the entire function. This likelihood implementation of the Gibbs sampler was used by Casella and Berger (1994) and is also described by Smith and Roberts (1993). A version of the EM algorithm, where the Markov chain connection is quite apparent, was given by Baum and Petrie (1966) and Baum et al. (1970).

9.5 Transition

As a natural extension of the (2D) slice sampler, the two-stage Gibbs sampler enjoys many optimality properties. In Chapter 8 we also developed the natural extension of the slice sampler to the case when the density f is easier to study when decomposed as a product of functions f_i (Section 8.2). Chapter 10 will provide the corresponding Gibbs generalization to this general slice sampler, which will cover cases when the decomposition of the simulation in two conditional simulations is not feasible any longer. This generalization obviously leads to a wider range of models, but also to fewer optimality properties.

As an introduction to the next chapter, note that if $(X, Y) = (X, (Y_1, Y_2))$, and if simulating from $f_{Y|X}$ is not directly possible, the two-stage Gibbs sampler could also be applied to this (conditional) density. This means that a sequence of successive simulations from $f(y_1|x, y_2)$ and from $f(y_2|x, y_1)$ (which is the translation of the two-stage Gibbs sampler for the conditional $f_{Y|X}$) converges to a simulation from $f_{Y|X}(y_1, y_2|x)$.

The fundamental property used in Chapter 10 is that stopping the recursion between $f(y_1|x, y_2)$ and $f(y_2|x, y_1)$ "before" convergence has no effect on the validation of the algorithm. In fact, a *single* simulation of each component is sufficient! Obviously, this feature will generalize (by a cascading argument) to an arbitrary number of components in Y.

9.6 Problems

9.1 For the Gibbs sampler [*A*.33]:
(a) Show that the sequence (X_i, Y_i) is a Markov chain, as is each sequence (X_i) and (Y_i).
(b) Show that $f_X(\cdot)$ and $f_Y(\cdot)$ are respectively the invariant densities of the X and Y sequences of [*A*.33].

9.2 Write a Gibbs sampler to generate standard bivariate normal random variables (with mean 0, variance 1 and correlation ρ). (Recall that if (X, Y) is standard bivariate normal, the conditional density of $X|Y = y$ is $N(\rho y, (1 - \rho^2))$. For $\rho = .3$, use the generated random variables to estimate the density of $X^2 + Y^2$ and calculate $P(X^2 + Y^2 > 2)$.

9.3 Referring to Problem 5.18, estimate p_A, p_B and p_O using a Gibbs sampler. Make a histogram of the samples.

9.4 In the case of the two-stage Gibbs sampler, the relationship between the Gibbs and Metropolis–Hastings algorithms becomes particularly clear. If we have the bivariate Gibbs sampler $X \sim f(x|y)$ and $Y \sim f(y|x)$, consider the X chain alone and show:

(a) $K(x, x') = g(x|x') = \int f(x'|y)f(y|x)dy$;
(b) $\varrho = \min\left\{\frac{f(x')/g(x'|x)}{f(x)/g(x|x')}, 1\right\}$, where $f(\cdot)$ is the marginal distribution;
(c) $f(x')/g(x'|x) = f(x)/g(x|x')$, so $\varrho = 1$ and the Metropolis–Hastings proposal is always accepted.

9.5 For the situation of Example 9.20:
(a) Verify the variance representations (9.14) and (9.15).
(b) For the bivariate normal sampler, show that $\text{cov}(X_1, X_k) = \rho^{2k}$, for all k, and $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = 1/\rho^2$.

9.6 The monotone decrease of the correlation seen in Section 9.3 does not hold uniformly for all Gibbs samplers as shown by the example of Liu et al. (1994): For the bivariate normal Gibbs sampler of Example 9.20, and $h(x, y) = x - y$, show that

$$\text{cov}[h(X_1, Y_1), h(X_2, Y_2)] = -\rho(1 - \rho)^2 < 0.$$

9.7 Refer to the Horse Kick data of Table 2.1. Fit the loglinear model of Example 9.7 using the bivariate Gibbs sampler, along with the ARS algorithm, and estimate both $\pi(a|\mathbf{x}, \mathbf{y})$ and $\pi(b|\mathbf{x}, \mathbf{y})$. Obtain both point estimates and error bounds for a and b. Take $\sigma^2 = \tau^2 = 5$.

9.8 The data of Example 9.7 can also be analyzed as a loglinear model, where we fit $\log\lambda = a + bt$, where $t =$ number of passages.
(a) Using the techniques of Example 2.26, find the posterior distributions of a and b. Compare you answer to that given in Example 9.7. (Ignore the "4 or more", and just use the category as 4.)
(b) Find the posterior distributions of a and b, but now take the "4 or more" censoring into account, as in Example 9.7.

9.9 The situation of Example 9.7 also lends itself to the EM algorithm, similar to the Gibbs treatment (Example 9.8) and EM treatment (Example 5.21) of the grouped multinomial data problem. For the data of Table 9.1:
(a) Use the EM algorithm to calculate the MLE of λ.
(b) Compare your answer in part (a) to that from the Gibbs sampler of Algorithm [*A*.35].
(c) Establish that the Rao–Blackwellized estimator is correct.

9.10 In the setup of Example 9.8, the (uncompleted) posterior distribution is available as

$$\begin{aligned}\pi(\eta, \mu|x) \propto\ & (a_1\mu + b_1)^{x_1}(a_2\mu + b_2)^{x_2}(a_3\eta + b_3)^{x_3}(a_4\eta + b_4)^{x_4} \\ & \times (1 - \mu - \eta)^{x_5+\alpha_3-1}\mu^{\alpha_1-1}\eta^{\alpha_2-1}.\end{aligned}$$

(a) Show that the marginal distributions $\pi(\mu|x)$ and $\pi(\eta|x)$ can be explicitly computed as polynomials when the α_i's are integers.
(b) Give the marginal posterior distribution of $\xi = \mu/(1 - \eta - \mu)$. (*Note:* See Robert 1995a for a solution.)
(c) Evaluate the Gibbs sampler proposed in Example 9.8 by comparing approximate moments of μ, η, and ξ with their exact counterpart, derived from the explicit marginal.

9.11 There is a connection between the EM algorithm and Gibbs sampling, in that both have their basis in Markov chain theory. One way of seeing this is to show that the incomplete-data likelihood is a solution to the integral equation of successive substitution sampling and that Gibbs sampling can then be used to

calculate the likelihood function. If $L(\theta|\mathbf{y})$ is the incomplete-data likelihood and $L(\theta|\mathbf{y},\mathbf{z})$ is the complete-data likelihood, define

$$L^*(\theta|\mathbf{y}) = \frac{L(\theta|\mathbf{y})}{\int L(\theta|\mathbf{y})d\theta}\,, \quad L^*(\theta|\mathbf{y},\mathbf{z}) = \frac{L(\theta|\mathbf{y},\mathbf{z})}{\int L(\theta|\mathbf{y},\mathbf{z})d\theta}\,,$$

assuming both integrals to be finite.

(a) Show that $L^*(\theta|\mathbf{y})$ is the solution to

$$L^*(\theta|\mathbf{y}) = \int \left[\int L^*(\theta|\mathbf{y},\mathbf{z})k(\mathbf{z}|\theta',\mathbf{y})d\mathbf{z}\right] L^*(\theta'|\mathbf{y})d\theta'\,,$$

where $k(\mathbf{z}|\theta,\mathbf{y}) = L(\theta|\mathbf{y},\mathbf{z})/L(\theta|\mathbf{y})$.

(b) Show that the sequence $\theta_{(j)}$ from the Gibbs iteration

$$\theta_{(j)} \sim L^*(\theta|\mathbf{y},\mathbf{z}_{(j-1)})\,,$$
$$\mathbf{z}_{(j)} \sim k(\mathbf{z}|\theta_{(j)},\mathbf{y})\,,$$

converges to a random variable with density $L^*(\theta|\mathbf{y})$ as j goes ∞. How can this be used to compute the likelihood function $L(\theta|\mathbf{y})$?

(*Note:* Based on the same functions $L(\theta|\mathbf{y},\mathbf{z})$ and $k(\mathbf{z}|\theta,\mathbf{y})$ the EM algorithm will get the ML estimator from $L(\theta|\mathbf{y})$, whereas the Gibbs sampler will get us the entire function. This likelihood implementation of the Gibbs sampler was used by Casella and Berger 1994 and is also described by Smith and Roberts 1993. A version of the EM algorithm, where the Markov chain connection is quite apparent, was given by Baum and Petrie 1966 and Baum et al. 1970.)

9.12 In the setup of Example 9.9, the posterior distribution of N can be evaluated by recursion.

(a) Show that

$$\pi(N) \propto \frac{(N-n_0)!\lambda^N}{N!(N-n_t)!}\,.$$

(b) Using the ratio $\pi(N)/\pi(N-1)$, derive a recursion relation to compute $\mathbb{E}^\pi[N|n_0,n_t]$.

(c) In the case $n_0 = 112$, $n_t = 79$, and $\lambda = 500$, compare the computation time of the above device with the computation time of the Gibbs sampler. (*Note:* See George and Robert 1992 for details.)

9.13 Recall that, in the setting of Example 5.22, animal i, $i = 1,2,\ldots,n$ may be captured at time j, $j = 1,2,\ldots,t$, in one of m locations, where the location is a multinomial random variable $H_{ij} \sim \mathcal{M}_m(\theta_1,\ldots,\theta_m)$. Given $H_{ij} = k$ ($k = 1,2,\ldots,m$), the animal is captured with probability p_k, represented by the random variable $X \sim \mathcal{B}(p_k)$. Define $y_{ijk} = \mathbb{I}(h_{ij} = k)\mathbb{I}(x_{ijk} = 1)$.

(a) Under the conjugate priors

$$(\theta_1,\ldots,\theta_m) \sim \mathcal{D}(\lambda_1,\ldots,\lambda_m) \quad \text{and} \quad p_k \sim \mathcal{B}e(\alpha,\beta)\,,$$

show that the full conditional posterior distributions are given by

$$\{\theta_1,\ldots,\theta_m\} \sim \mathcal{D}(\lambda_1 + \Sigma_{i=1}^n \Sigma_{j=1}^t y_{ij1},\ldots,\lambda_m + \Sigma_{i=1}^n \Sigma_{j=1}^t y_{ijm})$$

and

$$p_k \sim \mathcal{B}e(\alpha + \Sigma_{i=1}^n \Sigma_{j=1}^t x_{ijk}, \beta + n - \Sigma_{i=1}^n \Sigma_{j=1}^t x_{ijk}).$$

(b) Deduce that all of the full conditionals are conjugate and thus that the Gibbs sampler is straightforward to implement.
(c) For the data of Problem 5.28, estimate the θ_i's and the p_i's using the Gibbs sampler starting from the prior parameter values $\alpha = \beta = 5$ and $\lambda_i = 2$.

(*Note:* Dupuis 1995 and Scherrer 1997 discuss how to choose the prior parameter values to reflect the anticipated movement of the animals.)

9.14 Referring to Example 9.21:
(a) Show that, as a function of θ, the normalized complete-data likelihood is $\mathcal{N}((m\bar{x} + (n-m)\bar{z})/n, 1/n)$.
(b) Derive a Monte Carlo EM algorithm to estimate θ.
(c) Contrast the Gibbs sampler algorithm with the EM algorithm of Example 9.21 and the Monte Carlo EM algorithm of part (b).

9.15 Referring to Section 9.4:
(a) Show that the Gibbs sampler of (9.16) has $L(\theta|\mathbf{x})$ as stationary distribution.
(b) Show that if $L(\theta|\mathbf{x}, \mathbf{z})$ is integrable in θ, then so is $L(\theta|\mathbf{x})$, and hence the Markov chain of part(a) is positive.
(c) Complete the proof of ergodicity of the Markov chain. (*Hint:* In addition to Theorem 10.10, see Theorem 6.51. Theorem 9.12 may also be useful in some situations.)

9.16 Referring to Example 9.22:
(a) Verify the distributions in (9.17) for the Gibbs sampler.
(b) Compare the output of the Gibbs sampler to the EM algorithm of Example 5.18. Which algorithm do you prefer and why?

9.17 (Smith and Gelfand 1992) For $i = 1, 2, 3$, consider $Y_i = X_{1i} + X_{2i}$, with

$$X_{1i} \sim \mathcal{B}(n_{1i}, \theta_1), \qquad X_{2i} \sim \mathcal{B}(n_{2i}, \theta_2).$$

(1) Give the likelihood $L(\theta_1, \theta_2)$ for $n_{1i} = 5, 6, 4$, $n_{2i} = 5, 4, 6$, and $y_i = 7, 5, 6$.
(2) For a uniform prior on (θ_1, θ_2), derive the Gibbs sampler based on the natural parameterization.
(3) Examine whether an alternative parameterization or a Metropolis–Hastings algorithm may speed up convergence.

9.18 For the Gibbs sampler

$$X \mid y \sim \mathcal{N}(\rho y,\ 1 - \rho^2)\ ,$$
$$Y \mid x \sim \mathcal{N}(\rho x,\ 1 - \rho^2)\ ,$$

of Example 9.1:
(a) Show that for the X chain, the transition kernel is

$$K(x^*, x) = \frac{1}{2\pi(1-\rho^2)} \int e^{-\frac{1}{2(1-\rho^2)}(x-\rho y)^2} e^{-\frac{1}{2(1-\rho^2)}(y-\rho x^*)^2} dy.$$

(b) Show that $X \sim \mathcal{N}(0, 1)$ is the invariant distribution of the X chain.
(c) Show that $X|x^* \sim \mathcal{N}(\rho^2 x^*, 1 - \rho^4)$. (*Hint*: Complete the square in the exponent of part (a).)
(d) Show that we can write $X_k = \rho^2 X_{k-1} + U_k,\ k = 1, 2, \ldots$, where the U_k are iid $\mathcal{N}(0, 1 - \rho^4)$ and that $\text{cov}(X_0,\ X_k) = \rho^{2k}$, for all k. Deduce that the covariances go to zero.

9.19 In the setup of Example 9.23, show that the likelihood function is not bounded and deduce that, formally, there is no maximum likelihood estimator. (*Hint:* Take $\mu_{j_0} = x_{i_0}$ and let τ_{j_0} go to 0.)

9.20 A model consists in the partial observation of normal vectors $Z = (X, Y) \sim \mathcal{N}_2(0, \Sigma)$ according to a mechanism of random censoring. The corresponding data are given in Table 9.2.

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

Table 9.2. Independent observations of $Z = (X, Y) \sim \mathcal{N}_2(0, \Sigma)$ with missing data (denoted −).

(a) Show that inference can formally be based on the likelihood

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

(b) Show that the choice of the prior distribution $\pi(\Sigma) \propto |\Sigma|^{-1}$ leads to difficulties given that σ_1 and σ_2 are isolated in the likelihood.

(c) Show that the missing components can be simulated through the following algorithm.

Algorithm A.37 –Normal Completion–

1. Simulate

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

$$Y_i^\star \sim \mathcal{N}\left(\rho\frac{\sigma_2}{\sigma_1} x_i, \sigma_2^2(1-\rho^2)\right) \qquad (I = 4, 5).$$

2. Generate

$$\Sigma^{-1} \sim \mathcal{W}_2(8, X^{-1}),$$

with $X = \sum_{i=1}^{8} z_i^* z_i^{*t}$, the dispersion matrix of the completed data.

to derive the posterior distribution of the quantity of interest, ρ.

(d) Propose a Metropolis–Hastings alternative based on a slice sampler.

9.21 Roberts and Rosenthal (2003) derive the *polar slice sampler* from the decomposition $\pi(x) \propto f_0(x)\, f_1(x)$ of a target distribution $\pi(x)$.

(a) Show that the Gibbs sampler whose two steps are (i) to simulate U uniformly on $(0, f_1(x)$ and (ii) to simulate X from $f_0(x)$ restricted to the set of x's such that $f_1(x) > u$ is a valid MCMC algorithm with target distribution π.

(b) Show that, if f_0 is constant, this algorithm is simply the slice sampler of Algorithm [A.31] in Chapter 8. (It is also called the *uniform slice sampler* in Roberts and Rosenthal 2003.)

(c) When $x \in \mathbb{R}^d$, using the case $f_0(x) = |x|^{-(d-1)}$ and $f_1(x) = |x|^{d-1}\,\pi(x)$ is called the *polar slice sampler*. Show that using the polar slice sampler on a d-dimensional log-concave target density is equivalent to a uniform slice sampler on a corresponding one-dimensional log-concave density.

(d) Illustrate this property in the case when $\pi(x) = \exp -|x|$.

9.22 Consider the case of a mixture of normal distributions,

$$\tilde{f}(x) = \sum_{j=1}^{k} p_j \, \frac{e^{-(x-\mu_j)^2/(2\tau_j^2)}}{\sqrt{2\pi}\,\tau_j} \, .$$

(a) Show that the conjugate distribution on (μ_j, τ_j) is

$$\mu_j|\tau_j \sim \mathcal{N}\left(\alpha_j, \tau_j^2/\lambda_j\right), \qquad \tau_j^2 \sim \mathcal{IG}\left(\frac{\lambda_j+3}{2}, \frac{\beta_j}{2}\right) .$$

(b) Show that two valid steps of the Gibbs sampler are as follows.

Algorithm A.38 –Normal Mixture Posterior Simulation–

1. `Simulate` $(i = 1, \ldots, n)$

$$Z_i \sim P(Z_i = j) \propto p_j \, \exp\left\{-(x_i-\mu_j)^2/(2\tau_j^2)\right\} \, \tau_j^{-1}$$

`and compute the statistics` $(j = 1, \ldots, k)$ [*A*.38]

$$n_j = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}, \quad n_j\overline{x}_j = \sum_{i=1}^{n} \mathbb{I}_{z_i=j} x_i, \quad s_j^2 = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}(x_i - \overline{x}_j)^2 \, .$$

2. `Generate`

$$\mu_j|\tau_j \sim \mathcal{N}\left(\frac{\lambda_j\alpha_j + n_j\overline{x}_j}{\lambda_j + n_j}, \frac{\tau_j^2}{\lambda_j+n_j}\right),$$

$$\tau_j^2 \sim \mathcal{IG}\left(\frac{\lambda_j+n_j+3}{2}, \frac{\lambda_j\alpha_j^2 + \beta_j + s_j^2 + n_j i\overline{x}_j{}^2 - (\lambda_j+n_j)^{-1}(\lambda_j\alpha_j + n_j\overline{x}_j)^2}{2}\right),$$

$$p \sim \mathcal{D}_k(\gamma_1 + n_1, \ldots, \gamma_k + n_k) \, .$$

(*Note:* Robert and Soubiran 1993 use this algorithm to derive the maximum likelihood estimators by recursive integration (see Section 5.2.4), showing that the Bayes estimators converge to the local maximum, which is the closest to the initial Bayes estimator.)

9.23 Referring to Section 9.7.2, for the factor ARCH model of (9.18):

(a) Propose a noninformative prior distribution on the parameter $\theta = (\alpha, \beta, a, \Sigma)$ that leads to a proper posterior distribution.

(b) Propose a completion step for the latent variables based on $f(y_t^*|y_t, y_{t-1}^*, \theta)$.

(*Note:* See Diebold and Nerlove 1989, Gouriéroux et al. 1993, Kim et al. 1998, and Billio et al. 1998 for different estimation approaches to this model.)

9.24 Check whether a negative coefficient b in the random walk $Y_t = a + b(X^{(t)} - a) + Z_t$ induces a negative correlation between the $X^{(t)}$'s. Extend to the case where the random walk has an ARCH structure,

$$Y_t = a + b(X^{(t)} - a) + \exp(c + d(X^{(t)} - a)^2)Z_t.$$

9.25 (Diebolt and Robert 1990c,b)
(a) Show that, for a mixture of distributions from exponential families, there exist conjugate priors. (*Hint:* See Example 9.2.)
(b) For the conjugate priors, show that the posterior expectation of the mean parameters of the components can be written in a closed form.
(c) Show that the convergence to stationarity is geometric for all the chains involved in the Gibbs sampler for the mixture model.
(d) Show that Rao–Blackwellization applies in the setup of normal mixture models and that it theoretically improves upon the naïve average. (*Hint:* Use the Duality Principle.)

9.26 (Roeder and Wasserman 1997) In the setup of normal mixtures of Example 9.2:
(a) Derive the posterior distribution associated with the prior $\pi(\mu,\tau)$, where the τ_j^2's are inverted Gamma $\mathcal{IG}(\nu, A)$ and $\pi(\mu_j|\mu_{j-1},\tau)$ is a left-truncated normal distribution

$$\mathcal{N}\left(\mu_{j-1}, B(\tau_j^{-2}+\tau_{j-1}^{-2}))^{-1}\right) ,$$

except for $\pi(\mu_1) = 1/\mu_1$. Assume that the constant B is known and $\pi(A) = 1/A$.
(b) Show that the posterior is always proper.
(c) Derive the posterior distribution using the noninformative prior

$$\pi(\mu,\tau) = \tau^{-1}, \quad p, q_j \sim \mathcal{U}_{[0,1]}, \quad \sigma_j \sim \mathcal{U}_{[0,1]}, \quad \theta_j \sim \mathcal{N}(0,\zeta^2) .$$

and compare.

9.27 Consider the following mixture of uniforms,[1]

$$p\,\mathcal{U}_{[\lambda,\lambda+1]} + (1-p)\,\mathcal{U}_{[\mu,\mu+1]},$$

and an ordered sample $x_1 \le \cdots \le x_{n_1} \le \cdots \le x_n$ such that $x_{n_1} + 1 < x_{n_1+1}$.
(a) Show that the chain $(\lambda^{(t)}, \mu^{(t)})$ associated with the Gibbs sampler corresponding to [*A*.34] is not irreducible.
(b) Show that the above problem disappears if $\mathcal{U}_{[\mu,\mu+1]}$ is replaced with $\mathcal{U}_{[\mu,\mu+0.5]}$ (for n large enough).

9.28 (Billio et al. 1999) A *dynamic desequilibrium* model is defined as the observation of

$$Y_t = \min(Y_{1t}^*, Y_{2t}^*),$$

where the Y_{it}^* are distributed from a parametric joint model, $f(y_{1t}^*, y_{2t}^*)$.
(a) Give the distribution of (Y_{1t}^*, Y_{2t}^*) conditional on Y_t.
(b) Show that a possible completion of the model is to first draw the regime (1 versus 2) and then draw the missing component.
(c) Show that when $f(y_{1t}^*, y_{2t}^*)$ is Gaussian, the above steps can be implemented without approximation.

9.7 Notes

9.7.1 Inference for Mixtures

Although they may seem to apply only for some very particular sets of random phenomena, *mixtures of distributions* (9.5) are of wide use in practical modeling.

[1] This problem was suggested by Eric Moulines.