24
Simulation Methods

In this chapter we will show how simulation can be used to approximate integrals. Our leading example is the problem of computing integrals in Bayesian inference but the techniques are widely applicable. We will look at three integration methods: (i) basic Monte Carlo integration, (ii) importance sampling, and (iii) Markov chain Monte Carlo (MCMC).

24.1 Bayesian Inference Revisited

Simulation methods are especially useful in Bayesian inference so let us briefly review the main ideas in Bayesian inference. See Chapter 11 for more details.

Given a prior $f(\theta)$ and data $X^n = (X_1, \ldots, X_n)$ the posterior density is

$$f(\theta|X^n) = \frac{\mathcal{L}(\theta)f(\theta)}{c}$$

where $\mathcal{L}(\theta)$ is the likelihood function and

$$c = \int \mathcal{L}(\theta)f(\theta)\, d\theta$$

is the **normalizing constant**. The posterior mean is

$$\overline{\theta} = \int \theta f(\theta|X^n)d\theta = \frac{\int \theta\mathcal{L}(\theta)f(\theta)d\theta}{c}.$$

L. Wasserman, *All of Statistics*

If $\theta = (\theta_1, \ldots, \theta_k)$ is multidimensional, then we might be interested in the posterior for one of the components, θ_1, say. This marginal posterior density is

$$f(\theta_1|X^n) = \int \int \cdots \int f(\theta_1, \ldots, \theta_k|X^n) d\theta_2 \cdots d\theta_k$$

which involves high-dimensional integration.

When θ is high-dimensional, it may not be feasible to calculate these integrals analytically. Simulation methods will often be helpful.

24.2 Basic Monte Carlo Integration

Suppose we want to evaluate the integral

$$I = \int_a^b h(x)\, dx$$

for some function h. If h is an "easy" function like a polynomial or trigonometric function, then we can do the integral in closed form. If h is complicated there may be no known closed form expression for I. There are many numerical techniques for evaluating I such as Simpson's rule, the trapezoidal rule and Gaussian quadrature. Monte Carlo integration is another approach for approximating I which is notable for its simplicity, generality and scalability.

Let us begin by writing

$$I = \int_a^b h(x)dx = \int_a^b w(x)f(x)dx \tag{24.1}$$

where $w(x) = h(x)(b-a)$ and $f(x) = 1/(b-a)$. Notice that f is the probability density for a uniform random variable over (a, b). Hence,

$$I = \mathbb{E}_f(w(X))$$

where $X \sim \text{Unif}(a, b)$. If we generate $X_1, \ldots, X_N \sim \text{Unif}(a, b)$, then by the law of large numbers

$$\widehat{I} \equiv \frac{1}{N} \sum_{i=1}^{N} w(X_i) \xrightarrow{\text{P}} \mathbb{E}(w(X)) = I. \tag{24.2}$$

This is the basic **Monte Carlo integration method**. We can also compute the standard error of the estimate

$$\widehat{\mathsf{se}} = \frac{s}{\sqrt{N}}$$

where

$$s^2 = \frac{\sum_{i=1}^N (Y_i - \widehat{I})^2}{N-1}$$

where $Y_i = w(X_i)$. A $1-\alpha$ confidence interval for I is $\widehat{I} \pm z_{\alpha/2}\widehat{\mathsf{se}}$. We can take N as large as we want and hence make the length of the confidence interval very small.

24.1 Example. Let $h(x) = x^3$. Then, $I = \int_0^1 x^3 dx = 1/4$. Based on $N = 10,000$ observations from a Uniform$(0,1)$ we get $\widehat{I} = .248$ with a standard error of .0028. ■

A generalization of the basic method is to consider integrals of the form

$$I = \int h(x) f(x) dx \tag{24.3}$$

where $f(x)$ is a probability density function. Taking f to be a Uniform (a,b) gives us the special case above. Now we draw $X_1, \ldots, X_N \sim f$ and take

$$\widehat{I} \equiv \frac{1}{N} \sum_{i=1}^N h(X_i)$$

as before.

24.2 Example. Let

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

be the standard Normal PDF. Suppose we want to compute the CDF at some point x:

$$I = \int_{-\infty}^x f(s) ds = \Phi(x).$$

Write

$$I = \int h(s) f(s) ds$$

where

$$h(s) = \begin{cases} 1 & s < x \\ 0 & s \geq x. \end{cases}$$

Now we generate $X_1, \ldots, X_N \sim N(0,1)$ and set

$$\widehat{I} = \frac{1}{N} \sum_i h(X_i) = \frac{\text{number of observations } \leq x}{N}.$$

For example, with $x = 2$, the true answer is $\Phi(2) = .9772$ and the Monte Carlo estimate with $N = 10,000$ yields .9751. Using $N = 100,000$ we get .9771. ■

24.3 Example (Bayesian Inference for Two Binomials)**.** Let $X \sim \text{Binomial}(n, p_1)$ and $Y \sim \text{Binomial}(m, p_2)$. We would like to estimate $\delta = p_2 - p_1$. The MLE is $\widehat{\delta} = \widehat{p}_2 - \widehat{p}_1 = (Y/m) - (X/n)$. We can get the standard error $\widehat{\text{se}}$ using the delta method which yields

$$\widehat{\text{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$$

and then construct a 95 percent confidence interval $\widehat{\delta} \pm 2\,\widehat{\text{se}}$. Now consider a Bayesian analysis. Suppose we use the prior $f(p_1, p_2) = f(p_1)f(p_2) = 1$, that is, a flat prior on (p_1, p_2). The posterior is

$$f(p_1, p_2|X, Y) \propto p_1^X (1-p_1)^{n-X}\, p_2^Y (1-p_2)^{m-Y}.$$

The posterior mean of δ is

$$\overline{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2) f(p_1, p_2|X, Y) = \int_0^1 \int_0^1 (p_2 - p_1) f(p_1, p_2|X, Y).$$

If we want the posterior density of δ we can first get the posterior CDF

$$F(c|X, Y) = P(\delta \leq c|X, Y) = \int_A f(p_1, p_2|X, Y)$$

where $A = \{(p_1, p_2) : \ p_2 - p_1 \leq c\}$. The density can then be obtained by differentiating F.

To avoid all these integrals, let's use simulation. Note that $f(p_1, p_2|X, Y) = f(p_1|X)f(p_2|Y)$ which implies that p_1 and p_2 are independent under the posterior distribution. Also, we see that $p_1|X \sim \text{Beta}(X+1, n-X+1)$ and $p_2|Y \sim \text{Beta}(Y+1, m-Y+1)$. Hence, we can simulate $(P_1^{(1)}, P_2^{(1)}), \ldots, (P_1^{(N)}, P_2^{(N)})$ from the posterior by drawing

$$\begin{aligned} P_1^{(i)} &\sim \text{Beta}(X+1, n-X+1) \\ P_2^{(i)} &\sim \text{Beta}(Y+1, m-Y+1) \end{aligned}$$

for $i = 1, \ldots, N$. Now let $\delta^{(i)} = P_2^{(i)} - P_1^{(i)}$. Then,

$$\overline{\delta} \approx \frac{1}{N} \sum_i \delta^{(i)}.$$

We can also get a 95 percent posterior interval for δ by sorting the simulated values, and finding the .025 and .975 quantile. The posterior density $f(\delta|X, Y)$ can be obtained by applying density estimation techniques to $\delta^{(1)}, \ldots, \delta^{(N)}$ or, simply by plotting a histogram. For example, suppose that $n = m = 10$,

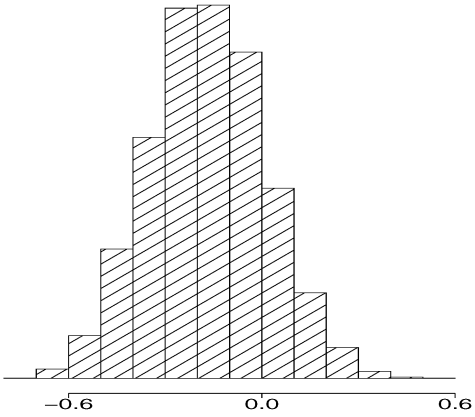

FIGURE 24.1. Posterior of δ from simulation.

$X = 8$ and $Y = 6$. From a posterior sample of size 1000 we get a 95 percent posterior interval of (-0.52,0.20). The posterior density can be estimated from a histogram of the simulated values as shown in Figure 24.1. ■

24.4 Example (Bayesian Inference for Dose Response)**.** Suppose we conduct an experiment by giving rats one of ten possible doses of a drug, denoted by $x_1 < x_2 < \ldots < x_{10}$. For each dose level x_i we use n rats and we observe Y_i, the number that survive. Thus we have ten independent binomials $Y_i \sim \text{Binomial}(n, p_i)$. Suppose we know from biological considerations that higher doses should have higher probability of death. Thus, $p_1 \leq p_2 \leq \cdots \leq p_{10}$. We want to estimate the dose at which the animals have a 50 percent chance of dying. This is called the LD50. Formally, $\delta = x_j$ where

$$j = \min\{i : \ p_i \geq .50\}.$$

Notice that δ is implicitly a (complicated) function of $p_1, \ldots, p_{10}$ so we can write $\delta = g(p_1, \ldots, p_{10})$ for some g. This just means that if we know $(p_1, \ldots, p_{10})$ then we can find δ. The posterior mean of δ is

$$\int \int \cdots \int_A g(p_1, \ldots, p_{10}) f(p_1, \ldots, p_{10} | Y_1, \ldots, Y_{10}) dp_1 dp_2 \ldots dp_{10}.$$

The integral is over the region

$$A = \{(p_1, \ldots, p_{10}) : \ p_1 \leq \cdots \leq p_{10}\}.$$

The posterior CDF of δ is

$$\begin{aligned} F(c|Y_1, \ldots, Y_{10}) &= \mathbb{P}(\delta \leq c | Y_1, \ldots, Y_{10}) \\ &= \int \int \cdots \int_B f(p_1, \ldots, p_{10} | Y_1, \ldots, Y_{10}) dp_1 dp_2 \ldots dp_{10} \end{aligned}$$

where

$$B = A \bigcap \left\{ (p_1, \ldots, p_{10}) : \ g(p_1, \ldots, p_{10}) \leq c \right\}.$$

We need to do a 10-dimensional integral over a restricted region A. Instead, we will use simulation. Let us take a flat prior truncated over A. Except for the truncation, each P_i has once again a Beta distribution. To draw from the posterior we do the following steps:

(1) Draw $P_i \sim \text{Beta}(Y_i + 1, n - Y_i + 1), i = 1, \ldots, 10$.

(2) If $P_1 \leq P_2 \leq \cdots \leq P_{10}$ keep this draw. Otherwise, throw it away and draw again until you get one you can keep.

(3) Let $\delta = x_j$ where

$$j = \min\{i : \ P_i > .50\}.$$

We repeat this N times to get $\delta^{(1)}, \ldots, \delta^{(N)}$ and take

$$\mathbb{E}(\delta | Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_i \delta^{(i)}.$$

δ is a discrete variable. We can estimate its probability mass function by

$$\mathbb{P}(\delta = x_j | Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_{i=1}^{N} I(\delta^{(i)} = j).$$

For example, consider the following data:

Dose	1	2	3	4	5	6	7	8	9	10
Number of animals n_i	15	15	15	15	15	15	15	15	15	15
Number of survivors Y_i	0	0	2	2	8	10	12	14	15	14

The posterior draws for $p_1, \ldots, p_{10}$ are shown in the second panel in the figure. We find that that $\overline{\delta} = 4.04$ with a 95 percent interval of (3,5). ■

24.3 Importance Sampling

Consider again the integral $I = \int h(x) f(x) dx$ where f is a probability density. The basic Monte Carlo method involves sampling from f. However, there are cases where we may not know how to sample from f. For example, in Bayesian inference, the posterior density density is is obtained by multiplying the likelihood $\mathcal{L}(\theta)$ times the prior $f(\theta)$. There is no guarantee that $f(\theta|x)$ will be a known distribution like a Normal or Gamma or whatever.

Importance sampling is a generalization of basic Monte Carlo which overcomes this problem. Let g be a probability density that we know how to simulate from. Then

$$I = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)}g(x)dx = \mathbb{E}_g(Y) \tag{24.4}$$

where $Y = h(X)f(X)/g(X)$ and the expectation $\mathbb{E}_g(Y)$ is with respect to g. We can simulate $X_1, \ldots, X_N \sim g$ and estimate I by

$$\widehat{I} = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_i \frac{h(X_i)f(X_i)}{g(X_i)}. \tag{24.5}$$

This is called **importance sampling.** By the law of large numbers, $\widehat{I} \xrightarrow{\text{P}} I$. However, there is a catch. It's possible that $\widehat{I}$ might have an infinite standard error. To see why, recall that I is the mean of $w(x) = h(x)f(x)/g(x)$. The second moment of this quantity is

$$\mathbb{E}_g(w^2(X)) = \int \left(\frac{h(x)f(x)}{g(x)}\right)^2 g(x)dx = \int \frac{h^2(x)f^2(x)}{g(x)}dx. \tag{24.6}$$

If g has thinner tails than f, then this integral might be infinite. To avoid this, a basic rule in importance sampling is to sample from a density g with thicker tails than f. Also, suppose that $g(x)$ is small over some set A where $f(x)$ is large. Again, the ratio of f/g could be large leading to a large variance. This implies that we should choose g to be similar in shape to f. In summary, a good choice for an importance sampling density g should be similar to f but with thicker tails. In fact, we can say what the optimal choice of g is.

24.5 Theorem. *The choice of g that minimizes the variance of $\widehat{I}$ is*

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}.$$

PROOF. The variance of $w = fh/g$ is

$$\begin{aligned}
\mathbb{E}_g(w^2) - (\mathbb{E}(w^2))^2 &= \int w^2(x)g(x)dx - \left(\int w(x)g(x)dx\right)^2 \\
&= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int \frac{h(x)f(x)}{g(x)}g(x)dx\right)^2 \\
&= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int h(x)f(x)dx\right)^2.
\end{aligned}$$

The second integral does not depend on g, so we only need to minimize the first integral. From Jensen's inequality (Theorem 4.9) we have

$$\mathbb{E}_g(W^2) \geq (\mathbb{E}_g(|W|))^2 = \left(\int |h(x)|f(x)dx\right)^2.$$

This establishes a lower bound on $\mathbb{E}_g(W^2)$. However, $\mathbb{E}_{g^*}(W^2)$ equals this lower bound which proves the claim. ∎

This theorem is interesting but it is only of theoretical interest. If we did not know how to sample from f then it is unlikely that we could sample from $|h(x)|f(x)/\int |h(s)|f(s)ds$. In practice, we simply try to find a thick-tailed distribution g which is similar to $f|h|$.

24.6 Example (Tail Probability)**.** Let's estimate $I = \mathbb{P}(Z > 3) = .0013$ where $Z \sim N(0,1)$. Write $I = \int h(x)f(x)dx$ where $f(x)$ is the standard Normal density and $h(x) = 1$ if $x > 3$, and 0 otherwise. The basic Monte Carlo estimator is $\widehat{I} = N^{-1}\sum_i h(X_i)$ where $X_1, \ldots, X_N \sim N(0,1)$. Using $N = 100$ we find (from simulating many times) that $\mathbb{E}(\widehat{I}) = .0015$ and $\mathbb{V}(\widehat{I}) = .0039$. Notice that most observations are wasted in the sense that most are not near the right tail. Now we will estimate this with importance sampling taking g to be a Normal(4,1) density. We draw values from g and the estimate is now $\widehat{I} = N^{-1}\sum_i f(X_i)h(X_i)/g(X_i)$. In this case we find that $\mathbb{E}(\widehat{I}) = .0011$ and $\mathbb{V}(\widehat{I}) = .0002$. We have reduced the standard deviation by a factor of 20. ∎

24.7 Example (Measurement Model With Outliers)**.** Suppose we have measurements $X_1, \ldots, X_n$ of some physical quantity θ. A reasonable model is

$$X_i = \theta + \epsilon_i.$$

If we assume that $\epsilon_i \sim N(0,1)$ then $X_i \sim N(\theta_i, 1)$. However, when taking measurements, it is often the case that we get the occasional wild observation, or outlier. This suggests that a Normal might be a poor model since Normals have thin tails which implies that extreme observations are rare. One way to improve the model is to use a density for ϵ_i with a thicker tail, for example, a t-distribution with ν degrees of freedom which has the form

$$t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\nu\pi}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Smaller values of ν correspond to thicker tails. For the sake of illustration we will take $\nu = 3$. Suppose we observe n $X_i = \theta + \epsilon_i$, $i = 1, \ldots, n$ where ϵ_i has

a t distribution with $\nu = 3$. We will take a flat prior on θ. The likelihood is $\mathcal{L}(\theta) = \prod_{i=1}^n t(X_i - \theta)$ and the posterior mean of θ is

$$\overline{\theta} = \frac{\int \theta \mathcal{L}(\theta) d\theta}{\int \mathcal{L}(\theta) d\theta}.$$

We can estimate the top and bottom integral using importance sampling. We draw $\theta_1, \ldots, \theta_N \sim g$ and then

$$\overline{\theta} \approx \frac{\frac{1}{N}\sum_{j=1}^N \frac{\theta_j \mathcal{L}(\theta_j)}{g(\theta_j)}}{\frac{1}{N}\sum_{j=1}^N \frac{\mathcal{L}(\theta_j)}{g(\theta_j)}}.$$

To illustrate the idea, we drew $n = 2$ observations. The posterior mean (computed numerically) is -0.54. Using a Normal importance sampler g yields an estimate of -0.74. Using a Cauchy (t-distribution with 1 degree of freedom) importance sampler yields an estimate of -0.53. ■

24.4 MCMC Part I: The Metropolis–Hastings Algorithm

Consider once more the problem of estimating the integral $I = \int h(x) f(x) dx$. Now we introduce Markov chain Monte Carlo (MCMC) methods. The idea is to construct a Markov chain $X_1, X_2, \ldots$, whose stationary distribution is f. Under certain conditions it will then follow that

$$\frac{1}{N}\sum_{i=1}^N h(X_i) \xrightarrow{\text{P}} \mathbb{E}_f(h(X)) = I.$$

This works because there is a law of large numbers for Markov chains; see Theorem 23.25.

The **Metropolis–Hastings** algorithm is a specific MCMC method that works as follows. Let $q(y|x)$ be an arbitrary, friendly distribution (i.e., we know how to sample from $q(y|x)$). The conditional density $q(y|x)$ is called the **proposal distribution.** The Metropolis–Hastings algorithm creates a sequence of observations $X_0, X_1, \ldots$, as follows.

Metropolis–Hastings Algorithm

Choose X_0 arbitrarily. Suppose we have generated $X_0, X_1, \ldots, X_i$. To generate X_{i+1} do the following:

(1) Generate a **proposal** or **candidate** value $Y \sim q(y|X_i)$.

(2) Evaluate $r \equiv r(X_i, Y)$ where

$$r(x,y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)},\ 1\right\}.$$

(3) Set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1-r. \end{cases}$$

24.8 Remark. A simple way to execute step (3) is to generate $U \sim (0,1)$. If $U < r$ set $X_{i+1} = Y$ otherwise set $X_{i+1} = X_i$.

24.9 Remark. A common choice for $q(y|x)$ is $N(x, b^2)$ for some $b > 0$. This means that the proposal is draw from a Normal, centered at the current value. In this case, the proposal density q is symmetric, $q(y|x) = q(x|y)$, and r simplifies to

$$r = \min\left\{\frac{f(Y)}{f(X_i)},\ 1\right\}.$$

By construction, $X_0, X_1, \ldots$ is a Markov chain. But why does this Markov chain have f as its stationary distribution? Before we explain why, let us first do an example.

24.10 Example. The Cauchy distribution has density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}.$$

Our goal is to simulate a Markov chain whose stationary distribution is f. As suggested in the remark above, we take $q(y|x)$ to be a $N(x, b^2)$. So in this case,

$$r(x,y) = \min\left\{\frac{f(y)}{f(x)},\ 1\right\} = \min\left\{\frac{1+x^2}{1+y^2},\ 1\right\}.$$

So the algorithm is to draw $Y \sim N(X_i, b^2)$ and set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y). \end{cases}$$

The simulator requires a choice of b. Figure 24.2 shows three chains of length $N = 1,000$ using $b = .1$, $b = 1$ and $b = 10$. Setting $b = .1$ forces the chain to take small steps. As a result, the chain doesn't "explore" much of the sample space. The histogram from the sample does not approximate the true density very well. Setting $b = 10$ causes the proposals to often be far in the

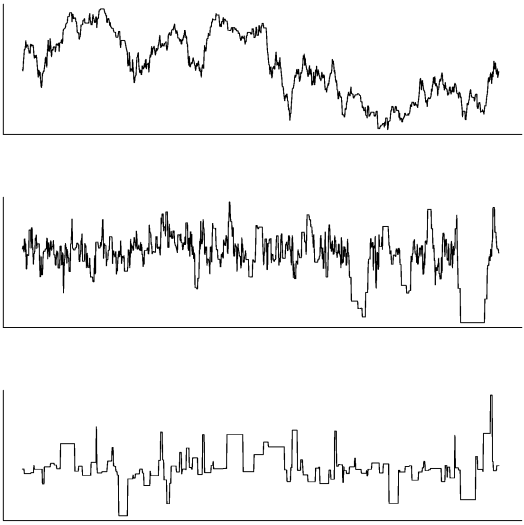

FIGURE 24.2. Three Metropolis chains corresponding to $b = .1$, $b = 1$, $b = 10$.

tails, making r small and hence we reject the proposal and keep the chain at its current position. The result is that the chain "gets stuck" at the same place quite often. Again, this means that the histogram from the sample does not approximate the true density very well. The middle choice avoids these extremes and results in a Markov chain sample that better represents the density sooner. In summary, there are tuning parameters and the efficiency of the chain depends on these parameters. We'll discuss this in more detail later. ■

If the sample from the Markov chain starts to "look like" the target distribution f quickly, then we say that the chain is "mixing well." Constructing a chain that mixes well is somewhat of an art.

Why It Works. Recall from Chapter 23 that a distribution π satisfies **detailed balance** for a Markov chain if

$$p_{ij}\pi_i = p_{ji}\pi_j.$$

We showed that if π satisfies detailed balance, then it is a stationary distribution for the chain.

Because we are now dealing with continuous state Markov chains, we will change notation a little and write $p(x, y)$ for the probability of making a transition from x to y. Also, let's use $f(x)$ instead of π for a distribution. In

this new notation, f is a stationary distribution if $f(x) = \int f(y)p(y,x)dy$ and detailed balance holds for f if

$$f(x)p(x,y) = f(y)p(y,x). \tag{24.7}$$

Detailed balance implies that f is a stationary distribution since, if detailed balance holds, then

$$\int f(y)p(y,x)dy = \int f(x)p(x,y)dy = f(x)\int p(x,y)dy = f(x)$$

which shows that $f(x) = \int f(y)p(y,x)dy$ as required. Our goal is to show that f satisfies detailed balance which will imply that f is a stationary distribution for the chain.

Consider two points x and y. Either

$$f(x)q(y|x) < f(y)q(x|y) \quad \text{or} \quad f(x)q(y|x) > f(y)q(x|y).$$

We will ignore ties (which occur with probability zero for continuous distributions). Without loss of generality, assume that $f(x)q(y|x) > f(y)q(x|y)$. This implies that

$$r(x,y) = \frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}$$

and that $r(y,x) = 1$. Now $p(x,y)$ is the probability of jumping from x to y. This requires two things: (i) the proposal distribution must generate y, and (ii) you must accept y. Thus,

$$p(x,y) = q(y|x)r(x,y) = q(y|x)\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)} = \frac{f(y)}{f(x)}q(x|y).$$

Therefore,

$$f(x)p(x,y) = f(y)q(x|y). \tag{24.8}$$

On the other hand, $p(y,x)$ is the probability of jumping from y to x. This requires two things: (i) the proposal distribution must generate x, and (ii) you must accept x. This occurs with probability $p(y,x) = q(x|y)r(y,x) = q(x|y)$. Hence,

$$f(y)p(y,x) = f(y)q(x|y). \tag{24.9}$$

Comparing (24.8) and (24.9), we see that we have shown that detailed balance holds.

24.5 MCMC Part II: Different Flavors

There are different types of MCMC algorithm. Here we will consider a few of the most popular versions.

Random-Walk-Metropolis–Hastings. In the previous section we considered drawing a proposal Y of the form

$$Y = X_i + \epsilon_i$$

where ϵ_i comes from some distribution with density g. In other words, $q(y|x) = g(y - x)$. We saw that in this case,

$$r(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

This is called a **random-walk-Metropolis–Hastings** method. The reason for the name is that, if we did not do the accept–reject step, we would be simulating a random walk. The most common choice for g is a $N(0, b^2)$. The hard part is choosing b so that the chain mixes well. A good rule of thumb is: choose b so that you accept the proposals about 50 percent of the time.

Warning! This method doesn't make sense unless X takes values on the whole real line. If X is restricted to some interval then it is best to transform X. For example, if $X \in (0, \infty)$ then you might take $Y = \log X$ and then simulate the distribution for Y instead of X.

Independence-Metropolis–Hastings. This is an importance-sampling version of MCMC. We draw the proposal from a fixed distribution g. Generally, g is chosen to be an approximation to f. The acceptance probability becomes

$$r(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{g(x)}{g(y)}\right\}.$$

Gibbs Sampling. The two previous methods can be easily adapted, in principle, to work in higher dimensions. In practice, tuning the chains to make them mix well is hard. Gibbs sampling is a way to turn a high-dimensional problem into several one-dimensional problems.

Here's how it works for a bivariate problem. Suppose that (X, Y) has density $f_{X,Y}(x, y)$. First, suppose that it is possible to simulate from the conditional distributions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Let (X_0, Y_0) be starting values. Assume we have drawn $(X_0, Y_0), \ldots, (X_n, Y_n)$. Then the Gibbs sampling algorithm for getting (X_{n+1}, Y_{n+1}) is:

Gibbs Sampling

$$\begin{aligned} X_{n+1} &\sim f_{X|Y}(x|Y_n) \\ Y_{n+1} &\sim f_{Y|X}(y|X_{n+1}) \\ &\text{repeat} \end{aligned}$$

This generalizes in the obvious way to higher dimensions.

24.11 Example (Normal Hierarchical Model)**.** Gibbs sampling is very useful for a class of models called **hierarchical models**. Here is a simple case. Suppose we draw a sample of k cities. From each city we draw n_i people and observe how many people Y_i have a disease. Thus, $Y_i \sim \text{Binomial}(n_i, p_i)$. We are allowing for different disease rates in different cities. We can also think of the $p_i's$ as random draws from some distribution F. We can write this model in the following way:

$$\begin{aligned} P_i &\sim F \\ Y_i|P_i = p_i &\sim \text{Binomial}(n_i, p_i). \end{aligned}$$

We are interested in estimating the $p_i's$ and the overall disease rate $\int p\, dF(p)$.

To proceed, it will simplify matters if we make some transformations that allow us to use some Normal approximations. Let $\widehat{p}_i = Y_i/n_i$. Recall that $\widehat{p}_i \approx N(p_i, s_i)$ where $s_i = \sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$. Let $\psi_i = \log(p_i/(1-p_i))$ and define $Z_i \equiv \widehat{\psi}_i = \log(\widehat{p}_i/(1-\widehat{p}_i))$. By the delta method,

$$\widehat{\psi}_i \approx N(\psi_i, \sigma_i^2)$$

where $\sigma_i^2 = 1/(n\widehat{p}_i(1-\widehat{p}_i))$. Experience shows that the Normal approximation for ψ is more accurate than the Normal approximation for p so we shall work with ψ. We shall treat σ_i as known. Furthermore, we shall take the distribution of the $\psi_i's$ to be Normal. The hierarchical model is now

$$\begin{aligned} \psi_i &\sim N(\mu, \tau^2) \\ Z_i|\psi_i &\sim N(\psi_i, \sigma_i^2). \end{aligned}$$

As yet another simplification we take $\tau = 1$. The unknown parameter are $\theta = (\mu, \psi_1, \ldots, \psi_k)$. The likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &\propto \prod_i f(\psi_i|\mu) \prod_i f(Z_i|\psi) \\ &\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2\right\}. \end{aligned}$$

If we use the prior $f(\mu) \propto 1$ then the posterior is proportional to the likelihood. To use Gibbs sampling, we need to find the conditional distribution of each parameter conditional on all the others. Let us begin by finding $f(\mu|\text{rest})$ where "rest" refers to all the other variables. We can throw away any terms that don't involve μ. Thus,

$$\begin{aligned} f(\mu|\text{rest}) &\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{k}{2}(\mu - b)^2\right\} \end{aligned}$$

where

$$b = \frac{1}{k}\sum_i \psi_i.$$

Hence we see that $\mu|\text{rest} \sim N(b, 1/k)$. Next we will find $f(\psi|\text{rest})$. Again, we can throw away any terms not involving ψ_i leaving us with

$$\begin{aligned} f(\psi_i|\text{rest}) &\propto \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2d_i^2}(\psi_i - e_i)^2\right\} \end{aligned}$$

where

$$e_i = \frac{\frac{Z_i}{\sigma_i^2} + \mu}{1 + \frac{1}{\sigma_i^2}} \quad \text{and} \quad d_i^2 = \frac{1}{1 + \frac{1}{\sigma_i^2}}$$

and so $\psi_i|\text{rest} \sim N(e_i, d_i^2)$. The Gibbs sampling algorithm then involves iterating the following steps N times:

$$\begin{aligned} \text{draw } \mu &\sim N(b, v^2) \\ \text{draw } \psi_1 &\sim N(e_1, d_1^2) \\ &\vdots \quad\quad \vdots \\ \text{draw } \psi_k &\sim N(e_k, d_k^2). \end{aligned}$$

It is understood that at each step, the most recently drawn version of each variable is used.

We generated a numerical example with $k = 20$ cities and $n = 20$ people from each city. After running the chain, we can convert each ψ_i back into p_i by way of $p_i = e^{\psi_i}/(1 + e^{\psi_i})$. The raw proportions are shown in Figure 24.4. Figure 24.3 shows "trace plots" of the Markov chain for p_1 and μ. Figure 24.4 shows the posterior for μ based on the simulated values. The second

panel of Figure 24.4 shows the raw proportions and the Bayes estimates. Note that the Bayes estimates are "shrunk" together. The parameter τ controls the amount of shrinkage. We set $\tau = 1$ but, in practice, we should treat τ as another unknown parameter and let the data determine how much shrinkage is needed. ■

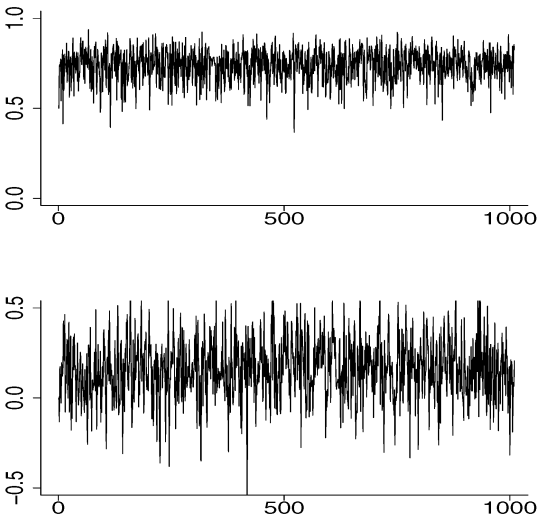

FIGURE 24.3. Posterior simulation for Example 24.11. The top panel shows simulated values of p_1. The top panel shows simulated values of μ.

So far we assumed that we know how to draw samples from the conditionals $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. If we don't know how, we can still use the Gibbs sampling algorithm by drawing each observation using a Metropolis–Hastings step. Let q be a proposal distribution for x and let $\widetilde{q}$ be a proposal distribution for y. When we do a Metropolis step for X, we treat Y as fixed. Similarly, when we do a Metropolis step for Y, we treat X as fixed. Here are the steps:

Metropolis within Gibbs

(1a) Draw a proposal $Z \sim q(z|X_n)$.

(1b) Evaluate

$$r = \min\left\{\frac{f(Z,Y_n)}{f(X_n,Y_n)}\frac{q(X_n|Z)}{q(Z|X_n)},\ 1\right\}.$$

(1c) Set

$$X_{n+1} = \begin{cases} Z & \text{with probability } r \\ X_n & \text{with probability } 1-r. \end{cases}$$

(2a) Draw a proposal $Z \sim \widetilde{q}(z|Y_n)$.

(2b) Evaluate

$$r = \min\left\{\frac{f(X_{n+1},Z)}{f(X_{n+1},Y_n)}\frac{\widetilde{q}(Y_n|Z)}{\widetilde{q}(Z|Y_n)},\ 1\right\}.$$

(2c) Set

$$Y_{n+1} = \begin{cases} Z & \text{with probability } r \\ Y_n & \text{with probability } 1-r. \end{cases}$$

Again, this generalizes to more than two dimensions.

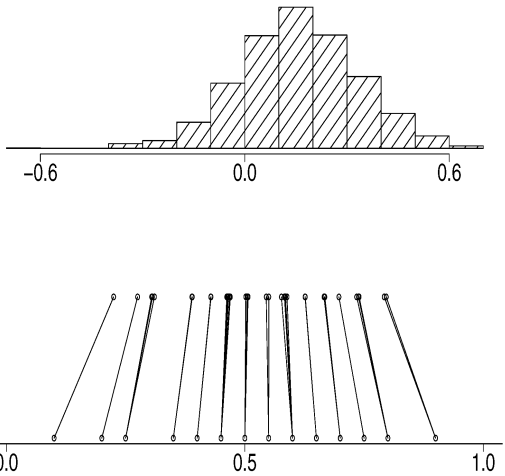

FIGURE 24.4. Example 24.11. Top panel: posterior histogram of μ. Lower panel: raw proportions and the Bayes posterior estimates. The Bayes estimates have been shrunk closer together than the raw proportions.

24.6 Bibliographic Remarks

MCMC methods go back to the effort to build the atomic bomb in World War II. They were used in various places after that, especially in spatial statistics. There was a new surge of interest in the 1990s that still continues. My main reference for this chapter was Robert and Casella (1999). See also Gelman et al. (1995) and Gilks et al. (1998).

24.7 Exercises

1. Let
$$I = \int_1^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$
(a) Estimate I using the basic Monte Carlo method. Use $N = 100,000$. Also, find the estimated standard error.

(b) Find an (analytical) expression for the standard error of your estimate in (a). Compare to the estimated standard error.

(c) Estimate I using importance sampling. Take g to be $N(1.5, v^2)$ with $v = .1$, $v = 1$ and $v = 10$. Compute the (true) standard errors in each case. Also, plot a histogram of the values you are averaging to see if there are any extreme values.

(d) Find the optimal importance sampling function g^*. What is the standard error using g^*?

2. Here is a way to use importance sampling to estimate a marginal density. Let $f_{X,Y}(x, y)$ be a bivariate density and let $(X_1, X_2), \ldots, (X_N, Y_N) \sim f_{X,Y}$.
(a) Let $w(x)$ be an arbitrary probability density function. Let
$$\widehat{f}_X(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{f_{X,Y}(x, Y_i) w(X_i)}{f_{X,Y}(X_i, Y_i)}.$$
Show that, for each x,
$$\widehat{f}_X(x) \xrightarrow{p} f_X(x).$$
Find an expression for the variance of this estimator.

(b) Let $Y \sim N(0, 1)$ and $X|Y = y \sim N(y, 1 + y^2)$. Use the method in (a) to estimate $f_X(x)$.

3. Here is a method called **accept–reject sampling** for drawing observations from a distribution.

   (a) Suppose that f is some probability density function. Let g be any other density and suppose that $f(x) \leq Mg(x)$ for all x, where M is a known constant. Consider the following algorithm:

   (step 1): Draw $X \sim g$ and $U \sim \text{Unif}(0, 1)$;

   (step 2): If $U \leq f(X)/(Mg(X))$ set $Y = X$, otherwise go back to step 1. (Keep repeating until you finally get an observation.)

   Show that the distribution of Y is f.

   (b) Let f be a standard Normal density and let $g(x) = 1/(1 + x^2)$ be the Cauchy density. Apply the method in (a) to draw 1,000 observations from the Normal distribution. Draw a histogram of the sample to verify that the sample appears to be Normal.

4. A random variable Z has a **inverse Gaussian distribution** if it has density

$$f(z) \propto z^{-3/2} \exp\left\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1\theta_2} + \log\left(\sqrt{2\theta_2}\right)\right\}, \quad z > 0$$

   where $\theta_1 > 0$ and $\theta_2 > 0$ are parameters. It can be shown that

$$\mathbb{E}(Z) = \sqrt{\frac{\theta_2}{\theta_1}} \quad \text{and} \quad \mathbb{E}\left(\frac{1}{Z}\right) = \sqrt{\frac{\theta_1}{\theta_2}} + \frac{1}{2\theta_2}.$$

   (a) Let $\theta_1 = 1.5$ and $\theta_2 = 2$. Draw a sample of size 1,000 using the independence-Metropolis–Hastings method. Use a Gamma distribution as the proposal density. To assess the accuracy, compare the mean of Z and $1/Z$ from the sample to the theoretical means Try different Gamma distributions to see if you can get an accurate sample.

   (b) Draw a sample of size 1,000 using the random-walk-Metropolis–Hastings method. Since $z > 0$ we cannot just use a Normal density. One strategy is this. Let $W = \log Z$. Find the density of W. Use the random-walk-Metropolis–Hastings method to get a sample $W_1, \ldots, W_N$ and let $Z_i = e^{W_i}$. Assess the accuracy of the simulation as in part (a).

5. Get the heart disease data from the book web site. Consider a Bayesian analysis of the logistic regression model

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}}.$$

Use the flat prior $f(\beta_0, \ldots, \beta_k) \propto 1$. Use the Gibbs–Metropolis algorithm to draw a sample of size 10,000 from the posterior $f(\beta_0, \beta_1|\text{data})$. Plot histograms of the posteriors for the β_j's. Get the posterior mean and a 95 percent posterior interval for each β_j.

(b) Compare your analysis to a frequentist approach using maximum likelihood.

Bibliography

Agresti, A. (1990). *Categorical Data Analysis.* Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* 267–281.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis (Second Edition).* Wiley.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27** 536–561.

Beecher, H. (1959). *Measurement of Subjective Responses.* Oxford University Press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological* **57** 289–300.

Beran, R. (2000). REACT scatterplot smoothers: Superefficiency through basis economy. *Journal of the American Statistical Association* **95** 155–171.

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *The Annals of Statistics* **26** 1826–1856.

Berger, J. and Wolpert, R. (1984). *The Likelihood Principle.* Institute of Mathematical Statistics.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis (Second Edition).* Springer-Verlag.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses (c/r: P335-352). *Statistical Science* **2** 317–335.

Berliner, L. M. (1983). Improving on inadmissible estimators in the control problem. *The Annals of Statistics* **11** 814–826.

Bickel, P. J. and Doksum, K. A. (2000). *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I (Second Edition).* Prentice Hall.

Billingsley, P. (1979). *Probability and Measure.* Wiley.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analyses: Theory and Practice.* MIT Press.

Breiman, L. (1992). *Probability.* Society for Industrial and Applied Mathematics.

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association* **58** 85–96.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Duxbury Press.

Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association* **94** 807–823.

Cox, D. and Lewis, P. (1966). *The Statistical Analysis of Series of Events.* Chapman & Hall.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics* **21** 903–923.

Cox, D. R. and Hinkley, D. V. (2000). *Theoretical statistics.* Chapman & Hall.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

DeGroot, M. and Schervish, M. (2002). *Probability and Statistics (Third Edition)*. Addison-Wesley.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *The Annals of Statistics* **14** 68–87.

Dobson, A. J. (2001). *An introduction to generalized linear models*. Chapman & Hall.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics* **26** 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (Disc: p 337–369). *Journal of the Royal Statistical Society, Series B, Methodological* **57** 301–337.

Dunsmore, I., Daly, F. et al. (1987). *M345 Statistical Methods, Unit 9: Categorical Data*. The Open University.

Edwards, D. (1995). *Introduction to graphical modelling*. Springer-Verlag.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer-Verlag.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7** 1–26.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96** 1151–1160.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Ferguson, T. (1967). *Mathematical Statistics : a Decision Theoretic Approach*. Academic Press.

Fisher, R. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1** 1–32.

Freedman, D. (1999). Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* **27** 1119–1141.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1** 55–77.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Ghosal, S., Ghosh, J. K. and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500–531.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Grimmett, G. and Stirzaker, D. (1982). *Probability and Random Processes*. Oxford University Press.

Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag.

Halverson, N., Leitch, E., Pryke, C., Kovac, J., Carlstrom, J., Holzapfel, W., Dragovan, M., Cartwright, J., Mason, B., Padin, S., Pearson, T., Shepherd, M. and Readhead, A. (2002). DASI first results: A measurement of the cosmic microwave background angular power spectrum. *Astrophysics Journal* **568** 38–45.

Hardle, W. (1990). *Applied nonparametric regression*. Cambridge University Press.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*. Springer-Verlag.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag.

Herbich, R. (2002). *Learning Kernel Classifiers: Theory and Algorithms.* MIT Press.

Johnson, R. A. and Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis.* Prentice-Hall.

Johnson, S. and Johnson, R. (1972). *New England Journal of Medicine* **287** 1122–1125.

Jordan, M. (2004). *Graphical models.* In Preparation.

Karr, A. (1993). *Probability.* Springer-Verlag.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90** 773–795.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules (corr: 1998 v93 p 412). *Journal of the American Statistical Association* **91** 1343–1370.

Larsen, R. J. and Marx, M. L. (1986). *An Introduction to Mathematical Statistics and Its Applications (Second Edition).* Prentice Hall.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Lee, A. T. et al. (2001). A high spatial resolution analysis of the maxima-1 cosmic microwave background anisotropy data. *Astrophys. J.* **561** L1–L6.

Lee, P. M. (1997). *Bayesian Statistics: An Introduction.* Edward Arnold.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses (Second Edition).* Wiley.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer-Verlag.

Loader, C. (1999). *Local regression and likelihood.* Springer-Verlag.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* **20** 712–736.

Morrison, A., Black, M., Lowe, C., Macmahon, B. and Yusa, S. (1973). Some international differences in histology and survival in breast cancer. *International Journal of Cancer* **11** 261–267.

Netterfield, C. B. et al. (2002). A measurement by boomerang of multiple peaks in the angular power spectrum of the cosmic microwave background. *Astrophys. J.* **571** 604–614.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser.

Pearl, J. (2000). *Casuality: models, reasoning, and inference*. Cambridge University Press.

Phillips, D. and King, E. (1988). Death takes a holiday: Mortality surrounding major social occasions. *Lancet* **2** 728–732.

Phillips, D. and Smith, D. (1990). Postponement of death until symbolically meaningful occasions. *Journal of the American Medical Association* **263** 1947–1961.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B* **11** 18–84.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis (Second Edition)*. Duxbury Press.

Robert, C. P. (1994). *The Bayesian Choice: A Decision-theoretic Motivation*. Springer-Verlag.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.

Robins, J., Scheines, R., Spirtes, P. and Wasserman, L. (2003). Uniform convergence in causal inference. *Biometrika* (to appear).

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16** 285–319.

Rosenbaum, P. (2002). *Observational Studies*. Springer-Verlag.

Ross, S. (2002). *Probability Models for Computer Science*. Academic Press.

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities. *South African Medical Journal* **64** 430–436.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

Scott, D., Gotto, A., Cole, J. and Gorry, G. (1978). Plasma lipids as collateral risk factors in coronary artery disease: a study of 371 males with chest pain. *Journal of Chronic Diseases* **31** 337–345.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap (German)*. Springer-Verlag.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics* **29** 687–714.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes With Applications to Statistics*. Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Taylor, H. M. and Karlin, S. (1994). *An Introduction to Stochastic Modeling*. Academic Press.

van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley.

Weisberg, S. (1985). *Applied Linear Regression.* Wiley.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics* **5** 161–215.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *The Annals of Statistics* **28** 532–552.

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association* **90** 151–156.

List of Symbols

General Symbols

$\mathbb{R}$	real numbers
$\inf_{x \in A} f(x)$	infimum: the largest number y such that $y \leq f(x)$ for all $x \in A$ think of this as the minimum of f
$\sup_{x \in A} f(x)$	supremum: the smallest number y such that $y \geq f(x)$ for all $x \in A$ think of this as the maximum of f
$n!$	$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$
$\binom{n}{k}$	$\frac{n!}{k!(n-k)!}$
$\Gamma(\alpha)$	Gamma function $\int_0^\infty y^{\alpha-1} e^{-y} dy$
Ω	sample space (set of outcomes)
ω	outcome, element, point
A	event (subset of Ω)
$I_A(\omega)$	indicator function; 1 if $\omega \in A$ and 0 otherwise
$\lvert A \rvert$	number of points in set A

Probability Symbols

$\mathbb{P}(A)$	probability of event A
$A \amalg B$	A and B are independent
$A \wr\!\!\!\wr\!\!\!\wr B$	A and B are dependent
F_X	cumulative distribution function $F_X(x) = \mathbb{P}(X \leq x)$
f_X	probability density (or mass) function
$X \sim F$	X has distribution F
$X \sim f$	X has density f
$X \stackrel{d}{=} Y$	X and Y have the same distribution
IID	independent and identically distributed
$X_1, \ldots, X_n \sim F$	IID sample of size n from F
ϕ	standard Normal probability density
Φ	standard Normal distribution function
z_α	upper α quantile of $N(0,1)$: $z_\alpha = \Phi^{-1}(1-\alpha)$
$\mathbb{E}(X) = \int x \, dF(x)$	expected value (mean) of random variable X
$\mathbb{E}(r(X)) = \int r(x) dF(x)$	expected value (mean) of $r(X)$
$\mathbb{V}(X)$	variance of random variable X
$\mathsf{Cov}(X, Y)$	covariance between X and Y
$X_1, \ldots, X_n$	data
n	sample size

Convergence Symbols

$\xrightarrow{\text{P}}$	convergence in probability
$\rightsquigarrow$	convergence in distribution
$\xrightarrow{\text{qm}}$	convergence in quadratic mean
$X_n \approx N(\mu, \sigma_n^2)$	$(X_n - \mu)/\sigma_n \rightsquigarrow N(0,1)$
$x_n = o(a_n)$	$x_n/a_n \to 0$
$x_n = O(a_n)$	$\lvert x_n/a_n \rvert$ is bounded for large n
$X_n = o_P(a_n)$	$X_n/a_n \xrightarrow{\text{P}} 0$
$X_n = O_P(a_n)$	$\lvert X_n/a_n \rvert$ is bounded in probability for large n

Statistical Models

$\mathfrak{F}$	statistical model; a set of distribution functions, density functions or regression functions
θ	parameter
$\widehat{\theta}$	estimate of parameter
$T(F)$	statistical functional (the mean, for example)
$\mathcal{L}_n(\theta)$	likelihood function

Useful Math Facts

$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \cdots$

$\sum_{j=k}^{\infty} r^j = \frac{r^k}{1-r} \quad$ for $0 < r < 1$

$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a$

Stirling's approximation: $n! \approx n^n e^{-n}\sqrt{2\pi n}$

The Gamma function. The Gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

for $\alpha \geq 0$. If $\alpha > 1$ then $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. If n is a positive integer then $\Gamma(n) = (n-1)!$. Some special values are: $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

Table of Distributions

Distribution	PDF or probability function	mean	variance	MGF
Point mass at a	$I(x = a)$	a	0	e^{at}
Bernoulli(p)	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$pe^t + (1-p)$
Binomial(n, p)	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$	$(pe^t + (1-p))^n$
Geometric(p)	$p(1-p)^{x-1} I(x \geq 1)$	$1/p$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$ $\left(t < -\log(1-p)\right)$
Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform(a, b)	$I(a < x < b)/(b-a)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$
Exponential(β)	$\frac{e^{-x/\beta}}{\beta}$	β	β^2	$\frac{1}{1-\beta t}$ $\left(t < 1/\beta\right)$
Gamma(α, β)	$\frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$	$\alpha\beta$	$\alpha\beta^2$	$\left(\frac{1}{1-\beta t}\right)^\alpha$ $(t < 1/\beta)$
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
t_ν	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}$.	0 (if $\nu > 1$)	$\frac{\nu}{\nu-2}$ (if $\nu > 2$)	does not exist
χ^2_p	$\frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}$	p	$2p$	$\left(\frac{1}{1-2t}\right)^{p/2}$ $\left(t < 1/2\right)$