?

(c) For the hierarchy

$$\begin{aligned} X|\theta &\sim \mathcal{N}(\theta, 1) \,, \\ \theta|\lambda &\sim \mathcal{N}(\lambda, 1) \,, \end{aligned}$$

with $\pi(\lambda) = 1$, show how to use the EM algorithm to calculate the posterior mode of $\pi(\theta|x)$.

5.34 Let $X \sim \mathcal{T}_\nu$ and consider the function

$$h(x) = \frac{\exp(-x^2/2)}{[1 + (x - \mu)^2]^\nu} \,.$$

(a) Show that $h(x)$ can be expressed as the conditional expectation $\mathbb{E}[H(x, Z)|x]$, when $Z \sim \mathcal{G}a(\nu, 1)$.

(b) Apply the direct Monte Carlo method of Section 5.5.4 to maximize (5.5.4) and determine whether or not the resulting sequence converges to the true maximum of h.

(c) Compare the implementation of (b) with an approach based on (5.26) for (i) $g = \mathcal{E}xp(\lambda)$ and (ii) $g = f(z|\mu)$, the conditional distribution of Z given $X = \mu$. For each choice, examine if the approximation (5.26) has a finite variance.

(d) Run [$A.22$] to see if the recursive scheme of Geyer (1996) improves the convergence speed.