Exercises

Chapter 2

2.1 Marginal and conditional probability: The social mobility data from Section 2.5 gives a joint probability distribution on (Y_1, Y_2)= (father's occupation, son's occupation). Using this joint distribution, calculate the following distributions:
   a) the marginal probability distribution of a father's occupation;
   b) the marginal probability distribution of a son's occupation;
   c) the conditional distribution of a son's occupation, given that the father is a farmer;
   d) the conditional distribution of a father's occupation, given that the son is a farmer.

2.2 Expectations and variances: Let Y_1 and Y_2 be two independent random variables, such that $\mathrm{E}[Y_i] = \mu_i$ and $\mathrm{Var}[Y_i] = \sigma_i^2$. Using the definition of expectation and variance, compute the following quantities, where a_1 and a_2 are given constants:
   a) $\mathrm{E}[a_1Y_1 + a_2Y_2]$, $\mathrm{Var}[a_1Y_1 + a_2Y_2]$;
   b) $\mathrm{E}[a_1Y_1 - a_2Y_2]$, $\mathrm{Var}[a_1Y_1 - a_2Y_2]$.

2.3 Full conditionals: Let X, Y, Z be random variables with joint density (discrete or continuous) $p(x, y, z) \propto f(x, z)g(y, z)h(z)$. Show that
   a) $p(x|y, z) \propto f(x, z)$, i.e. $p(x|y, z)$ is a function of x and z;
   b) $p(y|x, z) \propto g(y, z)$, i.e. $p(y|x, z)$ is a function of y and z;
   c) X and Y are conditionally independent given Z.

2.4 Symbolic manipulation: Prove the following form of Bayes' rule:

$$\Pr(H_j|E) = \frac{\Pr(E|H_j)\Pr(H_j)}{\sum_{k=1}^{K}\Pr(E|H_k)\Pr(H_k)}$$

where E is any event and $\{H_1, \ldots, H_K\}$ form a partition. Prove this using only axioms **P1**-**P3** from this chapter, by following steps a)-d) below:
   a) Show that $\Pr(H_j|E)\Pr(E) = \Pr(E|H_j)\Pr(H_j)$.

P.D. Hoff, *A First Course in Bayesian Statistical Methods*,
Springer Texts in Statistics, DOI 10.1007/978-0-387-92407-6_BM2,

b) Show that $\Pr(E) = \Pr(E \cap H_1) + \Pr(E \cap \{\cup_{k=2}^{K} H_k\})$.
c) Show that $\Pr(E) = \sum_{k=1}^{K} \Pr(E \cap H_k)$.
d) Put it all together to show Bayes' rule, as described above.

2.5 Urns: Suppose urn H is filled with 40% green balls and 60% red balls, and urn T is filled with 60% green balls and 40% red balls. Someone will flip a coin and then select a ball from urn H or urn T depending on whether the coin lands heads or tails, respectively. Let X be 1 or 0 if the coin lands heads or tails, and let Y be 1 or 0 if the ball is green or red.
a) Write out the joint distribution of X and Y in a table.
b) Find $\mathrm{E}[Y]$. What is the probability that the ball is green?
c) Find $\mathrm{Var}[Y|X=0]$, $\mathrm{Var}[Y|X=1]$ and $\mathrm{Var}[Y]$. Thinking of variance as measuring uncertainty, explain intuitively why one of these variances is larger than the others.
d) Suppose you see that the ball is green. What is the probability that the coin turned up tails?

2.6 Conditional independence: Suppose events A and B are conditionally independent given C, which is written $A \perp B|C$. Show that this implies that $A^c \perp B|C$, $A \perp B^c|C$, and $A^c \perp B^c|C$, where A^c means "not A." Find an example where $A \perp B|C$ holds but $A \perp B|C^c$ does not hold.

2.7 Coherence of bets: de Finetti thought of subjective probability as follows: Your probability $p(E)$ for event E is the amount you would be willing to pay or charge in exchange for a dollar on the occurrence of E. In other words, you must be willing to
- give $p(E)$ to someone, provided they give you \$1 if E occurs;
- take $p(E)$ from someone, and give them \$1 if E occurs.

Your probability for the event $E^c =$"not E" is defined similarly.
a) Show that it is a good idea to have $p(E) \leq 1$.
b) Show that it is a good idea to have $p(E) + p(E^c) = 1$.

2.8 Interpretations of probability: One abstract way to define probability is via measure theory, in that $\Pr(\cdot)$ is simply a "measure" that assigns mass to various events. For example, we can "measure" the number of times a particular event occurs in a potentially infinite sequence, or we can "measure" our information about the outcome of an unknown event. The above two types of measures are combined in de Finetti's theorem, which tells us that an exchangeable model for an infinite binary sequence $Y_1, Y_2, \ldots$ is equivalent to modeling the sequence as conditionally i.i.d. given a parameter θ, where $\Pr(\theta < c)$ represents our information that the long-run frequency of 1's is less than c. With this in mind, discuss the different ways in which probability could be interpreted in each of the following scenarios. Avoid using the word "probable" or "likely" when describing probability. Also discuss the different ways in which the events can be thought of as random.
a) The distribution of religions in Sri Lanka is 70% Buddhist, 15% Hindu, 8% Christian, and 7% Muslim. Suppose each person can be identified

by a number from 1 to K on a census roll. A number x is to be sampled from $\{1, \ldots, K\}$ using a pseudo-random number generator on a computer. Interpret the meaning of the following probabilities:
   i. Pr(person x is Hindu);
   ii. Pr($x = 6452859$);
   iii. Pr(Person x is Hindu|x=6452859).

b) A quarter which you got as change is to be flipped many times. Interpret the meaning of the following probabilities:
   i. Pr(θ, the long-run relative frequency of heads, equals 1/3);
   ii. Pr(the first coin flip will result in a heads);
   iii. Pr(the first coin flip will result in a heads $\mid \theta = 1/3$).

c) The quarter above has been flipped, but you have not seen the outcome. Interpret Pr(the flip has resulted in a heads).

Chapter 3

3.1 Sample survey: Suppose we are going to sample 100 individuals from a county (of size much larger than 100) and ask each sampled person whether they support policy Z or not. Let $Y_i = 1$ if person i in the sample supports the policy, and $Y_i = 0$ otherwise.

a) Assume $Y_1, \ldots, Y_{100}$ are, conditional on θ, i.i.d. binary random variables with expectation θ. Write down the joint distribution of $\Pr(Y_1 = y_1, \ldots, Y_{100} = y_{100}|\theta)$ in a compact form. Also write down the form of $\Pr(\sum Y_i = y|\theta)$.

b) For the moment, suppose you believed that $\theta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Given that the results of the survey were $\sum_{i=1}^{100} Y_i = 57$, compute $\Pr(\sum Y_i = 57|\theta)$ for each of these 11 values of θ and plot these probabilities as a function of θ.

c) Now suppose you originally had no prior information to believe one of these θ-values over another, and so $\Pr(\theta = 0.0) = \Pr(\theta = 0.1) = \cdots = \Pr(\theta = 0.9) = \Pr(\theta = 1.0)$. Use Bayes' rule to compute $p(\theta|\sum_{i=1}^{n} Y_i = 57)$ for each θ-value. Make a plot of this posterior distribution as a function of θ.

d) Now suppose you allow θ to be any value in the interval $[0, 1]$. Using the uniform prior density for θ, so that $p(\theta) = 1$, plot the posterior density $p(\theta) \times \Pr(\sum_{i=1}^{n} Y_i = 57|\theta)$ as a function of θ.

e) As discussed in this chapter, the posterior distribution of θ is beta$(1 + 57, 1 + 100 - 57)$. Plot the posterior density as a function of θ. Discuss the relationships among all of the plots you have made for this exercise.

3.2 Sensitivity analysis: It is sometimes useful to express the parameters a and b in a beta distribution in terms of $\theta_0 = a/(a + b)$ and $n_0 = a + b$, so that $a = \theta_0 n_0$ and $b = (1 - \theta_0)n_0$. Reconsidering the sample survey data in Exercise 3.1, for each combination of $\theta_0 \in \{0.1, 0.2, \ldots, 0.9\}$ and $n_0 \in \{1, 2, 8, 16, 32\}$ find the corresponding a, b values and compute $\Pr(\theta >$

$0.5|\sum Y_i = 57)$ using a beta(a, b) prior distribution for θ. Display the results with a contour plot, and discuss how the plot could be used to explain to someone whether or not they should believe that $\theta > 0.5$, based on the data that $\sum_{i=1}^{100} Y_i = 57$.

3.3 Tumor counts: A cancer laboratory is estimating the rate of tumorigenesis in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed with a mean of 12. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. The observed tumor counts for the two populations are

$$\boldsymbol{y}_A = (12, 9, 12, 14, 13, 13, 15, 8, 15, 6);$$
$$\boldsymbol{y}_B = (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7).$$

a) Find the posterior distributions, means, variances and 95% quantile-based confidence intervals for θ_A and θ_B, assuming a Poisson sampling distribution for each group and the following prior distribution:

$$\theta_A \sim \text{gamma}(120,10),\ \theta_B \sim \text{gamma}(12,1),\ p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B).$$

b) Compute and plot the posterior expectation of θ_B under the prior distribution $\theta_B \sim \text{gamma}(12 \times n_0, n_0)$ for each value of $n_0 \in \{1, 2, \ldots, 50\}$. Describe what sort of prior beliefs about θ_B would be necessary in order for the posterior expectation of θ_B to be close to that of θ_A.

c) Should knowledge about population A tell us anything about population B? Discuss whether or not it makes sense to have $p(\theta_A, \theta_B) = p(\theta_A) \times p(\theta_B)$.

3.4 Mixtures of beta priors: Estimate the probability θ of teen recidivism based on a study in which there were $n = 43$ individuals released from incarceration and $y = 15$ re-offenders within 36 months.

a) Using a beta(2,8) prior for θ, plot $p(\theta)$, $p(y|\theta)$ and $p(\theta|y)$ as functions of θ. Find the posterior mean, mode, and standard deviation of θ. Find a 95% quantile-based confidence interval.

b) Repeat a), but using a beta(8,2) prior for θ.

c) Consider the following prior distribution for θ:

$$p(\theta) = \frac{1}{4} \frac{\Gamma(10)}{\Gamma(2)\Gamma(8)} [3\theta(1-\theta)^7 + \theta^7(1-\theta)],$$

which is a 75-25% mixture of a beta(2,8) and a beta(8,2) prior distribution. Plot this prior distribution and compare it to the priors in a) and b). Describe what sort of prior opinion this may represent.

d) For the prior in c):

i. Write out mathematically $p(\theta) \times p(y|\theta)$ and simplify as much as possible.

ii. The posterior distribution is a mixture of two distributions you know. Identify these distributions.

iii. On a computer, calculate and plot $p(\theta) \times p(y|\theta)$ for a variety of θ values. Also find (approximately) the posterior mode, and discuss its relation to the modes in a) and b).

e) Find a general formula for the weights of the mixture distribution in d)ii, and provide an interpretation for their values.

3.5 Mixtures of conjugate priors: Let $p(y|\phi) = c(\phi)h(y)\exp\{\phi t(y)\}$ be an exponential family model and let $p_1(\phi), \ldots p_K(\phi)$ be K different members of the conjugate class of prior densities given in Section 3.3. A mixture of conjugate priors is given by $\tilde{p}(\theta) = \sum_{k=1}^{K} w_k p_k(\theta)$, where the w_k's are all greater than zero and $\sum w_k = 1$ (see also Diaconis and Ylvisaker (1985)).

a) Identify the general form of the posterior distribution of θ, based on n i.i.d. samples from $p(y|\theta)$ and the prior distribution given by $\tilde{p}$.

b) Repeat a) but in the special case that $p(y|\theta) = \text{dpois}(y, \theta)$ and $p_1, \ldots, p_K$ are gamma densities.

3.6 Exponential family expectations: Let $p(y|\phi) = c(\phi)h(y)\exp\{\phi t(y)\}$ be an exponential family model.

a) Take derivatives with respect to ϕ of both sides of the equation $\int p(y|\phi)\, dy = 1$ to show that $\text{E}[t(Y)|\phi] = -c'(\phi)/c(\phi)$.

b) Let $p(\phi) \propto c(\phi)^{n_0} e^{n_0 t_0 \phi}$ be the prior distribution for ϕ. Calculate $dp(\phi)/d\phi$ and, using the fundamental theorem of calculus, discuss what must be true so that $\text{E}[-c(\phi)/c(\phi)] = t_0$.

3.7 Posterior prediction: Consider a pilot study in which $n_1 = 15$ children enrolled in special education classes were randomly selected and tested for a certain type of learning disability. In the pilot study, $y_1 = 2$ children tested positive for the disability.

a) Using a uniform prior distribution, find the posterior distribution of θ, the fraction of students in special education classes who have the disability. Find the posterior mean, mode and standard deviation of θ, and plot the posterior density.

Researchers would like to recruit students with the disability to participate in a long-term study, but first they need to make sure they can recruit enough students. Let $n_2 = 278$ be the number of children in special education classes in this particular school district, and let Y_2 be the number of students with the disability.

b) Find $\Pr(Y_2 = y_2|Y_1 = 2)$, the posterior predictive distribution of Y_2, as follows:

i. Discuss what assumptions are needed about the joint distribution of (Y_1, Y_2) such that the following is true:

$$\Pr(Y_2 = y_2|Y_1 = 2) = \int_0^1 \Pr(Y_2 = y_2|\theta)p(\theta|Y_1 = 2)\, d\theta\,.$$

ii. Now plug in the forms for $\Pr(Y_2 = y_2|\theta)$ and $p(\theta|Y_1 = 2)$ in the above integral.

iii. Figure out what the above integral must be by using the calculus result discussed in Section 3.1.

c) Plot the function $\Pr(Y_2 = y_2|Y_1 = 2)$ as a function of y_2. Obtain the mean and standard deviation of Y_2, given $Y_1 = 2$.

d) The posterior mode and the MLE (maximum likelihood estimate; see Exercise 3.14) of θ, based on data from the pilot study, are both $\hat{\theta} = 2/15$. Plot the distribution $\Pr(Y_2 = y_2|\theta = \hat{\theta})$, and find the mean and standard deviation of Y_2 given $\theta = \hat{\theta}$. Compare these results to the plots and calculations in c) and discuss any differences. Which distribution for Y_2 would you use to make predictions, and why?

3.8 Coins: Diaconis and Ylvisaker (1985) suggest that coins spun on a flat surface display long-run frequencies of heads that vary from coin to coin. About 20% of the coins behave symmetrically, whereas the remaining coins tend to give frequencies of 1/3 or 2/3.

a) Based on the observations of Diaconis and Ylvisaker, use an appropriate mixture of beta distributions as a prior distribution for θ, the long-run frequency of heads for a particular coin. Plot your prior.

b) Choose a single coin and spin it at least 50 times. Record the number of heads obtained. Report the year and denomination of the coin.

c) Compute your posterior for θ, based on the information obtained in b).

d) Repeat b) and c) for a different coin, but possibly using a prior for θ that includes some information from the first coin. Your choice of a new prior may be informal, but needs to be justified. How the results from the first experiment influence your prior for the θ of the second coin may depend on whether or not the two coins have the same denomination, have a similar year, etc. Report the year and denomination of this coin.

3.9 Galenshore distribution: An unknown quantity Y has a Galenshore(a, θ) distribution if its density is given by

$$p(y) = \frac{2}{\Gamma(a)}\theta^{2a}y^{2a-1}e^{-\theta^2 y^2}$$

for $y > 0$, $\theta > 0$ and $a > 0$. Assume for now that a is known. For this density,

$$\mathrm{E}[Y] = \frac{\Gamma(a + 1/2)}{\theta\Gamma(a)}, \quad \mathrm{E}[Y^2] = \frac{a}{\theta^2}.$$

a) Identify a class of conjugate prior densities for θ. Plot a few members of this class of densities.

b) Let $Y_1, \ldots, Y_n \sim$ i.i.d. Galenshore(a, θ). Find the posterior distribution of θ given $Y_1, \ldots, Y_n$, using a prior from your conjugate class.

c) Write down $p(\theta_a|Y_1, \ldots, Y_n)/p(\theta_b|Y_1, \ldots, Y_n)$ and simplify. Identify a sufficient statistic.

d) Determine $\mathrm{E}[\theta|y_1, \ldots, y_n]$.

e) Determine the form of the posterior predictive density $p(\tilde{y}|y_1 \ldots, y_n)$.

3.10 Change of variables: Let $\psi = g(\theta)$, where g is a monotone function of θ, and let h be the inverse of g so that $\theta = h(\psi)$. If $p_\theta(\theta)$ is the probability density of θ, then the probability density of ψ induced by p_θ is given by $p_\psi(\psi) = p_\theta(h(\psi)) \times |\frac{dh}{d\psi}|$.

a) Let $\theta \sim \text{beta}(a, b)$ and let $\psi = \log[\theta/(1-\theta)]$. Obtain the form of p_ψ and plot it for the case that $a = b = 1$.

b) Let $\theta \sim \text{gamma}(a, b)$ and let $\psi = \log\theta$. Obtain the form of p_ψ and plot it for the case that $a = b = 1$.

3.12 Jeffreys' prior: Jeffreys (1961) suggested a default rule for generating a prior distribution of a parameter θ in a sampling model $p(y|\theta)$. Jeffreys' prior is given by $p_J(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta) = -\text{E}[\partial^2 \log p(Y|\theta)/\partial\theta^2|\theta]$ is the *Fisher information*.

a) Let $Y \sim \text{binomial}(n, \theta)$. Obtain Jeffreys' prior distribution $p_J(\theta)$ for this model.

b) Reparameterize the binomial sampling model with $\psi = \log\theta/(1-\theta)$, so that $p(y|\psi) = \binom{n}{y} e^{\psi y}(1+e^\psi)^{-n}$. Obtain Jeffreys' prior distribution $p_J(\psi)$ for this model.

c) Take the prior distribution from a) and apply the change of variables formula from Exercise 3.10 to obtain the induced prior density on ψ. This density should be the same as the one derived in part b) of this exercise. This consistency under reparameterization is the defining characteristic of Jeffrey's' prior.

3.13 Improper Jeffreys' prior: Let $Y \sim \text{Poisson}(\theta)$.

a) Apply Jeffreys' procedure to this model, and compare the result to the family of gamma densities. Does Jeffreys' procedure produce an actual probability density for θ? In other words, can $\sqrt{I(\theta)}$ be proportional to an actual probability density for $\theta \in (0, \infty)$?

b) Obtain the form of the function $f(\theta, y) = \sqrt{I(\theta)} \times p(y|\theta)$. What probability density for θ is $f(\theta, y)$ proportional to? Can we think of $f(\theta, y)/\int f(\theta, y)d\theta$ as a posterior density of θ given $Y = y$?

3.14 Unit information prior: Let $Y_1, \ldots, Y_n \sim$ i.i.d. $p(y|\theta)$. Having observed the values $Y_1 = y_1, \ldots, Y_n = y_n$, the *log likelihood* is given by $l(\theta|\boldsymbol{y}) = \sum \log p(y_i|\theta)$, and the value $\hat{\theta}$ of θ that maximizes $l(\theta|\boldsymbol{y})$ is called the *maximum likelihood estimator*. The negative of the curvature of the log-likelihood, $J(\theta) = -\partial^2 l(\theta|\boldsymbol{y})/\partial\theta^2$, describes the precision of the MLE $\hat{\theta}$ and is called the *observed Fisher information*. For situations in which it is difficult to quantify prior information in terms of a probability distribution, some have suggested that the "prior" distribution be based on the likelihood, for example, by centering the prior distribution around the MLE $\hat{\theta}$. To deal with the fact that the MLE is not really prior information, the curvature of the prior is chosen so that it has only "one nth" as much information as the likelihood, so that $-\partial^2 \log p(\theta)/\partial\theta^2 = J(\theta)/n$. Such a prior is called a *unit information prior* (Kass and Wasserman, 1995; Kass

and Raftery, 1995), as it has as much information as the average amount of information from a single observation. The unit information prior is not really a prior distribution, as it is computed from the observed data. However, it can be roughly viewed as the prior information of someone with weak but accurate prior information.

a) Let $Y_1, \ldots, Y_n \sim$ i.i.d. binary(θ). Obtain the MLE $\hat{\theta}$ and $J(\hat{\theta})/n$.
b) Find a probability density $p_U(\theta)$ such that $\log p_U(\theta) = l(\theta|\boldsymbol{y})/n + c$, where c is a constant that does not depend on θ. Compute the information $-\partial^2 \log p_U(\theta)/\partial\theta^2$ of this density.
c) Obtain a probability density for θ that is proportional to $p_U(\theta) \times p(y_1, \ldots, y_n|\theta)$. Can this be considered a posterior distribution for θ?
d) Repeat a), b) and c) but with $p(y|\theta)$ being the Poisson distribution.

Chapter 4

4.1 Posterior comparisons: Reconsider the sample survey in Exercise 3.1. Suppose you are interested in comparing the rate of support in that county to the rate in another county. Suppose that a survey of sample size 50 was done in the second county, and the total number of people in the sample who supported the policy was 30. Identify the posterior distribution of θ_2 assuming a uniform prior. Sample 5,000 values of each of θ_1 and θ_2 from their posterior distributions and estimate $\Pr(\theta_1 < \theta_2|\text{the data and prior})$.

4.2 Tumor count comparisons: Reconsider the tumor count data in Exercise 3.3:

a) For the prior distribution given in part a) of that exercise, obtain $\Pr(\theta_B < \theta_A|\boldsymbol{y}_A, \boldsymbol{y}_B)$ via Monte Carlo sampling.
b) For a range of values of n_0, obtain $\Pr(\theta_B < \theta_A|\boldsymbol{y}_A, \boldsymbol{y}_B)$ for $\theta_A \sim$ gamma$(120, 10)$ and $\theta_B \sim$ gamma$(12 \times n_0, n_0)$. Describe how sensitive the conclusions about the event $\{\theta_B < \theta_A\}$ are to the prior distribution on θ_B.
c) Repeat parts a) and b), replacing the event $\{\theta_B < \theta_A\}$ with the event $\{\tilde{Y}_B < \tilde{Y}_A\}$, where $\tilde{Y}_A$ and $\tilde{Y}_B$ are samples from the posterior predictive distribution.

4.3 Posterior predictive checks: Let's investigate the adequacy of the Poisson model for the tumor count data. Following the example in Section 4.4, generate posterior predictive datasets $\boldsymbol{y}_A^{(1)}, \ldots, \boldsymbol{y}_A^{(1000)}$. Each $\boldsymbol{y}_A^{(s)}$ is a sample of size $n_A = 10$ from the Poisson distribution with parameter $\theta_A^{(s)}$, $\theta_A^{(s)}$ is itself a sample from the posterior distribution $p(\theta_A|\boldsymbol{y}_A)$, and $\boldsymbol{y}_A$ is the observed data.

a) For each s, let $t^{(s)}$ be the sample average of the 10 values of $\boldsymbol{y}_A^{(s)}$, divided by the sample standard deviation of $\boldsymbol{y}_A^{(s)}$. Make a histogram of $t^{(s)}$ and compare to the observed value of this statistic. Based on this statistic, assess the fit of the Poisson model for these data.

b) Repeat the above goodness of fit evaluation for the data in population B.

4.4 Mixtures of conjugate priors: For the posterior density from Exercise 3.4:

a) Make a plot of $p(\theta|y)$ or $p(y|\theta)p(\theta)$ using the mixture prior distribution and a dense sequence of θ-values. Can you think of a way to obtain a 95% quantile-based posterior confidence interval for θ? You might want to try some sort of discrete approximation.

b) To sample a random variable z from the mixture distribution $wp_1(z) + (1-w)p_0(z)$, first toss a w-coin and let x be the outcome (this can be done in R with `x<-rbinom(1,1,w)`). Then if $x = 1$ sample z from p_1, and if $x = 0$ sample z from p_0. Using this technique, obtain a Monte Carlo approximation of the posterior distribution $p(\theta|y)$ and a 95% quantile-based confidence interval, and compare them to the results in part a).

4.5 Cancer deaths: Suppose for a set of counties $i \in \{1, \ldots, n\}$ we have information on the population size X_i = number of people in 10,000s, and Y_i = number of cancer fatalities. One model for the distribution of cancer fatalities is that, given the cancer rate θ, they are independently distributed with $Y_i \sim \text{Poisson}(\theta X_i)$.

a) Identify the posterior distribution of θ given data $(Y_1, X_1), \ldots, (Y_n, X_n)$ and a gamma(a, b) prior distribution.

The file `cancer_react.dat` contains 1990 population sizes (in 10,000s) and number of cancer fatalities for 10 counties in a Midwestern state that are near nuclear reactors. The file `cancer_noreact.dat` contains the same data on counties in the same state that are not near nuclear reactors. Consider these data as samples from two populations of counties: one is the population of counties with no neighboring reactors and a fatality rate of θ_1 deaths per 10,000, and the other is a population of counties having nearby reactors and a fatality rate of θ_2. In this exercise we will model beliefs about the rates as independent and such that $\theta_1 \sim \text{gamma}(a_1, b_1)$ and $\theta_2 \sim \text{gamma}(a_2, b_2)$.

b) Using the numerical values of the data, identify the posterior distributions for θ_1 and θ_2 for any values of (a_1, b_1, a_2, b_2).

c) Suppose cancer rates from previous years have been roughly $\tilde{\theta} = 2.2$ per 10,000 (and note that most counties are not near reactors). For each of the following three prior opinions, compute E$[\theta_1|\text{data}]$, E$[\theta_2|\text{data}]$, 95% quantile-based posterior intervals for θ_1 and θ_2, and $\Pr(\theta_2 > \theta_1|\text{data})$. Also plot the posterior densities (try to put $p(\theta_1|\text{data})$ and $p(\theta_2|\text{data})$ on the same plot). Comment on the differences across posterior opinions.

i. Opinion 1: $(a_1 = a_2 = 2.2 \times 100, b_1 = b_2 = 100)$. Cancer rates for both types of counties are similar to the average rates across all counties from previous years.

ii. Opinion 2: $(a_1 = 2.2 \times 100, b_1 = 100, a_2 = 2.2, b_1 = 1)$. Cancer rates in this year for nonreactor counties are similar to rates in previous years in nonreactor counties. We don't have much information on reactor counties, but perhaps the rates are close to those observed previously in nonreactor counties.

iii. Opinion 3: $(a_1 = a_2 = 2.2, b_1 = b_2 = 1)$. Cancer rates in this year could be different from rates in previous years, for both reactor and nonreactor counties.

d) In the above analysis we assumed that population size gives no information about fatality rate. Is this reasonable? How would the analysis have to change if this is not reasonable?

e) We encoded our beliefs about θ_1 and θ_2 such that they gave no information about each other (they were *a priori* independent). Think about why and how you might encode beliefs such that they were *a priori* dependent.

4.6 Non-informative prior distributions: Suppose for a binary sampling problem we plan on using a uniform, or beta(1,1), prior for the population proportion θ. Perhaps our reasoning is that this represents "no prior information about θ." However, some people like to look at proportions on the log-odds scale, that is, they are interested in $\gamma = \log \frac{\theta}{1-\theta}$. Via Monte Carlo sampling or otherwise, find the prior distribution for γ that is induced by the uniform prior for θ. Is the prior informative about γ?

4.7 Mixture models: After a posterior analysis on data from a population of squash plants, it was determined that the total vegetable weight of a given plant could be modeled with the following distribution:

$$p(y|\theta, \sigma^2) = .31\text{dnorm}(y, \theta, \sigma) + .46\text{dnorm}(2\theta_1, 2\sigma) + .23\text{dnorm}(y, 3\theta_1, 3\sigma)$$

where the posterior distributions of the parameters have been calculated as $1/\sigma^2 \sim \text{gamma}(10, 2.5)$, and $\theta|\sigma^2 \sim \text{normal}(4.1, \sigma^2/20)$.

a) Sample at least 5,000 y values from the posterior predictive distribution.

b) Form a 75% quantile-based confidence interval for a new value of Y.

c) Form a 75% HPD region for a new Y as follows:

i. Compute estimates of the posterior density of Y using the `density` command in R, and then normalize the density values so they sum to 1.

ii. Sort these discrete probabilities in decreasing order.

iii. Find the first probability value such that the cumulative sum of the sorted values exceeds 0.75. Your HPD region includes all values of y which have a discretized probability greater than this cutoff. Describe your HPD region, and compare it to your quantile-based region.

d) Can you think of a physical justification for the mixture sampling distribution of Y?

4.8 More posterior predictive checks: Let θ_A and θ_B be the average number of children of men in their 30s with and without bachelor's degrees, respectively.

a) Using a Poisson sampling model, a gamma(2,1) prior for each θ and the data in the files `menchild30bach.dat` and `menchild30nobach.dat`, obtain 5,000 samples of $\tilde{Y}_A$ and $\tilde{Y}_B$ from the posterior predictive distribution of the two samples. Plot the Monte Carlo approximations to these two posterior predictive distributions.

b) Find 95% quantile-based posterior confidence intervals for $\theta_B - \theta_A$ and $\tilde{Y}_B - \tilde{Y}_A$. Describe in words the differences between the two populations using these quantities and the plots in a), along with any other results that may be of interest to you.

c) Obtain the empirical distribution of the data in group B. Compare this to the Poisson distribution with mean $\hat{\theta} = 1.4$. Do you think the Poisson model is a good fit? Why or why not?

d) For each of the 5,000 θ_B-values you sampled, sample $n_B = 218$ Poisson random variables and count the number of 0s and the number of 1s in each of the 5,000 simulated datasets. You should now have two sequences of length 5,000 each, one sequence counting the number of people having zero children for each of the 5,000 posterior predictive datasets, the other counting the number of people with one child. Plot the two sequences against one another (one on the x-axis, one on the y-axis). Add to the plot a point marking how many people in the observed dataset had zero children and one child. Using this plot, describe the adequacy of the Poisson model.

Chapter 5

5.1 Studying: The files `school1.dat`, `school2.dat` and `school3.dat` contain data on the amount of time students from three high schools spent on studying or homework during an exam period. Analyze data from each of these schools separately, using the normal model with a conjugate prior distribution, in which $\{\mu_0 = 5, \sigma_0^2 = 4, \kappa_0 = 1, \nu_0 = 2\}$ and compute or approximate the following:

a) posterior means and 95% confidence intervals for the mean θ and standard deviation σ from each school;

b) the posterior probability that $\theta_i < \theta_j < \theta_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$;

c) the posterior probability that $\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$, where $\tilde{Y}_i$ is a sample from the posterior predictive distribution of school i.

d) Compute the posterior probability that θ_1 is bigger than both θ_2 and θ_3, and the posterior probability that $\tilde{Y}_1$ is bigger than both $\tilde{Y}_2$ and $\tilde{Y}_3$.

5.2 Sensitivity analysis: Thirty-two students in a science classroom were randomly assigned to one of two study methods, A and B, so that $n_A = n_B = 16$ students were assigned to each method. After several weeks of study, students were examined on the course material with an exam designed to give an average score of 75 with a standard deviation of 10. The scores for the two groups are summarized by $\{\bar{y}_A = 75.2, s_A = 7.3\}$ and $\{\bar{y}_B = 77.5, s_b = 8.1\}$. Consider independent, conjugate normal prior distributions for each of θ_A and θ_B, with $\mu_0 = 75$ and $\sigma_0^2 = 100$ for both groups. For each $(\kappa_0, \nu_0) \in \{(1,1),(2,2),(4,4),(8,8),(16,16),(32,32)\}$ (or more values), obtain $\Pr(\theta_A < \theta_B | \boldsymbol{y}_A, \boldsymbol{y}_B)$ via Monte Carlo sampling. Plot this probability as a function of $(\kappa_0 = \nu_0)$. Describe how you might use this plot to convey the evidence that $\theta_A < \theta_B$ to people of a variety of prior opinions.

5.3 Marginal distributions: Given observations $Y_1, \ldots, Y_n \sim$ i.i.d. normal(θ, σ^2) and using the conjugate prior distribution for θ and σ^2, derive the formula for $p(\theta | y_1, \ldots, y_n)$, the marginal posterior distribution of θ, conditional on the data but marginal over σ^2. Check your work by comparing your formula to a Monte Carlo estimate of the marginal distribution, using some values of $Y_1, \ldots, Y_n$, μ_0, σ_0^2, ν_0 and κ_0 that you choose. Also derive $p(\tilde{\sigma}^2 | y_1, \ldots, y_n)$, where $\tilde{\sigma}^2 = 1/\sigma^2$ is the precision.

5.4 Jeffreys' prior: For sampling models expressed in terms of a p-dimensional vector $\boldsymbol{\psi}$, Jeffreys' prior (Exercise 3.11) is defined as $p_J(\boldsymbol{\psi}) \propto \sqrt{|I(\boldsymbol{\psi})|}$, where $|I(\boldsymbol{\psi})|$ is the determinant of the $p \times p$ matrix $I(\boldsymbol{\psi})$ having entries $I(\boldsymbol{\psi})_{k,l} = -\mathrm{E}[\partial^2 \log p(Y|\boldsymbol{\psi}) / \partial\psi_k \partial\psi_l]$.

a) Show that Jeffreys' prior for the normal model is $p_J(\theta, \sigma^2) \propto (\sigma^2)^{-3/2}$.

b) Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the observed values of an i.i.d. sample from a normal(θ, σ^2) population. Find a probability density $p_J(\theta, \sigma^2 | \boldsymbol{y})$ such that $p_J(\theta, \sigma^2 | \boldsymbol{y}) \propto p_J(\theta, \sigma^2) p(\boldsymbol{y} | \theta, \sigma^2)$. It may be convenient to write this joint density as $p_J(\theta | \sigma^2, \boldsymbol{y}) \times p_J(\sigma^2 | \boldsymbol{y})$. Can this joint density be considered a posterior density?

5.5 Unit information prior: Obtain a unit information prior for the normal model as follows:

a) Reparameterize the normal model as $p(y | \theta, \psi)$, where $\psi = 1/\sigma^2$. Write out the log likelihood $l(\theta, \psi | \boldsymbol{y}) = \sum \log p(y_i | \theta, \psi)$ in terms of θ and ψ.

b) Find a probability density $p_U(\theta, \psi)$ so that $\log p_U(\theta, \psi) = l(\theta, \psi | \boldsymbol{y})/n + c$, where c is a constant that does not depend on θ or ψ. Hint: Write $\sum (y_i - \theta)^2$ as $\sum (y_i - \bar{y} + \bar{y} - \theta)^2 = \sum (y_i - \bar{y})^2 + n(\theta - \bar{y})^2$, and recall that $\log p_U(\theta, \psi) = \log p_U(\theta | \psi) + \log p_U(\psi)$.

c) Find a probability density $p_U(\theta, \psi | \boldsymbol{y})$ that is proportional to $p_U(\theta, \psi) \times p(y_1, \ldots, y_n | \theta, \psi)$. It may be convenient to write this joint density as $p_U(\theta | \psi, \boldsymbol{y}) \times p_U(\psi | \boldsymbol{y})$. Can this joint density be considered a posterior density?

Chapter 6

6.1 Poisson population comparisons: Let's reconsider the number of children data of Exercise 4.8. We'll assume Poisson sampling models for the two groups as before, but now we'll parameterize θ_A and θ_B as $\theta_A = \theta$, $\theta_B = \theta \times \gamma$. In this parameterization, γ represents the relative rate θ_B/θ_A. Let $\theta \sim \text{gamma}(a_\theta, b_\theta)$ and let $\gamma \sim \text{gamma}(a_\gamma, b_\gamma)$.

a) Are θ_A and θ_B independent or dependent under this prior distribution? In what situations is such a joint prior distribution justified?

b) Obtain the form of the full conditional distribution of θ given $\boldsymbol{y}_A$, $\boldsymbol{y}_B$ and γ.

c) Obtain the form of the full conditional distribution of γ given $\boldsymbol{y}_A$, $\boldsymbol{y}_B$ and θ.

d) Set $a_\theta = 2$ and $b_\theta = 1$. Let $a_\gamma = b_\gamma \in \{8, 16, 32, 64, 128\}$. For each of these five values, run a Gibbs sampler of at least 5,000 iterations and obtain $\text{E}[\theta_B - \theta_A | \boldsymbol{y}_A, \boldsymbol{y}_B]$. Describe the effects of the prior distribution for γ on the results.

6.2 Mixture model: The file `glucose.dat` contains the plasma glucose concentration of 532 females from a study on diabetes (see Exercise 7.6).

a) Make a histogram or kernel density estimate of the data. Describe how this empirical distribution deviates from the shape of a normal distribution.

b) Consider the following mixture model for these data: For each study participant there is an unobserved group membership variable X_i which is equal to 1 or 2 with probability p and $1 - p$. If $X_i = 1$ then $Y_i \sim \text{normal}(\theta_1, \sigma_1^2)$, and if $X_i = 2$ then $Y_i \sim \text{normal}(\theta_2, \sigma_2^2)$. Let $p \sim \text{beta}(a, b)$, $\theta_j \sim \text{normal}(\mu_0, \tau_0^2)$ and $1/\sigma_j \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$ for both $j = 1$ and $j = 2$. Obtain the full conditional distributions of $(X_1, \ldots, X_n)$, p, θ_1, θ_2, σ_1^2 and σ_2^2.

c) Setting $a = b = 1$, $\mu_0 = 120$, $\tau_0^2 = 200$, $\sigma_0^2 = 1000$ and $\nu_0 = 10$, implement the Gibbs sampler for at least 10,000 iterations. Let $\theta_{(1)}^{(s)} = \min\{\theta_1^{(s)}, \theta_2^{(s)}\}$ and $\theta_{(2)}^{(s)} = \max\{\theta_1^{(s)}, \theta_2^{(s)}\}$. Compute and plot the autocorrelation functions of $\theta_{(1)}^{(s)}$ and $\theta_{(2)}^{(s)}$, as well as their effective sample sizes.

d) For each iteration s of the Gibbs sampler, sample a value $x \sim \text{binary}(p^{(s)})$, then sample $\tilde{Y}^{(s)} \sim \text{normal}(\theta_x^{(s)}, \sigma_x^{2(s)})$. Plot a histogram or kernel density estimate for the empirical distribution of $\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(S)}$, and compare to the distribution in part a). Discuss the adequacy of this two-component mixture model for the glucose data.

6.3 Probit regression: A panel study followed 25 married couples over a period of five years. One item of interest is the relationship between divorce rates and the various characteristics of the couples. For example, the researchers would like to model the probability of divorce as a function of

age differential, recorded as the man's age minus the woman's age. The data can be found in the file `divorce.dat`. We will model these data with probit regression, in which a binary variable Y_i is described in terms of an explanatory variable x_i via the following latent variable model:

$$Z_i = \beta x_i + \epsilon_i$$
$$Y_i = \delta_{(c,\infty)}(Z_i),$$

where β and c are unknown coefficients, $\epsilon_1, \ldots, \epsilon_n \sim$ i.i.d. normal$(0, 1)$ and $\delta_{(c,\infty)}(z) = 1$ if $z > c$ and equals zero otherwise.

a) Assuming $\beta \sim$ normal$(0, \tau_\beta^2)$ obtain the full conditional distribution $p(\beta|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, c)$.
b) Assuming $c \sim$ normal$(0, \tau_c^2)$, show that $p(c|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \beta)$ is a constrained normal density, i.e. proportional to a normal density but constrained to lie in an interval. Similarly, show that $p(z_i|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}_{-i}, \beta, c)$ is proportional to a normal density but constrained to be either above c or below c, depending on y_i.
c) Letting $\tau_\beta^2 = \tau_c^2 = 16$, implement a Gibbs sampling scheme that approximates the joint posterior distribution of $\boldsymbol{Z}$, β, and c (a method for sampling from constrained normal distributions is outlined in Section 12.1.1). Run the Gibbs sampler long enough so that the effective sample sizes of all unknown parameters are greater than 1,000 (including the Z_i's). Compute the autocorrelation function of the parameters and discuss the mixing of the Markov chain.
d) Obtain a 95% posterior confidence interval for β, as well as $\Pr(\beta > 0|\boldsymbol{y}, \boldsymbol{x})$.

Chapter 7

7.1 Jeffreys' prior: For the multivariate normal model, Jeffreys' rule for generating a prior distribution on $(\boldsymbol{\theta}, \Sigma)$ gives $p_J(\boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-(p+2)/2}$.
   a) Explain why the function p_J cannot actually be a probability density for $(\boldsymbol{\theta}, \Sigma)$.
   b) Let $p_J(\boldsymbol{\theta}, \Sigma|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the probability density that is proportional to $p_J(\boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n|\boldsymbol{\theta}, \Sigma)$. Obtain the form of $p_J(\boldsymbol{\theta}, \Sigma|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $p_J(\boldsymbol{\theta}|\Sigma, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $p_J(\Sigma|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

7.2 Unit information prior: Letting $\Psi = \Sigma^{-1}$, show that a unit information prior for $(\boldsymbol{\theta}, \Psi)$ is given by $\boldsymbol{\theta}|\Psi \sim$ multivariate normal$(\bar{\boldsymbol{y}}, \Psi^{-1})$ and $\Psi \sim$ Wishart$(p + 1, \mathbf{S}^{-1})$, where $\mathbf{S} = \sum(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T/n$. This can be done by mimicking the procedure outlined in Exercise 5.6 as follows:
   a) Reparameterize the multivariate normal model in terms of the precision matrix $\Psi = \Sigma^{-1}$. Write out the resulting log likelihood, and find a probability density $p_U(\boldsymbol{\theta}, \Psi) = p_U(\boldsymbol{\theta}|\Psi)p_U(\Psi)$ such that $\log p(\boldsymbol{\theta}, \Psi) = l(\boldsymbol{\theta}, \Psi|\mathbf{Y})/n + c$, where c does not depend on $\boldsymbol{\theta}$ or Ψ.

Hint: Write $(\boldsymbol{y}_i - \boldsymbol{\theta})$ as $(\boldsymbol{y}_i - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} - \boldsymbol{\theta})$, and note that $\sum \boldsymbol{a}_i^T \mathbf{B} \boldsymbol{a}_i$ can be written as $\text{tr}(\mathbf{AB})$, where $\mathbf{A} = \sum \boldsymbol{a}_i \boldsymbol{a}_i^T$.

b) Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Psi)$. Obtain a density $p_U(\boldsymbol{\theta}, \Sigma | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \propto p_U(\boldsymbol{\theta}|\Sigma) p_U(\boldsymbol{\Sigma}) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma)$. Can this be interpreted as a posterior distribution for θ and Σ?

7.3 Australian crab data: The files `bluecrab.dat` and `orangecrab.dat` contain measurements of body depth (Y_1) and rear width (Y_2), in millimeters, made on 50 male crabs from each of two species, blue and orange. We will model these data using a bivariate normal distribution.

a) For each of the two species, obtain posterior distributions of the population mean $\boldsymbol{\theta}$ and covariance matrix Σ as follows: Using the semiconjugate prior distributions for $\boldsymbol{\theta}$ and Σ, set $\boldsymbol{\mu}_0$ equal to the sample mean of the data, Λ_0 and $\mathbf{S}_0$ equal to the sample covariance matrix and $\nu_0 = 4$. Obtain 10,000 posterior samples of $\boldsymbol{\theta}$ and Σ. Note that this "prior" distribution loosely centers the parameters around empirical estimates based on the observed data (and is very similar to the unit information prior described in the previous exercise). It cannot be considered as our true prior distribution, as it was derived from the observed data. However, it can be roughly considered as the prior distribution of someone with weak but unbiased information.

b) Plot values of $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ for each group and compare. Describe any size differences between the two groups.

c) From each covariance matrix obtained from the Gibbs sampler, obtain the corresponding correlation coefficient. From these values, plot posterior densities of the correlations ρ_{blue} and ρ_{orange} for the two groups. Evaluate differences between the two species by comparing these posterior distributions. In particular, obtain an approximation to $\Pr(\rho_{\text{blue}} < \rho_{\text{orange}} | \boldsymbol{y}_{\text{blue}}, \boldsymbol{y}_{\text{orange}})$. What do the results suggest about differences between the two populations?

7.4 Marriage data: The file `agehw.dat` contains data on the ages of 100 married couples sampled from the U.S. population.

a) Before you look at the data, use your own knowledge to formulate a semiconjugate prior distribution for $\boldsymbol{\theta} = (\theta_h, \theta_w)^T$ and Σ, where θ_h, θ_w are mean husband and wife ages, and Σ is the covariance matrix.

b) Generate a *prior predictive dataset* of size $n = 100$, by sampling $(\boldsymbol{\theta}, \Sigma)$ from your prior distribution and then simulating $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim$ i.i.d. multivariate normal$(\boldsymbol{\theta}, \Sigma)$. Generate several such datasets, make bivariate scatterplots for each dataset, and make sure they roughly represent your prior beliefs about what such a dataset would actually look like. If your prior predictive datasets do not conform to your beliefs, go back to part a) and formulate a new prior. Report the prior that you eventually decide upon, and provide scatterplots for at least three prior predictive datasets.

c) Using your prior distribution and the 100 values in the dataset, obtain an MCMC approximation to $p(\boldsymbol{\theta}, \Sigma | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{100})$. Plot the joint

posterior distribution of θ_h and θ_w, and also the marginal posterior density of the correlation between Y_h and Y_w, the ages of a husband and wife. Obtain 95% posterior confidence intervals for θ_h, θ_w and the correlation coefficient.

d) Obtain 95% posterior confidence intervals for θ_h, θ_w and the correlation coefficient using the following prior distributions:
   i. Jeffreys' prior, described in Exercise 7.1;
   ii. the unit information prior, described in Exercise 7.2;
   iii. a "diffuse prior" with $\mu_0 = \mathbf{0}$, $\Lambda_0 = 10^5 \times \mathbf{I}$, $\mathbf{S}_0 = 1000 \times \mathbf{I}$ and $\nu_0 = 3$.

e) Compare the confidence intervals from d) to those obtained in c). Discuss whether or not you think that your prior information is helpful in estimating $\boldsymbol{\theta}$ and Σ, or if you think one of the alternatives in d) is preferable. What about if the sample size were much smaller, say $n = 25$?

7.5 Imputation: The file `interexp.dat` contains data from an experiment that was interrupted before all the data could be gathered. Of interest was the difference in reaction times of experimental subjects when they were given stimulus A versus stimulus B. Each subject is tested under one of the two stimuli on their first day of participation in the study, and is tested under the other stimulus at some later date. Unfortunately the experiment was interrupted before it was finished, leaving the researchers with 26 subjects with both A and B responses, 15 subjects with only A responses and 17 subjects with only B responses.

a) Calculate empirical estimates of θ_A, θ_B, ρ, σ_A^2, σ_B^2 from the data using the commands `mean`, `cor` and `var`. Use *all* the A responses to get $\hat{\theta}_A$ and $\hat{\sigma}_A^2$, and use *all* the B responses to get $\hat{\theta}_B$ and $\hat{\sigma}_B^2$. Use only the complete data cases to get $\hat{\rho}$.

b) For each person i with only an A response, impute a B response as

$$\hat{y}_{i,B} = \hat{\theta}_B + (y_{i,A} - \hat{\theta}_A)\hat{\rho}\sqrt{\hat{\sigma}_B^2/\hat{\sigma}_A^2}.$$

For each person i with only a B response, impute an A response as

$$\hat{y}_{i,A} = \hat{\theta}_A + (y_{i,B} - \hat{\theta}_B)\hat{\rho}\sqrt{\hat{\sigma}_A^2/\hat{\sigma}_B^2}.$$

You now have two "observations" for each individual. Do a paired sample t-test and obtain a 95% confidence interval for $\theta_A - \theta_B$.

c) Using either Jeffreys' prior or a unit information prior distribution for the parameters, implement a Gibbs sampler that approximates the joint distribution of the parameters and the missing data. Compute a posterior mean for $\theta_A - \theta_B$ as well as a 95% posterior confidence interval for $\theta_A - \theta_B$. Compare these results with the results from b) and discuss.

7.6 Diabetes data: A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age. This information appears in the file `azdiabetes.dat`. Model the joint distribution of these variables for the diabetics and non-diabetics separately, using a multivariate normal distribution:

a) For both groups separately, use the following type of unit information prior, where $\hat{\Sigma}$ is the sample covariance matrix.
   i. $\boldsymbol{\mu}_0 = \bar{\boldsymbol{y}}$, $\Lambda_0 = \hat{\Sigma}$;
   ii. $\mathbf{S}_0 = \hat{\Sigma}$, $\nu_0 = p + 2 = 9$.
   Generate at least 10,000 Monte Carlo samples for $\{\boldsymbol{\theta}_d, \Sigma_d\}$ and $\{\boldsymbol{\theta}_n, \Sigma_n\}$, the model parameters for diabetics and non-diabetics respectively. For each of the seven variables $j \in \{1, \ldots, 7\}$, compare the marginal posterior distributions of $\theta_{d,j}$ and $\theta_{n,j}$. Which variables seem to differ between the two groups? Also obtain $\Pr(\theta_{d,j} > \theta_{n,j}|\mathbf{Y})$ for each $j \in \{1, \ldots, 7\}$.

b) Obtain the posterior means of Σ_d and Σ_n, and plot the entries versus each other. What are the main differences, if any?

Chapter 8

8.1 Components of variance: Consider the hierarchical model where

$$\begin{aligned} \theta_1, \ldots, \theta_m | \mu, \tau^2 &\sim \text{i.i.d. normal}(\mu, \tau^2) \\ y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma^2 &\sim \text{i.i.d. normal}(\theta_j, \sigma^2)\,. \end{aligned}$$

For this problem, we will eventually compute the following:

$\text{Var}[y_{i,j}|\theta_i, \sigma^2]$, $\text{Var}[\bar{y}_{\cdot,j}|\theta_i, \sigma^2]$, $\text{Cov}[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2]$
$\text{Var}[y_{i,j}|\mu, \tau^2]$, $\text{Var}[\bar{y}_{\cdot,j}|\mu, \tau^2]$, $\text{Cov}[y_{i_1,j}, y_{i_2,j}|\mu, \tau^2]$

First, lets use our intuition to guess at the answers:

a) Which do you think is bigger, $\text{Var}[y_{i,j}|\theta_i, \sigma^2]$ or $\text{Var}[y_{i,j}|\mu, \tau^2]$? To guide your intuition, you can interpret the first as the variability of the Y's when sampling from a fixed group, and the second as the variability in first sampling a group, then sampling a unit from within the group.

b) Do you think $\text{Cov}[y_{i_1,j}, y_{i_2,j}|\theta_j, \sigma^2]$ is negative, positive, or zero? Answer the same for $\text{Cov}[y_{i_1,j}, y_{i_2,j}|\mu, \tau]$. You may want to think about what $y_{i_2,j}$ tells you about $y_{i_1,j}$ if θ_j is known, and what it tells you when θ_j is unknown.

c) Now compute each of the six quantities above and compare to your answers in a) and b).

d) Now assume we have a prior $p(\mu)$ for μ. Using Bayes' rule, show that

$$p(\mu|\theta_1, \ldots, \theta_m, \sigma^2, \tau^2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) = p(\mu|\theta_1, \ldots, \theta_m, \tau^2).$$

Interpret in words what this means.

8.2 Sensitivity analysis: In this exercise we will revisit the study from Exercise 5.2, in which 32 students in a science classroom were randomly assigned to one of two study methods, A and B, with $n_A = n_B = 16$. After several weeks of study, students were examined on the course material, and the scores are summarized by $\{\bar{y}_A = 75.2, s_A = 7.3\}$, $\{\bar{y}_B = 77.5, s_b = 8.1\}$. We will estimate $\theta_A = \mu + \delta$ and $\theta_B = \mu - \delta$ using the two-sample model and prior distributions of Section 8.1.

a) Let $\mu \sim \text{normal}(75, 100)$, $1/\sigma^2 \sim \text{gamma}(1, 100)$ and $\delta \sim \text{normal}(\delta_0, \tau_0^2)$. For each combination of $\delta_0 \in \{-4, -2, 0, 2, 4\}$ and $\tau_0^2 \in \{10, 50, 100, 500\}$, obtain the posterior distribution of μ, δ and σ^2 and compute
   i. $\Pr(\delta < 0|\mathbf{Y})$;
   ii. a 95% posterior confidence interval for δ;
   iii. the prior and posterior correlation of θ_A and θ_B.

b) Describe how you might use these results to convey evidence that $\theta_A < \theta_B$ to people of a variety of prior opinions.

8.3 Hierarchical modeling: The files `school1.dat` through `school8.dat` give weekly hours spent on homework for students sampled from eight different schools. Obtain posterior distributions for the true means for the eight different schools using a hierarchical normal model with the following prior parameters:

$$\mu_0 = 7, \gamma_0^2 = 5, \quad \tau_0^2 = 10, \eta_0 = 2, \quad \sigma_0^2 = 15, \nu_0 = 2\,.$$

a) Run a Gibbs sampling algorithm to approximate the posterior distribution of $\{\boldsymbol{\theta}, \sigma^2, \mu, \tau^2\}$. Assess the convergence of the Markov chain, and find the effective sample size for $\{\sigma^2, \mu, \tau^2\}$. Run the chain long enough so that the effective sample sizes are all above 1,000.

b) Compute posterior means and 95% confidence regions for $\{\sigma^2, \mu, \tau^2\}$. Also, compare the posterior densities to the prior densities, and discuss what was learned from the data.

c) Plot the posterior density of $R = \frac{\tau^2}{\sigma^2+\tau^2}$ and compare it to a plot of the prior density of R. Describe the evidence for between-school variation.

d) Obtain the posterior probability that θ_7 is smaller than θ_6, as well as the posterior probability that θ_7 is the smallest of all the θ's.

e) Plot the sample averages $\bar{y}_1, \ldots, \bar{y}_8$ against the posterior expectations of $\theta_1, \ldots, \theta_8$, and describe the relationship. Also compute the sample mean of all observations and compare it to the posterior mean of μ.

Chapter 9

9.1 Extrapolation: The file `swim.dat` contains data on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

a) Perform the following data analysis for each swimmer separately:
   i. Fit a linear regression model of swimming time as the response and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds.
   ii. For each swimmer j, obtain a posterior predictive distribution for Y_j^*, their time if they were to swim two weeks from the last recorded time.

b) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Using your predictive distributions, compute $\Pr(Y_j^* = \max\{Y_1^*, \ldots, Y_4^*\}|\mathbf{Y})$ for each swimmer j, and based on this make a recommendation to the coach.

9.2 Model selection: As described in Example 6 of Chapter 7, The file `azdiabetes.dat` contains data on health-related variables of a population of 532 women. In this exercise we will be modeling the conditional distribution of glucose level (`glu`) as a linear combination of the other variables, excluding the variable `diabetes`.

a) Fit a regression model using the g-prior with $g = n$, $\nu_0 = 2$ and $\sigma_0^2 = 1$. Obtain posterior confidence intervals for all of the parameters.

b) Perform the model selection and averaging procedure described in Section 9.3. Obtain $\Pr(\beta_j \neq 0|\boldsymbol{y})$, as well as posterior confidence intervals for all of the parameters. Compare to the results in part a).

9.3 Crime: The file `crime.dat` contains crime rates and data on 15 explanatory variables for 47 U.S. states, in which both the crime rates and the explanatory variables have been centered and scaled to have variance 1. A description of the variables can be obtained by typing `library(MASS);?UScrime` in R.

a) Fit a regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using the g-prior with $g = n$, $\nu_0 = 2$ and $\sigma_0^2 = 1$. Obtain marginal posterior means and 95% confidence intervals for $\boldsymbol{\beta}$, and compare to the least squares estimates. Describe the relationships between crime and the explanatory variables. Which variables seem strongly predictive of crime rates?

b) Lets see how well regression models can predict crime rates based on the $\mathbf{X}$-variables. Randomly divide the crime roughly in half, into a training set $\{\boldsymbol{y}_{\rm tr}, \mathbf{X}_{\rm tr}\}$ and a test set $\{\boldsymbol{y}_{\rm te}, \mathbf{X}_{\rm te}\}$
   i. Using only the training set, obtain least squares regression coefficients $\hat{\boldsymbol{\beta}}_{\rm ols}$. Obtain predicted values for the test data by computing $\hat{\boldsymbol{y}}_{\rm ols} = \mathbf{X}_{\rm te}\hat{\boldsymbol{\beta}}_{\rm ols}$. Plot $\hat{\boldsymbol{y}}_{\rm ols}$ versus $\boldsymbol{y}_{\rm te}$ and compute the prediction error $\frac{1}{n_{\rm te}}\sum(y_{i,\rm te} - \hat{y}_{i,\rm ols})^2$.
   ii. Now obtain the posterior mean $\hat{\boldsymbol{\beta}}_{\rm Bayes} = \mathrm{E}[\boldsymbol{\beta}|\boldsymbol{y}_{\rm tr}]$ using the g-prior described above and the training data only. Obtain predictions for the test set $\hat{\boldsymbol{y}}_{\rm Bayes} = \mathbf{X}_{\rm test}\hat{\boldsymbol{\beta}}_{\rm Bayes}$. Plot versus the test data, compute the prediction error, and compare to the OLS prediction error. Explain the results.

c) Repeat the procedures in b) many times with different randomly generated test and training sets. Compute the average prediction error for both the OLS and Bayesian methods.

Chapter 10

10.1 Reflecting random walks: It is often useful in MCMC to have a proposal distribution which is both symmetric and has support only on a certain region. For example, if we know $\theta > 0$, we would like our proposal distribution $J(\theta_1|\theta_0)$ to have support on positive θ values. Consider the following proposal algorithm:

- sample $\tilde{\theta} \sim \text{uniform}(\theta_0 - \delta, \theta_0 + \delta)$;
- if $\tilde{\theta} < 0$, set $\theta_1 = -\tilde{\theta}$;
- if $\tilde{\theta} \geq 0$, set $\theta_1 = \tilde{\theta}$.

In other words, $\theta_1 = |\tilde{\theta}|$. Show that the above algorithm draws samples from a symmetric proposal distribution which has support on positive values of θ. It may be helpful to write out the associated proposal density $J(\theta_1|\theta_0)$ under the two conditions $\theta_0 \leq \delta$ and $\theta_0 > \delta$ separately.

10.2 Nesting success: Younger male sparrows may or may not nest during a mating season, perhaps depending on their physical characteristics. Researchers have recorded the nesting success of 43 young male sparrows of the same age, as well as their wingspan, and the data appear in the file `msparrownest.dat`. Let Y_i be the binary indicator that sparrow i successfully nests, and let x_i denote their wingspan. Our model for Y_i is $\text{logit}\Pr(Y_i = 1|\alpha, \beta, x_i) = \alpha + \beta x_i$, where the logit function is given by $\text{logit}\ \theta = \log[\theta/(1-\theta)]$.

a) Write out the joint sampling distribution $\prod_{i=1}^{n} p(y_i|\alpha, \beta, x_i)$ and simplify as much as possible.

b) Formulate a prior probability distribution over α and β by considering the range of $\Pr(Y = 1|\alpha, \beta, x)$ as x ranges over 10 to 15, the approximate range of the observed wingspans.

c) Implement a Metropolis algorithm that approximates $p(\alpha, \beta|\boldsymbol{y}, \boldsymbol{x})$. Adjust the proposal distribution to achieve a reasonable acceptance rate, and run the algorithm long enough so that the effective sample size is at least 1,000 for each parameter.

d) Compare the posterior densities of α and β to their prior densities.

e) Using output from the Metropolis algorithm, come up with a way to make a confidence band for the following *function* $f_{\alpha\beta}(x)$ of wingspan:

$$f_{\alpha\beta}(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

where α and β are the parameters in your sampling model. Make a plot of such a band.

10.3 Tomato plants: The file `tplant.dat` contains data on the heights of ten tomato plants, grown under a variety of soil pH conditions. Each plant was measured twice. During the first measurement, each plant's height was recorded and a reading of soil pH was taken. During the second measurement only plant height was measured, although it is assumed that pH levels did not vary much from measurement to measurement.

a) Using ordinary least squares, fit a linear regression to the data, modeling plant height as a function of time (measurement period) and pH level. Interpret your model parameters.

b) Perform model diagnostics. In particular, carefully analyze the residuals and comment on possible violations of assumptions. In particular, assess (graphically or otherwise) whether or not the residuals within a plant are independent. What parts of your ordinary linear regression model do you think are sensitive to any violations of assumptions you may have detected?

c) Hypothesize a new model for your data which allows for observations within a plant to be correlated. Fit the model using a MCMC approximation to the posterior distribution, and present diagnostics for your approximation.

d) Discuss the results of your data analysis. In particular, discuss similarities and differences between the ordinary linear regression and the model fit with correlated responses. Are the conclusions different?

10.4 Gibbs sampling: Consider the general Gibbs sampler for a vector of parameters $\boldsymbol{\phi}$. Suppose $\boldsymbol{\phi}^{(s)}$ is sampled from the target distribution $p(\boldsymbol{\phi})$ and then $\boldsymbol{\phi}^{(s+1)}$ is generated using the Gibbs sampler by iteratively updating each component of the parameter vector. Show that the marginal probability $\Pr(\boldsymbol{\phi}^{(s+1)} \in A)$ equals the target distribution $\int_A p(\boldsymbol{\phi})\, d\phi$.

10.5 Logistic regression variable selection: Consider a logistic regression model for predicting diabetes as a function of x_1 = number of pregnancies, x_2 = blood pressure, x_3 = body mass index, x_4 = diabetes pedigree and x_5 = age. Using the data in `azdiabetes.dat`, center and scale each of the x-variables by subtracting the sample average and dividing by the sample standard deviation for each variable. Consider a logistic regression model of the form $\Pr(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}) = e^{\theta_i}/(1 + e^{\theta_i})$ where

$$\theta_i = \beta_0 + \beta_1\gamma_1 x_{i,1} + \beta_2\gamma_2 x_{i,2} + \beta_3\gamma_3 x_{i,3} + \beta_4\gamma_4 x_{i,4} + \beta_5\gamma_5 x_{i,5}.$$

In this model, each γ_j is either 0 or 1, indicating whether or not variable j is a predictor of diabetes. For example, if it were the case that $\boldsymbol{\gamma} = (1, 1, 0, 0, 0)$, then $\theta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$. Obtain posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, using independent prior distributions for the parameters, such that $\gamma_j \sim \text{binary}(1/2)$, $\beta_0 \sim \text{normal}(0, 16)$ and $\beta_j \sim \text{normal}(0, 4)$ for each $j > 0$.

a) Implement a Metropolis-Hastings algorithm for approximating the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Examine the sequences $\beta_j^{(s)}$ and $\beta_j^{(s)} \times \gamma_j^{(s)}$ for each j and discuss the mixing of the chain.
b) Approximate the posterior probability of the top five most frequently occurring values of $\boldsymbol{\gamma}$. How good do you think the MCMC estimates of these posterior probabilities are?
c) For each j, plot posterior densities and obtain posterior means for $\beta_j\gamma_j$. Also obtain $\Pr(\gamma_j = 1|\boldsymbol{x}, \boldsymbol{y})$.

Chapter 11

11.1 Full conditionals: Derive formally the full conditional distributions of $\boldsymbol{\theta}, \Sigma, \sigma^2$ and the $\boldsymbol{\beta}_j$'s as given in Section 11.2.

11.2 Randomized block design: Researchers interested in identifying the optimal planting density for a type of perennial grass performed the following randomized experiment: Ten different plots of land were each divided into eight subplots, and planting densities of 2, 4, 6 and 8 plants per square meter were randomly assigned to the subplots, so that there are two subplots at each density in each plot. At the end of the growing season the amount of plant matter yield was recorded in metric tons per hectare. These data appear in the file `pdensity.dat`. The researchers want to fit a model like $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$, where y is yield and x is planting density, but worry that since soil conditions vary across plots they should allow for some across-plot heterogeneity in this relationship. To accommodate this possibility we will analyze these data using the hierarchical linear model described in Section 11.1.

a) Before we do a Bayesian analysis we will get some ad hoc estimates of these parameters via least squares regression. Fit the model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$ using OLS for each group, and make a plot showing the heterogeneity of the least squares regression lines. From the least squares coefficients find ad hoc estimates of $\boldsymbol{\theta}$ and Σ. Also obtain an estimate of σ^2 by combining the information from the residuals across the groups.
b) Now we will perform an analysis of the data using the following distributions as prior distributions:

$$\begin{aligned} \Sigma^{-1} &\sim \text{Wishart}(4, \hat{\Sigma}^{-1}) \\ \boldsymbol{\theta} &\sim \text{multivariate normal}(\hat{\boldsymbol{\theta}}, \hat{\Sigma}) \\ \sigma^2 &\sim \text{inverse} - \text{gamma}(1, \hat{\sigma}^2) \end{aligned}$$

where $\hat{\boldsymbol{\theta}}, \hat{\Sigma}, \hat{\sigma}^2$ are the estimates you obtained in a). Note that this analysis is not combining prior information with information from the data, as the"prior" distribution is based on the observed data.

However, such an analysis can be roughly interpreted as the Bayesian analysis of an individual who has weak but unbiased prior information.

c) Use a Gibbs sampler to approximate posterior expectations of $\boldsymbol{\beta}$ for each group j, and plot the resulting regression lines. Compare to the regression lines in a) above and describe why you see any differences between the two sets of regression lines.

d) From your posterior samples, plot marginal posterior and prior densities of $\boldsymbol{\theta}$ and the elements of Σ. Discuss the evidence that the slopes or intercepts vary across groups.

e) Suppose we want to identify the planting density that maximizes average yield over a random sample of plots. Find the value $x_{\max}$ of x that maximizes expected yield, and provide a 95% posterior predictive interval for the yield of a randomly sampled plot having planting density $x_{\max}$.

11.3 Hierarchical variances:. The researchers in Exercise 11.2 are worried that the plots are not just heterogeneous in their regression lines, but also in their variances. In this exercise we will consider the same hierarchical model as above except that the sampling variability within a group is given by $y_{i,j} \sim \text{normal}(\beta_{1,j} + \beta_{2,j}x_{i,j} + \beta_{3,j}x_{i,j}^2, \sigma_j^2)$, that is, the variances are allowed to differ across groups. As in Section 8.5, we will model $\sigma_1^2, \ldots, \sigma_m^2 \sim$ i.i.d. inverse gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$, with $\sigma_0^2 \sim \text{gamma}(2, 2)$ and $p(\nu_0)$ uniform on the integers $\{1, 2, \ldots, 100\}$.

a) Obtain the full conditional distribution of σ_0^2.

b) Obtain the full conditional distribution of σ_j^2.

c) Obtain the full conditional distribution of $\boldsymbol{\beta}_j$.

d) For two values $\nu_0^{(s)}$ and ν_0^* of ν_0, obtain the ratio $p(\nu_0^*|\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)$ divided by $p(\nu_0^{(s)}|\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)$, and simplify as much as possible.

e) Implement a Metropolis-Hastings algorithm for obtaining the joint posterior distribution of all of the unknown parameters. Plot values of σ_0^2 and ν_0 versus iteration number and describe the mixing of the Markov chain in terms of these parameters.

f) Compare the prior and posterior distributions of ν_0. Comment on any evidence there is that the variances differ across the groups.

11.4 Hierarchical logistic regression: The Washington Assessment of Student Learning (WASL) is a standardized test given to students in the state of Washington. Letting j index the counties within the state of Washington and i index schools within counties, the file `mathstandard.dat` includes data on the following variables:

$y_{i,j}$ = the indicator that more than half the 10th graders in school i, j passed the WASL math exam;

$x_{i,j}$ = the percentage of teachers in school i, j who have a masters degree.

In this exercise we will construct an algorithm to approximate the posterior distribution of the parameters in a generalized linear mixed-effects

model for these data. The model is a mixed effects version of logistic regression:

$y_{i,j} \sim$ binomial$(e^{\gamma_{i,j}}/[1+e^{\gamma_{i,j}}])$, where $\gamma_{i,j} = \beta_{0,j} + \beta_{1,j}x_{i,j}$

$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim$ i.i.d. multivariate normal $(\boldsymbol{\theta}, \Sigma)$, where $\boldsymbol{\beta}_j = (\beta_{0,j}, \beta_{1,j})$

a) The unknown parameters in the model include population-level parameters $\{\boldsymbol{\theta}, \Sigma\}$ and the group-level parameters $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m\}$. Draw a diagram that describes the relationships between these parameters, the data $\{y_{i,j}, x_{i,j}, i = 1 \ldots, n_j, j = 1, \ldots, m\}$, and prior distributions.

b) Before we do a Bayesian analysis, we will get some ad hoc estimates of these parameters via maximum likelihood: Fit a separate logistic regression model for each group, possibly using the `glm` command in R via `beta.j <- glm(y.j~X.j,family=binomial)$coef`. Explain any problems you have with obtaining estimates for each county. Plot $\exp\{\hat{\beta}_{0,j} + \hat{\beta}_{1,j}x\}/(1+\exp\{\hat{\beta}_{0,j} + \hat{\beta}_{1,j}x\})$ as a function of x for each county and describe what you see. Using maximum likelihood estimates only from those counties with 10 or more schools, obtain ad hoc estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\Sigma}$ of $\boldsymbol{\theta}$ and Σ. Note that these estimates may not be representative of patterns from schools with small sample sizes.

c) Formulate a unit information prior distribution for $\boldsymbol{\theta}$ and Σ based on the observed data. Specifically, let $\boldsymbol{\theta} \sim$ multivariate normal$(\hat{\boldsymbol{\theta}}, \hat{\Sigma})$ and let $\Sigma^{-1} \sim$ Wishart$(4, \hat{\Sigma}^{-1})$. Use a Metropolis-Hastings algorithm to approximate the joint posterior distribution of all parameters.

d) Make plots of the samples of $\boldsymbol{\theta}$ and Σ (5 parameters) versus MCMC iteration number. Make sure you run the chain long enough so that your MCMC samples are likely to be a reasonable approximation to the posterior distribution.

e) Obtain posterior expectations of $\boldsymbol{\beta}_j$ for each group j, plot $\mathrm{E}[\beta_{0,j}|\boldsymbol{y}] + \mathrm{E}[\beta_{1,j}|\boldsymbol{y}]x$ as a function of x for each county, compare to the plot in b) and describe why you see any differences between the two sets of regression lines.

f) From your posterior samples, plot marginal posterior and prior densities of $\boldsymbol{\theta}$ and the elements of Σ. Include your ad hoc estimates from b) in the plots. Discuss the evidence that the slopes or intercepts vary across groups.

11.5 Disease rates: The number of occurrences of a rare, nongenetic birth defect in a five-year period for six neighboring counties is $\boldsymbol{y} = (1, 3, 2, 12, 1, 1)$. The counties have populations of $\boldsymbol{x} = (33, 14, 27, 90, 12, 17)$, given in thousands. The second county has higher rates of toxic chemicals (PCBs) present in soil samples, and it is of interest to know if this town has a high disease rate as well. We will use the following hierarchical model to analyze these data:

- $Y_i|\theta_i, x_i \sim$ Poisson$(\theta_i x_i)$;
- $\theta_1, \ldots, \theta_6|a, b \sim$ gamma(a, b);
- $a \sim$ gamma(1,1) ; $b \sim$ gamma(10,1).

a) Describe in words what the various components of the hierarchical model represent in terms of observed and expected disease rates.
b) Identify the form of the conditional distribution of $p(\theta_1, \ldots, \theta_6 | a, b, \boldsymbol{x}, \boldsymbol{y})$, and from this identify the full conditional distribution of the rate for each county $p(\theta_i | \boldsymbol{\theta}_{-i}, a, b, \boldsymbol{x}, \boldsymbol{y})$.
c) Write out the ratio of the posterior densities comparing a set of proposal values $(a^*, b^*, \boldsymbol{\theta})$ to values $(a, b, \boldsymbol{\theta})$. Note the value of $\boldsymbol{\theta}$, the vector of county-specific rates, is unchanged.
d) Construct a Metropolis-Hastings algorithm which generates samples of $(a, b, \boldsymbol{\theta})$ from the posterior. Do this by iterating the following steps:
   1. Given a current value $(a, b, \boldsymbol{\theta})$, generate a proposal $(a^*, b^*, \boldsymbol{\theta})$ by sampling a^* and b^* from a symmetric proposal distribution centered around a and b, but making sure all proposals are positive (see Exercise 10.1). Accept the proposal with the appropriate probability.
   2. Sample new values of the θ_j's from their full conditional distributions.

   Perform diagnostic tests on your chain and modify if necessary.
e) Make posterior inference on the infection rates using the samples from the Markov chain. In particular,
   i. Compute marginal posterior distributions of $\theta_1, \ldots, \theta_6$ and compare them to $y_1/x_1, \ldots y_6/x_6$.
   ii. Examine the posterior distribution of a/b, and compare it to the corresponding prior distribution as well as to the average of y_i/x_i across the six counties.
   iii. Plot samples of θ_2 versus θ_j for each $j \neq 2$, and draw a 45 degree line on the plot as well. Also estimate $\Pr(\theta_2 > \theta_j | \boldsymbol{x}, \boldsymbol{y})$ for each j and $\Pr(\theta_2 = \max\{\theta_1, \ldots, \theta_6\} | \boldsymbol{x}, \boldsymbol{y})$. Interpret the results of these calculations, and compare them to the conclusions one might obtain if they just examined y_j/x_j for each county j.

Chapter 12

12.1 Rank regression: The 1996 General Social Survey gathered a wide variety of information on the adult U.S. population, including each survey respondent's sex, their self-reported frequency of religious prayer (on a six-level ordinal scale), and the number of items correct out of 10 on a short vocabulary test. These data appear in the file `prayer.dat`. Using the rank regression procedure described in Section 12.1.2, estimate the parameters in a regression model for Y_i= prayer as a function of $x_{i,1}$ = sex of respondent (0-1 indicator of being female) and $x_{i,2}$ = vocabulary score, as well as their interaction $x_{i,3} = x_{i,1} \times x_{i,2}$. Compare marginal prior distributions of the three regression parameters to their posterior

distributions, and comment on the evidence that the relationship between prayer and score differs across the sexes.

12.2 Copula modeling: The file `azdiabetes_alldata.dat` contains data on eight variables for 632 women in a study on diabetes (see Exercise 7.6 for a description of the variables). Data on subjects labeled 201-300 have missing values for some variables, mostly for the skin fold thickness measurement.

a) Using only the data from subjects 1-200, implement the Gaussian copula model for the eight variables in this dataset. Obtain posterior means and 95% posterior confidence intervals for all $\binom{8}{2} = 28$ parameters.

b) Now use the data from subjects 1-300, thus including data from subjects who are missing some variables. Implement the Gaussian copula model and obtain posterior means and 95% posterior confidence intervals for all parameters. How do the results differ from those in a)?

12.3 Constrained normal: Let $p(z) \propto \text{dnorm}(z, \theta, \sigma) \times \delta_{(a,b)}(z)$, the normal density constrained to the interval (a, b). Prove that the inverse-cdf method outlined in Section 12.1.1 generates a sample from this distribution.

12.4 Categorical data and the Dirichlet distribution: Consider again the data on the number of children of men in their 30s from Exercise 4.8. These data could be considered as categorical data, as each sample Y lies in the discrete set $\{1, \ldots, 8\}$ (8 here actually denotes "8 or more" children). Let $\boldsymbol{\theta}_A = (\theta_{A,1}, \ldots, \theta_{A,8})$ be the proportion in each of the eight categories from the population of men with bachelor's degrees, and let the vector $\boldsymbol{\theta}_B$ be defined similarly for the population of men without bachelor's degrees.

a) Write in a compact form the conditional probability given $\boldsymbol{\theta}_A$ of observing a particular sequence $\{y_{A,1}, \ldots, y_{A,n_1}\}$ for a random sample from the A population.

b) Identify the sufficient statistic. Show that the Dirichlet family of distributions, with densities of the form $p(\boldsymbol{\theta}|\boldsymbol{a}) \propto \theta_1^{a_1-1} \times \cdots \theta_K^{a_K-1}$, are a conjugate class of prior distributions for this sampling model.

c) The function `rdir()` below samples from the Dirichlet distribution:

```
rdir<-function(nsamp=1,a)  # a is a vector
{
  Z<-matrix( rgamma(length(a)*nsamp,a,1),
             nsamp,length(a),byrow=T)
  Z/apply(Z,1,sum)
}
```

Using this function, generate 5,000 or more samples of $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ from their posterior distributions. Using a Monte Carlo approximation, obtain and plot the posterior distributions of $\text{E}[Y_A|\boldsymbol{\theta}_A]$ and $\text{E}[Y_B|\boldsymbol{\theta}_B]$, as well as of $\tilde{Y}_A$ and $\tilde{Y}_B$.

d) Compare the results above to those in Exercise 4.8. Perform the goodness of fit test from that exercise on this model, and compare to the fit of the Poisson model.

Common distributions

The binomial distribution

A random variable $X \in \{0, 1, \ldots, n\}$ has a binomial(n, θ) distribution if $\theta \in [0, 1]$ and

$$\Pr(X = x|\theta, n) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } x \in \{0, 1, \ldots, n\}.$$

For this distribution,

$$\begin{aligned} \mathrm{E}[X|\theta] &= n\theta, \\ \mathrm{Var}[X|\theta] &= n\theta(1-\theta), \\ \mathrm{mode}[X|\theta] &= \lfloor (n+1)\theta \rfloor, \\ p(x|\theta, n) &= \texttt{dbinom(x,n,theta)}\,. \end{aligned}$$

If $X_1 \sim$ binomial(n_1, θ) and $X_2 \sim$ binomial(n_2, θ) are independent, then $X = X_1 + X_2 \sim$ binomial$(n_1 + n_2, \theta)$. When $n = 1$ this distribution is called the *binary* or *Bernoulli* distribution. The binomial(n, θ) model assumes that X is (equal in distribution to) a sum of independent binary(θ) random variables.

The beta distribution

A random variable $X \in [0, 1]$ has a beta(a, b) distribution if $a > 0$, $b > 0$ and

$$p(x|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \quad \text{for } 0 \le x \le 1.$$

For this distribution,

$$\begin{aligned}
\mathrm{E}[X|a,b] &= \frac{a}{a+b}, \\
\mathrm{Var}[X|a,b] &= \frac{ab}{(a+b+1)(a+b)^2} = \mathrm{E}[X] \times \mathrm{E}[1-X] \times \frac{1}{a+b+1}, \\
\mathrm{mode}[X|a,b] &= \frac{a-1}{(a-1)+(b-1)} \text{ if } a>1 \text{ and } b>1, \\
p(x|a,b) &= \texttt{dbeta(x,a,b)}\,.
\end{aligned}$$

The beta distribution is closely related to the gamma distribution. See the paragraph on the gamma distribution below for details. A multivariate version of the beta distribution is the Dirichlet distribution, described in Exercise 12.4.

The Poisson distribution

A random variable $X \in \{0, 1, 2, \ldots\}$ has a Poisson(θ) distribution if $\theta > 0$ and

$$\Pr(X = x|\theta) = \theta^x e^{-\theta}/x! \quad \text{for } x \in \{0, 1, 2, \ldots\}.$$

For this distribution,

$$\begin{aligned}
\mathrm{E}[X|\theta] &= \theta, \\
\mathrm{Var}[X|\theta] &= \theta, \\
\mathrm{mode}[X|\theta] &= \lfloor \theta \rfloor, \\
p(x|\theta) &= \texttt{dpois(x,theta)}\,.
\end{aligned}$$

If $X_1 \sim \mathrm{Poisson}(\theta_1)$ and $X_2 \sim \mathrm{Poisson}(\theta_2)$ are independent, then $X_1 + X_2 \sim \mathrm{Poisson}(\theta_1+\theta_2)$. The Poisson family has a "mean-variance relationship," which describes the fact that $\mathrm{E}[X|\theta] = \mathrm{Var}[X|\theta] = \theta$. If it is observed that a sample mean is very different than the sample variance, then the Poisson model may not be appropriate. If the variance is larger than the sample mean, then a negative binomial model (Section 3.2.1) might be a better fit.

The gamma and inverse-gamma distributions

A random variable $X \in (0, \infty)$ has a gamma(a, b) distribution if $a > 0$, $b > 0$ and

$$p(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \quad \text{for } x > 0.$$

For this distribution,

$$\begin{aligned}
\mathrm{E}[X|a,b] &= a/b, \\
\mathrm{Var}[X|a,b] &= a/b^2, \\
\mathrm{mode}[X|a,b] &= (a-1)/b \text{ if } a \geq 1,\ 0 \text{ if } 0 < a < 1\ , \\
p(x|a,b) &= \ \texttt{dgamma(x,a,b)}\ .
\end{aligned}$$

If $X_1 \sim \text{gamma}(a_1, b)$ and $X_1 \sim \text{gamma}(a_2, b)$ are independent, then $X_1 + X_2 \sim \text{gamma}(a_1+a_2, b)$ and $X_1/(X_1+X_2) \sim \text{beta}(a_1, a_2)$. If $X \sim \text{normal}(0, \sigma^2)$ then $X^2 \sim \text{gamma}(1/2, 1/[2\sigma^2])$. The chi-square distribution with ν degrees of freedom is the same as a $\text{gamma}(\nu/2, 1/2)$ distribution.

A random variable $X \in (0, \infty)$ has an inverse-gamma(a, b) distribution if $1/X$ has a gamma(a, b) distribution. In other words, if $Y \sim \text{gamma}(a, b)$ and $X = 1/Y$, then $X \sim \text{inverse-gamma}(a, b)$. The density of X is

$$p(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x} \quad \text{for } x > 0.$$

For this distribution,

$$\begin{aligned}
\mathrm{E}[X|a,b] &= b/(a-1) \text{ if } a \geq 1,\ \infty \text{ if } 0 < a < 1, \\
\mathrm{Var}[X|a,b] &= b^2/[(a-1)^2(a-2)] \text{ if } a \geq 2,\ \infty \text{ if } 0 < a < 2, \\
\mathrm{mode}[X|a,b] &= b/(a+1).
\end{aligned}$$

Note that the inverse-gamma density is not simply the gamma density with x replaced by $1/x$: There is an additional factor of x^{-2} due to the Jacobian in the change-of-variables formula (see Exercise 10.3).

The univariate normal distribution

A random variable $X \in \mathbb{R}$ has a normal(θ, σ^2) distribution if $\sigma^2 > 0$ and

$$p(x|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\theta)^2/\sigma^2} \quad \text{for } -\infty < x < \infty.$$

For this distribution,

$$\begin{aligned}
\mathrm{E}[X|\theta,\sigma^2] &= \theta, \\
\mathrm{Var}[X|\theta,\sigma^2] &= \sigma^2, \\
\mathrm{mode}[X|\theta,\sigma^2] &= \theta, \\
p(x|\theta,\sigma^2) &= \ \texttt{dnorm(x,theta,sigma)}\ .
\end{aligned}$$

Remember that R parameterizes things in terms of the standard deviation σ, and not the variance σ^2. If $X_1 \sim \text{normal}(\theta_1, \sigma_1^2)$ and $X_2 \sim \text{normal}(\theta_2, \sigma_2^2)$ are independent, then $aX_1 + bX_2 + c \sim \text{normal}(a\theta_1 + b\theta_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$.

A normal sampling model is often useful even if the underlying population does not have a normal distribution. This is because statistical procedures that assume a normal model will generally provide good estimates of the population mean and variance, regardless of whether or not the population is normal (see Section 5.5 for a discussion).

The multivariate normal distribution

A random vector $\boldsymbol{X} \in \mathbb{R}^p$ has a multivariate normal$(\boldsymbol{\theta}, \Sigma)$ distribution if Σ is a positive definite $p \times p$ matrix and

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\right\} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^p.$$

For this distribution,

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}, \Sigma] &= \boldsymbol{\theta}, \\ \mathrm{Var}[\boldsymbol{X}|\boldsymbol{\theta}, \Sigma] &= \Sigma, \\ \mathrm{mode}[\boldsymbol{X}|\boldsymbol{\theta}, \Sigma] &= \boldsymbol{\theta}. \end{aligned}$$

Just like the univariate normal distribution, if $\boldsymbol{X}_1 \sim \text{normal}(\boldsymbol{\theta}_1, \Sigma_1)$ and $\boldsymbol{X}_2 \sim \text{normal}(\boldsymbol{\theta}_2, \Sigma_2)$ are independent, then $a\boldsymbol{X}_1 + b\boldsymbol{X}_2 + \boldsymbol{c} \sim \text{normal}(a\boldsymbol{\theta}_1 + b\boldsymbol{\theta}_2 + \boldsymbol{c}, a^2\Sigma_1 + b^2\Sigma_2)$. Marginal and conditional distributions of subvectors of $\boldsymbol{X}$ also have multivariate normal distributions: Let $\boldsymbol{a} \subset \{1, \ldots, p\}$ be a subset of variable indices, and let $\boldsymbol{b} = \boldsymbol{a}^c$ be the remaining indices. Then $\boldsymbol{X}_{[a]} \sim$ multivariate normal$(\boldsymbol{\theta}_{[a]}, \Sigma_{[a,a]})$ and $\{\boldsymbol{X}_{[b]}|\boldsymbol{X}_{[a]}\} \sim$ multivariate normal$(\boldsymbol{\theta}_{b|a}, \Sigma_{b|a})$, where

$$\begin{aligned} \boldsymbol{\theta}_{b|a} &= \boldsymbol{\theta}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\boldsymbol{X}_{[a]} - \boldsymbol{\theta}_{[a]}) \\ \Sigma_{b|a} &= \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}. \end{aligned}$$

Simulating a multivariate normal random variable can be achieved by a linear transformation of a vector of i.i.d. standard normal random variables. If $\boldsymbol{Z}$ is the vector with elements $Z_1, \ldots, Z_p \sim$ i.i.d. normal$(0, 1)$ and $\mathbf{A}\mathbf{A}^T = \Sigma$, then $\boldsymbol{X} = \boldsymbol{\theta} + \mathbf{A}\boldsymbol{Z} \sim$ multivariate normal$(\boldsymbol{\theta}, \Sigma)$. Usually $\mathbf{A}$ is the Choleski factorization of Σ. The following R-code will generate an $n \times p$ matrix such that the rows are i.i.d. samples from a multivariate normal distribution:

```
Z<-matrix(rnorm(n*p),nrow=n,ncol=p)
X<-t(   t(Z%*%chol(Sigma)) + c(theta)  )
```

The Wishart and inverse-Wishart distributions

A random $p \times p$ symmetric positive definite matrix $\mathbf{X}$ has a Wishart$(\nu, \mathbf{M})$ distribution if the integer $\nu \geq p$, $\mathbf{M}$ is a $p \times p$ symmetric positive definite matrix and

$$p(\mathbf{X}|\nu, \mathbf{M}) = [2^{\nu p/2} \Gamma_p(\nu/2) |\mathbf{M}|^{\nu/2}]^{-1} \times |\mathbf{X}|^{(\nu-p-1)/2} \text{etr}(-\mathbf{M}^{-1}\mathbf{X}/2),$$

where

$\Gamma_p(\nu/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma[(\nu + 1 - j)/2]$, and
$\text{etr}(\mathbf{A}) = \exp(\sum a_{j,j})$, the exponent of the sum of the diagonal elements.

For this distribution,

$$\begin{aligned} \text{E}[\mathbf{X}|\nu, \mathbf{M}] &= \nu\mathbf{M}, \\ \text{Var}[X_{i,j}|\nu, \mathbf{M}] &= \nu \times (m_{i,j}^2 + m_{i,i} m_{j,j}), \\ \text{mode}[\mathbf{X}|\nu, \mathbf{M}] &= (\nu - p - 1)\mathbf{M}. \end{aligned}$$

The Wishart distribution is a multivariate version of the gamma distribution. Just as the sum of squares of i.i.d. univariate normal variables has a gamma distribution, the sums of squares of i.i.d. multivariate normal vectors has a Wishart distribution. Specifically, if $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_\nu \sim$ i.i.d. multivariate normal$(\mathbf{0}, \mathbf{M})$, then $\sum \boldsymbol{Y}_i \boldsymbol{Y}_i^T \sim$ Wishart$(\nu, \mathbf{M})$. This relationship can be used to generate a Wishart-distributed random matrix:

```
Z<-matrix(rnorm(nu*p),nrow=nu,ncol=p)  # standard normal
Y<-Z%*%chol(M)                         # rows have cov=M
X<-t(Y)%*%Y                            # Wishart matrix
```

A random $p \times p$ symmetric positive definite matrix $\mathbf{X}$ has an inverse-Wishart$(\nu, \mathbf{M})$ distribution if $\mathbf{X}^{-1}$ has a Wishart$(\nu, \mathbf{M})$ distribution. In other words, if $\mathbf{Y} \sim$ Wishart$(\nu, \mathbf{M})$ and $\mathbf{X} = \mathbf{Y}^{-1}$, then $\mathbf{X} \sim$ inverse-Wishart$(\nu, \mathbf{M})$. The density of $\mathbf{X}$ is

$$p(\mathbf{X}|\nu, \mathbf{M}) = [2^{\nu p/2} \Gamma_p(\nu/2) |\mathbf{M}|^{\nu/2}]^{-1} \times |\mathbf{X}|^{-(\nu+p+1)/2} \text{etr}(-\mathbf{M}^{-1}\mathbf{X}^{-1}/2).$$

For this distribution,

$$\begin{aligned} \text{E}[\mathbf{X}|\nu, \mathbf{M}] &= (\nu - p - 1)^{-1}\mathbf{M}^{-1}, \\ \text{mode}[\mathbf{X}|\nu, \mathbf{M}] &= (\nu + p + 1)^{-1}\mathbf{M}^{-1}. \end{aligned}$$

The second moments (i.e. the variances) of the elements of $\mathbf{X}$ are given in Press (1972). Since we often use the inverse-Wishart distribution as a prior distribution for a covariance matrix Σ, it is sometimes useful to parameterize the distribution in terms of $\mathbf{S} = \mathbf{M}^{-1}$. Then if $\Sigma \sim$ inverse-Wishart$(\nu, \mathbf{S}^{-1})$, we have mode$[\mathbf{X}|\nu, \mathbf{S}] = (\nu + p + 1)^{-1}\mathbf{S}$. If Σ_0 were the most probable value

of Σ *a priori*, then we would set $\mathbf{S} = (\nu_0 + p + 1)\Sigma_0$, so that $\Sigma \sim$ inverse-Wishart$(\nu, [(\nu + p - 1)\Sigma_0]^{-1})$ and mode$[\Sigma|\nu, \mathbf{S}] = \Sigma_0$.

For more on the Wishart distribution and its relationship to the multivariate normal distribution, see Press (1972) or Mardia et al (1979).

References

Agresti A, Coull BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. Amer Statist 52(2):119–126

Aldous DJ (1985) Exchangeability and related topics. In: École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Math., vol 1117, Springer, Berlin, pp 1–198

Arcese P, Smith JNM, Hochachka WM, Rogers CM, Ludwig D (1992) Stability, regulation, and the determination of abundance in an insular song sparrow population. Ecology 73(3):805–822

Atchadé YF, Rosenthal JS (2005) On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11(5):815–828

Bayarri MJ, Berger JO (2000) p values for composite null models. J Amer Statist Assoc 95(452):1127–1142, 1157–1170, with comments and a rejoinder by the authors

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. R Soc Lond Philos Trans 5(3):370–418

Berger JO (1980) Statistical decision theory: foundations, concepts, and methods. Springer-Verlag, New York, springer Series in Statistics

Berk KN (1978) Comparing subset regression procedures. Technometrics 20:1–6

Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Ltd., Chichester

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J Roy Statist Soc Ser B 36:192–236, with discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author

Bickel PJ, Ritov Y (1997) Local asymptotic normality of ranks and covariates in transformation models. In: Festschrift for Lucien Le Cam, Springer, New York, pp 43–54

Box GEP, Draper NR (1987) Empirical model-building and response surfaces. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York

Bunke O, Milhaud X (1998) Asymptotic behavior of Bayes estimates under possibly incorrect models. Ann Statist 26(2):617–644

Carlin BP, Louis TA (1996) Bayes and empirical Bayes methods for data analysis, Monographs on Statistics and Applied Probability, vol 69. Chapman & Hall, London

Casella G (1985) An introduction to empirical Bayes data analysis. Amer Statist 39(2):83–87

Chib S, Winkelmann R (2001) Markov chain Monte Carlo analysis of correlated count data. J Bus Econom Statist 19(4):428–435

Congdon P (2003) Applied Bayesian modelling. Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester

Cox RT (1946) Probability, frequency and reasonable expectation. Amer J Phys 14:1–13

Cox RT (1961) The algebra of probable inference. The Johns Hopkins Press, Baltimore, Md

Dawid AP, Lauritzen SL (1993) Hyper-Markov laws in the statistical analysis of decomposable graphical models. Ann Statist 21(3):1272–1317

Diaconis P (1988) Recent progress on de Finetti's notions of exchangeability. In: Bayesian statistics, 3 (Valencia, 1987), Oxford Sci. Publ., Oxford Univ. Press, New York, pp 111–125

Diaconis P, Freedman D (1980) Finite exchangeable sequences. Ann Probab 8(4):745–764

Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. Ann Statist 7(2):269–281

Diaconis P, Ylvisaker D (1985) Quantifying prior opinion. In: Bayesian statistics, 2 (Valencia, 1983), North-Holland, Amsterdam, pp 133–156, with discussion and a reply by Diaconis

Dunson DB (2000) Bayesian latent variable models for clustered mixed outcomes. J R Stat Soc Ser B Stat Methodol 62(2):355–366

Efron B (2005) Bayesians, frequentists, and scientists. J Amer Statist Assoc 100(469):1–5

Efron B, Morris C (1973) Stein's estimation rule and its competitors—an empirical Bayes approach. J Amer Statist Assoc 68:117–130

de Finetti B (1931) Funzione caratteristica di un fenomeno aleatorio. Atti della R Academia Nazionale dei Lincei, Serie 6 Memorie, Classe di Scienze Fisiche, Mathematice e Naturale 4:251–299

de Finetti B (1937) La prévision : ses lois logiques, ses sources subjectives. Ann Inst H Poincaré 7(1):1–68

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Amer Statist Assoc 85(410):398–409

Gelfand AE, Sahu SK, Carlin BP (1995) Efficient parameterisations for normal linear mixed models. Biometrika 82(3):479–488

Gelman A, Hill J (2007) Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences (Disc: P483-501, 503-511). Statistical Science 7:457–472

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statist Sinica 6(4):733–807, with comments and a rejoinder by the authors

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton, FL

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721–741

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88:881–889

Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (Disc: P189-193). In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting, Clarendon Press [Oxford University Press], pp 169–188

Geyer CJ (1992) Practical Markov chain Monte Carlo (Disc: P483-503). Statistical Science 7:473–483

Gilks WR, Roberts GO, Sahu SK (1998) Adaptive Markov chain Monte Carlo through regeneration. J Amer Statist Assoc 93(443):1045–1054

Grogan W, Wirth W (1981) A new American genus of predaceous midges related to Palpomyia and Bezzia (Diptera: Ceratopogonidae). Proceedings of the Biological Society of Washington 94:1279–1305

Guttman I (1967) The use of the concept of a future observation in goodness-of-fit problems. J Roy Statist Soc Ser B 29:83–100

Haario H, Saksman E, Tamminen J (2001) An adaptive metropolis algorithm. Bernoulli 7(2):223–242

Haigis KM, Hoff PD, White A, Shoemaker AR, Halberg RB, Dove WF (2004) Tumor regionality in the mouse intestine reflects the mechanism of loss of apc function. PNAS 101(26):9769–9773

Hartigan JA (1966) Note on the confidence-prior of Welch and Peers. J Roy Statist Soc Ser B 28:55–56

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Hewitt E, Savage LJ (1955) Symmetric measures on Cartesian products. Trans Amer Math Soc 80:470–501

Hoff PD (2007) Extending the rank likelihood for semiparametric copula estimation. Ann Appl Statist 1(1):265–283

Jaynes ET (2003) Probability theory. Cambridge University Press, Cambridge, the logic of science, Edited and with a foreword by G. Larry Bretthorst

Jeffreys H (1961) Theory of probability. Third edition, Clarendon Press, Oxford

Johnson VE (2007) Bayesian model assessment using pivotal quantities. Bayesian Anal 2(4):719–733

Jordan MIe (1998) Learning in Graphical Models. Kluwer Academic Publishers Group

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90:773–795

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Amer Statist Assoc 90(431):928–934

Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules (Corr: 1998V93 p412). Journal of the American Statistical Association 91:1343–1370

Kelley TL (1927) The Interpretation of Educational Measurement. World Book Company, New York

Key JT, Pericchi LR, Smith AFM (1999) Bayesian model choice: what and why? In: Bayesian statistics, 6 (Alcoceber, 1998), Oxford Univ. Press, New York, pp 343–370

Kleijn BJK, van der Vaart AW (2006) Misspecification in infinite-dimensional Bayesian statistics. Ann Statist 34(2):837–877

Kuehl RO (2000) Design of Experiments: Statistical Principles of Research Design and Analysis. Duxbury Press

Laplace PS (1995) A philosophical essay on probabilities, english edn. Dover Publications Inc., New York, translated from the sixth French edition by Frederick William Truscott and Frederick Lincoln Emory, With an introductory note by E. T. Bell

Lauritzen SL (1996) Graphical models, Oxford Statistical Science Series, vol 17. The Clarendon Press Oxford University Press, New York, oxford Science Publications

Lawrence E, Bingham D, Liu C, Nair V (2008) Bayesian inference for multivariate ordinal data using parameter expansion. Technometrics 50(2):182–191

Leamer EE (1978) Specification searches. Wiley-Interscience [John Wiley & Sons], New York, ad hoc inference with nonexperimental data, Wiley Series in Probability and Mathematical Statistics

Letac G, Massam H (2007) Wishart distributions for decomposable graphs. Ann Statist 35(3):1278–1323

Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g Priors for Bayesian Variable Selection. Journal of the American Statistical Association 103(481):410–423

Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. J Roy Statist Soc Ser B 34:1–41, with discussions by J. A. Nelder, V. D. Barnett, C. A. B. Smith, T. Leonard, M. R. Novick, D. R. Cox, R. L. Plackett, P. Sprent, J. B. Copas, D. V. Hinkley, E. F. Harding, A. P. Dawid, C.

Chatfield, S. E. Fienberg, B. M. Hill, R. Thompson, B. de Finetti, and O. Kempthorne

Little RJ (2006) Calibrated Bayes: a Bayes/frequentist roadmap. Amer Statist 60(3):213–223

Liu JS, Wu YN (1999) Parameter expansion for data augmentation. J Amer Statist Assoc 94(448):1264–1274

Logan JA (1983) A multivariate model for mobility tables. The American Journal of Sociology 89(2):324–349

Lukacs E (1942) A characterization of the normal distribution. Ann Math Statistics 13:91–93

Madigan D, Raftery AE (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 89:1535–1546

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press [Harcourt Brace Jovanovich Publishers], London, probability and Mathematical Statistics: A Series of Monographs and Textbooks

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. Journal of Chemical Physics 21(6):1087–1092

Monahan JF, Boos DD (1992) Proper likelihoods for Bayesian analysis. Biometrika 79(2):271–278

Moore KA, Waite LJ (1977) Early childbearing and educational attainment. Family Planning Perspectives 9(5):220–225

Papaspiliopoulos O, Roberts GO, Sköld M (2007) A general framework for the parametrization of hierarchical models. Statist Sci 22(1):59–73

Petit J, , Jouzel J, Raynaud D, Barkov N, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delayque G, Delmotte M, Kotlyakov V, Legrand M, Lipenkov V, Lorius C, Pepin L, Ritz C, Saltzman E, Stievenard M (1999) Climate and atmospheric history of the past 420,000 years from the vostok ice core, antarctica. Nature 399:429–436

Pettitt AN (1982) Inference for the linear model using a likelihood based on ranks. J Roy Statist Soc Ser B 44(2):234–243

Pitt M, Chan D, Kohn R (2006) Efficient Bayesian inference for Gaussian copula regression models. Biometrika 93(3):537–554

Press SJ (1972) Applied multivariate analysis. Holt, Rinehart and Winston, Inc., New York, series in Quantitative Methods for Decision-Making, International Series in Decision Processes

Press SJ (1982) Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. Krieger Publishing Company, Inc.

Quinn K (2004) Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. Political Analysis 12(4):338–353

Raftery AE, Lewis SM (1992) How many iterations in the Gibbs sampler? In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting, Clarendon Press [Oxford University Press], pp 763–773

Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. J Amer Statist Assoc 92(437):179–191

Raiffa H, Schlaifer R (1961) Applied statistical decision theory. Studies in Managerial Economics, Division of Research, Graduate School of Business Administration, Harvard University, Boston, Mass.

Rao JNK (1958) A characterization of the normal distribution. Ann Math Statist 29:914–919

Ripley BD (1979) [Algorithm AS 137] Simulating spatial patterns: Dependent samples from a multivariate density. Applied Statistics 28:109–112

Robert C, Casella G (2008) A history of markov chain monte carlo–subjective recollections from incomplete data `arXiv:0808.2902 [stat.CO]`, `arxiv:0808.2902`

Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer Texts in Statistics, Springer-Verlag, New York

Roberts GO, Rosenthal JS (2007) Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J Appl Probab 44(2):458–475

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Statist 12(4):1151–1172

Rubinstein RY, Kroese DP (2008) Simulation and the Monte Carlo method, 2nd edn. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ

Savage LJ (1954) The foundations of statistics. John Wiley & Sons Inc., New York

Savage LJ (1962) The foundations of statistical inference. Methuen & Co. Ltd., London

Savage LJ (1972) The foundations of statistics, revised edn. Dover Publications Inc., New York

Severini TA (1991) On the relationship between Bayesian and non-Bayesian interval estimates. J Roy Statist Soc Ser B 53(3):611–618

Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS (1988) Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: Greenes RA (ed) Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988), IEEE Computer Society Press, pp 261–265

Stein C (1955) A necessary and sufficient condition for admissibility. Ann Math Statist 26:518–522

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I, University of California Press, Berkeley and Los Angeles, pp 197–206

Stein CM (1981) Estimation of the mean of a multivariate normal distribution. Ann Statist 9(6):1135–1151

Sweeting TJ (1999) On the construction of Bayes-confidence regions. J R Stat Soc Ser B Stat Methodol 61(4):849–861

Sweeting TJ (2001) Coverage probability bias, objective Bayes and the likelihood principle. Biometrika 88(3):657–675

Tibshirani R (1989) Noninformative priors for one parameter of many. Biometrika 76(3):604–608

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B 58(1):267–288

Welch BL, Peers HW (1963) On formulae for confidence points based on integrals of weighted likelihoods. J Roy Statist Soc Ser B 25:318–329

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50(1):1–25

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian inference and decision techniques, Stud. Bayesian Econometrics Statist., vol 6, North-Holland, Amsterdam, pp 233–243

Index