the case in hypothesis testing. Also, one can't use improper priors in testing because this leads to an undefined constant in the denominator of the expression above. Thus, if you use Bayesian testing you must choose the prior $f(\theta)$ very carefully. It is possible to get a prior-free bound on $\mathbb{P}(H_0|X^n = x^n)$. Notice that $0 \leq \int \mathcal{L}(\theta)f(\theta)d\theta \leq \mathcal{L}(\widehat{\theta})$. Hence,

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \mathcal{L}(\widehat{\theta})} \ \leq \ \mathbb{P}(H_0|X^n = x^n) \ \leq \ 1.$$

The upper bound is not very interesting, but the lower bound is non-trivial.

## 11.9 Strengths and Weaknesses of Bayesian Inference

Bayesian inference is appealing when prior information is available since Bayes' theorem is a natural way to combine prior information with data. Some people find Bayesian inference psychologically appealing because it allows us to make probability statements about parameters. In contrast, frequentist inference provides confidence sets $C_n$ which trap the parameter 95 percent of the time, but we cannot say that $\mathbb{P}(\theta \in C_n|X^n)$ is .95. In the frequentist approach we can make probability statements about $C_n$, not $\theta$. However, psychological appeal is not a compelling scientific argument for using one type of inference over another.

In parametric models, with large samples, Bayesian and frequentist methods give approximately the same inferences. In general, they need not agree.

Here are three examples that illustrate the strengths and weakness of Bayesian inference. The first example is Example 6.14 revisited. This example shows the psychological appeal of Bayesian inference. The second and third show that Bayesian methods can fail.

**11.8 Example** (Example 6.14 revisited)**.** We begin by reviewing the example. Let $\theta$ be a fixed, known real number and let $X_1, X_2$ be independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Now define $Y_i = \theta + X_i$ and suppose that you only observe $Y_1$ and $Y_2$. Let

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2 \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

This is a 75 percent confidence set since, no matter what $\theta$ is, $\mathbb{P}_\theta(\theta \in C) = 3/4$.

Suppose we observe $Y_1 = 15$ and $Y_2 = 17$. Then our 75 percent confidence interval is $\{16\}$. However, we are certain, in this case, that $\theta = 16$. So calling

this a 75 percent confidence set, bothers many people. Nonetheless, $C$ is a valid 75 percent confidence set. It will trap the true value 75 percent of the time.

The Bayesian solution is more satisfying to many. For simplicity, assume that $\theta$ is an integer. Let $f(\theta)$ be a prior mass function such that $f(\theta) > 0$ for every integer $\theta$. When $Y = (Y_1, Y_2) = (15, 17)$, the likelihood function is

$$\mathcal{L}(\theta) = \begin{cases} 1/4 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Applying Bayes' theorem we see that

$$\mathbb{P}(\Theta = \theta | Y = (15, 17)) = \begin{cases} 1 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\mathbb{P}(\theta \in C | Y = (15, 17)) = 1$. There is nothing wrong with saying that $\{16\}$ is a 75 percent confidence interval. But is it not a probability statement about $\theta$. ■

**11.9 Example.** This is a simplified version of the example in Robins and Ritov (1997). The data consist of $n$ IID triples

$$(X_1, R_1, Y_1), \ldots, (X_n, Y_n, R_n).$$

Let $B$ be a finite but very large number, like $B = 100^{100}$. Any realistic sample size $n$ will be small compared to $B$. Let

$$\theta = (\theta_1, \ldots, \theta_B)$$

be a vector of unknown parameters such that $0 \leq \theta_j \leq 1$ for $1 \leq j \leq B$. Let

$$\xi = (\xi_1, \ldots, \xi_B)$$

be a vector of **known** numbers such that

$$0 < \delta \leq \xi_j \leq 1 - \delta < 1, \quad 1 \leq j \leq B,$$

where $\delta$ is some, small, positive number. Each data point $(X_i, R_i, Y_i)$ is drawn in the following way:

1. Draw $X_i$ uniformly from $\{1, \ldots, B\}$.
2. Draw $R_i \sim \text{Bernoulli}(\xi_{X_i})$.
3. If $R_i = 1$, then draw $Y_i \sim \text{Bernoulli}(\theta_{X_i})$. If $R_i = 0$, do not draw $Y_i$.

The model may seem a little artificial but, in fact, it is caricature of some real **missing data** problems in which some data points are not observed. In this example, $R_i = 0$ can be thought of as meaning "missing." Our goal is to estimate

$$\psi = \mathbb{P}(Y_i = 1).$$

Note that

$$\begin{aligned} \psi &= \mathbb{P}(Y_i = 1) = \sum_{j=1}^{B} \mathbb{P}(Y_i = 1 | X = j)\mathbb{P}(X = j) \\ &= \frac{1}{B}\sum_{j=1}^{B} \theta_j \equiv g(\theta) \end{aligned}$$

so $\psi = g(\theta)$ is a function of $\theta$.

Let us consider a Bayesian analysis first. The likelihood of a single observation is

$$f(X_i, R_i, Y_i) = f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i}.$$

The last term is raised to the power $R_i$ since, if $R_i = 0$, then $Y_i$ is not observed and hence that term drops out of the likelihood. Since $f(X_i) = 1/B$ and that $Y_i$ and $R_i$ are Bernoulli,

$$f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} = \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i}.$$

Thus, the likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{n} f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} \\ &= \prod_{i=1}^{n} \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i} \\ &\propto \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i}. \end{aligned}$$

We have dropped all the terms involving $B$ and the $\xi_j$'s since these are known constants, not parameters. The log-likelihood is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} Y_i\, R_i \log \theta_{X_i} + (1 - Y_i)\, R_i\, \log(1 - \theta_{X_i}) \\ &= \sum_{j=1}^{B} n_j \log \theta_j + \sum_{j=1}^{B} m_j \log(1 - \theta_j) \end{aligned}$$

where

$$\begin{aligned} n_j &= \#\{i:\ Y_i = 1, R_i = 1, X_i = j\} \\ m_j &= \#\{i:\ Y_i = 0, R_i = 1, X_i = j\}. \end{aligned}$$

Now, $n_j = m_j = 0$ for most $j$ since $B$ is so much larger than $n$. This has several implications. First, the MLE for most $\theta_j$ is not defined. Second, for most $\theta_j$, the posterior distribution is equal to the prior distribution, since those $\theta_j$ do not appear in the likelihood. Hence, $f(\theta|\text{Data}) \approx f(\theta)$. It follows that $f(\psi|\text{Data}) \approx f(\psi)$. In other words, the data provide little information about $\psi$ in a Bayesian analysis.

Now we consider a frequentist solution. Define

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\xi_{X_i}}. \tag{11.10}$$

We will now show that this estimator is unbiased and has small mean-squared error. It can be shown (see Exercise 7) that

$$\mathbb{E}(\widehat{\psi}) = \psi \quad \text{and} \quad \mathbb{V}(\widehat{\psi}) \leq \frac{1}{n\delta^2}. \tag{11.11}$$

Therefore, the MSE is of order $1/n$ which goes to 0 fairly quickly as we collect more data, no matter how large $B$ is. The estimator defined in (11.10) is called the **Horwitz-Thompson** estimator. It cannot be derived from a Bayesian or likelihood point of view since it involves the terms $\xi_{X_i}$. These terms drop out of the log-likelihood and hence will not show up in any likelihood-based method including Bayesian estimators.

The moral of the story is this. Bayesian methods are tied to the likelihood function. But in high dimensional (and nonparametric) problems, the likelihood may not yield accurate inferences. ■

**11.10 Example.** Suppose that $f$ is a probability density function and that

$$f(x) = cg(x)$$

where $g(x) > 0$ is a known function and $c$ is unknown. In principle we can compute $c$ since $\int f(x)\,dx = 1$ implies that $c = 1/\int g(x)\,dx$. But in many cases we can't do the integral $\int g(x)\,dx$ since $g$ might be a complicated function and $x$ could be high dimensional. Despite the fact that $c$ is not known, it is often possible to draw a sample $X_1, \ldots, X_n$ from $f$; see Chapter 24. Can we use the sample to estimate the normalizing constant $c$? Here is a frequentist solution:

Let $\widehat{f}_n(x)$ be a consistent estimate of the density $f$. Chapter 20 explains how to construct such an estimate. Choose any point $x$ and note that $c = f(x)/g(x)$. Hence, $\widehat{c} = \widehat{f}(x)/g(x)$ is a consistent estimate of $c$. Now let us try to solve this problem from a Bayesian approach. Let $\pi(c)$ be a prior such that $\pi(c) > 0$ for all $c > 0$. The likelihood function is

$$\mathcal{L}_n(c) = \prod_{i=1}^n f(X_i) = \prod_{i=1}^n cg(X_i) = c^n \prod_{i=1}^n g(X_i) \ \propto \ c^n.$$

Hence the posterior is proportional to $c^n\pi(c)$. The posterior does not depend on $X_1, \ldots, X_n$, so we come to the startling conclusion that, from the Bayesian point of view, there is no information in the data about $c$. Moreover, the posterior mean is

$$\frac{\int_0^\infty c^{n+1}\pi(c)\,dc}{\int_0^\infty c^n\pi(c)\,dc}$$

which tends to infinity as $n$ increases. ■

These last two examples illustrate an important point. Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference.

What should we conclude from all this? The important thing is to understand that frequentist and Bayesian methods are answering different questions. To combine prior beliefs with data in a principled way, use Bayesian inference. To construct procedures with guaranteed long run performance, such as confidence intervals, use frequentist methods. Generally, Bayesian methods run into problems when the parameter space is high dimensional. In particular, 95 percent posterior intervals need not contain the true value 95 percent of the time (in the frequency sense).

## 11.10 Bibliographic Remarks

Some references on Bayesian inference include Carlin and Louis (1996), Gelman et al. (1995), Lee (1997), Robert (1994), and Schervish (1995). See Cox (1993), Diaconis and Freedman (1986), Freedman (1999), Barron et al. (1999), Ghosal et al. (2000), Shen and Wasserman (2001), and Zhao (2000) for discussions of some of the technicalities of nonparametric Bayesian inference. The Robins-Ritov example is discussed in detail in Robins and Ritov (1997) where it is cast more properly as a nonparametric problem. Example 11.10 is due to Edward George (personal communication). See Berger and Delampady (1987)

and Kass and Raftery (1995) for a discussion of Bayesian testing. See Kass and Wasserman (1996) for a discussion of noninformative priors.

## 11.11 Appendix

**Proof of Theorem 11.5.**

It can be shown that the effect of the prior diminishes as $n$ increases so that $f(\theta|X^n) \propto \mathcal{L}_n(\theta)f(\theta) \approx \mathcal{L}_n(\theta)$. Hence, $\log f(\theta|X^n) \approx \ell(\theta)$. Now, $\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta}) = \ell(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta})$ since $\ell'(\widehat{\theta}) = 0$. Exponentiating, we get approximately that

$$f(\theta|X^n) \propto \exp\left\{-\frac{1}{2}\frac{(\theta - \widehat{\theta})^2}{\sigma_n^2}\right\}$$

where $\sigma_n^2 = -1/\ell''(\widehat{\theta}_n)$. So the posterior of $\theta$ is approximately Normal with mean $\widehat{\theta}$ and variance $\sigma_n^2$. Let $\ell_i = \log f(X_i|\theta)$, then

$$\begin{aligned}
\frac{1}{\sigma_n^2} &= -\ell''(\widehat{\theta}_n) = \sum_i -\ell_i''(\widehat{\theta}_n) \\
&= n\left(\frac{1}{n}\right)\sum_i -\ell_i''(\widehat{\theta}_n) \approx n\mathbb{E}_\theta\left[-\ell_i''(\widehat{\theta}_n)\right] \\
&= nI(\widehat{\theta}_n)
\end{aligned}$$

and hence $\sigma_n \approx \mathsf{se}(\widehat{\theta})$. ∎

## 11.12 Exercises

1. Verify (11.7).
2. Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$.
   (a) Simulate a data set (using $\mu = 5$) consisting of n=100 observations.
   (b) Take $f(\mu) = 1$ and find the posterior density. Plot the density.
   (c) Simulate 1,000 draws from the posterior. Plot a histogram of the simulated values and compare the histogram to the answer in (b).
   (d) Let $\theta = e^\mu$. Find the posterior density for $\theta$ analytically and by simulation.
   (e) Find a 95 percent posterior interval for $\mu$.
   (f) Find a 95 percent confidence interval for $\theta$.

3. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $f(\theta) \propto 1/\theta$. Find the posterior density.

4. Suppose that 50 people are given a placebo and 50 are given a new treatment. 30 placebo patients show improvement while 40 treated patients show improvement. Let $\tau = p_2 - p_1$ where $p_2$ is the probability of improving under treatment and $p_1$ is the probability of improving under placebo.

   (a) Find the MLE of $\tau$. Find the standard error and 90 percent confidence interval using the delta method.

   (b) Find the standard error and 90 percent confidence interval using the parametric bootstrap.

   (c) Use the prior $f(p_1, p_2) = 1$. Use simulation to find the posterior mean and posterior 90 percent interval for $\tau$.

   (d) Let
   $$\psi = \log\left(\left(\frac{p_1}{1-p_1}\right) \div \left(\frac{p_2}{1-p_2}\right)\right)$$
   be the log-odds ratio. Note that $\psi = 0$ if $p_1 = p_2$. Find the MLE of $\psi$. Use the delta method to find a 90 percent confidence interval for $\psi$.

   (e) Use simulation to find the posterior mean and posterior 90 percent interval for $\psi$.

5. Consider the Bernoulli(p) observations

   0 1 0 1 0 0 0 0 0 0

   Plot the posterior for $p$ using these priors: Beta(1/2,1/2), Beta(1,1), Beta(10,10), Beta(100,100).

6. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

   (a) Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior. Show that the posterior is also a Gamma. Find the posterior mean.

   (b) Find the Jeffreys' prior. Find the posterior.

7. In Example 11.9, verify (11.11).

8. Let $X \sim N(\mu, 1)$. Consider testing
   $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

Take $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. Let the prior for $\mu$ under $H_1$ be $\mu \sim N(0, b^2)$. Find an expression for $\mathbb{P}(H_0|X = x)$. Compare $\mathbb{P}(H_0|X = x)$ to the p-value of the Wald test. Do the comparison numerically for a variety of values of $x$ and $b$. Now repeat the problem using a sample of size $n$. You will see that the posterior probability of $H_0$ can be large even when the p-value is small, especially when $n$ is large. This disagreement between Bayesian and frequentist testing is called the Jeffreys-Lindley paradox.

# 12
# Statistical Decision Theory

## 12.1 Preliminaries

We have considered several point estimators such as the maximum likelihood estimator, the method of moments estimator, and the posterior mean. In fact, there are many other ways to generate estimators. How do we choose among them? The answer is found in **decision theory** which is a formal theory for comparing statistical procedures.

Consider a parameter $\theta$ which lives in a parameter space $\Theta$. Let $\widehat{\theta}$ be an estimator of $\theta$. In the language of decision theory, an estimator is sometimes called a **decision rule** and the possible values of the decision rule are called **actions**.

We shall measure the discrepancy between $\theta$ and $\widehat{\theta}$ using a **loss function** $L(\theta, \widehat{\theta})$. Formally, $L$ maps $\Theta \times \Theta$ into $\mathbb{R}$. Here are some examples of loss functions:

| | |
|---|---|
| $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ | squared error loss, |
| $L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$ | absolute error loss, |
| $L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$ | $L_p$ loss, |
| $L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$ | zero–one loss, |
| $L(\theta, \widehat{\theta}) = \int \log\left(\frac{f(x;\theta)}{f(x;\widehat{\theta})}\right) f(x;\theta)dx$ | Kullback–Leibler loss. |

Bear in mind in what follows that an estimator $\widehat{\theta}$ is a function of the data. To emphasize this point, sometimes we will write $\widehat{\theta}$ as $\widehat{\theta}(X)$. To assess an estimator, we evaluate the average loss or risk.

**12.1 Definition.** *The* **risk** *of an estimator* $\widehat{\theta}$ *is*

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\left(L(\theta, \widehat{\theta})\right) = \int L(\theta, \widehat{\theta}(x)) f(x;\theta) dx.$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \text{MSE} = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2_\theta(\widehat{\theta}).$$

In the rest of the chapter, if we do not state what loss function we are using, assume the loss function is squared error.

## 12.2 Comparing Risk Functions

To compare two estimators we can compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

**12.2 Example.** Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ and $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. If $2 < \theta < 4$ then $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, otherwise, $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 12.1. ■

**12.3 Example.** Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has 0 bias, we have that

$$R(p, \widehat{p}_1) = \mathbb{V}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_{i=1}^n X_i$ and $\alpha$ and $\beta$ are positive constants. This is the posterior mean using a Beta $(\alpha, \beta)$ prior. Now,

$$R(p, \widehat{p}_2) \quad = \quad \mathbb{V}_p(\widehat{p}_2) + (\mathsf{bias}_p(\widehat{p}_2))^2$$

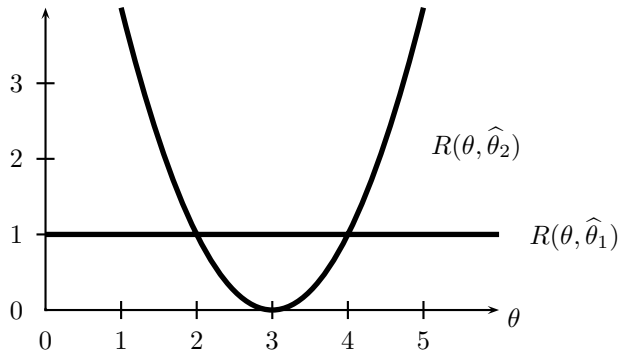



FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of $\theta$.

$$= \mathbb{V}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2$$
$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

Let $\alpha = \beta = \sqrt{n/4}$. (In Example 12.12 we will explain this choice.) The resulting estimator is

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

The risk functions are plotted in figure 12.2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

**12.4 Definition.** *The* **maximum risk** *is*

$$\overline{R}(\widehat{\theta}) = \sup_\theta R(\theta, \widehat{\theta}) \tag{12.1}$$

*and* **the Bayes risk** *is*

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta \tag{12.2}$$

*where $f(\theta)$ is a prior for $\theta$.*

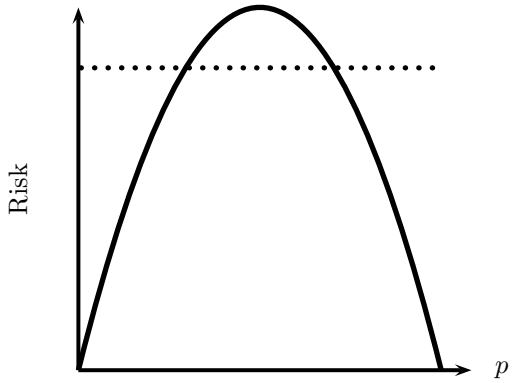



FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

**12.5 Example.** Consider again the two estimators in Example 12.3. We have

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

and

$$\overline{R}(\widehat{p}_2) = \max_{p} \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when $n$ is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like maximum risk are imperfect. Now consider the Bayes risk. For illustration, let us take $f(p) = 1$. Then

$$r(f, \widehat{p}_1) = \int R(p, \widehat{p}_1)dp = \int \frac{p(1-p)}{n} dp = \frac{1}{6n}$$

and

$$r(f, \widehat{p}_2) = \int R(p, \widehat{p}_2)dp = \frac{n}{4(n+\sqrt{n})^2}.$$

For $n \geq 20$, $r(f, \widehat{p}_2) > r(f, \widehat{p}_1)$ which suggests that $\widehat{p}_1$ is a better estimator. This might seem intuitively reasonable but this answer depends on the choice of prior. The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior. ■

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to

minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

**12.6 Definition.** *A decision rule that minimizes the Bayes risk is called a* **Bayes rule**. *Formally, $\widehat{\theta}$ is a Bayes rule with respect to the prior $f$ if*

$$r(f,\widehat{\theta}) = \inf_{\widetilde{\theta}} r(f,\widetilde{\theta}) \tag{12.3}$$

*where the infimum is over all estimators $\widetilde{\theta}$. An estimator that minimizes the maximum risk is called a* **minimax rule**. *Formally, $\widehat{\theta}$ is minimax if*

$$\sup_{\theta} R(\theta,\widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta,\widetilde{\theta}) \tag{12.4}$$

*where the infimum is over all estimators $\widetilde{\theta}$.*

## 12.3 Bayes Estimators

Let $f$ be a prior. From Bayes' theorem, the posterior density is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{12.5}$$

where $m(x) = \int f(x,\theta)d\theta = \int f(x|\theta)f(\theta)d\theta$ is the **marginal distribution** of $X$. Define the **posterior risk** of an estimator $\widehat{\theta}(x)$ by

$$r(\widehat{\theta}|x) = \int L(\theta,\widehat{\theta}(x))f(\theta|x)d\theta. \tag{12.6}$$

**12.7 Theorem.** *The Bayes risk $r(f,\widehat{\theta})$ satisfies*

$$r(f,\widehat{\theta}) = \int r(\widehat{\theta}|x)m(x)\,dx.$$

*Let $\widehat{\theta}(x)$ be the value of $\theta$ that minimizes $r(\widehat{\theta}|x)$. Then $\widehat{\theta}$ is the Bayes estimator.*

Proof. We can rewrite the Bayes risk as follows:

$$\begin{aligned} r(f,\widehat{\theta}) &= \int R(\theta,\widehat{\theta})f(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x))f(x|\theta)dx\right)f(\theta)d\theta \\ &= \int\int L(\theta,\widehat{\theta}(x))f(x,\theta)dxd\theta = \int\int L(\theta,\widehat{\theta}(x))f(\theta|x)m(x)dxd\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x))f(\theta|x)d\theta\right)m(x)\,dx = \int r(\widehat{\theta}|x)m(x)\,dx. \end{aligned}$$

If we choose $\widehat{\theta}(x)$ to be the value of $\theta$ that minimizes $r(\widehat{\theta}|x)$ then we will minimize the integrand at every $x$ and thus minimize the integral $\int r(\widehat{\theta}|x)m(x)dx$. ■

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

**12.8 Theorem.** *If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x) = \int \theta f(\theta|x)d\theta = \mathbb{E}(\theta|X = x). \tag{12.7}$$

*If $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$ then the Bayes estimator is the median of the posterior $f(\theta|x)$. If $L(\theta, \widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $f(\theta|x)$.*

PROOF. We will prove the theorem for squared error loss. The Bayes rule $\widehat{\theta}(x)$ minimizes $r(\widehat{\theta}|x) = \int(\theta - \widehat{\theta}(x))^2 f(\theta|x)d\theta$. Taking the derivative of $r(\widehat{\theta}|x)$ with respect to $\widehat{\theta}(x)$ and setting it equal to 0 yields the equation $2\int(\theta - \widehat{\theta}(x))f(\theta|x)d\theta = 0$. Solving for $\widehat{\theta}(x)$ we get 12.7. ■

**12.9 Example.** Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where $\sigma^2$ is known. Suppose we use a $N(a, b^2)$ prior for $\mu$. The Bayes estimator with respect to squared error loss is the posterior mean, which is

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{b^2}{b^2 + \frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}}a. \quad ■$$

## 12.4 Minimax Rules

Finding minimax rules is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: Bayes estimators with a constant risk function are minimax.

**12.10 Theorem.** *Let $\widehat{\theta}^f$ be the Bayes rule for some prior $f$:*

$$r(f, \widehat{\theta}^f) = \inf_{\widehat{\theta}} r(f, \widehat{\theta}). \tag{12.8}$$

*Suppose that*

$$R(\theta, \widehat{\theta}^f) \leq r(f, \widehat{\theta}^f) \text{ for all } \theta. \tag{12.9}$$

*Then $\widehat{\theta}^f$ is minimax and $f$ is called a* **least favorable prior**.

PROOF. Suppose that $\widehat{\theta}^f$ is not minimax. Then there is another rule $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f)$. Since the average of a function is always less than or equal to its maximum, we have that $r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f) \leq r(f, \widehat{\theta}^f)$$

which contradicts (12.8). ■

**12.11 Theorem.** *Suppose that $\widehat{\theta}$ is the Bayes rule with respect to some prior $f$. Suppose further that $\widehat{\theta}$ has constant risk: $R(\theta, \widehat{\theta}) = c$ for some $c$. Then $\widehat{\theta}$ is minimax.*

PROOF. The Bayes risk is $r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all $\theta$. Now apply the previous theorem. ■

**12.12 Example.** Consider the Bernoulli model with squared error loss. In example 12.3 we showed that the estimator

$$\widehat{p}(X^n) = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes rule, for the prior Beta$(\alpha, \beta)$ with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax. ■

**12.13 Example.** Consider again the Bernoulli but with loss function

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

Let

$$\widehat{p}(X^n) = \widehat{p} = \frac{\sum_{i=1}^n X_i}{n}.$$

The risk is

$$R(p, \widehat{p}) = E\left(\frac{(\widehat{p} - p)^2}{p(1-p)}\right) = \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) = \frac{1}{n}$$

which, as a function of $p$, is constant. It can be shown that, for this loss function, $\widehat{p}(X^n)$ is the Bayes estimator under the prior $f(p) = 1$. Hence, $\widehat{p}$ is minimax. ■

A natural question to ask is: what is the minimax estimator for a Normal model?

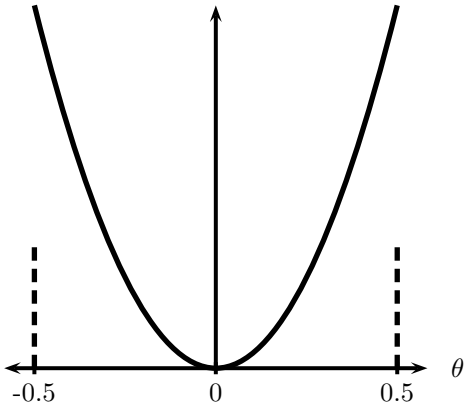



FIGURE 12.3. Risk function for constrained Normal with m=.5. The two short dashed lines show the least favorable prior which puts its mass at two points.

**12.14 Theorem.** *Let $X_1, \ldots, X_n \sim N(\theta, 1)$ and let $\widehat{\theta} = \overline{X}$. Then $\widehat{\theta}$ is minimax with respect to any well-behaved loss function.* [1] *It is the only estimator with this property.*

If the parameter space is restricted, then the theorem above does not apply as the next example shows.

**12.15 Example.** Suppose that $X \sim N(\theta, 1)$ and that $\theta$ is known to lie in the interval $[-m, m]$ where $0 < m < 1$. The unique, minimax estimator under squared error loss is

$$\widehat{\theta}(X) = m \tanh(mX)$$

where $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$. It can be shown that this is the Bayes rule with respect to the prior that puts mass $1/2$ at $m$ and mass $1/2$ at $-m$. Moreover, it can be shown that the risk is not constant but it does satisfy $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all $\theta$; see Figure 12.3. Hence, Theorem 12.10 implies that $\widehat{\theta}$ is minimax. ■

[1] "Well-behaved" means that the level sets must be convex and symmetric about the origin. The result holds up to sets of measure 0.

## 12.5 Maximum Likelihood, Minimax, and Bayes

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the MLE $\widehat{\theta}$ roughly equals the variance:[2]

$$R(\theta, \widehat{\theta}) = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2 \approx \mathbb{V}_\theta(\widehat{\theta}).$$

As we saw in Chapter 9, the variance of the MLE is approximately

$$\mathbb{V}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{12.10}$$

For any other estimator $\theta'$, it can be shown that for large $n$, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. More precisely,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} n\, R(\theta', \widehat{\theta}) \geq \frac{1}{I(\theta)}. \tag{12.11}$$

This says that, in a local, large sample sense, the MLE is minimax. It can also be shown that the MLE is approximately the Bayes rule.

In summary:

> In most parametric models, with large samples, the MLE is approximately minimax and Bayes.

There is a caveat: these results break down when the number of parameters is large as the next example shows.

**12.16 Example** (Many Normal means)**.** Let $Y_i \sim N(\theta_i, \sigma^2/n)$, $i = 1, \ldots, n$. Let $Y = (Y_1, \ldots, Y_n)$ denote the data and let $\theta = (\theta_1, \ldots, \theta_n)$ denote the unknown parameters. Assume that

$$\theta \in \Theta_n \equiv \left\{ (\theta_1, \ldots, \theta_n) : \sum_{i=1}^n \theta_i^2 \leq c^2 \right\}$$

[2] Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

for some $c > 0$. In this model, there are as many parameters as observations. [3] The MLE is $\widehat{\theta} = Y = (Y_1, \ldots, Y_n)$. Under the loss function $L(\theta, \widehat{\theta}) = \sum_{i=1}^n (\widehat{\theta}_i - \theta_i)^2$, the risk of the MLE is $R(\theta, \widehat{\theta}) = \sigma^2$. It can be shown that the minimax risk is approximately $\sigma^2/(\sigma^2 + c^2)$ and one can find an estimator $\widetilde{\theta}$ that achieves this risk. Since $\sigma^2/(\sigma^2 + c^2) < \sigma^2$, we see that $\widetilde{\theta}$ has smaller risk than the MLE. In practice, the difference between the risks can be substantial. This shows that maximum likelihood is not an optimal estimator in high dimensional problems. ■

## 12.6 Admissibility

Minimax estimators and Bayes estimators are "good estimators" in the sense that they have small risk. It is also useful to characterize bad estimators.

**12.17 Definition.** *An estimator $\widehat{\theta}$ is* **inadmissible** *if there exists another rule $\widehat{\theta}'$ such that*

$$\begin{aligned} R(\theta, \widehat{\theta}') &\leq R(\theta, \widehat{\theta}) \text{ for all } \theta \text{ and} \\ R(\theta, \widehat{\theta}') &< R(\theta, \widehat{\theta}) \text{ for at least one } \theta. \end{aligned}$$

*Otherwise, $\widehat{\theta}$ is* **admissible**.

**12.18 Example.** Let $X \sim N(\theta, 1)$ and consider estimating $\theta$ with squared error loss. Let $\widehat{\theta}(X) = 3$. We will show that $\widehat{\theta}$ is admissible. Suppose not. Then there exists a different rule $\widehat{\theta}'$ with smaller risk. In particular, $R(3, \widehat{\theta}') \leq R(3, \widehat{\theta}) = 0$. Hence, $0 = R(3, \widehat{\theta}') = \int (\widehat{\theta}'(x) - 3)^2 f(x; 3) dx$. Thus, $\widehat{\theta}'(x) = 3$. So there is no rule that beats $\widehat{\theta}$. Even though $\widehat{\theta}$ is admissible it is clearly a bad decision rule. ■

**12.19 Theorem** (Bayes Rules Are Admissible)**.** *Suppose that $\Theta \subset \mathbb{R}$ and that $R(\theta, \widehat{\theta})$ is a continuous function of $\theta$ for every $\widehat{\theta}$. Let $f$ be a prior density with full support, meaning that, for every $\theta$ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} f(\theta) d\theta > 0$. Let $\widehat{\theta}^f$ be the Bayes' rule. If the Bayes risk is finite then $\widehat{\theta}^f$ is admissible.*

PROOF. Suppose $\widehat{\theta}^f$ is inadmissible. Then there exists a better rule $\widehat{\theta}$ such that $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}^f)$ for all $\theta$ and $R(\theta_0, \widehat{\theta}) < R(\theta_0, \widehat{\theta}^f)$ for some $\theta_0$. Let

[3] The many Normal means problem is more general than it looks. Many nonparametric estimation problems are mathematically equivalent to this model.

$\nu = R(\theta_0, \widehat{\theta}^f) - R(\theta_0, \widehat{\theta}) > 0$. Since $R$ is continuous, there is an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) > \nu/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Now,

$$\begin{aligned} r(f, \widehat{\theta}^f) - r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}^f) f(\theta) d\theta - \int R(\theta, \widehat{\theta}) f(\theta) d\theta \\ &= \int \left[ R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} \left[ R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \frac{\nu}{2} \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} f(\theta) d\theta \\ &> 0. \end{aligned}$$

Hence, $r(f, \widehat{\theta}^f) > r(f, \widehat{\theta})$. This implies that $\widehat{\theta}^f$ does not minimize $r(f, \widehat{\theta})$ which contradicts the fact that $\widehat{\theta}^f$ is the Bayes rule. ∎

**12.20 Theorem.** *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Under squared error loss, $\overline{X}$ is admissible.*

The proof of the last theorem is quite technical and is omitted but the idea is as follows: The posterior mean is admissible for any strictly positive prior. Take the prior to be $N(a, b^2)$. When $b^2$ is very large, the posterior mean is approximately equal to $\overline{X}$.

How are minimaxity and admissibility linked? In general, a rule may be one, both, or neither. But here are some facts linking admissibility and minimaxity.

**12.21 Theorem.** *Suppose that $\widehat{\theta}$ has constant risk and is admissible. Then it is minimax.*

PROOF. The risk is $R(\theta, \widehat{\theta}) = c$ for some $c$. If $\widehat{\theta}$ were not minimax then there exists a rule $\widehat{\theta}'$ such that

$$R(\theta, \widehat{\theta}') \leq \sup_\theta R(\theta, \widehat{\theta}') < \sup_\theta R(\theta, \widehat{\theta}) = c.$$

This would imply that $\widehat{\theta}$ is inadmissible. ∎

Now we can prove a restricted version of Theorem 12.14 for squared error loss.

**12.22 Theorem.** *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Then, under squared error loss, $\widehat{\theta} = \overline{X}$ is minimax.*

PROOF. According to Theorem 12.20, $\widehat{\theta}$ is admissible. The risk of $\widehat{\theta}$ is $1/n$ which is constant. The result follows from Theorem 12.21. ∎

Although minimax rules are not guaranteed to be admissible they are "close to admissible." Say that $\widehat{\theta}$ is **strongly inadmissible** if there exists a rule $\widehat{\theta}'$ and an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}') < R(\theta, \widehat{\theta}) - \epsilon$ for all $\theta$.

**12.23 Theorem.** *If $\widehat{\theta}$ is minimax, then it is not strongly inadmissible.*

## 12.7 Stein's Paradox

Suppose that $X \sim N(\theta, 1)$ and consider estimating $\theta$ with squared error loss. From the previous section we know that $\widehat{\theta}(X) = X$ is admissible. Now consider estimating two, unrelated quantities $\theta = (\theta_1, \theta_2)$ and suppose that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, with loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{2}(\theta_j - \widehat{\theta}_j)^2$. Not surprisingly, $\widehat{\theta}(X) = X$ is again admissible where $X = (X_1, X_2)$. Now consider the generalization to $k$ normal means. Let $\theta = (\theta_1, \ldots, \theta_k)$, $X = (X_1, \ldots, X_k)$ with $X_i \sim N(\theta_i, 1)$ (independent) and loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{k}(\theta_j - \widehat{\theta}_j)^2$. Stein astounded everyone when he proved that, if $k \geq 3$, then $\widehat{\theta}(X) = X$ is inadmissible. It can be shown that the **James-Stein estimator** $\widehat{\theta}^S$ has smaller risk, where $\widehat{\theta}^S = (\widehat{\theta}_1^S, \ldots, \widehat{\theta}_k^S)$,

$$\widehat{\theta}_i^S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i \tag{12.12}$$

and $(z)^+ = \max\{z, 0\}$. This estimator shrinks the $X_i$'s towards 0. The message is that, when estimating many parameters, there is great value in shrinking the estimates. This observation plays an important role in modern nonparametric function estimation.

## 12.8 Bibliographic Remarks

Aspects of decision theory can be found in Casella and Berger (2002), Berger (1985), Ferguson (1967), and Lehmann and Casella (1998).

## 12.9 Exercises

1. In each of the following models, find the Bayes risk and the Bayes estimator, using squared error loss.

   (a) $X \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$.

(b) $X \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(\alpha, \beta)$.

(c) $X \sim N(\theta, \sigma^2)$ where $\sigma^2$ is known and $\theta \sim N(a, b^2)$.

2. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and suppose we estimate $\theta$ with loss function $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2/\sigma^2$. Show that $\overline{X}$ is admissible and minimax.

3. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a finite parameter space. Prove that the posterior mode is the Bayes estimator under zero–one loss.

4. (Casella and Berger (2002).) Let $X_1, \ldots, X_n$ be a sample from a distribution with variance $\sigma^2$. Consider estimators of the form $bS^2$ where $S^2$ is the sample variance. Let the loss function for estimating $\sigma^2$ be

$$L(\sigma^2, \widehat{\sigma}^2) = \frac{\widehat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\widehat{\sigma}^2}{\sigma^2}\right).$$

Find the optimal value of $b$ that minimizes the risk for all $\sigma^2$.

5. (Berliner (1983).) Let $X \sim \text{Binomial}(n, p)$ and suppose the loss function is

$$L(p, \widehat{p}) = \left(1 - \frac{\widehat{p}}{p}\right)^2$$

where $0 < p < 1$. Consider the estimator $\widehat{p}(X) = 0$. This estimator falls outside the parameter space $(0, 1)$ but we will allow this. Show that $\widehat{p}(X) = 0$ is the unique, minimax rule.

6. (Computer Experiment.) Compare the risk of the MLE and the James-Stein estimator (12.12) by simulation. Try various values of $n$ and various vectors $\theta$. Summarize your results.

# Part III

# Statistical Models and Methods

# 13
# Linear and Logistic Regression

**Regression** is a method for studying the relationship between a **response variable** $Y$ and a **covariate** $X$. The covariate is also called a **predictor variable** or a **feature.** [1] One way to summarize the relationship between $X$ and $Y$ is through the **regression function**

$$r(x) = \mathbb{E}(Y|X = x) = \int y\, f(y|x)dy. \tag{13.1}$$

Our goal is to estimate the regression function $r(x)$ from data of the form

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}.$$

In this Chapter, we take a parametric approach and assume that $r$ is linear. In Chapters 20 and 21 we discuss nonparametric regression.

## 13.1 Simple Linear Regression

The simplest version of regression is when $X_i$ is simple (one-dimensional) and $r(x)$ is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x.$$

[1] The term "regression" is due to Sir Francis Galton (1822-1911) who noticed that tall and short men tend to have sons with heights closer to the mean. He called this "regression towards the mean."

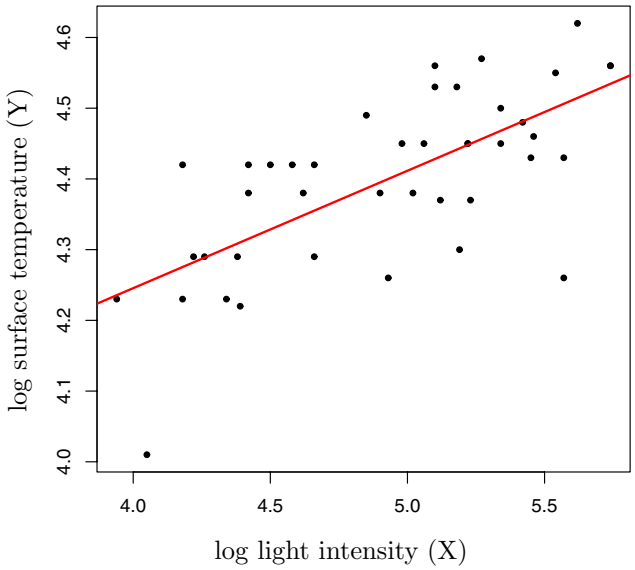



FIGURE 13.1. Data on nearby stars. The solid line is the least squares line.

This model is called the **the simple linear regression model**. We will make the further simplifying assumption that $\mathbb{V}(\epsilon_i|X = x) = \sigma^2$ does not depend on $x$. We can thus write the linear regression model as follows.

**13.1 Definition.** The Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{13.2}$$

*where* $\mathbb{E}(\epsilon_i|X_i) = 0$ *and* $\mathbb{V}(\epsilon_i|X_i) = \sigma^2$.

**13.2 Example.** Figure 13.1 shows a plot of log surface temperature (Y) versus log light intensity (X) for some nearby stars. Also on the plot is an estimated linear regression line which will be explained shortly. ■

The unknown parameters in the model are the intercept $\beta_0$ and the slope $\beta_1$ and the variance $\sigma^2$. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote estimates of $\beta_0$ and $\beta_1$. The **fitted line** is

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{13.3}$$

The **predicted values** or **fitted values** are $\widehat{Y}_i = \widehat{r}(X_i)$ and the **residuals** are defined to be

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right). \tag{13.4}$$

The **residual sums of squares** or RSS, which measures how well the line fits the data, is defined by $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

**13.3 Definition.** *The* **least squares estimates** *are the values* $\widehat{\beta}_0$ *and* $\widehat{\beta}_1$ *that minimize* $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

**13.4 Theorem.** *The least squares estimates are given by*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}, \tag{13.5}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \tag{13.6}$$

*An unbiased estimate of* $\sigma^2$ *is*

$$\widehat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2. \tag{13.7}$$

**13.5 Example.** Consider the star data from Example 13.2. The least squares estimates are $\widehat{\beta}_0 = 3.58$ and $\widehat{\beta}_1 = 0.166$. The fitted line $\widehat{r}(x) = 3.58 + 0.166\, x$ is shown in Figure 13.1. ■

**13.6 Example** (The 2001 Presidential Election)**.** Figure 13.2 shows the plot of votes for Buchanan (Y) versus votes for Bush (X) in Florida. The least squares estimates (omitting Palm Beach County) and the standard errors are

$$\begin{aligned} \widehat{\beta}_0 &= 66.0991 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 17.2926 \\ \widehat{\beta}_1 &= 0.0035 \quad\;\; \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0002. \end{aligned}$$

The fitted line is

$$\text{Buchanan} = 66.0991 + 0.0035\, \text{Bush}.$$

(We will see later how the standard errors were computed.) Figure 13.2 also shows the residuals. The inferences from linear regression are most accurate when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. If we repeat the analysis replacing votes with log(votes) we get

$$\begin{aligned} \widehat{\beta}_0 &= -2.3298 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 0.3529 \\ \widehat{\beta}_1 &= 0.730300 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0358. \end{aligned}$$

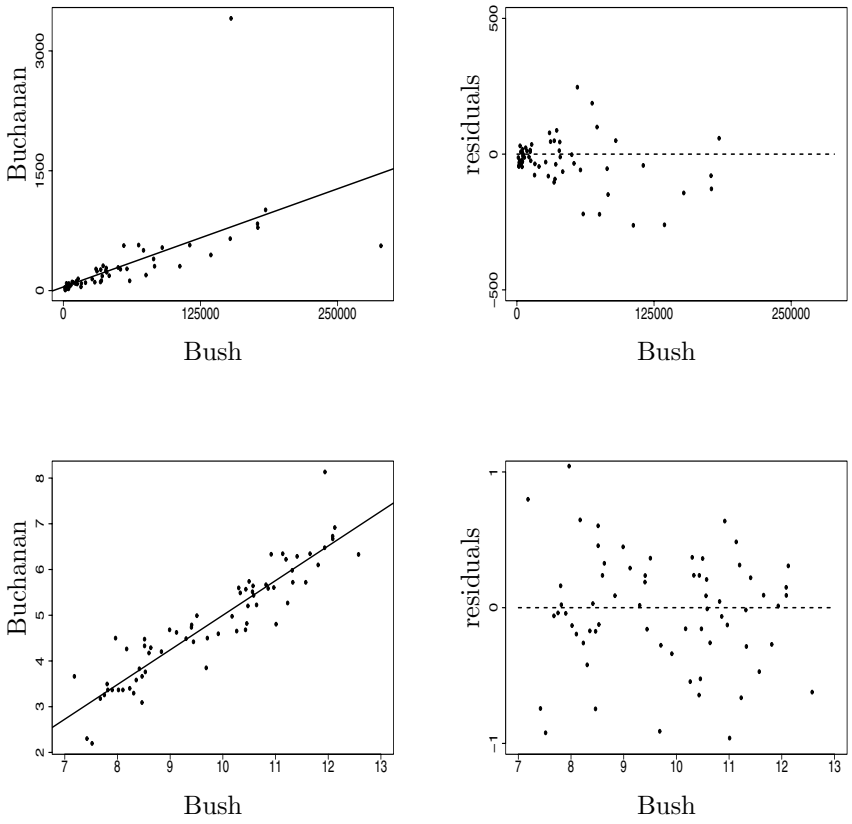



FIGURE 13.2. Voting Data for Election 2000. See example 13.6.

This gives the fit

$$\log(\text{Buchanan}) = -2.3298 + 0.7303\ \log(\text{Bush}).$$

The residuals look much healthier. Later, we shall address the following question: how do we see if Palm Beach County has a statistically plausible outcome? ■

## 13.2 Least Squares and Maximum Likelihood

Suppose we add the assumption that $\epsilon_i | X_i \sim N(0, \sigma^2)$, that is,

$$Y_i | X_i \sim N(\mu_i, \sigma^2)$$

where $\mu_i = \beta_0 + \beta_1 X_i$. The likelihood function is

$$\begin{aligned}\prod_{i=1}^n f(X_i, Y_i) &= \prod_{i=1}^n f_X(X_i) f_{Y|X}(Y_i|X_i) \\ &= \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \\ &= \mathcal{L}_1 \times \mathcal{L}_2\end{aligned}$$

where $\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$ and

$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i|X_i). \tag{13.8}$$

The term $\mathcal{L}_1$ does not involve the parameters $\beta_0$ and $\beta_1$. We shall focus on the second term $\mathcal{L}_2$ which is called the **conditional likelihood**, given by

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2\right\}.$$

The conditional log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2. \tag{13.9}$$

To find the MLE of $(\beta_0, \beta_1)$ we maximize $\ell(\beta_0, \beta_1, \sigma)$. From (13.9) we see that maximizing the likelihood is the same as minimizing the RSS $\sum_{i=1}^n \Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2$. Therefore, we have shown the following:

**13.7 Theorem.** *Under the assumption of Normality, the least squares estimator is also the maximum likelihood estimator.*

We can also maximize $\ell(\beta_0, \beta_1, \sigma)$ over $\sigma$, yielding the MLE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\epsilon}_i^2. \tag{13.10}$$

This estimator is similar to, but not identical to, the unbiased estimator. Common practice is to use the unbiased estimator (13.7).

## 13.3 Properties of the Least Squares Estimators

We now record the standard errors and limiting distribution of the least squares estimator. In regression problems, we usually focus on the properties of the estimators conditional on $X^n = (X_1, \ldots, X_n)$. Thus, we state the means and variances as conditional means and variances.

**13.8 Theorem.** *Let $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ denote the least squares estimators. Then,*

$$\begin{aligned} \mathbb{E}(\widehat{\beta}|X^n) &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ \mathbb{V}(\widehat{\beta}|X^n) &= \frac{\sigma^2}{n\, s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} \end{aligned} \tag{13.11}$$

*where $s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.*

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by taking the square roots of the corresponding diagonal terms of $\mathbb{V}(\widehat{\beta}|X^n)$ and inserting the estimate $\widehat{\sigma}$ for $\sigma$. Thus,

$$\widehat{\mathsf{se}}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \tag{13.12}$$

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}. \tag{13.13}$$

We should really write these as $\widehat{\mathsf{se}}(\widehat{\beta}_0|X^n)$ and $\widehat{\mathsf{se}}(\widehat{\beta}_1|X^n)$ but we will use the shorter notation $\widehat{\mathsf{se}}(\widehat{\beta}_0)$ and $\widehat{\mathsf{se}}(\widehat{\beta}_1)$.

**13.9 Theorem.** *Under appropriate conditions we have:*

1. *(Consistency): $\widehat{\beta}_0 \xrightarrow{\mathrm{P}} \beta_0$ and $\widehat{\beta}_1 \xrightarrow{\mathrm{P}} \beta_1$.*

2. *(Asymptotic Normality):*

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathsf{se}}(\widehat{\beta}_0)} \rightsquigarrow N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}(\widehat{\beta}_1)} \rightsquigarrow N(0,1).$$

3. *Approximate $1-\alpha$ confidence intervals for $\beta_0$ and $\beta_1$ are*

$$\widehat{\beta}_0 \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_1). \tag{13.14}$$

4. *The Wald test* [2] *for testing* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *is: reject* $H_0$ *if* $|W| > z_{\alpha/2}$ *where* $W = \widehat{\beta}_1/\widehat{\mathsf{se}}(\widehat{\beta}_1)$.

**13.10 Example.** For the election data, on the log scale, a 95 percent confidence interval is $.7303 \pm 2(.0358) = (.66, .80)$. The Wald statistics for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is $|W| = |.7303 - 0|/.0358 = 20.40$ with a p-value of $\mathbb{P}(|Z| > 20.40) \approx 0$. This is strong evidence that that the true slope is not 0. ■

## 13.4 Prediction

Suppose we have estimated a regression model $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ from data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We observe the value $X = x_*$ of the covariate for a new subject and we want to predict their outcome $Y_*$. An estimate of $Y_*$ is

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*. \tag{13.15}$$

Using the formula for the variance of the sum of two random variables,

$$\mathbb{V}(\widehat{Y}_*) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = \mathbb{V}(\widehat{\beta}_0) + x_*^2 \mathbb{V}(\widehat{\beta}_1) + 2x_* \mathsf{Cov}(\widehat{\beta}_0, \widehat{\beta}_1).$$

Theorem 13.8 gives the formulas for all the terms in this equation. The estimated standard error $\widehat{\mathsf{se}}(\widehat{Y}_*)$ is the square root of this variance, with $\widehat{\sigma}^2$ in place of $\sigma^2$. However, the confidence interval for $Y_*$ is **not** of the usual form $\widehat{Y}_* \pm z_{\alpha/2}\widehat{\mathsf{se}}$. The reason for this is explained in Exercise 10. The correct form of the confidence interval is given in the following theorem.

**13.11 Theorem** (Prediction Interval). *Let*

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left( \frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_i (X_i - \overline{X})^2} + 1 \right). \tag{13.16}$$

*An approximate* $1 - \alpha$ *prediction interval for* $Y_*$ *is*

$$\widehat{Y}_* \pm z_{\alpha/2}\, \widehat{\xi}_n. \tag{13.17}$$

[2] Recall from equation (10.5) that the Wald statistic for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ is $W = (\widehat{\beta} - \beta_0)/\widehat{\mathsf{se}}(\widehat{\beta})$.

**13.12 Example** (Election Data Revisited)**.** On the log scale, our linear regression gives the following prediction equation:

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

In Palm Beach, Bush had 152,954 votes and Buchanan had 3,467 votes. On the log scale this is 11.93789 and 8.151045. How likely is this outcome, assuming our regression model is appropriate? Our prediction for log Buchanan votes -2.3298 + .7303 (11.93789)=6.388441. Now, 8.151045 is bigger than 6.388441 but is it "significantly" bigger? Let us compute a confidence interval. We find that $\widehat{\xi}_n = .093775$ and the approximate 95 percent confidence interval is (6.200,6.578) which clearly excludes 8.151. Indeed, 8.151 is nearly 20 standard errors from $\widehat{Y}_*$. Going back to the vote scale by exponentiating, the confidence interval is (493,717) compared to the actual number of votes which is 3,467. ■

## 13.5 Multiple Regression

Now suppose that the covariate is a vector of length $k$. The data are of the form

$$(Y_1, X_1), \ldots, (Y_i, X_i), \ldots, (Y_n, X_n)$$

where

$$X_i = (X_{i1}, \ldots, X_{ik}).$$

Here, $X_i$ is the vector of $k$ covariate values for the $i^{\text{th}}$ observation. The linear regression model is

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{13.18}$$

for $i = 1, \ldots, n$, where $\mathbb{E}(\epsilon_i | X_{1i}, \ldots, X_{ki}) = 0$. Usually we want to include an intercept in the model which we can do by setting $X_{i1} = 1$ for $i = 1, \ldots, n$. At this point it will be more convenient to express the model in matrix notation. The outcomes will be denoted by

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and the covariates will be denoted by

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}.$$

Each row is one observation; the columns correspond to the $k$ covariates. Thus, $X$ is a $(n \times k)$ matrix. Let

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then we can write (13.18) as

$$Y = X\beta + \epsilon. \tag{13.19}$$

The form of the least squares estimate is given in the following theorem.

**13.13 Theorem.** *Assuming that the $(k \times k)$ matrix $X^T X$ is invertible,*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{13.20}$$

$$\mathbb{V}(\widehat{\beta}|X^n) = \sigma^2 (X^T X)^{-1} \tag{13.21}$$

$$\widehat{\beta} \approx N(\beta, \sigma^2 (X^T X)^{-1}). \tag{13.22}$$

The estimate regression function is $\widehat{r}(x) = \sum_{j=1}^k \widehat{\beta}_j x_j$. An unbiased estimate of $\sigma^2$ is

$$\widehat{\sigma}^2 = \left(\frac{1}{n-k}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon} = X\widehat{\beta} - Y$ is the vector of residuals. An approximate $1-\alpha$ confidence interval for $\beta_j$ is

$$\widehat{\beta}_j \pm z_{\alpha/2} \widehat{\mathsf{se}}(\widehat{\beta}_j) \tag{13.23}$$

where $\widehat{\mathsf{se}}^2(\widehat{\beta}_j)$ is the $j^{\text{th}}$ diagonal element of the matrix $\widehat{\sigma}^2 (X^T X)^{-1}$.

**13.14 Example.** Crime data on 47 states in 1960 can be obtained from
http://lib.stat.cmu.edu/DASL/Stories/USCrime.html.
If we fit a linear regression of crime rate on 10 variables we get the following:

| Covariate | $\widehat{\beta}_j$ | $\widehat{\mathsf{se}}(\widehat{\beta}_j)$ | t value | p-value |
|---|---|---|---|---|
| (Intercept) | -589.39 | 167.59 | -3.51 | 0.001 ** |
| Age | 1.04 | 0.45 | 2.33 | 0.025 * |
| Southern State | 11.29 | 13.24 | 0.85 | 0.399 |
| Education | 1.18 | 0.68 | 1.7 | 0.093 |
| Expenditures | 0.96 | 0.25 | 3.86 | 0.000 *** |
| Labor | 0.11 | 0.15 | 0.69 | 0.493 |
| Number of Males | 0.30 | 0.22 | 1.36 | 0.181 |
| Population | 0.09 | 0.14 | 0.65 | 0.518 |
| Unemployment (14–24) | -0.68 | 0.48 | -1.4 | 0.165 |
| Unemployment (25–39) | 2.15 | 0.95 | 2.26 | 0.030 * |
| Wealth | -0.08 | 0.09 | -0.91 | 0.367 |

This table is typical of the output of a multiple regression program. The "t-value" is the Wald test statistic for testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. The asterisks denote "degree of significance" and more asterisks denote smaller p-values. The example raises several important questions: (1) should we eliminate some variables from this model? (2) should we interpret these relationships as causal? For example, should we conclude that low crime prevention expenditures cause high crime rates? We will address question (1) in the next section. We will not address question (2) until Chapter 16. ■

## 13.6 Model Selection

Example 13.14 illustrates a problem that often arises in multiple regression. We may have data on many covariates but we may not want to include all of them in the model. A smaller model with fewer covariates has two advantages: it might give better predictions than a big model and it is more parsimonious (simpler). Generally, as you add more variables to a regression, the bias of the predictions decreases and the variance increases. Too few covariates yields high bias; this called **underfitting**. Too many covariates yields high variance; this called **overfitting**. Good predictions result from achieving a good balance between bias and variance.

In model selection there are two problems: (i) assigning a "score" to each model which measures, in some sense, how good the model is, and (ii) searching through all the models to find the model with the best score.

Let us first discuss the problem of scoring models. Let $S \subset \{1, \ldots, k\}$ and let $\mathcal{X}_S = \{X_j : \ j \in S\}$ denote a subset of the covariates. Let $\beta_S$ denote the coefficients of the corresponding set of covariates and let $\widehat{\beta}_S$ denote the least squares estimate of $\beta_S$. Also, let $X_S$ denote the $X$ matrix for this subset of

covariates and define $\widehat{r}_S(x)$ to be the estimated regression function. The predicted values from model $S$ are denoted by $\widehat{Y}_i(S) = \widehat{r}_S(X_i)$. The **prediction risk** is defined to be

$$R(S) = \sum_{i=1}^{n} \mathbb{E}(\widehat{Y}_i(S) - Y_i^*)^2 \tag{13.24}$$

where $Y_i^*$ denotes the value of a future observation of $Y_i$ at covariate value $X_i$. Our goal is to choose $S$ to make $R(S)$ small.

The **training error** is defined to be

$$\widehat{R}_{\text{tr}}(S) = \sum_{i=1}^{n} (\widehat{Y}_i(S) - Y_i)^2.$$

This estimate is very biased as an estimate of $R(S)$.

**13.15 Theorem.** *The training error is a downward-biased estimate of the prediction risk:*

$$\mathbb{E}(\widehat{R}_{\text{tr}}(S)) < R(S).$$

*In fact,*

$$\mathsf{bias}(\widehat{R}_{\text{tr}}(S)) = \mathbb{E}(\widehat{R}_{\text{tr}}(S)) - R(S) = -2\sum_{i=1} \mathsf{Cov}(\widehat{Y}_i, Y_i). \tag{13.25}$$

The reason for the bias is that the data are being used twice: to estimate the parameters and to estimate the risk. When we fit a complex model with many parameters, the covariance $\mathsf{Cov}(\widehat{Y}_i, Y_i)$ will be large and the bias of the training error gets worse. Here are some better estimates of risk.

**Mallow's $C_p$ statistic** is defined by

$$\widehat{R}(S) = \widehat{R}_{\text{tr}}(S) + 2|S|\widehat{\sigma}^2 \tag{13.26}$$

where $|S|$ denotes the number of terms in $S$ and $\widehat{\sigma}^2$ is the estimate of $\sigma^2$ obtained from the full model (with all covariates in the model). This is simply the training error plus a bias correction. This estimate is named in honor of Colin Mallows who invented it. The first term in (13.26) measures the fit of the model while the second measure the complexity of the model. Think of the $C_p$ statistic as:

$$\text{lack of fit } + \text{ complexity penalty.}$$

Thus, **finding a good model involves trading off fit and complexity.**

A related method for estimating risk is **AIC (Akaike Information Criterion).** The idea is to choose $S$ to maximize

$$\ell_S - |S| \tag{13.27}$$

where $\ell_S$ is the log-likelihood of the model evaluated at the MLE. [3] This can be thought of "goodness of fit" minus "complexity." In linear regression with Normal errors (and taking $\sigma$ equal to its estimate from the largest model), maximizing AIC is equivalent to minimizing Mallow's $C_p$; see Exercise 8. The appendix contains more explanation about AIC.

Yet another method for estimating risk is **leave-one-out cross-validation**. In this case, the risk estimator is

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2 \tag{13.28}$$

where $\widehat{Y}_{(i)}$ is the prediction for $Y_i$ obtained by fitting the model with $Y_i$ omitted. It can be shown that

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)}\right)^2 \tag{13.29}$$

where $U_{ii}(S)$ is the $i^{\text{th}}$ diagonal element of the matrix

$$U(S) = X_S(X_S^T X_S)^{-1} X_S^T. \tag{13.30}$$

Thus, one need not actually drop each observation and re-fit the model. A generalization is **k-fold cross-validation.** Here we divide the data into $k$ groups; often people take $k = 10$. We omit one group of data and fit the models to the remaining data. We use the fitted model to predict the data in the group that was omitted. We then estimate the risk by $\sum_i (Y_i - \widehat{Y}_i)^2$ where the sum is over the the data points in the omitted group. This process is repeated for each of the $k$ groups and the resulting risk estimates are averaged.

For linear regression, Mallows $C_p$ and cross-validation often yield essentially the same results so one might as well use Mallows' method. In some of the more complex problems we will discuss later, cross-validation will be more useful.

Another scoring method is BIC (Bayesian information criterion). Here we choose a model to maximize

$$\text{BIC}(S) = \ell_S - \frac{|S|}{2}\log n. \tag{13.31}$$

[3] Some texts use a slightly different definition of AIC which involves multiplying the definition here by 2 or -2. This has no effect on which model is selected.

The BIC score has a Bayesian interpretation. Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote a set of models. Suppose we assign the prior $\mathbb{P}(S_j) = 1/m$ over the models. Also, assume we put a smooth prior on the parameters within each model. It can be shown that the posterior probability for a model is approximately,

$$\mathbb{P}(S_j|\text{data}) \approx \frac{e^{BIC(S_j)}}{\sum_r e^{BIC(S_r)}}.$$

Hence, choosing the model with highest BIC is like choosing the model with highest posterior probability. The BIC score also has an information-theoretic interpretation in terms of something called minimum description length. The BIC score is identical to Mallows $C_p$ except that it puts a more severe penalty for complexity. It thus leads one to choose a smaller model than the other methods.

Now let us turn to the problem of model search. If there are $k$ covariates then there are $2^k$ possible models. We need to search through all these models, assign a score to each one, and choose the model with the best score. If $k$ is not too large we can do a complete search over all the models. When $k$ is large, this is infeasible. In that case we need to search over a subset of all the models. Two common methods are **forward and backward stepwise regression.** In forward stepwise regression, we start with no covariates in the model. We then add the one variable that leads to the best score. We continue adding variables one at a time until the score does not improve. Backwards stepwise regression is the same except that we start with the biggest model and drop one variable at a time. Both are greedy searches; nether is guaranteed to find the model with the best score. Another popular method is to do random searching through the set of all models. However, there is no reason to expect this to be superior to a deterministic search.

**13.16 Example.** We applied backwards stepwise regression to the crime data using AIC. The following was obtained from the program R. This program uses a slightly different definition of AIC. With their definition, we seek the smallest (not largest) possible AIC. This is the same is minimizing Mallows $C_p$.

The full model (which includes all covariates) has AIC= 310.37. In ascending order, the AIC scores for deleting one variable are as follows:

| variable | Pop | Labor | South | Wealth | Males | U1 | Educ. | U2 | Age | Expend |
|---|---|---|---|---|---|---|---|---|---|---|
| AIC | 308 | 309 | 309 | 309 | 310 | 310 | 312 | 314 | 315 | 324 |

For example, if we dropped Pop from the model and kept the other terms, then the AIC score would be 308. Based on this information we drop "pop-

ulation" from the model and the current AIC score is 308. Now we consider dropping a variable from the current model. The AIC scores are:

| variable | South | Labor | Wealth | Males | U1 | Education | U2 | Age | Expend |
|---|---|---|---|---|---|---|---|---|---|
| AIC | 308 | 308 | 308 | 309 | 309 | 310 | 313 | 313 | 329 |

We then drop "Southern" from the model. This process is continued until there is no gain in AIC by dropping any variables. In the end, we are left with the following model:

$$\begin{aligned}\text{Crime} \;=\; & 1.2\ \text{Age} \;+\; .75\ \text{Education} \;+\; .87\ \text{Expenditure} \\ & +\; .34\ \text{Males} \;-\; .86\ \text{U1} \;+\; 2.31\ \text{U2}.\end{aligned}$$

**Warning!** This does not yet address the question of which variables are **causes** of crime. ■

There is another method for model selection that avoids having to search through all possible models. This method, which is due to Zheng and Loh (1995), does not seek to minimize prediction errors. Rather, it assumes some subset of the $\beta_j$'s are exactly equal to 0 and tries to find the true model, that is, the smallest sub-model consisting of nonzero $\beta_j$ terms. The method is carried out as follows.

**Zheng-Loh Model Selection Method** [4]

1. Fit the full model with all $k$ covariates and let $W_j = \widehat{\beta}_j / \widehat{\mathsf{se}}(\widehat{\beta}_j)$ denote the Wald test statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
2. Order the test statistics from largest to smallest in absolute value:
$$|W_{(1)}| \geq |W_{(2)}| \geq \cdots \geq |W_{(k)}|.$$
3. Let $\widehat{j}$ be the value of $j$ that minimizes
$$\text{RSS}(j) + j\,\widehat{\sigma}^2\,\log n$$
where $\text{RSS}(j)$ is the residual sums of squares from the model with the $j$ largest Wald statistics.
4. Choose, as the final model, the regression with the $\widehat{j}$ terms with the largest absolute Wald statistics.

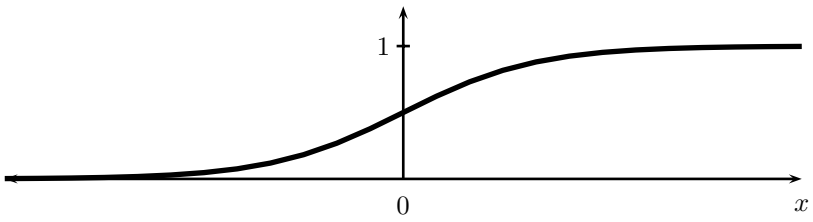



FIGURE 13.3. The logistic function $p = e^x/(1+e^x)$.

Zheng and Loh showed that, under appropriate conditions, this method chooses the true model with probability tending to one as the sample size increases.

## 13.7 Logistic Regression

So far we have assumed that $Y_i$ is real valued. **Logistic regression** is a parametric method for regression when $Y_i \in \{0, 1\}$ is binary. For a $k$-dimensional covariate $X$, the model is

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1|X = x) = \frac{e^{\beta_0+\sum_{j=1}^k \beta_j x_{ij}}}{1 + e^{\beta_0+\sum_{j=1}^k \beta_j x_{ij}}} \tag{13.32}$$

or, equivalently,

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \tag{13.33}$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{13.34}$$

The name "logistic regression" comes from the fact that $e^x/(1+e^x)$ is called the logistic function. A plot of the logistic for a one-dimensional covariate is shown in Figure 13.3.

Because the $Y_i$'s are binary, the data are Bernoulli:

$$Y_i|X_i = x_i \sim \text{Bernoulli}(p_i).$$

Hence the (conditional) likelihood function is

$$\mathcal{L}(\beta) = \prod_{i=1}^n p_i(\beta)^{Y_i}(1 - p_i(\beta))^{1-Y_i}. \tag{13.35}$$

[4]This is just one version of their method. In particular, the penalty $j \log n$ is only one choice from a set of possible penalty functions.

The MLE $\widehat{\beta}$ has to be obtained by maximizing $\mathcal{L}(\beta)$ numerically. There is a fast numerical algorithm called reweighted least squares. The steps are as follows:

**Reweighted Least Squares Algorithm**

Choose starting values $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$ and compute $p_i^0$ using equation (13.32), for $i = 1, \ldots, n$. Set $s = 0$ and iterate the following steps until convergence.

1. Set
$$Z_i = \text{logit}(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1-p_i^s)}, \quad i = 1, \ldots, n.$$
2. Let $W$ be a diagonal matrix with $(i,i)$ element equal to $p_i^s(1-p_i^s)$.
3. Set
$$\widehat{\beta}^s = (X^T W X)^{-1} X^T W Y.$$
This corresponds to doing a (weighted) linear regression of $Z$ on $Y$.
4. Set $s = s + 1$ and go back to the first step.

The Fisher information matrix $I$ can also be obtained numerically. The estimate standard error of $\widehat{\beta}_j$ is the $(j,j)$ element of $J = I^{-1}$. Model selection is usually done using the AIC score $\ell_S - |S|$.

**13.17 Example.** The Coronary Risk-Factor Study (CORIS) data involve 462 males between the ages of 15 and 64 from three rural areas in South Africa, (Rousseauw et al. (1983)). The outcome $Y$ is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease. There are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. A logistic regression yields the following estimates and Wald statistics $W_j$ for the coefficients:

| Covariate | $\widehat{\beta}_j$ | $\widehat{\mathsf{se}}$ | $W_j$ | p-value |
|---|---|---|---|---|
| Intercept | -6.145 | 1.300 | -4.738 | 0.000 |
| sbp | 0.007 | 0.006 | 1.138 | 0.255 |
| tobacco | 0.079 | 0.027 | 2.991 | 0.003 |
| ldl | 0.174 | 0.059 | 2.925 | 0.003 |
| adiposity | 0.019 | 0.029 | 0.637 | 0.524 |
| famhist | 0.925 | 0.227 | 4.078 | 0.000 |
| typea | 0.040 | 0.012 | 3.233 | 0.001 |
| obesity | -0.063 | 0.044 | -1.427 | 0.153 |
| alcohol | 0.000 | 0.004 | 0.027 | 0.979 |
| age | 0.045 | 0.012 | 3.754 | 0.000 |

Are you surprised by the fact that systolic blood pressure is not significant or by the minus sign for the obesity coefficient? If yes, then you are confusing association and causation. This issue is discussed in Chapter 16. The fact that blood pressure is not significant does not mean that blood pressure is not an important *cause* of heart disease. It means that it is not an important *predictor* of heart disease relative to the other variables in the model. ■

## 13.8 Bibliographic Remarks

A succinct book on linear regression is Weisberg (1985). A data-mining view of regression is given in Hastie et al. (2001). The Akaike Information Criterion (AIC) is due to Akaike (1973). The Bayesian Information Criterion (BIC) is due to Schwarz (1978). References on logistic regression include Agresti (1990) and Dobson (2001).

## 13.9 Appendix

The Akaike Information Criterion (AIC). Consider a set of models $\{M_1, M_2, \ldots\}$. Let $\widehat{f}_j(x)$ denote the estimated probability function obtained by using the maximum likelihood estimator of model $M_j$. Thus, $\widehat{f}_j(x) = \widehat{f}(x; \widehat{\beta}_j)$ where $\widehat{\beta}_j$ is the MLE of the set of parameters $\beta_j$ for model $M_j$. We will use the loss function $D(f, \widehat{f})$ where

$$D(f, g) = \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right)$$

is the Kullback-Leibler distance between two probability functions. The corresponding risk function is $R(f, \widehat{f}) = \mathbb{E}(D(f, \widehat{f})$. Notice that $D(f, \widehat{f}) = c -$

$A(f, \widehat{f})$ where $c = \sum_x f(x) \log f(x)$ does not depend on $\widehat{f}$ and

$$A(f, \widehat{f}) = \sum_x f(x) \log \widehat{f}(x).$$

Thus, minimizing the risk is equivalent to maximizing $a(f, \widehat{f}) \equiv \mathbb{E}(A(f, \widehat{f}))$.

It is tempting to estimate $a(f, \widehat{f})$ by $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ but, just as the training error in regression is a highly biased estimate of prediction risk, it is also the case that $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ is a highly biased estimate of $a(f, \widehat{f})$. In fact, the bias is approximately equal to $|M_j|$. Thus:

**13.18 Theorem.** *$AIC(M_j)$ is an approximately unbiased estimate of $a(f, \widehat{f})$.*

## 13.10 Exercises

1. Prove Theorem 13.4.

2. Prove the formulas for the standard errors in Theorem 13.8. You should regard the $X_i$'s as fixed constants.

3. Consider the **regression through the origin** model:

   $$Y_i = \beta X_i + \epsilon.$$

   Find the least squares estimate for $\beta$. Find the standard error of the estimate. Find conditions that guarantee that the estimate is consistent.

4. Prove equation (13.25).

5. In the simple linear regression model, construct a Wald test for $H_0 : \beta_1 = 17\beta_0$ versus $H_1 : \beta_1 \neq 17\beta_0$.

6. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a simple linear regression model to predict MPG (miles per gallon) from HP (horsepower). Summarize your analysis including a plot of the data with the fitted line.

   (b) Repeat the analysis but use log(MPG) as the response. Compare the analyses.

7. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a multiple linear regression model to predict MPG (miles per gallon) from the other variables. Summarize your analysis.

   (b) Use Mallow $C_p$ to select a best sub-model. To search through the models try (i) forward stepwise, (ii) backward stepwise. Summarize your findings.

   (c) Use the Zheng-Loh model selection method and compare to (b).

   (d) Perform all possible regressions. Compare $C_p$ and BIC. Compare the results.

8. Assume a linear regression model with Normal errors. Take $\sigma$ known. Show that the model with highest AIC (equation (13.27)) is the model with the lowest Mallows $C_p$ statistic.

9. In this question we will take a closer look at the AIC method. Let $X_1, \ldots, X_n$ be IID observations. Consider two models $\mathcal{M}_0$ and $\mathcal{M}_1$. Under $\mathcal{M}_0$ the data are assumed to be $N(0,1)$ while under $\mathcal{M}_1$ the data are assumed to be $N(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$:

$$\begin{aligned} \mathcal{M}_0 : \; X_1, \ldots, X_n &\sim N(0,1) \\ \mathcal{M}_1 : \; X_1, \ldots, X_n &\sim N(\theta, 1), \quad \theta \in \mathbb{R}. \end{aligned}$$

   This is just another way to view the hypothesis testing problem: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let $\ell_n(\theta)$ be the log-likelihood function. The AIC score for a model is the log-likelihood at the MLE minus the number of parameters. (Some people multiply this score by 2 but that is irrelevant.) Thus, the AIC score for $\mathcal{M}_0$ is $AIC_0 = \ell_n(0)$ and the AIC score for $\mathcal{M}_1$ is $AIC_1 = \ell_n(\widehat{\theta}) - 1$. Suppose we choose the model with the highest AIC score. Let $J_n$ denote the selected model:

$$J_n = \begin{cases} 0 & \text{if } AIC_0 > AIC_1 \\ 1 & \text{if } AIC_1 > AIC_0. \end{cases}$$

   (a) Suppose that $\mathcal{M}_0$ is the true model, i.e. $\theta = 0$. Find

$$\lim_{n \to \infty} \mathbb{P}\left(J_n = 0\right).$$

   Now compute $\lim_{n\to\infty} \mathbb{P}\left(J_n = 0\right)$ when $\theta \neq 0$.

(b) The fact that $\lim_{n\to\infty} \mathbb{P}(J_n = 0) \neq 1$ when $\theta = 0$ is why some people say that AIC "overfits." But this is not quite true as we shall now see. Let $\phi_\theta(x)$ denote a Normal density function with mean $\theta$ and variance 1. Define

$$\widehat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } J_n = 0 \\ \phi_{\widehat{\theta}}(x) & \text{if } J_n = 1. \end{cases}$$

If $\theta = 0$, show that $D(\phi_0, \widehat{f}_n) \xrightarrow{p} 0$ as $n \to \infty$ where

$$D(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

is the Kullback-Leibler distance. Show also that $D(\phi_\theta, \widehat{f}_n) \xrightarrow{p} 0$ if $\theta \neq 0$. Hence, AIC consistently estimates the true density even if it "overshoots" the correct model.

(c) Repeat this analysis for BIC which is the log-likelihood minus $(p/2)\log n$ where $p$ is the number of parameters and $n$ is sample size.

10. In this question we take a closer look at prediction intervals. Let $\theta = \beta_0 + \beta_1 X_*$ and let $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$. Thus, $\widehat{Y}_* = \widehat{\theta}$ while $Y_* = \theta + \epsilon$. Now, $\widehat{\theta} \approx N(\theta, \mathsf{se}^2)$ where

$$\mathsf{se}^2 = \mathbb{V}(\widehat{\theta}) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*).$$

Note that $\mathbb{V}(\widehat{\theta})$ is the same as $\mathbb{V}(\widehat{Y}_*)$. Now, $\widehat{\theta} \pm 2\sqrt{\mathbb{V}(\widehat{\theta})}$ is an approximate 95 percent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual argument for a confidence interval. But, as you shall now show, it is not a valid confidence interval for $Y_*$.

(a) Let $s = \sqrt{\mathbb{V}(\widehat{Y}_*)}$. Show that

$$\begin{aligned} \mathbb{P}(\widehat{Y}_* - 2s < Y_* < \widehat{Y}_* + 2s) &\approx \mathbb{P}\left(-2 < N\left(0, 1 + \frac{\sigma^2}{s^2}\right) < 2\right) \\ &\neq 0.95. \end{aligned}$$

(b) The problem is that the quantity of interest $Y_*$ is equal to a parameter $\theta$ plus a random variable. We can fix this by defining

$$\xi_n^2 = \mathbb{V}(\widehat{Y}_*) + \sigma^2 = \left[\frac{\sum_i (x_i - x_*)^2}{n\sum_i (x_i - \overline{x})^2} + 1\right]\sigma^2.$$

In practice, we substitute $\widehat{\sigma}$ for $\sigma$ and we denote the resulting quantity by $\widehat{\xi}_n$. Now consider the interval $\widehat{Y}_* \pm 2\,\widehat{\xi}_n$. Show that

$$\mathbb{P}(\widehat{Y}_* - 2\widehat{\xi}_n < Y_* < \widehat{Y}_* + 2\widehat{\xi}_n) \approx \mathbb{P}(-2 < N(0,1) < 2) \approx 0.95.$$

11. Get the Coronary Risk-Factor Study (CORIS) data from the book web site. Use backward stepwise logistic regression based on AIC to select a model. Summarize your results.

# 14
# Multivariate Models

In this chapter we revisit the Multinomial model and the multivariate Normal. Let us first review some notation from linear algebra. In what follows, $x$ and $y$ are vectors and $A$ is a matrix.

Linear Algebra Notation

| | |
|---|---|
| $x^T y$ | inner product $\sum_j x_j y_j$ |
| $\|A\|$ | determinant |
| $A^T$ | transpose of $A$ |
| $A^{-1}$ | inverse of $A$ |
| $I$ | the identity matrix |
| $\text{tr}(A)$ | trace of a square matrix; sum of its diagonal elements |
| $A^{1/2}$ | square root matrix |

The trace satisfies $\text{tr}(AB) = \text{tr}(BA)$ and $\text{tr}(A) + \text{tr}(B)$. Also, $\text{tr}(a) = a$ if $a$ is a scalar. A matrix is **positive definite** if $x^T \Sigma x > 0$ for all nonzero vectors $x$. If a matrix $A$ is symmetric and positive definite, its square root $A^{1/2}$ exists and has the following properties: (1) $A^{1/2}$ is symmetric; (2) $A = A^{1/2} A^{1/2}$; (3) $A^{1/2} A^{-1/2} = A^{-1/2} A^{1/2} = I$ where $A^{-1/2} = (A^{1/2})^{-1}$.

## 14.1 Random Vectors

Multivariate models involve a random vector $X$ of the form

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.$$

The mean of a random vector $X$ is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix}. \tag{14.1}$$

The **covariance matrix** $\Sigma$, also written $\mathbb{V}(X)$, is defined to be

$$\Sigma = \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_k) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_k) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_k, X_1) & \mathsf{Cov}(X_k, X_2) & \cdots & \mathbb{V}(X_k) \end{bmatrix}. \tag{14.2}$$

This is also called the variance matrix or the variance–covariance matrix. The inverse $\Sigma^{-1}$ is called the **precision matrix**.

**14.1 Theorem.** *Let $a$ be a vector of length $k$ and let $X$ be a random vector of the same length with mean $\mu$ and variance $\Sigma$. Then $\mathbb{E}(a^T X) = a^T \mu$ and $\mathbb{V}(a^T X) = a^T \Sigma a$. If $A$ is a matrix with $k$ columns, then $\mathbb{E}(AX) = A\mu$ and $\mathbb{V}(AX) = A\Sigma A^T$.*

Now suppose we have a random sample of $n$ vectors:

$$\begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{k1} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{k2} \end{pmatrix}, \ \ldots, \ \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{kn} \end{pmatrix}. \tag{14.3}$$

The sample mean $\overline{X}$ is a vector defined by

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_k \end{pmatrix}$$

where $\overline{X}_i = n^{-1}\sum_{j=1}^n X_{ij}$. The sample variance matrix, also called the covariance matrix or the variance–covariance matrix, is

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{12} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{1k} & s_{2k} & \cdots & s_{kk} \end{bmatrix} \tag{14.4}$$

where

$$s_{ab} = \frac{1}{n-1}\sum_{j=1}^n (X_{aj} - \overline{X}_a)(X_{bj} - \overline{X}_b).$$

It follows that $\mathbb{E}(\overline{X}) = \mu$. and $\mathbb{E}(S) = \Sigma$.

## 14.2 Estimating the Correlation

Consider $n$ data points from a bivariate distribution:

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \cdots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}.$$

Recall that the correlation between $X_1$ and $X_2$ is

$$\rho = \frac{\mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2))}{\sigma_1 \sigma_2} \tag{14.5}$$

where $\sigma_j^2 = \mathbb{V}(X_{ji})$, $j = 1, 2$. The nonparametric plug-in estimator is the sample correlation [1]

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{s_1 s_2} \tag{14.6}$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^n (X_{ji} - \overline{X}_j)^2.$$

We can construct a confidence interval for $\rho$ by applying the delta method. However, it turns out that we get a more accurate confidence interval by first constructing a confidence interval for a function $\theta = f(\rho)$ and then applying

[1] More precisely, the plug-in estimator has $n$ rather than $n-1$ in the formula for $s_j$ but this difference is small.

the inverse function $f^{-1}$. The method, due to Fisher, is as follows: Define $f$ and its inverse by

$$\begin{aligned} f(r) &= \frac{1}{2}\Big(\log(1+r) - \log(1-r)\Big) \\ f^{-1}(z) &= \frac{e^{2z}-1}{e^{2z}+1}. \end{aligned}$$

**Approximate Confidence Interval for The Correlation**

1. Compute
$$\widehat{\theta} = f(\widehat{\rho}) = \frac{1}{2}\Big(\log(1+\widehat{\rho}) - \log(1-\widehat{\rho})\Big).$$
2. Compute the approximate standard error of $\widehat{\theta}$ which can be shown to be
$$\widehat{\mathsf{se}}(\widehat{\theta}) = \frac{1}{\sqrt{n-3}}.$$
3. An approximate $1-\alpha$ confidence interval for $\theta = f(\rho)$ is
$$(a,b) \equiv \left(\widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, \widehat{\theta} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right).$$
4. Apply the inverse transformation $f^{-1}(z)$ to get a confidence interval for $\rho$:
$$\left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right).$$

Yet another method for getting a confidence interval for $\rho$ is to use the bootstrap.

## 14.3 Multivariate Normal

Recall that a vector $X$ has a multivariate Normal distribution, denoted by $X \sim N(\mu, \Sigma)$, if its density is

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{14.7}$$

where $\mu$ is a vector of length $k$ and $\Sigma$ is a $k \times k$ symmetric, positive definite matrix. Then $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \Sigma$.

**14.2 Theorem.** *The following properties hold:*

1. *If $Z \sim N(0,1)$ and $X = \mu + \Sigma^{1/2} Z$, then $X \sim N(\mu, \Sigma)$.*
2. *If $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0,1)$.*
3. *If $X \sim N(\mu, \Sigma)$ $a$ is a vector of the same length as $X$, then $a^T X \sim N(a^T \mu, a^T \Sigma a)$.*
4. *Let*
$$V = (X - \mu)^T \Sigma^{-1} (X - \mu).$$
*Then $V \sim \chi^2_k$.*

**14.3 Theorem.** *Given a random sample of size $n$ from a $N(\mu, \Sigma)$, the log-likelihood is (up to a constant not depending on $\mu$ or $\Sigma$) given by*

$$\ell(\mu, \Sigma) = -\frac{n}{2}(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) - \frac{n}{2} \text{tr}(\Sigma^{-1} S) - \frac{n}{2} \log |\Sigma|.$$

*The* MLE *is*

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\Sigma} = \left( \frac{n-1}{n} \right) S. \tag{14.8}$$

## 14.4 Multinomial

Let us now review the Multinomial distribution. The data take the form $X = (X_1, \ldots, X_k)$ where each $X_j$ is a count. Think of drawing $n$ balls (with replacement) from an urn which has balls with $k$ different colors. In this case, $X_j$ is the number of balls of the $k^{\text{th}}$ color. Let $p = (p_1, \ldots, p_k)$ where $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$ and suppose that $p_j$ is the probability of drawing a ball of color $j$.

**14.4 Theorem.** *Let $X \sim$ Multinomial$(n, p)$. Then the marginal distribution of $X_j$ is $X_j \sim$ Binomial$(n, p_j)$. The mean and variance of $X$ are*

$$\mathbb{E}(X) = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$$

*and*

$$\mathbb{V}(X) = \begin{pmatrix} np_1(1-p_1) & -np_1 p_2 & \cdots & -np_1 p_k \\ -np_1 p_2 & np_2(1-p_2) & \cdots & -np_2 p_k \\ \vdots & \vdots & \vdots & \vdots \\ -np_1 p_k & -np_2 p_k & \cdots & np_k(1-p_k) \end{pmatrix}.$$

PROOF. That $X_j \sim \text{Binomial}(n, p_j)$ follows easily. Hence, $\mathbb{E}(X_j) = np_j$ and $\mathbb{V}(X_j) = np_j(1-p_j)$. To compute $\mathsf{Cov}(X_i, X_j)$ we proceed as follows: Notice that $X_i + X_j \sim \text{Binomial}(n, p_i + p_j)$ and so $\mathbb{V}(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j)$. On the other hand,

$$\begin{aligned} \mathbb{V}(X_i + X_j) &= \mathbb{V}(X_i) + \mathbb{V}(X_j) + 2\mathsf{Cov}(X_i, X_j) \\ &= np_i(1-p_i) + np_j(1-p_j) + 2\mathsf{Cov}(X_i, X_j). \end{aligned}$$

Equating this last expression with $n(p_i + p_j)(1 - p_i - p_j)$ implies that $\mathsf{Cov}(X_i, X_j) = -np_ip_j$. ∎

**14.5 Theorem.** *The maximum likelihood estimator of $p$ is*

$$\widehat{p} = \begin{pmatrix} \widehat{p}_1 \\ \vdots \\ \widehat{p}_k \end{pmatrix} = \begin{pmatrix} \frac{X_1}{n} \\ \vdots \\ \frac{X_k}{n} \end{pmatrix} = \frac{X}{n}.$$

PROOF. The log-likelihood (ignoring a constant) is

$$\ell(p) = \sum_{j=1}^{k} X_j \log p_j.$$

When we maximize $\ell$ we have to be careful since we must enforce the constraint that $\sum_j p_j = 1$. We use the method of Lagrange multipliers and instead maximize

$$A(p) = \sum_{j=1}^{k} X_j \log p_j + \lambda\left(\sum_j p_j - 1\right).$$

Now

$$\frac{\partial A(p)}{\partial p_j} = \frac{X_j}{p_j} + \lambda.$$

Setting $\frac{\partial A(p)}{\partial p_j} = 0$ yields $\widehat{p}_j = -X_j/\lambda$. Since $\sum_j \widehat{p}_j = 1$ we see that $\lambda = -n$ and hence $\widehat{p}_j = X_j/n$ as claimed. ∎

Next we would like to know the variability of the MLE. We can either compute the variance matrix of $\widehat{p}$ directly or we can approximate the variability of the MLE by computing the Fisher information matrix. These two approaches give the same answer in this case. The direct approach is easy: $\mathbb{V}(\widehat{p}) = \mathbb{V}(X/n) = n^{-2}\mathbb{V}(X)$, and so

$$\mathbb{V}(\widehat{p}) = \frac{1}{n}\Sigma$$

where

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}.$$

For large $n$, $\widehat{p}$ has approximately a multivariate Normal distribution.

**14.6 Theorem.** *As $n \to \infty$,*

$$\sqrt{n}(\widehat{p} - p) \rightsquigarrow N(0, \Sigma).$$

## 14.5 Bibliographic Remarks

Some references on multivariate analysis are Johnson and Wichern (1982) and Anderson (1984). The method for constructing the confidence interval for the correlation described in this chapter is due to Fisher (1921).

## 14.6 Appendix

PROOF of Theorem 14.3. Denote the $i^{\text{th}}$ random vector by $X^i$. The log-likelihood is

$$\begin{aligned} \ell(\mu, \Sigma) &= \sum_{i=1}^{n} f(X^i;\ \mu, \Sigma) \\ &= -\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu). \end{aligned}$$

Now,

$$\begin{aligned} &\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu) \\ &\quad= \sum_{i=1}^{n}[(X^i - \overline{X}) + (\overline{X} - \mu)]^T\Sigma^{-1}[(X^i - \overline{X}) + (\overline{X} - \mu)] \\ &\quad= \sum_{i=1}^{n}\left[(X^i - \overline{X})^T\Sigma^{-1}(X^i - \overline{X})\right] + n(\overline{X} - \mu)^T\Sigma^{-1}(\overline{X} - \mu) \end{aligned}$$

since $\sum_{i=1}^{n}(X^i - \overline{X})\Sigma^{-1}(\overline{X} - \mu) = 0$. Also, notice that $(X^i - \mu)^T\Sigma^{-1}(X^i - \mu)$ is a scalar, so

$$\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu) = \sum_{i=1}^{n}\text{tr}\left[(X^i - \mu)^T\Sigma^{-1}(X^i - \mu)\right]$$

$$\begin{aligned} &= \sum_{i=1}^{n} \text{tr}\left[\Sigma^{-1}(X^i - \mu)(X^i - \mu)^T\right] \\ &= \text{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(X^i - \mu)(X^i - \mu)^T\right] \\ &= n \text{ tr}\left[\Sigma^{-1}S\right] \end{aligned}$$

and the conclusion follows. ■

## 14.7 Exercises

1. Prove Theorem 14.1.
2. Find the Fisher information matrix for the MLE of a Multinomial.
3. (Computer Experiment.) Write a function to generate `nsim` observations from a Multinomial$(n, p)$ distribution.
4. (Computer Experiment.) Write a function to generate `nsim` observations from a Multivariate normal with given mean $\mu$ and covariance matrix $\Sigma$.
5. (Computer Experiment.) Generate 100 random vectors from a $N(\mu, \Sigma)$ distribution where

$$\mu = \begin{pmatrix} 3 \\ 8 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

   Plot the simulation as a scatterplot. Estimate the mean and covariance matrix $\Sigma$. Find the correlation $\rho$ between $X_1$ and $X_2$. Compare this with the sample correlations from your simulation. Find a 95 percent confidence interval for $\rho$. Use two methods: the bootstrap and Fisher's method. Compare.
6. (Computer Experiment.) Repeat the previous exercise 1000 times. Compare the coverage of the two confidence intervals for $\rho$.

# 15
# Inference About Independence

In this chapter we address the following questions:

(1) How do we test if two random variables are independent?

(2) How do we estimate the strength of dependence between two random variables?

When $Y$ and $Z$ are not independent, we say that they are **dependent** or **associated** or **related.** If $Y$ and $Z$ are associated, it does **not** imply that $Y$ causes $Z$ or that $Z$ causes $Y$. Causation is discussed in Chapter 16.

Recall that we write $Y \amalg Z$ to mean that $Y$ and $Z$ are independent and we write $Y \not\amalg Z$ to mean that $Y$ and $Z$ are dependent.

## 15.1 Two Binary Variables

Suppose that $Y$ and $Z$ are both binary and consider data $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. We can represent the data as a two-by-two table:

| | $Y=0$ | $Y=1$ | |
|---|---|---|---|
| $Z=0$ | $X_{00}$ | $X_{01}$ | $X_{0\cdot}$ |
| $Z=1$ | $X_{10}$ | $X_{11}$ | $X_{1\cdot}$ |
| | $X_{\cdot 0}$ | $X_{\cdot 1}$ | $n = X_{\cdot\cdot}$ |

where

$$X_{ij} = \text{number of observations for which } Y = i \text{ and } Z = j.$$

The dotted subscripts denote sums. Thus,

$$X_{i\cdot} = \sum_j X_{ij}, \quad X_{\cdot j} = \sum_i X_{ij}, \quad n = X_{\cdot\cdot} = \sum_{i,j} X_{ij}.$$

This is a convention we use throughout the remainder of the book. Denote the corresponding probabilities by:

| | $Y=0$ | $Y=1$ | |
|---|---|---|---|
| $Z=0$ | $p_{00}$ | $p_{01}$ | $p_{0\cdot}$ |
| $Z=1$ | $p_{10}$ | $p_{11}$ | $p_{1\cdot}$ |
| | $p_{\cdot 0}$ | $p_{\cdot 1}$ | 1 |

where $p_{ij} = \mathbb{P}(Z = i, Y = j)$. Let $X = (X_{00}, X_{01}, X_{10}, X_{11})$ denote the vector of counts. Then $X \sim \text{Multinomial}(n, p)$ where $p = (p_{00}, p_{01}, p_{10}, p_{11})$. It is now convenient to introduce two new parameters.

**15.1 Definition.** *The* **odds ratio** *is defined to be*

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}. \tag{15.1}$$

*The* **log odds ratio** *is defined to be*

$$\gamma = \log(\psi). \tag{15.2}$$

**15.2 Theorem.** *The following statements are equivalent:*

1. $Y \amalg Z$.
2. $\psi = 1$.
3. $\gamma = 0$.
4. *For* $i, j \in \{0, 1\}$, $p_{ij} = p_{i\cdot}p_{\cdot j}$.

Now consider testing

$$H_0 : \ Y \amalg Z \ \text{ versus } H_1 : \ Y \not\amalg Z. \tag{15.3}$$

First we consider the likelihood ratio test. Under $H_1$, $X \sim \text{Multinomial}(n, p)$ and the MLE is the vector $\widehat{p} = X/n$. Under $H_0$, we again have that $X \sim \text{Multinomial}(n, p)$ but the restricted MLE is computed under the constraint $p_{ij} = p_{i\cdot}p_{\cdot j}$ This leads to the following test:

**15.3 Theorem.** *The likelihood ratio test statistic for (15.3) is*

$$T = 2\sum_{i=0}^{1}\sum_{j=0}^{1} X_{ij}\log\left(\frac{X_{ij}X_{\cdot\cdot}}{X_{i\cdot}X_{\cdot j}}\right). \tag{15.4}$$

*Under $H_0$, $T \rightsquigarrow \chi^2_1$. Thus, an approximate level $\alpha$ test is obtained by rejecting $H_0$ when $T > \chi^2_{1,\alpha}$.*

Another popular test for independence is Pearson's $\chi^2$ test.

**15.4 Theorem.** *Pearson's $\chi^2$ test statistic for independence is*

$$U = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(X_{ij} - E_{ij})^2}{E_{ij}} \tag{15.5}$$

*where*

$$E_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

*Under $H_0$, $U \rightsquigarrow \chi^2_1$. Thus, an approximate level $\alpha$ test is obtained by rejecting $H_0$ when $U > \chi^2_{1,\alpha}$.*

Here is the intuition for the Pearson test. Under $H_0$, $p_{ij} = p_{i\cdot}p_{\cdot j}$, so the maximum likelihood estimator of $p_{ij}$ under $H_0$ is

$$\widehat{p}_{ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}.$$

Thus, the expected number of observations in the (i,j) cell is

$$E_{ij} = n\widehat{p}_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

The statistic $U$ compares the observed and expected counts.

**15.5 Example.** The following data from Johnson and Johnson (1972) relate tonsillectomy and Hodgkins disease. [1]

| | Hodgkins Disease | No Disease | |
|---|---|---|---|
| Tonsillectomy | 90 | 165 | 255 |
| No Tonsillectomy | 84 | 307 | 391 |
| Total | 174 | 472 | 646 |

[1] The data are actually from a case-control study; see the appendix for an explanation of case-control studies.

We would like to know if tonsillectomy is related to Hodgkins disease. The likelihood ratio statistic is $T = 14.75$ and the p-value is $\mathbb{P}(\chi^2_1 > 14.75) = .0001$. The $\chi^2$ statistic is $U = 14.96$ and the p-value is $\mathbb{P}(\chi^2_1 > 14.96) = .0001$. We reject the null hypothesis of independence and conclude that tonsillectomy is associated with Hodgkins disease. This does not mean that tonsillectomies cause Hodgkins disease. Suppose, for example, that doctors gave tonsillectomies to the most seriously ill patients. Then the association between tonsillectomies and Hodgkins disease may be due to the fact that those with tonsillectomies were the most ill patients and hence more likely to have a serious disease. ■

We can also estimate the strength of dependence by estimating the odds ratio $\psi$ and the log-odds ratio $\gamma$.

**15.6 Theorem.** *The* MLE*'s of $\psi$ and $\gamma$ are*

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}, \quad \widehat{\gamma} = \log \widehat{\psi}. \tag{15.6}$$

*The asymptotic standard errors (computed using the delta method) are*

$$\widehat{\mathsf{se}}(\widehat{\gamma}) = \sqrt{\frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}} \tag{15.7}$$

$$\widehat{\mathsf{se}}(\widehat{\psi}) = \widehat{\psi}\, \widehat{\mathsf{se}}(\widehat{\gamma}). \tag{15.8}$$

**15.7 Remark.** For small sample sizes, $\widehat{\psi}$ and $\widehat{\gamma}$ can have a very large variance. In this case, we often use the modified estimator

$$\widehat{\psi} = \frac{\left(X_{00} + \frac{1}{2}\right)\left(X_{11} + \frac{1}{2}\right)}{\left(X_{01} + \frac{1}{2}\right)\left(X_{10} + \frac{1}{2}\right)}. \tag{15.9}$$

Another test for independence is the Wald test for $\gamma = 0$ given by $W = (\widehat{\gamma} - 0)/\widehat{\mathsf{se}}(\widehat{\gamma})$. A $1 - \alpha$ confidence interval for $\gamma$ is $\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\gamma})$.

A $1 - \alpha$ confidence interval for $\psi$ can be obtained in two ways. First, we could use $\widehat{\psi} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\psi})$. Second, since $\psi = e^{\gamma}$ we could use

$$\exp\left\{\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\gamma})\right\}. \tag{15.10}$$

This second method is usually more accurate.

**15.8 Example.** In the previous example,

$$\widehat{\psi} = \frac{90 \times 307}{165 \times 84} = 1.99$$

and

$$\widehat{\gamma} = \log(1.99) = .69.$$

So tonsillectomy patients were twice as likely to have Hodgkins disease. The standard error of $\widehat{\gamma}$ is

$$\sqrt{\frac{1}{90}+\frac{1}{84}+\frac{1}{165}+\frac{1}{307}} = .18.$$

The Wald statistic is $W = .69/.18 = 3.84$ whose p-value is $\mathbb{P}(|Z| > 3.84) = .0001$, the same as the other tests. A 95 per cent confidence interval for $\gamma$ is $\widehat{\gamma} \pm 2(.18) = (.33, 1.05)$. A 95 per cent confidence interval for $\psi$ is $(e^{.33}, e^{1.05}) = (1.39, 2.86)$. ■

## 15.2 Two Discrete Variables

Now suppose that $Y \in \{1, \ldots, I\}$ and $Z \in \{1, \ldots, J\}$ are two discrete variables. The data can be represented as an $I \times J$ table of counts:

| | $Y=1$ | $Y=2$ | $\cdots$ | $Y=j$ | $\cdots$ | $Y=J$ | |
|---|---|---|---|---|---|---|---|
| $Z=1$ | $X_{11}$ | $X_{12}$ | $\cdots$ | $X_{1j}$ | $\cdots$ | $X_{1J}$ | $X_{1\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Z=i$ | $X_{i1}$ | $X_{i2}$ | $\cdots$ | $X_{ij}$ | $\cdots$ | $X_{iJ}$ | $X_{i\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $Z=I$ | $X_{I1}$ | $X_{I2}$ | $\cdots$ | $X_{Ij}$ | $\cdots$ | $X_{IJ}$ | $X_{I\cdot}$ |
| | $X_{\cdot 1}$ | $X_{\cdot 2}$ | $\cdots$ | $X_{\cdot j}$ | $\cdots$ | $X_{\cdot J}$ | n |

where

$$X_{ij} = \text{number of observations for which } Z = i \text{ and } Y = j.$$

Consider testing

$$H_0 : Y \amalg Z \quad \text{versus} \quad H_1 : Y \not\amalg Z. \tag{15.11}$$

**15.9 Theorem.** *The likelihood ratio test statistic for (15.11) is*

$$T = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij} \log\left(\frac{X_{ij}\, X_{\cdot\cdot}}{X_{i\cdot}\, X_{\cdot j}}\right). \tag{15.12}$$

*The limiting distribution of $T$ under the null hypothesis of independence is $\chi^2_\nu$ where $\nu = (I-1)(J-1)$. Pearson's $\chi^2$ test statistic is*

$$U = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}. \tag{15.13}$$

*Asymptotically, under $H_0$, $U$ has a $\chi^2_\nu$ distribution where $\nu = (I-1)(J-1)$.*

**15.10 Example.** These data are from Dunsmore et al. (1987). Patients with Hodgkins disease are classified by their response to treatment and by histological type.

| Type | Positive Response | Partial Response | No Response | |
|---|---|---|---|---|
| LP | 74 | 18 | 12 | 104 |
| NS | 68 | 16 | 12 | 96 |
| MC | 154 | 54 | 58 | 266 |
| LD | 18 | 10 | 44 | 72 |

The $\chi^2$ test statistic is 75.89 with $2 \times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 75.89) \approx 0$. The likelihood ratio test statistic is 68.30 with $2 \times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 68.30) \approx 0$. Thus there is strong evidence that response to treatment and histological type are associated. ■

## 15.3 Two Continuous Variables

Now suppose that $Y$ and $Z$ are both continuous. If we assume that the joint distribution of $Y$ and $Z$ is bivariate Normal, then we measure the dependence between $Y$ and $Z$ by means of the correlation coefficient $\rho$. Tests, estimates, and confidence intervals for $\rho$ in the Normal case are given in the previous chapter in Section 14.2. If we do not assume Normality then we can still use the methods in Section 14.2 to draw inferences about the correlation $\rho$. However, if we conclude that $\rho$ is 0, we cannot conclude that $Y$ and $Z$ are independent, only that they are uncorrelated. Fortunately, the reverse direction is valid: if we conclude that $Y$ and $Z$ are correlated than we can conclude they are dependent.

## 15.4 One Continuous Variable and One Discrete

Suppose that $Y \in \{1, \ldots, I\}$ is discrete and $Z$ is continuous. Let $F_i(z) = \mathbb{P}(Z \leq z | Y = i)$ denote the CDF of $Z$ conditional on $Y = i$.

**15.11 Theorem.** *When $Y \in \{1, \ldots, I\}$ is discrete and $Z$ is continuous, then $Y \amalg Z$ if and only if $F_1 = \cdots = F_I$.*

It follows from the previous theorem that to test for independence, we need to test

$$H_0 : F_1 = \cdots = F_I \quad \text{versus} \quad H_1 : \text{not } H_0.$$

For simplicity, we consider the case where $I = 2$. To test the null hypothesis that $F_1 = F_2$ we will use the **two sample Kolmogorov-Smirnov test**. Let $n_1$ denote the number of observations for which $Y_i = 1$ and let $n_2$ denote the number of observations for which $Y_i = 2$. Let

$$\widehat{F}_1(z) = \frac{1}{n_1} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 1)$$

and

$$\widehat{F}_2(z) = \frac{1}{n_2} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 2)$$

denote the empirical distribution function of $Z$ given $Y = 1$ and $Y = 2$ respectively. Define the test statistic

$$D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)|.$$

**15.12 Theorem.** *Let*

$$H(t) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 t^2}. \tag{15.14}$$

*Under the null hypothesis that $F_1 = F_2$,*

$$\lim_{n \to \infty} \mathbb{P}\left( \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D \leq t \right) = H(t).$$

It follows from the theorem that an approximate level $\alpha$ test is obtained by rejecting $H_0$ when

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D > H^{-1}(1 - \alpha).$$

## 15.5 Appendix

Interpreting The Odds Ratios. Suppose event $A$ as probability $\mathbb{P}(A)$. The odds of $A$ are defined as $\text{odds}(A) = \mathbb{P}(A)/(1 - \mathbb{P}(A))$. It follows that

$\mathbb{P}(A) = \text{odds}(A)/(1 + \text{odds}(A))$. Let $E$ be the event that someone is exposed to something (smoking, radiation, etc) and let $D$ be the event that they get a disease. The odds of getting the disease given that you are exposed are:

$$\text{odds}(D|E) = \frac{\mathbb{P}(D|E)}{1 - \mathbb{P}(D|E)}$$

and the odds of getting the disease given that you are not exposed are:

$$\text{odds}(D|E^c) = \frac{\mathbb{P}(D|E^c)}{1 - \mathbb{P}(D|E^c)}.$$

The *odds ratio* is defined to be

$$\psi = \frac{\text{odds}(D|E)}{\text{odds}(D|E^c)}.$$

If $\psi = 1$ then disease probability is the same for exposed and unexposed. This implies that these events are independent. Recall that the log-odds ratio is defined as $\gamma = \log(\psi)$. Independence corresponds to $\gamma = 0$.

Consider this table of probabilities and corresponding table of data:

| | $D^c$ | $D$ | |
|---|---|---|---|
| $E^c$ | $p_{00}$ | $p_{01}$ | $p_{0\cdot}$ |
| $E$ | $p_{10}$ | $p_{11}$ | $p_{1\cdot}$ |
| | $p_{\cdot 0}$ | $p_{\cdot 1}$ | 1 |

| | $D^c$ | $D$ | |
|---|---|---|---|
| $E^c$ | $X_{00}$ | $X_{01}$ | $X_{0\cdot}$ |
| $E$ | $X_{10}$ | $X_{11}$ | $X_{1\cdot}$ |
| | $X_{\cdot 0}$ | $X_{\cdot 1}$ | $X_{\cdot\cdot}$ |

Now

$$\mathbb{P}(D|E) = \frac{p_{11}}{p_{10} + p_{11}} \quad \text{and} \quad \mathbb{P}(D|E^c) = \frac{p_{01}}{p_{00} + p_{01}},$$

and so

$$\text{odds}(D|E) = \frac{p_{11}}{p_{10}} \quad \text{and} \quad \text{odds}(D|E^c) = \frac{p_{01}}{p_{00}},$$

and therefore,

$$\psi = \frac{p_{11}p_{00}}{p_{01}p_{10}}.$$

To estimate the parameters, we have to first consider how the data were collected. There are three methods.

Multinomial Sampling. We draw a sample from the population and, for each person, record their exposure and disease status. In this case, $X = (X_{00}, X_{01}, X_{10}, X_{11}) \sim \text{Multinomial}(n, p)$. We then estimate the probabilities in the table by $\widehat{p}_{ij} = X_{ij}/n$ and

$$\widehat{\psi} = \frac{\widehat{p}_{11}\widehat{p}_{00}}{\widehat{p}_{01}\widehat{p}_{10}} = \frac{X_{11}X_{00}}{X_{01}X_{10}}.$$

Prospective Sampling. (Cohort Sampling). We get some exposed and unexposed people and count the number with disease in each group. Thus,

$$\begin{aligned} X_{01} &\sim \text{Binomial}(X_{0\cdot}, \mathbb{P}(D|E^c)) \\ X_{11} &\sim \text{Binomial}(X_{1\cdot}, \mathbb{P}(D|E)). \end{aligned}$$

We should really write $x_{0\cdot}$ and $x_{1\cdot}$ instead of $X_{0\cdot}$ and $X_{1\cdot}$ since in this case, these are fixed not random, but for notational simplicity I'll keep using capital letters. We can estimate $\mathbb{P}(D|E)$ and $\mathbb{P}(D|E^c)$ but we cannot estimate all the probabilities in the table. Still, we can estimate $\psi$ since $\psi$ is a function of $\mathbb{P}(D|E)$ and $\mathbb{P}(D|E^c)$. Now

$$\widehat{\mathbb{P}}(D|E) = \frac{X_{11}}{X_{1\cdot}} \quad \text{and} \quad \widehat{\mathbb{P}}(D|E^c) = \frac{X_{01}}{X_{0\cdot}}.$$

Thus,

$$\widehat{\psi} = \frac{X_{11}X_{00}}{X_{01}X_{10}}$$

just as before.

Case-Control (Retrospective) Sampling. Here we get some diseased and non-diseased people and we observe how many are exposed. This is much more efficient if the disease is rare. Hence,

$$\begin{aligned} X_{10} &\sim \text{Binomial}(X_{\cdot 0}, \mathbb{P}(E|D^c)) \\ X_{11} &\sim \text{Binomial}(X_{\cdot 1}, \mathbb{P}(E|D)). \end{aligned}$$

From these data we can estimate $\mathbb{P}(E|D)$ and $\mathbb{P}(E|D^c)$. Surprisingly, we can also still estimate $\psi$. To understand why, note that

$$\mathbb{P}(E|D) = \frac{p_{11}}{p_{01}+p_{11}}, \quad 1-\mathbb{P}(E|D) = \frac{p_{01}}{p_{01}+p_{11}}, \quad \text{odds}(E|D) = \frac{p_{11}}{p_{01}}.$$

By a similar argument,

$$\text{odds}(E|D^c) = \frac{p_{10}}{p_{00}}.$$

Hence,

$$\frac{\text{odds}(E|D)}{\text{odds}(E|D^c)} = \frac{p_{11}p_{00}}{p_{01}p_{10}} = \psi.$$

From the data, we form the following estimates:

$$\widehat{P}(E|D) = \frac{X_{11}}{X_{\cdot 1}}, \ 1-\widehat{P}(E|D) = \frac{X_{01}}{X_{\cdot 1}}, \ \widehat{\text{odds}}(E|D) = \frac{X_{11}}{X_{01}}, \ \widehat{\text{odds}}(E|D^c) = \frac{X_{10}}{X_{00}}.$$

Therefore,

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}.$$

So in all three data collection methods, the estimate of $\psi$ turns out to be the same.

It is tempting to try to estimate $\mathbb{P}(D|E) - \mathbb{P}(D|E^c)$. In a case-control design, this quantity is not estimable. To see this, we apply Bayes' theorem to get

$$\mathbb{P}(D|E) - \mathbb{P}(D|E^c) = \frac{\mathbb{P}(E|D)\mathbb{P}(D)}{\mathbb{P}(E)} - \frac{\mathbb{P}(E^c|D)\mathbb{P}(D)}{\mathbb{P}(E^c)}.$$

Because of the way we obtained the data, $\mathbb{P}(D)$ is not estimable from the data. However, we can estimate $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$, which is called the **relative risk**, under the **rare disease assumption.**

**15.13 Theorem.** *Let* $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$. *Then*

$$\frac{\psi}{\xi} \to 1$$

*as* $\mathbb{P}(D) \to 0$.

Thus, under the rare disease assumption, the relative risk is approximately the same as the odds ratio and, as we have seen, we can estimate the odds ratio.

## 15.6 Exercises

1. Prove Theorem 15.2.
2. Prove Theorem 15.3.
3. Prove Theorem 15.6.
4. The *New York Times* (January 8, 2003, page A12) reported the following data on death sentencing and race, from a study in Maryland: [2]

| | Death Sentence | No Death Sentence |
|---|---|---|
| Black Victim | 14 | 641 |
| White Victim | 62 | 594 |

   Analyze the data using the tools from this chapter. Interpret the results. Explain why, based only on this information, you can't make causal conclusions. (The authors of the study did use much more information in their full report.)

[2] The data here are an approximate re-creation using the information in the article.

5. Analyze the data on the variables Age and Financial Status from:

   http://lib.stat.cmu.edu/DASL/Datafiles/montanadat.html

6. Estimate the correlation between temperature and latitude using the data from

   http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html

   Use the correlation coefficient. Provide estimates, tests, and confidence intervals.

7. Test whether calcium intake and drop in blood pressure are associated. Use the data in

   http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html

# 16
# Causal Inference

Roughly speaking, the statement "$X$ causes $Y$" means that changing the value of $X$ will change the distribution of $Y$. When $X$ causes $Y$, $X$ and $Y$ will be associated but the reverse is not, in general, true. Association does not necessarily imply causation. We will consider two frameworks for discussing causation. The first uses **counterfactual** random variables. The second, presented in the next chapter, uses **directed acyclic graphs.**

## 16.1 The Counterfactual Model

Suppose that $X$ is a binary treatment variable where $X = 1$ means "treated" and $X = 0$ means "not treated." We are using the word "treatment" in a very broad sense. Treatment might refer to a medication or something like smoking. An alternative to "treated/not treated" is "exposed/not exposed" but we shall use the former.

Let $Y$ be some outcome variable such as presence or absence of disease. To distinguish the statement "$X$ is associated $Y$" from the statement "$X$ causes $Y$" we need to enrich our probabilistic vocabulary. Specifically, we will decompose the response $Y$ into a more fine-grained object.

We introduce two new random variables $(C_0, C_1)$, called **potential outcomes** with the following interpretation: $C_0$ is the outcome if the subject is

not treated ($X = 0$) and $C_1$ is the outcome if the subject is treated ($X = 1$). Hence,

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1. \end{cases}$$

We can express the relationship between $Y$ and $(C_0, C_1)$ more succinctly by

$$Y = C_X. \tag{16.1}$$

Equation (16.1) is called the **consistency relationship**.

Here is a toy dataset to make the idea clear:

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 4 | 4 | * |
| 0 | 7 | 7 | * |
| 0 | 2 | 2 | * |
| 0 | 8 | 8 | * |
| 1 | 3 | * | 3 |
| 1 | 5 | * | 5 |
| 1 | 8 | * | 8 |
| 1 | 9 | * | 9 |

The asterisks denote unobserved values. When $X = 0$ we don't observe $C_1$, in which case we say that $C_1$ is a **counterfactual** since it is the outcome you would have had if, counter to the fact, you had been treated ($X = 1$). Similarly, when $X = 1$ we don't observe $C_0$, and we say that $C_0$ is **counterfactual**. There are four types of subjects:

| Type | $C_0$ | $C_1$ |
|---|---|---|
| Survivors | 1 | 1 |
| Responders | 0 | 1 |
| Anti-responders | 1 | 0 |
| Doomed | 0 | 0 |

Think of the potential outcomes $(C_0, C_1)$ as hidden variables that contain all the relevant information about the subject.

Define the **average causal effect** or **average treatment effect** to be

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0). \tag{16.2}$$

The parameter $\theta$ has the following interpretation: $\theta$ is the mean if everyone were treated ($X = 1$) minus the mean if everyone were not treated ($X = 0$). There are other ways of measuring the causal effect. For example, if $C_0$ and $C_1$ are binary, we define the **causal odds ratio**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_1 = 0)} \div \frac{\mathbb{P}(C_0 = 1)}{\mathbb{P}(C_0 = 0)}$$

and the **causal relative risk**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_0 = 1)}.$$

The main ideas will be the same whatever causal effect we use. For simplicity, we shall work with the average causal effect $\theta$.

Define the **association** to be

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0). \tag{16.3}$$

Again, we could use odds ratios or other summaries if we wish.

**16.1 Theorem** (Association Is Not Causation). *In general,* $\theta \neq \alpha$.

**16.2 Example.** Suppose the whole population is as follows:

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 0 | 0 | 0 | $0^*$ |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |

Again, the asterisks denote unobserved values. Notice that $C_0 = C_1$ for every subject, thus, this treatment has no effect. Indeed,

$$\begin{aligned}
\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) = \frac{1}{8}\sum_{i=1}^{8} C_{1i} - \frac{1}{8}\sum_{i=1}^{8} C_{0i} \\
&= \frac{0+0+0+0+1+1+1+1}{8} - \frac{0+0+0+0+1+1+1+1}{8} \\
&= 0.
\end{aligned}$$

Thus, the average causal effect is 0. The observed data are only the $X$'s and $Y$'s, from which we can estimate the association:

$$\begin{aligned}
\alpha &= \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) \\
&= \frac{1+1+1+1}{4} - \frac{0+0+0+0}{4} = 1.
\end{aligned}$$

Hence, $\theta \neq \alpha$.

To add some intuition to this example, imagine that the outcome variable is 1 if "healthy" and 0 if "sick". Suppose that $X = 0$ means that the subject

does not take vitamin C and that $X = 1$ means that the subject does take vitamin C. Vitamin C has no causal effect since $C_0 = C_1$ for each subject. In this example there are two types of people: healthy people $(C_0, C_1) = (1, 1)$ and unhealthy people $(C_0, C_1) = (0, 0)$. Healthy people tend to take vitamin C while unhealthy people don't. It is this association between $(C_0, C_1)$ and $X$ that creates an association between $X$ and $Y$. If we only had data on $X$ and $Y$ we would conclude that $X$ and $Y$ are associated. Suppose we wrongly interpret this causally and conclude that vitamin C prevents illness. Next we might encourage everyone to take vitamin C. If most people comply with our advice, the population will look something like this:

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 0 | 0 | $0^*$ |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |
| 1 | 1 | $1^*$ | 1 |

Now $\alpha = (4/7) - (0/1) = 4/7$. We see that $\alpha$ went down from 1 to $4/7$. Of course, the causal effect never changed but the naive observer who does not distinguish association and causation will be confused because his advice seems to have made things worse instead of better. ■

In the last example, $\theta = 0$ and $\alpha = 1$. It is not hard to create examples in which $\alpha > 0$ and yet $\theta < 0$. The fact that the association and causal effects can have different signs is very confusing to many people.

The example makes it clear that, in general, we cannot use the association to estimate the causal effect $\theta$. The reason that $\theta \neq \alpha$ is that $(C_0, C_1)$ was not independent of $X$. That is, treatment assignment was not independent of person type.

Can we ever estimate the causal effect? The answer is: sometimes. In particular, random assignment to treatment makes it possible to estimate $\theta$.

**16.3 Theorem.** *Suppose we randomly assign subjects to treatment and that* $\mathbb{P}(X = 0) > 0$ *and* $\mathbb{P}(X = 1) > 0$. *Then* $\alpha = \theta$. *Hence, any consistent estimator of* $\alpha$ *is a consistent estimator of* $\theta$. *In particular, a consistent estimator is*

$$\widehat{\theta} \;=\; \widehat{\mathbb{E}}(Y|X = 1) - \widehat{\mathbb{E}}(Y|X = 0)$$

$$= \quad \overline{Y}_1 - \overline{Y}_0$$

*is a consistent estimator of* $\theta$*, where*

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} Y_i X_i, \qquad \overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-X_i),$$

$n_1 = \sum_{i=1}^n X_i$*, and* $n_0 = \sum_{i=1}^n (1-X_i)$.

PROOF. Since $X$ is randomly assigned, $X$ is independent of $(C_0, C_1)$. Hence,

$$\begin{aligned}
\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) \\
&= \mathbb{E}(C_1|X=1) - \mathbb{E}(C_0|X=0) \qquad \text{since } X \amalg (C_0, C_1) \\
&= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \qquad \text{since } Y = C_X \\
&= \alpha.
\end{aligned}$$

The consistency follows from the law of large numbers. ■

If $Z$ is a covariate, we define the **conditional causal effect** by

$$\theta_z = \mathbb{E}(C_1|Z=z) - \mathbb{E}(C_0|Z=z).$$

For example, if $Z$ denotes gender with values $Z = 0$ (women) and $Z = 1$ (men), then $\theta_0$ is the causal effect among women and $\theta_1$ is the causal effect among men. In a randomized experiment, $\theta_z = \mathbb{E}(Y|X=1, Z=z) - \mathbb{E}(Y|X=0, Z=z)$ and we can estimate the conditional causal effect using appropriate sample averages.

**Summary of the Counterfactual Model**

Random variables: $(C_0, C_1, X, Y)$.
Consistency relationship: $Y = C_X$.
Causal Effect: $\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0)$.
Association: $\alpha = \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0)$.
Random assignment $\implies$ $(C_0, C_1) \amalg X$ $\implies$ $\theta = \alpha$.

## 16.2 Beyond Binary Treatments

Let us now generalize beyond the binary case. Suppose that $X \in \mathcal{X}$. For example, $X$ could be the dose of a drug in which case $X \in \mathbb{R}$. The counterfactual vector $(C_0, C_1)$ now becomes the **counterfactual function** $C(x)$ where

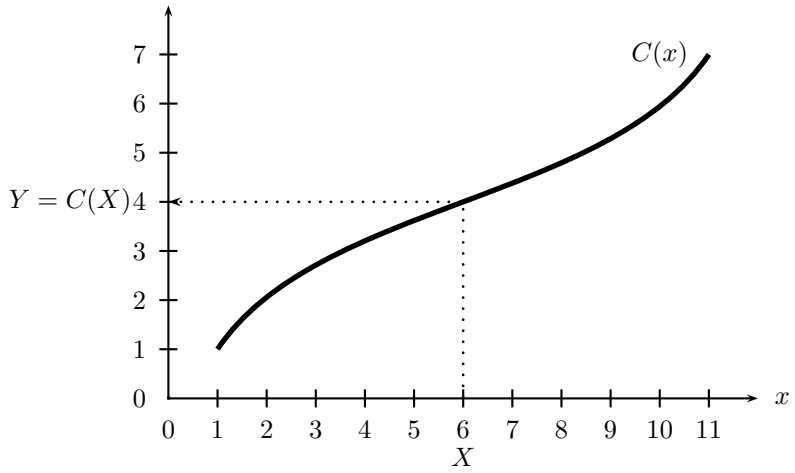



FIGURE 16.1. A counterfactual function $C(x)$. The outcome $Y$ is the value of the curve $C(x)$ evaluated at the observed dose $X$.

$C(x)$ is the outcome a subject would have if he received dose $x$. The observed response is given by the consistency relation

$$Y \equiv C(X). \tag{16.4}$$

See Figure 16.1. The **causal regression function** is

$$\theta(x) = \mathbb{E}(C(x)). \tag{16.5}$$

The regression function, which measures association, is $r(x) = \mathbb{E}(Y|X = x)$.

**16.4 Theorem.** *In general, $\theta(x) \neq r(x)$. However, when $X$ is randomly assigned, $\theta(x) = r(x)$.*

**16.5 Example.** An example in which $\theta(x)$ is constant but $r(x)$ is not constant is shown in Figure 16.2. The figure shows the counterfactual functions for four subjects. The dots represent their $X$ values $X_1, X_2, X_3, X_4$. Since $C_i(x)$ is constant over $x$ for all $i$, there is no causal effect and hence

$$\theta(x) = \frac{C_1(x) + C_2(x) + C_3(x) + C_4(x)}{4}$$

is constant. Changing the dose $x$ will not change anyone's outcome. The four dots in the lower plot represent the observed data points $Y_1 = C_1(X_1), Y_2 = C_2(X_2), Y_3 = C_3(X_3), Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. Although there is no causal effect, there is an association since the regression curve $r(x)$ is not constant. ■

## 16.3 Observational Studies and Confounding

A study in which treatment (or exposure) is not randomly assigned is called an **observational study.** In these studies, subjects select their own value of the exposure $X$. Many of the health studies you read about in the newspaper are like this. As we saw, association and causation could in general be quite different. This discrepancy occurs in non-randomized studies because the potential outcome $C$ is not independent of treatment $X$. However, suppose we could find groupings of subjects such that, within groups, $X$ and $\{C(x) : \ x \in \mathcal{X}\}$ are independent. This would happen if the subjects are very similar within groups. For example, suppose we find people who are very similar in age, gender, educational background, and ethnic background. Among these people we might feel it is reasonable to assume that the choice of $X$ is essentially random. These other variables are called **confounding variables.**[1] If we denote these other variables collectively as $Z$, then we can express this idea by saying that

$$\{C(x) : \ x \in \mathcal{X}\} \amalg X|Z. \tag{16.6}$$

Equation (16.6) means that, within groups of $Z$, the choice of treatment $X$ does not depend on type, as represented by $\{C(x) : \ x \in \mathcal{X}\}$. If (16.6) holds and we observe $Z$ then we say that there is **no unmeasured confounding**.

**16.6 Theorem.** *Suppose that (16.6) holds. Then,*

$$\theta(x) = \int \mathbb{E}(Y|X = x, Z = z) dF_Z(z) dz. \tag{16.7}$$

*If $\widehat{r}(x, z)$ is a consistent estimate of the regression function $\mathbb{E}(Y|X = x, Z = z)$, then a consistent estimate of $\theta(x)$ is*

$$\widehat{\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}(x, Z_i).$$

[1] A more precise definition of confounding is given in the next chapter.

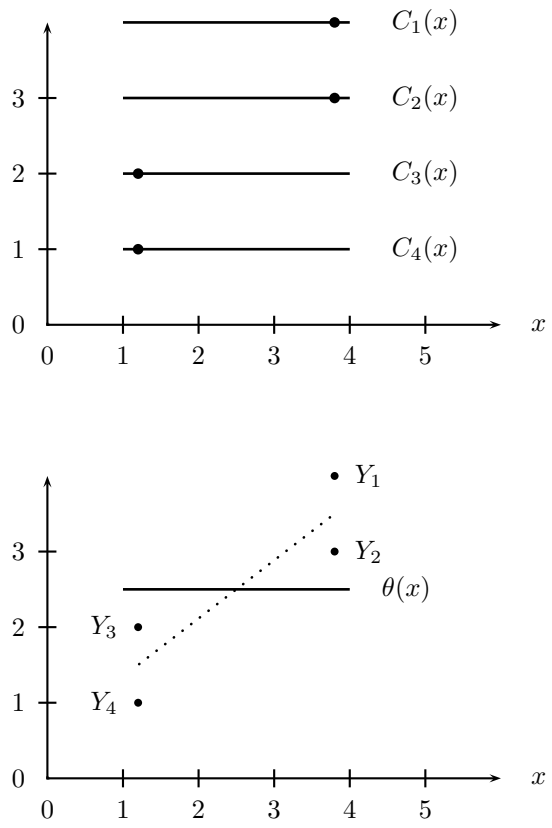



FIGURE 16.2. The top plot shows the counterfactual function $C(x)$ for four subjects. The dots represent their $X$ values. Since $C_i(x)$ is constant over $x$ for all $i$, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x) + C_2(x) + C_3(x) + C_4(x))/4$. The four dots represent the observed data points $Y_1 = C_1(X_1)$, $Y_2 = C_2(X_2)$, $Y_3 = C_3(X_3)$, $Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. There is no causal effect since $C_i(x)$ is constant for all $i$. But there is an association since the regression curve $r(x)$ is not constant.

*In particular, if* $r(x,z) = \beta_0 + \beta_1 x + \beta_2 z$ *is linear, then a consistent estimate of* $\theta(x)$ *is*

$$\widehat{\theta}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \overline{Z}_n \tag{16.8}$$

*where* $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ *are the least squares estimators.*

**16.7 Remark.** It is useful to compare equation (16.7) to $\mathbb{E}(Y|X=x)$ which can be written as $\mathbb{E}(Y|X=x) = \int \mathbb{E}(Y|X=x, Z=z) dF_{Z|X}(z|x)$.

Epidemiologists call (16.7) the **adjusted treatment effect**. The process of computing adjusted treatment effects is called **adjusting (or controlling) for confounding.** The selection of what confounders $Z$ to measure and control for requires scientific insight. Even after adjusting for confounders, we cannot be sure that there are not other confounding variables that we missed. This is why observational studies must be treated with healthy skepticism. Results from observational studies start to become believable when: (i) the results are replicated in many studies, (ii) each of the studies controlled for plausible confounding variables, (iii) there is a plausible scientific explanation for the existence of a causal relationship.

A good example is smoking and cancer. Numerous studies have shown a relationship between smoking and cancer even after adjusting for many confounding variables. Moreover, in laboratory studies, smoking has been shown to damage lung cells. Finally, a causal link between smoking and cancer has been found in randomized animal studies. It is this collection of evidence over many years that makes this a convincing case. One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper.

## 16.4 Simpson's Paradox

Simpson's paradox is a puzzling phenomenon that is discussed in most statistics texts. Unfortunately, most explanations are confusing (and in some cases incorrect). The reason is that it is nearly impossible to explain the paradox without using counterfactuals (or directed acyclic graphs).

Let $X$ be a binary treatment variable, $Y$ a binary outcome, and $Z$ a third binary variable such as gender. Suppose the joint distribution of $X, Y, Z$ is

| | $Y=1$ | $Y=0$ | $Y=1$ | $Y=0$ |
|---|---|---|---|---|
| $X=1$ | .1500 | .2250 | .1000 | .0250 |
| $X=0$ | .0375 | .0875 | .2625 | .1125 |
| | $Z=1$ (men) | | $Z=0$ (women) | |

The marginal distribution for $(X, Y)$ is

| | $Y=1$ | $Y=0$ | |
|---|---|---|---|
| $X=1$ | .25 | .25 | .50 |
| $X=0$ | .30 | .20 | .50 |
| | .55 | .45 | 1 |

From these tables we find that,

$$\begin{aligned}
\mathbb{P}(Y=1|X=1) - \mathbb{P}(Y=1|X=0) &= -0.1 \\
\mathbb{P}(Y=1|X=1, Z=1) - \mathbb{P}(Y=1|X=0, Z=1) &= 0.1 \\
\mathbb{P}(Y=1|X=1, Z=0) - \mathbb{P}(Y=1|X=0, Z=0) &= 0.1.
\end{aligned}$$

To summarize, we *seem* to have the following information:

| Mathematical Statement | English Statement? |
|---|---|
| $\mathbb{P}(Y=1\|X=1) < \mathbb{P}(Y=1\|X=0)$ | treatment is harmful |
| $\mathbb{P}(Y=1\|X=1, Z=1) > \mathbb{P}(Y=1\|X=0, Z=1)$ | treatment is beneficial to men |
| $\mathbb{P}(Y=1\|X=1, Z=0) > \mathbb{P}(Y=1\|X=0, Z=0)$ | treatment is beneficial to women |

Clearly, something is amiss. There can't be a treatment which is good for men, good for women, but bad overall. This is nonsense. The problem is with the set of English statements in the table. Our translation from math into English is specious.

> **The inequality $\mathbb{P}(Y=1|X=1) < \mathbb{P}(Y=1|X=0)$ does not mean that treatment is harmful.**

The phrase "treatment is harmful" should be written mathematically as $\mathbb{P}(C_1 = 1) < \mathbb{P}(C_0 = 1)$. The phrase "treatment is harmful for men" should be written $\mathbb{P}(C_1 = 1|Z = 1) < \mathbb{P}(C_0 = 1|Z = 1)$. The three mathematical statements in the table are not at all contradictory. It is only the translation into English that is wrong.

Let us now show that a real Simpson's paradox cannot happen, that is, there cannot be a treatment that is beneficial for men and women but harmful overall. Suppose that treatment is beneficial for both sexes. Then

$$\mathbb{P}(C_1 = 1|Z = z) > \mathbb{P}(C_0 = 1|Z = z)$$

for all $z$. It then follows that

$$\begin{aligned}\mathbb{P}(C_1 = 1) &= \sum_z \mathbb{P}(C_1 = 1|Z = z)\mathbb{P}(Z = z) \\ &> \sum_z \mathbb{P}(C_0 = 1|Z = z)\mathbb{P}(Z = z) \\ &= \mathbb{P}(C_0 = 1).\end{aligned}$$

Hence, $\mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1)$, so treatment is beneficial overall. No paradox.

## 16.5 Bibliographic Remarks

The use of potential outcomes to clarify causation is due mainly to Jerzy Neyman and Donald Rubin. Later developments are due to Jamie Robins, Paul Rosenbaum, and others. A parallel development took place in econometrics by various people including James Heckman and Charles Manski. Texts on causation include Pearl (2000), Rosenbaum (2002), Spirtes et al. (2000), and van der Laan and Robins (2003).

## 16.6 Exercises

1. Create an example like Example 16.2 in which $\alpha > 0$ and $\theta < 0$.

2. Prove Theorem 16.4.

3. Suppose you are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an observational study, where $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Although it is not possible to estimate the causal effect $\theta$, it is possible to put bounds on $\theta$. Find upper and lower bounds on $\theta$ that can be consistently estimated from the data. Show that the bounds have width 1.

   Hint: Note that $\mathbb{E}(C_1) = \mathbb{E}(C_1|X = 1)\mathbb{P}(X = 1) + \mathbb{E}(C_1|X = 0)\mathbb{P}(X = 0)$.

4. Suppose that $X \in \mathbb{R}$ and that, for each subject $i$, $C_i(x) = \beta_{1i}x$. Each subject has their own slope $\beta_{1i}$. Construct a joint distribution on $(\beta_1, X)$ such that $\mathbb{P}(\beta_1 > 0) = 1$ but $\mathbb{E}(Y|X = x)$ is a decreasing function of $x$, where $Y = C(X)$. Interpret.

5. Let $X \in \{0, 1\}$ be a binary treatment variable and let $(C_0, C_1)$ denote the corresponding potential outcomes. Let $Y = C_X$ denote the observed

response. Let $F_0$ and $F_1$ be the cumulative distribution functions for $C_0$ and $C_1$. Assume that $F_0$ and $F_1$ are both continuous and strictly increasing. Let $\theta = m_1 - m_0$ where $m_0 = F_0^{-1}(1/2)$ is the median of $C_0$ and $m_1 = F_1^{-1}(1/2)$ is the median of $C_1$. Suppose that the treatment $X$ is assigned randomly. Find an expression for $\theta$ involving only the joint distribution of $X$ and $Y$.

# 17

# Directed Graphs and Conditional Independence

## 17.1 Introduction

A directed graph consists of a set of nodes with arrows between some nodes. An example is shown in Figure 17.1.

Graphs are useful for representing independence relations between variables. They can also be used as an alternative to counterfactuals to represent causal relationships. Some people use the phrase **Bayesian network** to refer to a directed graph endowed with a probability distribution. This is a poor choice of terminology. Statistical inference for directed graphs can be performed using

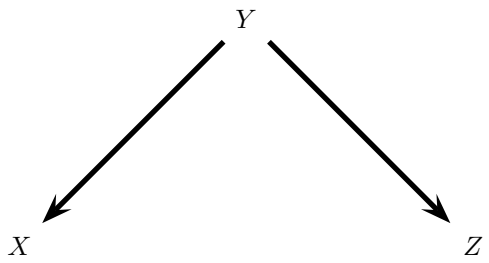



FIGURE 17.1. A directed graph with vertices $V = \{X, Y, Z\}$ and edges $E = \{(Y, X), (Y, Z)\}$.

frequentist or Bayesian methods, so it is misleading to call them Bayesian networks.

Before getting into details about directed acyclic graphs (DAGs), we need to discuss conditional independence.

## 17.2 Conditional Independence

**17.1 Definition.** *Let $X$, $Y$ and $Z$ be random variables. $X$ and $Y$ are* **conditionally independent given** *$Z$, written $X \amalg Y \mid Z$, if*

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z). \tag{17.1}$$

*for all $x$, $y$ and $z$.*

Intuitively, this means that, once you know $Z$, $Y$ provides no extra information about $X$. An equivalent definition is that

$$f(x|y,z) = f(x|z). \tag{17.2}$$

The conditional independence relation satisfies some basic properties.

**17.2 Theorem.** *The following implications hold:* [1]

$$\begin{aligned}
X \amalg Y \mid Z &\implies Y \amalg X \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies U \amalg Y \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies X \amalg Y \mid (Z,U) \\
X \amalg Y \mid Z \text{ and } X \amalg W|(Y,Z) &\implies X \amalg (W,Y) \mid Z \\
X \amalg Y \mid Z \text{ and } X \amalg Z \mid Y &\implies X \amalg (Y,Z).
\end{aligned}$$

## 17.3 DAGs

A **directed graph** $\mathcal{G}$ consists of a set of vertices $V$ and an edge set $E$ of ordered pairs of vertices. For our purposes, each vertex will correspond to a random variable. If $(X,Y) \in E$ then there is an arrow pointing from $X$ to $Y$. See Figure 17.1.

[1] The last property requires the assumption that all events have positive probability; the first four do not.

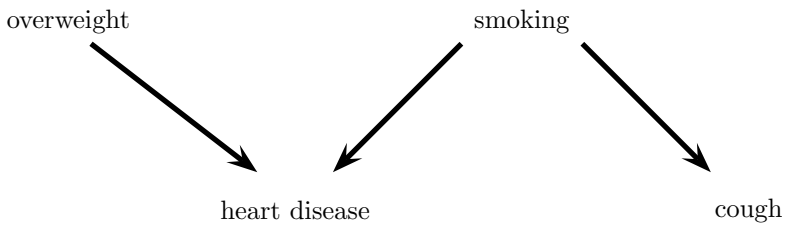
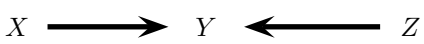



FIGURE 17.2. DAG for Example 17.4.

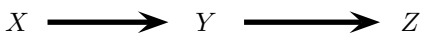
If an arrow connects two variables $X$ and $Y$ (in either direction) we say that $X$ and $Y$ are **adjacent.** If there is an arrow from $X$ to $Y$ then $X$ is a **parent** of $Y$ and $Y$ is a **child** of $X$. The set of all parents of $X$ is denoted by $\pi_X$ or $\pi(X)$. A **directed path** between two variables is a set of arrows all pointing in the same direction linking one variable to the other such as:


A sequence of adjacent vertices staring with $X$ and ending with $Y$ but ignoring the direction of the arrows is called an **undirected path.** The sequence $\{X, Y, Z\}$ in Figure 17.1 is an undirected path. $X$ is an **ancestor** of $Y$ if there is a directed path from $X$ to $Y$ (or $X = Y$). We also say that $Y$ is a **descendant** of $X$.

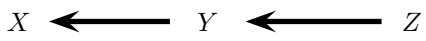
A configuration of the form:




is called a **collider** at $Y$. A configuration not of that form is called a **non-collider**, for example,



or




The collider property is path dependent. In Figure 17.7, $Y$ is a collider on the path $\{X, Y, Z\}$ but it is a non-collider on the path $\{X, Y, W\}$. When the variables pointing into the collider are not adjacent, we say that the collider is **unshielded**. A directed path that starts and ends at the same variable is called a **cycle.** A directed graph is **acyclic** if it has no cycles. In this case we say that the graph is a **directed acyclic graph** or **DAG**. From now on, we only deal with acyclic graphs.

## 17.4 Probability and DAGs

Let $\mathcal{G}$ be a DAG with vertices $V = (X_1, \ldots, X_k)$.

**17.3 Definition.** *If $\mathbb{P}$ is a distribution for $V$ with probability function $f$, we say that $\mathbb{P}$ is* **Markov to $\mathcal{G}$**, *or that $\mathcal{G}$* **represents** *$\mathbb{P}$, if*

$$f(v) = \prod_{i=1}^{k} f(x_i \mid \pi_i) \tag{17.3}$$

*where $\pi_i$ are the parents of $X_i$. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.*

**17.4 Example.** Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} & f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\ &\quad = f(\text{overweight}) \times f(\text{smoking}) \\ &\quad \times f(\text{heart disease} \,|\, \text{overweight}, \text{smoking}) \\ &\quad \times f(\text{cough} \,|\, \text{smoking}). \quad \blacksquare \end{aligned}$$

**17.5 Example.** For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function $f$ has the form

$$f(x, y, z, w) = f(x) f(y) f(z \mid x, y) f(w \mid z). \quad \blacksquare$$

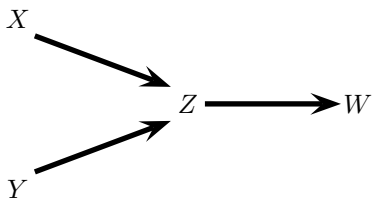



FIGURE 17.3. Another DAG.

The following theorem says that $\mathbb{P} \in M(\mathcal{G})$ if and only if the **Markov Condition** holds. Roughly speaking, the Markov Condition means that every variable $W$ is independent of the "past" given its parents.

**17.6 Theorem.** *A distribution $\mathbb{P} \in M(\mathcal{G})$ if and only if the following* **Markov Condition** *holds: for every variable $W$,*

$$W \amalg \widetilde{W} \mid \pi_W \tag{17.4}$$

*where $\widetilde{W}$ denotes all the other variables except the parents and descendants of $W$.*

**17.7 Example.** In Figure 17.3, the Markov Condition implies that

$$X \amalg Y \quad \text{and} \quad W \amalg \{X, Y\} \mid Z. \quad \blacksquare$$

**17.8 Example.** Consider the DAG in Figure 17.4. In this case probability function must factor like

$$f(a, b, c, d, e) = f(a)f(b|a)f(c|a)f(d|b, c)f(e|d).$$

The Markov Condition implies the following independence relations:

$$D \amalg A \mid \{B, C\}, \quad E \amalg \{A, B, C\} \mid D \quad \text{and} \quad B \amalg C \mid A \quad \blacksquare$$

## 17.5 More Independence Relations

The Markov Condition allows us to list some independence relations implied by a DAG. These relations might imply other independence relations. Con-

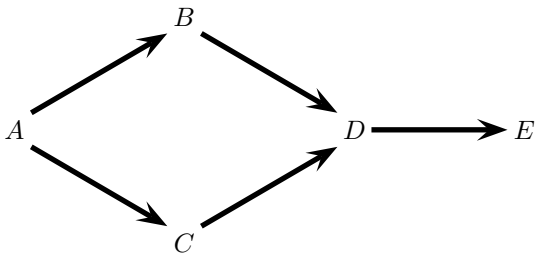



FIGURE 17.4. Yet another DAG.

sider the DAG in Figure 17.5. The Markov Condition implies:

$$X_1 \amalg X_2, \qquad X_2 \amalg \{X_1, X_4\}, \qquad X_3 \amalg X_4 \mid \{X_1, X_2\},$$

$$X_4 \amalg \{X_2, X_3\} \mid X_1, \qquad X_5 \amalg \{X_1, X_2\} \mid \{X_3, X_4\}$$

It turns out (but it is not obvious) that these conditions imply that

$$\{X_4, X_5\} \amalg X_2 \mid \{X_1, X_3\}.$$

How do we find these extra independence relations? The answer is "d-separation" which means "directed separation." d-separation can be summarized by three rules. Consider the four DAG's in Figure 17.6 and the DAG in Figure 17.7. The first 3 DAG's in Figure 17.6 have no colliders. The DAG in the lower right of Figure 17.6 has a collider. The DAG in Figure 17.7 has a collider with a descendant.

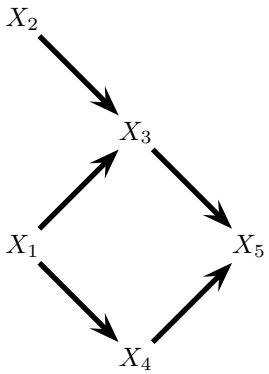



FIGURE 17.5. And yet another DAG.


FIGURE 17.6. The first three DAG's have no colliders. The fourth DAG in the lower right corner has a collider at $Y$.

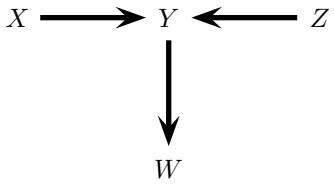



FIGURE 17.7. A collider with a descendant.

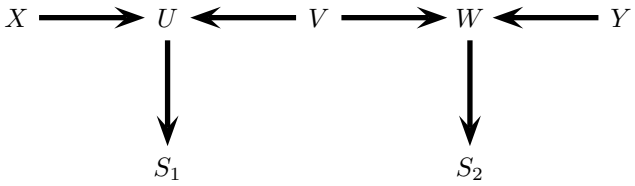



FIGURE 17.8. d-separation explained.

The Rules of d-Separation

Consider the DAGs in Figures 17.6 and 17.7.

1. When $Y$ is not a collider, $X$ and $Z$ are **d-connected**, but they are **d-separated** given $Y$.
2. If $X$ and $Z$ collide at $Y$, then $X$ and $Z$ are **d-separated**, but they are **d-connected** given $Y$.
3. Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 17.7, $X$ and $Z$ are **d-separated** but they are **d-connected** given $W$.

Here is a more formal definition of d-separation. Let $X$ and $Y$ be distinct vertices and let $W$ be a set of vertices not containing $X$ or $Y$. Then $X$ and $Y$ are **d-separated given** $W$ if there exists no undirected path $U$ between $X$ and $Y$ such that (i) every collider on $U$ has a descendant in $W$, and (ii) no other vertex on $U$ is in $W$. If $A, B$, and $W$ are distinct sets of vertices and $A$ and $B$ are not empty, then $A$ and $B$ are d-separated given $W$ if for every $X \in A$ and $Y \in B$, $X$ and $Y$ are d-separated given $W$. Sets of vertices that are not d-separated are said to be d-connected.

**17.9 Example.** Consider the DAG in Figure 17.8. From the d-separation rules we conclude that:

$X$ and $Y$ are d-separated (given the empty set);
$X$ and $Y$ are d-connected given $\{S_1, S_2\}$;
$X$ and $Y$ are d-separated given $\{S_1, S_2, V\}$.

**17.10 Theorem.** [2] *Let $A$, $B$, and $C$ be disjoint sets of vertices. Then $A \amalg B \mid C$ if and only if $A$ and $B$ are d-separated by $C$.*

[2] We implicitly assume that $\mathbb{P}$ is **faithful** to $\mathcal{G}$ which means that $\mathbb{P}$ has no extra independence relations other than those logically implied by the Markov Condition.

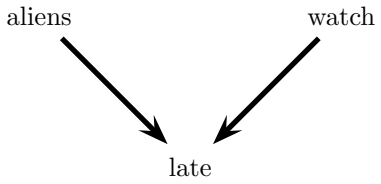



FIGURE 17.9. Jordan's alien example (Example 17.11). Was your friend kidnapped by aliens or did you forget to set your watch?

**17.11 Example.** The fact that conditioning on a collider creates dependence might not seem intuitive. Here is a whimsical example from Jordan (2004) that makes this idea more palatable. Your friend appears to be late for a meeting with you. There are two explanations: she was abducted by aliens or you forgot to set your watch ahead one hour for daylight savings time. (See Figure 17.9.) Aliens and Watch are blocked by a collider which implies they are marginally independent. This seems reasonable since — before we know anything about your friend being late — we would expect these variables to be independent. We would also expect that $\mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}) > \mathbb{P}(\mathsf{Aliens} = \mathsf{yes})$; learning that your friend is late certainly increases the probability that she was abducted. But when we learn that you forgot to set your watch properly, we would lower the chance that your friend was abducted. Hence, $\mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}) \neq \mathbb{P}(\mathsf{Aliens} = \mathsf{yes}|\mathsf{Late} = \mathsf{yes}, \mathsf{Watch} = \mathsf{no})$. Thus, Aliens and Watch are dependent given Late. ■

**17.12 Example.** Consider the DAG in Figure 17.2. In this example, overweight and smoking are marginally independent but they are dependent given heart disease. ■

Graphs that look different may actually imply the same independence relations. If $\mathcal{G}$ is a DAG, we let $\mathcal{I}(\mathcal{G})$ denote all the independence statements implied by $\mathcal{G}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ for the same variables $V$ are **Markov equivalent** if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. Given a DAG $\mathcal{G}$, let skeleton($\mathcal{G}$) denote the undirected graph obtained by replacing the arrows with undirected edges.

**17.13 Theorem.** *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if (i)* skeleton($\mathcal{G}_1$) = skeleton($\mathcal{G}_2$) *and (ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders.*

**17.14 Example.** The first three DAGs in Figure 17.6 are Markov equivalent. The DAG in the lower right of the Figure is not Markov equivalent to the others. ■

## 17.6 Estimation for DAGs

Two estimation questions arise in the context of DAGs. First, given a DAG $\mathcal{G}$ and data $V_1, \ldots, V_n$ from a distribution $f$ consistent with $\mathcal{G}$, how do we estimate $f$? Second, given data $V_1, \ldots, V_n$ how do we estimate $\mathcal{G}$? The first question is pure estimation while the second involves model selection. These are very involved topics and are beyond the scope of this book. We will just briefly mention the main ideas.

Typically, one uses some parametric model $f(x|\pi_x; \theta_x)$ for each conditional density. The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(V_i; \theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}|\pi_j; \theta_j),$$

where $X_{ij}$ is the value of $X_j$ for the $i^{\text{th}}$ data point and $\theta_j$ are the parameters for the $j^{\text{th}}$ conditional density. We can then estimate the parameters by maximum likelihood.

To estimate the structure of the DAG itself, we could fit every possible DAG using maximum likelihood and use AIC (or some other method) to choose a DAG. However, there are many possible DAGs so you would need much data for such a method to be reliable. Also, searching through all possible DAGs is a serious computational challenge. Producing a valid, accurate confidence set for the DAG structure would require astronomical sample sizes. If prior information is available about part of the DAG structure, the computational and statistical problems are at least partly ameliorated.

## 17.7 Bibliographic Remarks

There are a number of texts on DAGs including Edwards (1995) and Jordan (2004). The first use of DAGs for representing causal relationships was by Wright (1934). Modern treatments are contained in Spirtes et al. (2000) and Pearl (2000). Robins et al. (2003) discuss the problems with estimating causal structure from data.

## 17.8 Appendix

Causation Revisited. We discussed causation in Chapter 16 using the idea of counterfactual random variables. A different approach to causation uses

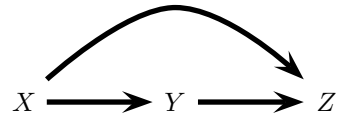



FIGURE 17.10. Conditioning versus intervening.

DAGs. The two approaches are mathematically equivalent though they appear to be quite different. In the DAG approach, the extra element is the idea of **intervention**. Consider the DAG in Figure 17.10.

The probability function for a distribution consistent with this DAG has the form $f(x,y,z) = f(x)f(y|x)f(z|x,y)$. The following is pseudocode for generating from this distribution.

$$\begin{aligned} \text{For } i &= 1,\ldots,n: \\ x_i &\leftarrow p_X(x_i) \\ y_i &\leftarrow p_{Y|X}(y_i|x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y_i) \end{aligned}$$

Suppose we repeat this code many times, yielding data $(x_1,y_1,z_1),\ldots,(x_n,y_n,z_n)$. Among all the times that we observe $Y=y$, how often is $Z=z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$\begin{aligned} \mathbb{P}(Z=z|Y=y) &= \frac{\mathbb{P}(Y=y,Z=z)}{\mathbb{P}(Y=y)} = \frac{f(y,z)}{f(y)} \\ &= \frac{\sum_x f(x,y,z)}{f(y)} = \frac{\sum_x f(x)\,f(y|x)\,f(z|x,y)}{f(y)} \\ &= \sum_x f(z|x,y)\frac{f(y|x)\,f(x)}{f(y)} = \sum_x f(z|x,y)\frac{f(x,y)}{f(y)} \\ &= \sum_x f(z|x,y)\,f(x|y). \end{aligned}$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix $Y$ at the value $y$. The code now looks like this:

$$\begin{aligned} \text{set } Y &= y \\ \text{for } i &= 1,\ldots,n \\ x_i &\leftarrow p_X(x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y) \end{aligned}$$

Having **set** $Y = y$, how often was $Z = z$? To answer, note that the intervention has changed the joint probability to be

$$f^*(x, z) = f(x)f(z|x, y).$$

The answer to our question is given by the marginal distribution

$$f^*(z) = \sum_x f^*(x, z) = \sum_x f(x)f(z|x, y).$$

We shall denote this as $\mathbb{P}(Z = z|Y := y)$ or $f(z|Y := y)$. We call $\mathbb{P}(Z = z|Y = y)$ **conditioning by observation** or **passive conditioning.** We call $\mathbb{P}(Z = z|Y := y)$ **conditioning by intervention** or **active conditioning.**

Passive conditioning is used to answer a predictive question like:

"Given that Joe smokes, what is the probability he will get lung cancer?"

Active conditioning is used to answer a causal question like:

"If Joe quits smoking, what is the probability he will get lung cancer?"

Consider a pair $(\mathcal{G}, \mathbb{P})$ where $\mathcal{G}$ is a DAG and $\mathbb{P}$ is a distribution for the variables $V$ of the DAG. Let $p$ denote the probability function for $\mathbb{P}$. Consider intervening and fixing a variable $X$ to be equal to $x$. We represent the intervention by doing two things:

(1) Create a new DAG $\mathcal{G}^*$ by removing all arrows pointing into $X$;

(2) Create a new distribution $f^*(v) = \mathbb{P}(V = v|X := x)$ by removing the term $f(x|\pi_X)$ from $f(v)$.

The new pair $(\mathcal{G}^*, f^*)$ represents the intervention "set $X = x$."

**17.15 Example.** You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence $p_{R,W}(r, w) \neq p_R(r)p_W(w)$ where $R$ denotes Rain and W denotes Wet Lawn. Consider the following two DAGs:

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn}.$$

The first DAG implies that $f(w, r) = f(r)f(w|r)$ while the second implies that $f(w, r) = f(w)f(r|w)$ No matter what the joint distribution $f(w, r)$ is, both graphs are correct. Both imply that $R$ and $W$ are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.

Look at the first graph and form the intervention $W = 1$ where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into $W$

to get the modified graph:


with distribution $f^*(r) = f(r)$. Thus $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ tells us that "wet lawn" does not cause rain.

Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention $W = 1$ on the second graph. There are no arrows into $W$ that need to be broken so the intervention graph is the same as the original graph. Thus $f^*(r) = f(r|w)$ which would imply that changing "wet" changes "rain." Clearly, this is nonsense.

Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.

**17.16 Remark.** We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable.

We can use DAGs to represent confounding variables. If $X$ is a treatment and $Y$ is an outcome, a confounding variable $Z$ is a variable with arrows into both $X$ and $Y$; see Figure 17.11. It is easy to check, using the formalism of interventions, that the following facts are true:

In a randomized study, the arrow between $Z$ and $X$ is broken. In this case, even with $Z$ unobserved (represented by enclosing $Z$ in a circle), the causal relationship between $X$ and $Y$ is estimable because it can be shown that $\mathbb{E}(Y|X := x) = \mathbb{E}(Y|X = x)$ which does not involve the unobserved $Z$. In an observational study, with all confounders observed, we get $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ as in formula (16.7). If $Z$ is unobserved then we cannot estimate the causal effect because $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ involves the unobserved $Z$. We can't just use $X$ and $Y$ since in this case. $\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|X := x)$ which is just another way of saying that causation is not association.

In fact, we can make a precise connection between DAGs and counterfactuals as follows. Suppose that $X$ and $Y$ are binary. Define the confounding

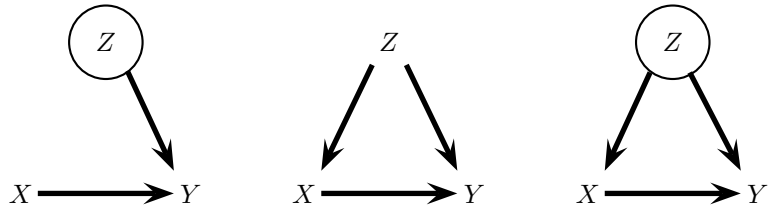



FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

variable $Z$ by

$$Z = \begin{cases} 1 & \text{if } (C_0, C_1) = (0,0) \\ 2 & \text{if } (C_0, C_1) = (0,1) \\ 3 & \text{if } (C_0, C_1) = (1,0) \\ 4 & \text{if } (C_0, C_1) = (1,1). \end{cases}$$

From this, you can make the correspondence between the DAG approach and the counterfactual approach explicit. I leave this for the interested reader.

## 17.9 Exercises

1. Show that (17.1) and (17.2) are equivalent.

2. Prove Theorem 17.2.

3. Let $X$, $Y$ and $Z$ have the following joint distribution:

| | $Y=0$ | $Y=1$ |
|---|---|---|
| $X=0$ | .405 | .045 |
| $X=1$ | .045 | .005 |

$Z = 0$

| | $Y=0$ | $Y=1$ |
|---|---|---|
| $X=0$ | .125 | .125 |
| $X=1$ | .125 | .125 |

$Z = 1$

   (a) Find the conditional distribution of $X$ and $Y$ given $Z = 0$ and the conditional distribution of $X$ and $Y$ given $Z = 1$.

   (b) Show that $X \amalg Y | Z$.

   (c) Find the marginal distribution of $X$ and $Y$.

   (d) Show that $X$ and $Y$ are not marginally independent.

4. Consider the three DAGs in Figure 17.6 without a collider. Prove that $X \amalg Z | Y$.

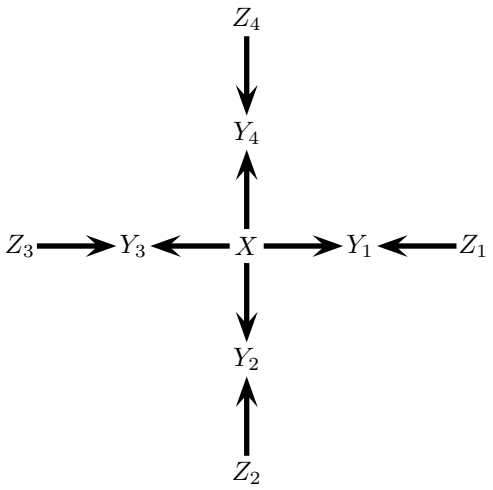



FIGURE 17.12. DAG for exercise 7.

5. Consider the DAG in Figure 17.6 with a collider. Prove that $X \amalg Z$ and that $X$ and $Z$ are dependent given $Y$.

6. Let $X \in \{0, 1\}$, $Y \in \{0, 1\}$, $Z \in \{0, 1, 2\}$. Suppose the distribution of $(X, Y, Z)$ is Markov to:

$$X \longrightarrow Y \longrightarrow Z$$

Create a joint distribution $f(x, y, z)$ that is Markov to this DAG. Generate 1000 random vectors from this distribution. Estimate the distribution from the data using maximum likelihood. Compare the estimated distribution to the true distribution. Let $\theta = (\theta_{000}, \theta_{001}, \ldots, \theta_{112})$ where $\theta_{rst} = \mathbb{P}(X = r, Y = s, Z = t)$. Use the bootstrap to get standard errors and 95 percent confidence intervals for these 12 parameters.

7. Consider the DAG in Figure 17.12.

   (a) Write down the factorization of the joint density.

   (b) Prove that $X \amalg Z_j$.

8. Let $V = (X, Y, Z)$ have the following joint distribution

$$X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

$$\begin{aligned} Y \mid X = x &\sim \text{Bernoulli}\left(\frac{e^{4x-2}}{1+e^{4x-2}}\right) \\ Z \mid X = x, Y = y &\sim \text{Bernoulli}\left(\frac{e^{2(x+y)-2}}{1+e^{2(x+y)-2}}\right). \end{aligned}$$

(a) Find an expression for $\mathbb{P}(Z = z \mid Y = y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y = 1)$.

(b) Write a program to simulate the model. Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y = 1)$ empirically. Plot this as a function of the simulation size $N$. It should converge to the theoretical value you computed in (a).

(c) (Refers to material in the appendix.) Write down an expression for $\mathbb{P}(Z = 1 \mid Y := y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y := 1)$.

(d) (Refers to material in the appendix.) Modify your program to simulate the intervention "set $Y = 1$." Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y := 1)$ empirically. Plot this as a function of the simulation size $N$. It should converge to the theoretical value you computed in (c).

9. This is a continuous, Gaussian version of the last question. Let $V = (X, Y, Z)$ have the following joint distribution

$$\begin{aligned} X &\sim \text{Normal } (0, 1) \\ Y \mid X = x &\sim \text{Normal } (\alpha x, 1) \\ Z \mid X = x, Y = y &\sim \text{Normal } (\beta y + \gamma x, 1). \end{aligned}$$

Here, $\alpha, \beta$ and $\gamma$ are fixed parameters. economists refer to models like this as **structural equation models.**

(a) Find an explicit expression for $f(z \mid y)$ and $\mathbb{E}(Z \mid Y = y) = \int z f(z \mid y) dz$.

(b) (Refers to material in the appendix.) Find an explicit expression for $f(z \mid Y := y)$ and then find $\mathbb{E}(Z \mid Y := y) \equiv \int z f(z \mid Y := y) dy$. Compare to (b).

(c) Find the joint distribution of $(Y, Z)$. Find the correlation $\rho$ between $Y$ and $Z$.

(d) (Refers to material in the appendix.) Suppose that $X$ is not observed and we try to make causal conclusions from the marginal distribution of $(Y, Z)$. (Think of $X$ as unobserved confounding variables.) In particular,

suppose we declare that $Y$ causes $Z$ if $\rho \neq 0$ and we declare that $Y$ does not cause $Z$ if $\rho = 0$. Show that this will lead to erroneous conclusions.

(e) (Refers to material in the appendix.) Suppose we conduct a randomized experiment in which $Y$ is randomly assigned. To be concrete, suppose that

$$\begin{aligned} X &\sim \text{Normal}(0, 1) \\ Y &\sim \text{Normal}(\alpha, 1) \\ Z \mid X = x, Y = y &\sim \text{Normal}(\beta y + \gamma x, 1). \end{aligned}$$

Show that the method in (d) now yields correct conclusions (i.e., $\rho = 0$ if and only if $f(z \mid Y := y)$ does not depend on $y$).

# 18
# Undirected Graphs

Undirected graphs are an alternative to directed graphs for representing independence relations. Since both directed and undirected graphs are used in practice, it is a good idea to be facile with both. The main difference between the two is that the rules for reading independence relations from the graph are different.

## 18.1 Undirected Graphs

An **undirected graph** $\mathcal{G} = (V, E)$ has a finite set $V$ of **vertices (or nodes)** and a set $E$ of **edges (or arcs)** consisting of pairs of vertices. The vertices correspond to random variables $X, Y, Z, \ldots$ and edges are written as unordered pairs. For example, $(X, Y) \in E$ means that $X$ and $Y$ are joined by an edge. An example of a graph is in Figure 18.1.

Two vertices are **adjacent**, written $X \sim Y$, if there is an edge between them. In Figure 18.1, $X$ and $Y$ are adjacent but $X$ and $Z$ are not adjacent. A sequence $X_0, \ldots, X_n$ is called a **path** if $X_{i-1} \sim X_i$ for each $i$. In Figure 18.1, $X, Y, Z$ is a path. A graph is **complete** if there is an edge between every pair of vertices. A subset $U \subset V$ of vertices together with their edges is called a **subgraph.**

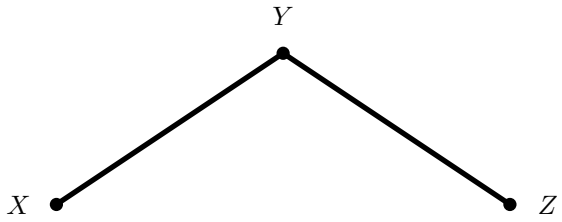



FIGURE 18.1. A graph with vertices $V = \{X, Y, Z\}$. The edge set is $E = \{(X, Y), (Y, Z)\}$.

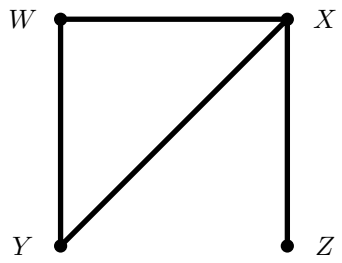



FIGURE 18.2. $\{Y, W\}$ and $\{Z\}$ are separated by $\{X\}$. Also, $W$ and $Z$ are separated by $\{X, Y\}$.

If $A, B$ and $C$ are three distinct subsets of $V$, we say that $C$ **separates** $A$ **and** $B$ if every path from a variable in $A$ to a variable in $B$ intersects a variable in $C$. In Figure 18.2 $\{Y, W\}$ and $\{Z\}$ are separated by $\{X\}$. Also, $W$ and $Z$ are separated by $\{X, Y\}$.

## 18.2 Probability and Graphs

Let $V$ be a set of random variables with distribution $\mathbb{P}$. Construct a graph with one vertex for each random variable in $V$. Omit the edge between a pair of variables if they are independent given the rest of the variables:

$$\text{no edge between } X \text{ and } Y \iff X \amalg Y | \text{rest}$$

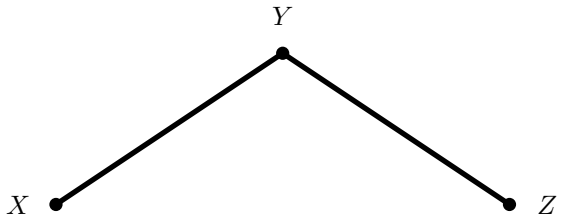



FIGURE 18.3. $X \amalg Z|Y$.

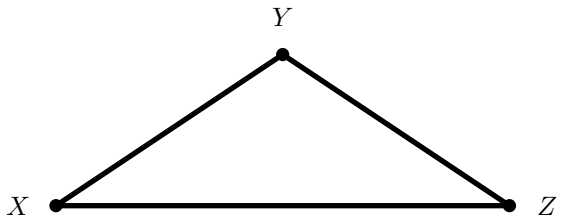



FIGURE 18.4. No implied independence relations.

where "rest" refers to all the other variables besides $X$ and $Y$. The resulting graph is called a **pairwise Markov graph**. Some examples are shown in Figures 18.3, 18.4, 18.5, and 18.6.

The graph encodes a set of pairwise conditional independence relations. These relations imply other conditional independence relations. How can we figure out what they are? Fortunately, we can read these other conditional independence relations directly from the graph as well, as is explained in the next theorem.

**18.1 Theorem.** *Let $\mathcal{G} = (V, E)$ be a pairwise Markov graph for a distribution $\mathbb{P}$. Let $A, B$ and $C$ be distinct subsets of $V$ such that $C$ separates $A$ and $B$. Then $A \amalg B|C$.*

**18.2 Remark.** If $A$ and $B$ are not connected (i.e., there is no path from $A$ to $B$) then we may regard $A$ and $B$ as being separated by the empty set. Then Theorem 18.1 implies that $A \amalg B$.

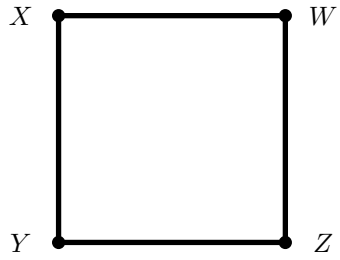



FIGURE 18.5. $X \amalg Z|\{Y, W\}$ and $Y \amalg W|\{X, Z\}$.

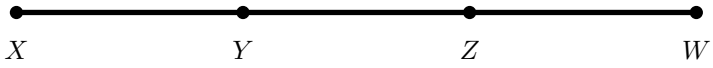



FIGURE 18.6. Pairwise independence implies that $X \amalg Z|\{Y, W\}$. But is $X \amalg Z|Y$?

The independence condition in Theorem 18.1 is called the **global Markov property.** We thus see that the pairwise and global Markov properties are equivalent. Let us state this more precisely. Given a graph $\mathcal{G}$, let $M_{\text{pair}}(\mathcal{G})$ be the set of distributions which satisfy the pairwise Markov property: thus $\mathbb{P} \in M_{\text{pair}}(\mathcal{G})$ if, under $\mathbb{P}$, $X \amalg Y|\text{rest}$ if and only if there is no edge between $X$ and $Y$. Let $M_{\text{global}}(\mathcal{G})$ be the set of distributions which satisfy the global Markov property: thus $\mathbb{P} \in M_{\text{pair}}(\mathcal{G})$ if, under $\mathbb{P}$, $A \amalg B|C$ if and only if $C$ separates $A$ and $B$.

**18.3 Theorem.** *Let $\mathcal{G}$ be a graph. Then, $M_{\text{pair}}(\mathcal{G}) = M_{\text{global}}(\mathcal{G})$.*

Theorem 18.3 allows us to construct graphs using the simpler pairwise property and then we can deduce other independence relations using the global Markov property. Think how hard this would be to do algebraically. Returning to 18.6, we now see that $X \amalg Z|Y$ and $Y \amalg W|Z$.

**18.4 Example.** Figure 18.7 implies that $X \amalg Y$, $X \amalg Z$ and $X \amalg (Y, Z)$. ■

**18.5 Example.** Figure 18.8 implies that $X \amalg W|(Y, Z)$ and $X \amalg Z|Y$. ■

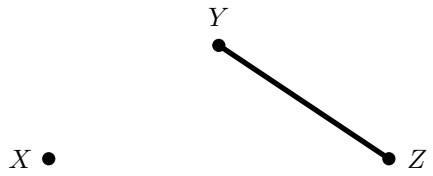



FIGURE 18.7. $X \amalg Y$, $X \amalg Z$ and $X \amalg (Y, Z)$.

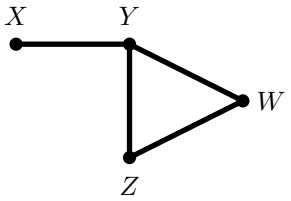



FIGURE 18.8. $X \amalg W|(Y, Z)$ and $X \amalg Z|Y$.

## 18.3 Cliques and Potentials

A **clique** is a set of variables in a graph that are all adjacent to each other. A set of variables is a **maximal clique** if it is a clique and if it is not possible to include another variable and still be a clique. A **potential** is any positive function. Under certain conditions, it can be shown that $\mathbb{P}$ is Markov $\mathcal{G}$ if and only if its probability function $f$ can be written as

$$f(x) = \frac{\prod_{C \in \mathcal{C}} \psi_C(x_C)}{Z} \tag{18.1}$$

where $\mathcal{C}$ is the set of maximal cliques and

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

**18.6 Example.** The maximal cliques for the graph in Figure 18.1 are $C_1 = \{X, Y\}$ and $C_2 = \{Y, Z\}$. Hence, if $\mathbb{P}$ is Markov to the graph, then its probability function can be written

$$f(x, y, z) \propto \psi_1(x, y)\psi_2(y, z)$$

for some positive functions $\psi_1$ and $\psi_2$. ■

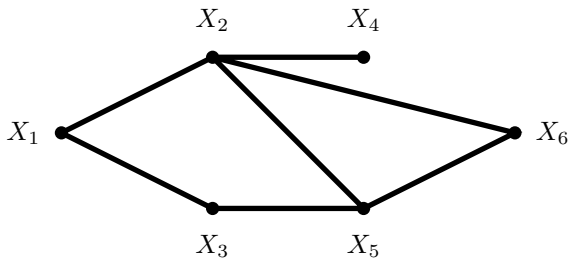



FIGURE 18.9. The maximumly cliques of this graph are $\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_4\}, \{X_3, X_5\}, \{X_2, X_5, X_6\}$.

**18.7 Example.** The maximal cliques for the graph in Figure 18.9 are

$$\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_4\}, \{X_3, X_5\}, \{X_2, X_5, X_6\}.$$

Thus we can write the probability function as

$$\begin{aligned} f(x_1, x_2, x_3, x_4, x_5, x_6) \quad &\propto \quad \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4) \\ &\quad \times \psi_{35}(x_3, x_5)\psi_{256}(x_2, x_5, x_6). \quad \blacksquare \end{aligned}$$

## 18.4 Fitting Graphs to Data

Given a data set, how do we find a graphical model that fits the data? As with directed graphs, this is a big topic that we will not treat here. However, in the discrete case, one way to fit a graph to data is to use a **log-linear model**, which is the subject of the next chapter.

## 18.5 Bibliographic Remarks

Thorough treatments of undirected graphs can be found in Whittaker (1990) and Lauritzen (1996). Some of the exercises below are from Whittaker (1990).

## 18.6 Exercises

1. Consider random variables $(X_1, X_2, X_3)$. In each of the following cases, draw a graph that has the given independence relations.

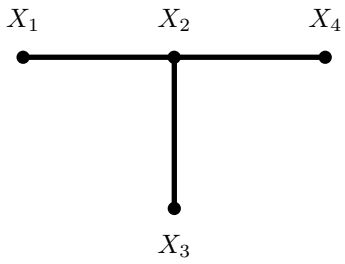



FIGURE 18.10.


FIGURE 18.11.

(a) $X_1 \amalg X_3 \mid X_2$.

(b) $X_1 \amalg X_2 \mid X_3$ and $X_1 \amalg X_3 \mid X_2$.

(c) $X_1 \amalg X_2 \mid X_3$ and $X_1 \amalg X_3 \mid X_2$ and $X_2 \amalg X_3 \mid X_1$.

2. Consider random variables $(X_1, X_2, X_3, X_4)$. In each of the following cases, draw a graph that has the given independence relations.

   (a) $X_1 \amalg X_3 \mid X_2, X_4$ and $X_1 \amalg X_4 \mid X_2, X_3$ and $X_2 \amalg X_4 \mid X_1, X_3$.

   (b) $X_1 \amalg X_2 \mid X_3, X_4$ and $X_1 \amalg X_3 \mid X_2, X_4$ and $X_2 \amalg X_3 \mid X_1, X_4$.

   (c) $X_1 \amalg X_3 \mid X_2, X_4$ and $X_2 \amalg X_4 \mid X_1, X_3$.

3. A conditional independence between a pair of variables is **minimal** if it is not possible to use the Separation Theorem to eliminate any variable from the conditioning set, i.e. from the right hand side of the bar Whittaker (1990). Write down the minimal conditional independencies from: (a) Figure 18.10; (b) Figure 18.11; (c) Figure 18.12; (d) Figure 18.13.

4. Let $X_1, X_2, X_3$ be binary random variables. Construct the likelihood ratio test for

$$H_0 : X_1 \amalg X_2 | X_3 \quad \text{versus} \quad H_1 : X_1 \text{is not independent of } X_2 | X_3.$$

5. Here are breast cancer data from Morrison et al. (1973) on diagnostic center ($X_1$), nuclear grade ($X_2$), and survival ($X_3$):

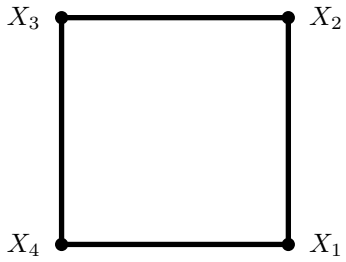



FIGURE 18.12.

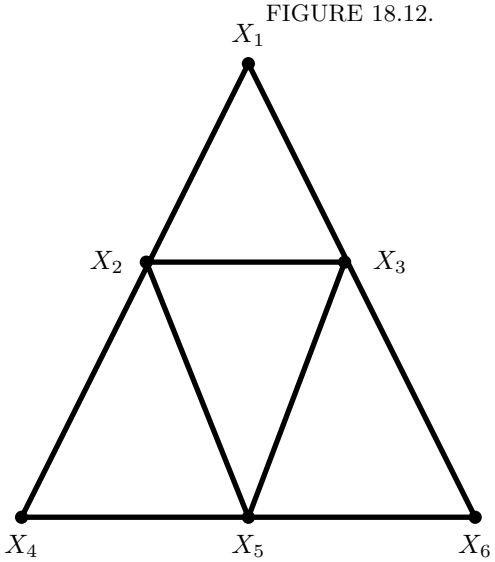



FIGURE 18.13.

| | $X_2$<br>$X_3$ | malignant<br>died | malignant<br>survived | benign<br>died | benign<br>survived |
|---|---|---|---|---|---|
| $X_1$ | Boston | 35 | 59 | 47 | 112 |
| | Glamorgan | 42 | 77 | 26 | 76 |

(a) Treat this as a multinomial and find the maximum likelihood estimator.

(b) If someone has a tumor classified as benign at the Glamorgan clinic, what is the estimated probability that they will die? Find the standard error for this estimate.

(c) Test the following hypotheses:

$$\begin{array}{lll} X_1 \perp\!\!\!\perp X_2|X_3 & \text{versus} & X_1 \not\perp\!\!\!\perp X_2|X_3 \\ X_1 \perp\!\!\!\perp X_3|X_2 & \text{versus} & X_1 \not\perp\!\!\!\perp X_3|X_2 \\ X_2 \perp\!\!\!\perp X_3|X_1 & \text{versus} & X_2 \not\perp\!\!\!\perp X_3|X_1 \end{array}$$

Use the test from question 4. Based on the results of your tests, draw and interpret the resulting graph.

# 19
# Log-Linear Models

In this chapter we study **log-linear models** which are useful for modeling multivariate discrete data. There is a strong connection between log-linear models and undirected graphs.

## 19.1 The Log-Linear Model

Let $X = (X_1, \ldots, X_m)$ be a discrete random vector with probability function

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \ldots, X_m = x_m)$$

where $x = (x_1, \ldots, x_m)$. Let $r_j$ be the number of values that $X_j$ takes. Without loss of generality, we can assume that $X_j \in \{0, 1, \ldots, r_j - 1\}$. Suppose now that we have $n$ such random vectors. We can think of the data as a sample from a Multinomial with $N = r_1 \times r_2 \times \cdots \times r_m$ categories. The data can be represented as counts in a $r_1 \times r_2 \times \cdots \times r_m$ table. Let $p = (p_1, \ldots, p_N)$ denote the multinomial parameter.

Let $S = \{1, \ldots, m\}$. Given a vector $x = (x_1, \ldots, x_m)$ and a subset $A \subset S$, let $x_A = (x_j : \ j \in A)$. For example, if $A = \{1, 3\}$ then $x_A = (x_1, x_3)$.

**19.1 Theorem.** *The joint probability function $f(x)$ of a single random vector $X = (X_1, \ldots, X_m)$ can be written as*

$$\log f(x) = \sum_{A \subset S} \psi_A(x) \tag{19.1}$$

*where the sum is over all subsets $A$ of $S = \{1, \ldots, m\}$ and the $\psi$'s satisfy the following conditions:*

1. *$\psi_\emptyset(x)$ is a constant;*
2. *For every $A \subset S$, $\psi_A(x)$ is only a function of $x_A$ and not the rest of the $x_j's$.*
3. *If $i \in A$ and $x_i = 0$, then $\psi_A(x) = 0$.*

The formula in equation (19.1) is called the **log-linear expansion** of $f$. Each $\psi_A(x)$ may depend on some unknown parameters $\beta_A$. Let $\beta = (\beta_A : A \subset S)$ be the set of all these parameters. We will write $f(x) = f(x; \beta)$ when we want to emphasize the dependence on the unknown parameters $\beta$.

In terms of the multinomial, the parameter space is

$$\mathcal{P} = \left\{ p = (p_1, \ldots, p_N) : \ p_j \geq 0, \ \sum_{j=1}^{N} p_j = 1 \right\}.$$

This is an $N - 1$ dimensional space. In the log-linear representation, the parameter space is

$$\Theta = \left\{ \beta = (\beta_1, \ldots, \beta_N) : \ \beta = \beta(p), p \in \mathcal{P} \right\}$$

where $\beta(p)$ is the set of $\beta$ values associated with $p$. The set $\Theta$ is a $N - 1$ dimensional surface in $\mathbb{R}^N$. We can always go back and forth between the two parameterizations we can write $\beta = \beta(p)$ and $p = p(\beta)$.

**19.2 Example.** Let $X \sim$ Bernoulli$(p)$ where $0 < p < 1$. We can write the probability mass function for $X$ as

$$f(x) = p^x (1-p)^{1-x} = p_1^x \, p_2^{1-x}$$

for $x = 0, 1$, where $p_1 = p$ and $p_2 = 1 - p$. Hence,

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x)$$

where

$$\begin{aligned}\psi_\emptyset(x) &= \log(p_2)\\ \psi_1(x) &= x\log\left(\frac{p_1}{p_2}\right).\end{aligned}$$

Notice that $\psi_\emptyset(x)$ is a constant (as a function of $x$) and $\psi_1(x) = 0$ when $x = 0$. Thus the three conditions of Theorem 19.1 hold. The log-linear parameters are

$$\beta_0 = \log(p_2), \quad \beta_1 = \log\left(\frac{p_1}{p_2}\right).$$

The original, multinomial parameter space is $\mathcal{P} = \{(p_1, p_2) : \ p_j \geq 0, p_1 + p_2 = 1\}$. The log-linear parameter space is

$$\Theta = \left\{(\beta_0, \beta_1) \in \mathbb{R}^2 : \ e^{\beta_0+\beta_1} + e^{\beta_0} = 1.\right\}$$

Given $(p_1, p_2)$ we can solve for $(\beta_0, \beta_1)$. Conversely, given $(\beta_0, \beta_1)$ we can solve for $(p_1, p_2)$. ■

**19.3 Example.** Let $X = (X_1, X_2)$ where $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1, 2\}$. The joint distribution of $n$ such random vectors is a multinomial with 6 categories. The multinomial parameters can be written as a 2-by-3 table as follows:

| multinomial | | $x_2$ | 0 | 1 | 2 |
|---|---|---|---|---|---|
| $x_1$ | 0 | | $p_{00}$ | $p_{01}$ | $p_{02}$ |
| | 1 | | $p_{10}$ | $p_{11}$ | $p_{12}$ |

The $n$ data vectors can be summarized as counts:

| data | | $x_2$ | 0 | 1 | 2 |
|---|---|---|---|---|---|
| $x_1$ | 0 | | $C_{00}$ | $C_{01}$ | $C_{02}$ |
| | 1 | | $C_{10}$ | $C_{11}$ | $C_{12}$ |

For $x = (x_1, x_2)$, the log-linear expansion takes the form

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_{12}(x)$$

where

$$\begin{aligned}\psi_\emptyset(x) &= \log p_{00}\\ \psi_1(x) &= x_1\log\left(\frac{p_{10}}{p_{00}}\right)\\ \psi_2(x) &= I(x_2 = 1)\log\left(\frac{p_{01}}{p_{00}}\right) + I(x_2 = 2)\log\left(\frac{p_{02}}{p_{00}}\right)\\ \psi_{12}(x) &= I(x_1 = 1, x_2 = 1)\log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) + I(x_1 = 1, x_2 = 2)\log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right).\end{aligned}$$

Convince yourself that the three conditions on the $\psi$'s of the theorem are satisfied. The six parameters of this model are:

$$\beta_1 = \log p_{00} \qquad \beta_2 = \log\left(\frac{p_{10}}{p_{00}}\right) \qquad \beta_3 = \log\left(\frac{p_{01}}{p_{00}}\right)$$
$$\beta_4 = \log\left(\frac{p_{02}}{p_{00}}\right) \quad \beta_5 = \log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) \quad \beta_6 = \log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right).$$

■

The next theorem gives an easy way to check for conditional independence in a log-linear model.

**19.4 Theorem.** *Let $(X_a, X_b, X_c)$ be a partition of a vectors $(X_1, \ldots, X_m)$. Then $X_b \amalg X_c | X_a$ if and only if all the $\psi$-terms in the log-linear expansion that have at least one coordinate in $b$ and one coordinate in $c$ are 0.*

To prove this theorem, we will use the following lemma whose proof follows easily from the definition of conditional independence.

**19.5 Lemma.** *A partition $(X_a, X_b, X_c)$ satisfies $X_b \amalg X_c | X_a$ if and only if $f(x_a, x_b, x_c) = g(x_a, x_b)h(x_a, x_c)$ for some functions $g$ and $h$*

PROOF. (Theorem 19.4.) Suppose that $\psi_t$ is 0 whenever $t$ has coordinates in $b$ and $c$. Hence, $\psi_t$ is 0 if $t \not\subset a \bigcup b$ or $t \not\subset a \bigcup c$. Therefore

$$\log f(x) = \sum_{t \subset a \bigcup b} \psi_t(x) + \sum_{t \subset a \bigcup c} \psi_t(x) - \sum_{t \subset a} \psi_t(x).$$

Exponentiating, we see that the joint density is of the form $g(x_a, x_b)h(x_a, x_c)$. By Lemma 19.5, $X_b \amalg X_c | X_a$. The converse follows by reversing the argument. ■

## 19.2 Graphical Log-Linear Models

A log-linear model is **graphical** if missing terms correspond only to conditional independence constraints.

**19.6 Definition.** *Let $\log f(x) = \sum_{A \subset S} \psi_A(x)$ be a log-linear model. Then $f$ is* **graphical** *if all $\psi$-terms are nonzero except for any pair of coordinates not in the edge set for some graph $\mathcal{G}$. In other words, $\psi_A(x) = 0$ if and only if $\{i, j\} \subset A$ and $(i, j)$ is not an edge.*

Here is a way to think about the definition above:

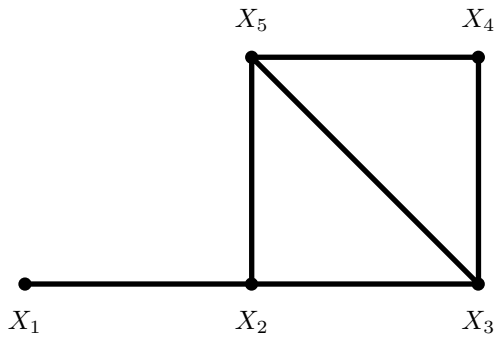



FIGURE 19.1. Graph for Example 19.7.

If you can add a term to the model and the graph does not change, then the model is not graphical.

**19.7 Example.** Consider the graph in Figure 19.1.

The graphical log-linear model that corresponds to this graph is

$$\begin{aligned}\log f(x) &= \psi_\emptyset + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x) + \psi_{235}(x) + \psi_{345}(x).\end{aligned}$$

Let's see why this model is graphical. The edge $(1, 5)$ is missing in the graph. Hence any term containing that pair of indices is omitted from the model. For example,

$$\psi_{15},\ \psi_{125},\ \psi_{135},\ \psi_{145},\ \psi_{1235},\ \psi_{1245},\ \psi_{1345},\ \psi_{12345}$$

are all omitted. Similarly, the edge $(2, 4)$ is missing and hence

$$\psi_{24},\ \psi_{124},\ \psi_{234},\ \psi_{245},\ \psi_{1234},\ \psi_{1245},\ \psi_{2345},\ \psi_{12345}$$

are all omitted. There are other missing edges as well. You can check that the model omits all the corresponding $\psi$ terms. Now consider the model

$$\begin{aligned}\log f(x) &= \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x).\end{aligned}$$

This is the same model except that the three way interactions were removed. If we draw a graph for this model, we will get the same graph. For example, no $\psi$ terms contain $(1, 5)$ so we omit the edge between $X_1$ and $X_5$. But this is not graphical since it has extra terms omitted. The independencies and graphs

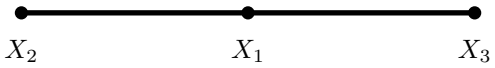



FIGURE 19.2. Graph for Example 19.10.

for the two models are the same but the latter model has other constraints besides conditional independence constraints. This is not a bad thing. It just means that if we are only concerned about presence or absence of conditional independences, then we need not consider such a model. The presence of the three-way interaction $\psi_{235}$ means that the strength of association between $X_2$ and $X_3$ varies as a function of $X_5$. Its absence indicates that this is not so. ■

## 19.3 Hierarchical Log-Linear Models

There is a set of log-linear models that is larger than the set of graphical models and that are used quite a bit. These are the hierarchical log-linear models.

**19.8 Definition.** *A log-linear model is* **hierarchical** *if* $\psi_A = 0$ *and* $A \subset B$ *implies that* $\psi_B = 0$.

**19.9 Lemma.** *A graphical model is hierarchical but the reverse need not be true.*

**19.10 Example.** Let

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_{12}(x) + \psi_{13}(x).$$

The model is hierarchical; its graph is given in Figure 19.2. The model is graphical because all terms involving (2,3) are omitted. It is also hierarchical. ■

**19.11 Example.** Let

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{23}(x).$$

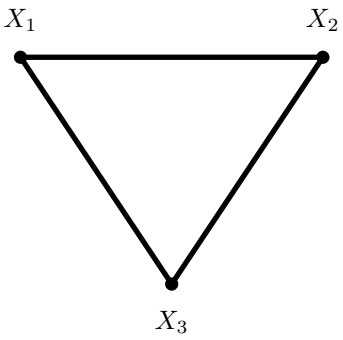



FIGURE 19.3. The graph is complete. The model is hierarchical but not graphical.

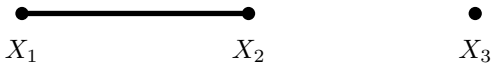



FIGURE 19.4. The model for this graph is not hierarchical.

The model is hierarchical. It is not graphical. The graph corresponding to this model is complete; see Figure 19.3. It is not graphical because $\psi_{123}(x) = 0$ which does not correspond to any pairwise conditional independence. ∎

**19.12 Example.** Let

$$\log f(x) = \psi_\emptyset(x) + \psi_3(x) + \psi_{12}(x).$$

The graph corresponding is in Figure 19.4. This model is not hierarchical since $\psi_2 = 0$ but $\psi_{12}$ is not. Since it is not hierarchical, it is not graphical either. ∎

## 19.4 Model Generators

Hierarchical models can be written succinctly using **generators**. This is most easily explained by example. Suppose that $X = (X_1, X_2, X_3)$. Then, $M = 1.2 + 1.3$ stands for

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13}.$$

The formula $M = 1.2{+}1.3$ says: "include $\psi_{12}$ and $\psi_{13}$." We have to also include the lower order terms or it won't be hierarchical. The generator $M = 1.2.3$ is the **saturated** model

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13} + \psi_{23} + \psi_{123}.$$

The saturated models corresponds to fitting an unconstrained multinomial. Consider $M = 1 + 2 + 3$ which means

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3.$$

This is the mutual independence model. Finally, consider $M = 1.2$ which has log-linear expansion

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_{12}.$$

This model makes $X_3|X_2 = x_2, X_1 = x_1$ a uniform distribution.

## 19.5 Fitting Log-Linear Models to Data

Let $\beta$ denote all the parameters in a log-linear model $M$. The loglikelihood for $\beta$ is

$$\ell(\beta) = \sum_{i=1}^{n} \log f(X_i; \beta)$$

where $f(X_i; \beta)$ is the probability function for the $i^{\text{th}}$ random vector $X_i = (X_{i1}, \ldots, X_{im})$ as give by equation (19.1). The MLE $\widehat{\beta}$ generally has to be found numerically. The Fisher information matrix is also found numerically and we can then get the estimated standard errors from the inverse Fisher information matrix.

When fitting log-linear models, one has to address the following model selection problem: which $\psi$ terms should we include in the model? This is essentially the same as the model selection problem in linear regression.

One approach is is to use AIC. Let $M$ denote some log-linear model. Different models correspond to setting different $\psi$ terms to 0. Now we choose the model $M$ which maximizes

$$\text{AIC}(M) = \widehat{\ell}(M) - |M| \tag{19.2}$$

where $|M|$ is the number of parameters in model $M$ and $\widehat{\ell}(M)$ is the value of the log-likelihood evaluated at the MLE for that model. Usually the model search is restricted to hierarchical models. This reduces the search space. Some

also claim that we should only search through the hierarchical models because other models are less interpretable.

A different approach is based on hypothesis testing. The model that includes all possible $\psi$-terms is called the **saturated model** and we denote it by $M_{sat}$. Now for each $M$ we test the hypothesis

$$H_0 : \text{the true model is } M \quad \text{versus} \quad H_1 : \text{the true model is } M_{sat}.$$

The likelihood ratio test for this hypothesis is called the deviance.

**19.13 Definition.** *For any submodel $M$, define the* **deviance** $\text{dev}(M)$ *by*

$$\text{dev}(M) = 2(\widehat{\ell}_{sat} - \widehat{\ell}_M)$$

*where $\widehat{\ell}_{sat}$ is the log-likelihood of the saturated model evaluated at the* MLE *and $\widehat{\ell}_M$ is the log-likelihood of the model $M$ evaluated at its* MLE.

**19.14 Theorem.** *The deviance is the likelihood ratio test statistic for*

$$H_0 : \text{the model is } M \quad \text{versus} \quad H_1 : \text{the model is } M_{sat}.$$

*Under $H_0$, $dev(M) \xrightarrow{d} \chi^2_\nu$ with $\nu$ degrees of freedom equal to the difference in the number of parameters between the saturated model and $M$.*

One way to find a good model is to use the deviance to test every sub-model. Every model that is not rejected by this test is then considered a plausible model. However, this is not a good strategy for two reasons. First, we will end up doing many tests which means that there is ample opportunity for making Type I and Type II errors. Second, we will end up using models where we failed to reject $H_0$. But we might fail to reject $H_0$ due to low power. The result is that we end up with a bad model just due to low power.

After finding a "best model" this way we can draw the corresponding graph.

**19.15 Example.** The following breast cancer data are from Morrison et al. (1973). The data are on diagnostic center ($X_1$), nuclear grade ($X_2$), and survival ($X_3$):

| | $X_2$ | malignant | malignant | benign | benign |
|---|---|---|---|---|---|
| | $X_3$ | died | survived | died | survived |
| $X_1$ | Boston | 35 | 59 | 47 | 112 |
| | Glamorgan | 42 | 77 | 26 | 76 |

The saturated log-linear model is:

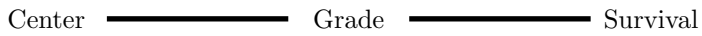



FIGURE 19.5. The graph for Example 19.15.

| Variable | $\widehat{\beta}_j$ | $\widehat{\mathsf{se}}$ | $W_j$ | p-value |
|---|---|---|---|---|
| (Intercept) | 3.56 | 0.17 | 21.03 | 0.00 *** |
| center | 0.18 | 0.22 | 0.79 | 0.42 |
| grade | 0.29 | 0.22 | 1.32 | 0.18 |
| survival | 0.52 | 0.21 | 2.44 | 0.01 * |
| center×grade | -0.77 | 0.33 | -2.31 | 0.02 * |
| center×survival | 0.08 | 0.28 | 0.29 | 0.76 |
| grade×survival | 0.34 | 0.27 | 1.25 | 0.20 |
| center×grade×survival | 0.12 | 0.40 | 0.29 | 0.76 |

The best sub-model, selected using AIC and backward searching is:

| Variable | $\widehat{\beta}_j$ | $\widehat{\mathsf{se}}$ | $W_j$ | p-value |
|---|---|---|---|---|
| (Intercept) | 3.52 | 0.13 | 25.62 | $<$ 0.00 *** |
| center | 0.23 | 0.13 | 1.70 | 0.08 |
| grade | 0.26 | 0.18 | 1.43 | 0.15 |
| survival | 0.56 | 0.14 | 3.98 | 6.65e-05 *** |
| center×grade | -0.67 | 0.18 | -3.62 | 0.00 *** |
| grade×survival | 0.37 | 0.19 | 1.90 | 0.05 |

The graph for this model $M$ is shown in Figure 19.5. To test the fit of this model, we compute the deviance of $M$ which is 0.6. The appropriate $\chi^2$ has $8 - 6 = 2$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_2 > .6) = .74$. So we have no evidence to suggest that the model is a poor fit. ■

## 19.6 Bibliographic Remarks

For this chapter, I drew heavily on Whittaker (1990) which is an excellent text on log-linear models and graphical models. Some of the exercises are from Whittaker. A classic reference on log-linear models is Bishop et al. (1975).

## 19.7 Exercises

1. Solve for the $p'_{ij}s$ in terms of the $\beta$'s in Example 19.3.

2. Prove Lemma 19.5.

3. Prove Lemma 19.9.

4. Consider random variables $(X_1, X_2, X_3, X_4)$. Suppose the log-density is

$$\log f(x) = \psi_\emptyset(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{24}(x) + \psi_{34}(x).$$

   (a) Draw the graph $G$ for these variables.

   (b) Write down all independence and conditional independence relations implied by the graph.

   (c) Is this model graphical? Is it hierarchical?

5. Suppose that parameters $p(x_1, x_2, x_3)$ are proportional to the following values:

| | $x_2$ | 0 | 0 | 1 | 1 |
|---|---|---|---|---|---|
| | $x_3$ | 0 | 1 | 0 | 1 |
| $x_1$ | 0 | 2 | 8 | 4 | 16 |
| | 1 | 16 | 128 | 32 | 256 |

   Find the $\psi$-terms for the log-linear expansion. Comment on the model.

6. Let $X_1, \ldots, X_4$ be binary. Draw the independence graphs corresponding to the following log-linear models. Also, identify whether each is graphical and/or hierarchical (or neither).

   (a) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4$

   (b) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$

   (c) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$

   (d) $\log f = 7 + 5055x_1x_2x_3x_4$

# 20
# Nonparametric Curve Estimation

In this Chapter we discuss nonparametric estimation of probability density functions and regression functions which we refer to as **curve estimation** or **smoothing**.

In Chapter 7 we saw that it is possible to consistently estimate a cumulative distribution function $F$ without making any assumptions about $F$. If we want to estimate a probability density function $f(x)$ or a regression function $r(x) = \mathbb{E}(Y|X = x)$ the situation is different. We cannot estimate these functions consistently without making some smoothness assumptions. Correspondingly, we need to perform some sort of smoothing operation on the data.

An example of a density estimator is a **histogram**, which we discuss in detail in Section 20.2. To form a histogram estimator of a density $f$, we divide the real line to disjoint sets called **bins**. The histogram estimator is a piecewise constant function where the height of the function is proportional to number of observations in each bin; see Figure 20.3. The number of bins is an example of a **smoothing parameter**. If we smooth too much (large bins) we get a highly biased estimator while if we smooth too little (small bins) we get a highly variable estimator. Much of curve estimation is concerned with trying to optimally balance variance and bias.

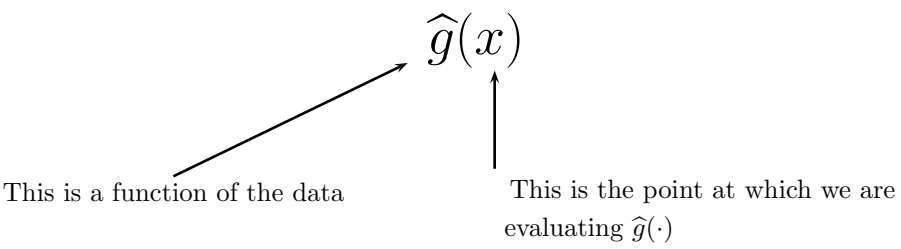



FIGURE 20.1. A curve estimate $\widehat{g}$ is random because it is a function of the data. The point $x$ at which we evaluate $\widehat{g}$ is not a random variable.

## 20.1 The Bias-Variance Tradeoff

Let $g$ denote an unknown function such as a density function or a regression function. Let $\widehat{g}_n$ denote an estimator of $g$. Bear in mind that $\widehat{g}_n(x)$ is a random function evaluated at a point $x$. The estimator is random because it depends on the data. See Figure 20.1.

As a loss function, we will use the **integrated squared error (ISE)**: [1]

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 \, du. \tag{20.1}$$

The **risk** or **mean integrated squared error (MISE)** with respect to squared error loss is

$$R(f, \widehat{f}) = \mathbb{E}\Big(L(g, \widehat{g})\Big). \tag{20.2}$$

**20.1 Lemma.** *The risk can be written as*

$$R(g, \widehat{g}_n) = \int b^2(x) \, dx + \int v(x) \, dx \tag{20.3}$$

*where*

$$b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x) \tag{20.4}$$

*is the bias of $\widehat{g}_n(x)$ at a fixed $x$ and*

$$v(x) = \mathbb{V}(\widehat{g}_n(x)) = \mathbb{E}\bigg(\big(\widehat{g}_n(x) - \mathbb{E}\big(\widehat{g}_n(x)\big)^2\big)\bigg) \tag{20.5}$$

*is the variance of $\widehat{g}_n(x)$ at a fixed $x$.*

[1] We could use other loss functions. The results are similar but the analysis is much more complicated.

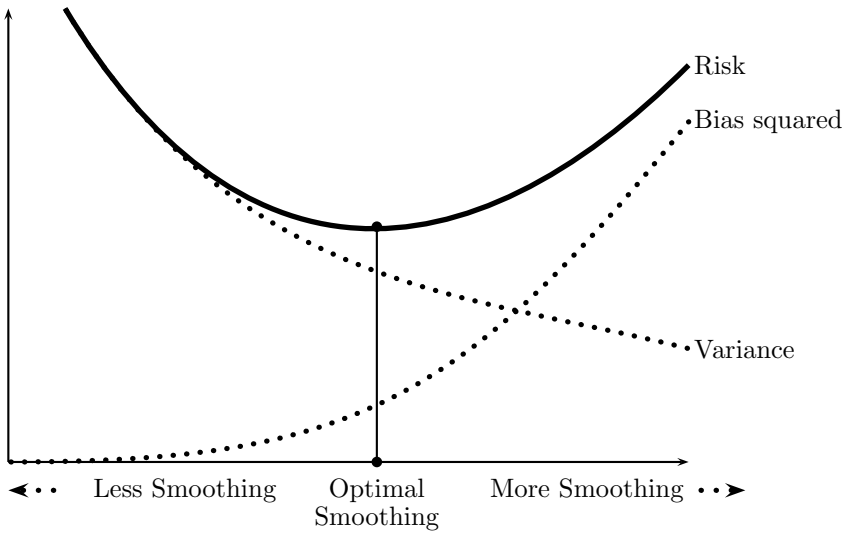



FIGURE 20.2. The Bias-Variance trade-off. The bias increases and the variance decreases with the amount of smoothing. The optimal amount of smoothing, indicated by the vertical line, minimizes the risk = bias$^2$ + variance.

In summary,

$$\text{RISK} = \text{BIAS}^2 + \text{VARIANCE}. \tag{20.6}$$

When the data are oversmoothed, the bias term is large and the variance is small. When the data are undersmoothed the opposite is true; see Figure 20.2. This is called the **bias-variance tradeoff**. Minimizing risk corresponds to balancing bias and variance.

## 20.2 Histograms

Let $X_1, \ldots, X_n$ be IID on $[0,1]$ with density $f$. The restriction to $[0,1]$ is not crucial; we can always rescale the data to be on this interval. Let $m$ be an

integer and define **bins**

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ \ldots, \ B_m = \left[\frac{m-1}{m}, 1\right]. \tag{20.7}$$

Define the **binwidth** $h = 1/m$, let $\nu_j$ be the number of observations in $B_j$, let $\widehat{p}_j = \nu_j/n$ and let $p_j = \int_{B_j} f(u)du$.

The **histogram estimator** is defined by

$$\widehat{f}_n(x) = \begin{cases} \widehat{p}_1/h & x \in B_1 \\ \widehat{p}_2/h & x \in B_2 \\ \vdots & \vdots \\ \widehat{p}_m/h & x \in B_m \end{cases}$$

which we can write more succinctly as

$$\widehat{f}_n(x) = \sum_{j=1}^{n} \frac{\widehat{p}_j}{h} I(x \in B_j). \tag{20.8}$$

To understand the motivation for this estimator, let $p_j = \int_{B_j} f(u)du$ and note that, for $x \in B_j$ and $h$ small,

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{\mathbb{E}(\widehat{p}_j)}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{f(x)} = f(x).$$

**20.2 Example.** Figure 20.3 shows three different histograms based on $n = 1,266$ data points from an astronomical sky survey. Each data point represents the distance from us to a galaxy. The galaxies lie on a "pencilbeam" pointing directly from the Earth out into space. Because of the finite speed of light, looking at galaxies farther and farther away corresponds to looking back in time. Choosing the right number of bins involves finding a good tradeoff between bias and variance. We shall see later that the top left histogram has too few bins resulting in oversmoothing and too much bias. The bottom left histogram has too many bins resulting in undersmoothing and too few bins. The top right histogram is just right. The histogram reveals the presence of clusters of galaxies. Seeing how the size and number of galaxy clusters varies with time, helps cosmologists understand the evolution of the universe. ■

The mean and variance of $\widehat{f}_n(x)$ are given in the following Theorem.

**20.3 Theorem.** *Consider fixed $x$ and fixed $m$, and let $B_j$ be the bin containing $x$. Then,*

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}. \tag{20.9}$$

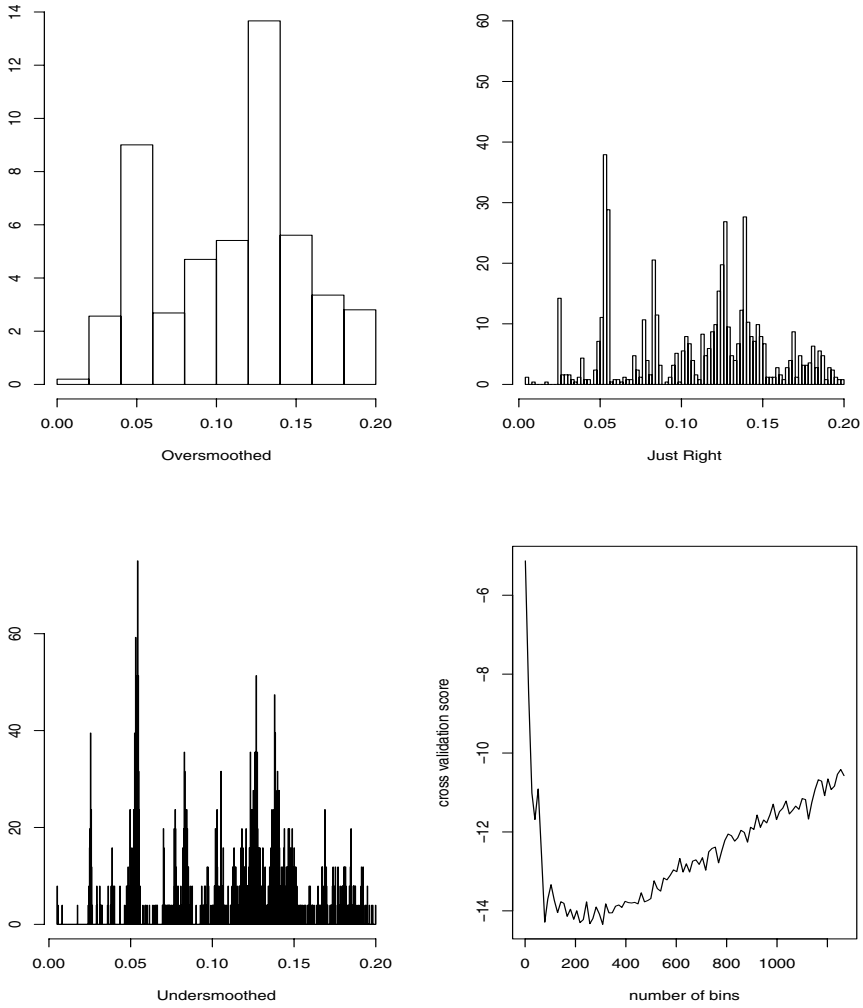



FIGURE 20.3. Three versions of a histogram for the astronomy data. The top left histogram has too few bins. The bottom left histogram has too many bins. The top right histogram is just right. The lower, right plot shows the estimated risk versus the number of bins.

Let's take a closer look at the bias-variance tradeoff using equation (20.9). Consider some $x \in B_j$. For any other $u \in B_j$,

$$f(u) \approx f(x) + (u-x)f'(x)$$

and so

$$\begin{aligned} p_j = \int_{B_j} f(u)du &\approx \int_{B_j} \Big( f(x) + (u-x)f'(x) \Big) du \\ &= f(x)h + hf'(x) \left( h\left(j - \frac{1}{2}\right) - x \right). \end{aligned}$$

Therefore, the bias $b(x)$ is

$$\begin{aligned} b(x) &= \mathbb{E}(\widehat{f}_n(x)) - f(x) = \frac{p_j}{h} - f(x) \\ &\approx \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right) - x\right)}{h} - f(x) \\ &= f'(x)\left( h\left(j - \frac{1}{2}\right) - x \right). \end{aligned}$$

If $\widetilde{x}_j$ is the center of the bin, then

$$\begin{aligned} \int_{B_j} b^2(x)\, dx &\approx \int_{B_j} (f'(x))^2 \left( h\left(j - \frac{1}{2}\right) - x \right)^2 dx \\ &\approx (f'(\widetilde{x}_j))^2 \int_{B_j} \left( h\left(j - \frac{1}{2}\right) - x \right)^2 dx \\ &= (f'(\widetilde{x}_j))^2 \frac{h^3}{12}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 b^2(x)dx &= \sum_{j=1}^m \int_{B_j} b^2(x)dx \approx \sum_{j=1}^m (f'(\widetilde{x}_j))^2 \frac{h^3}{12} \\ &= \frac{h^2}{12} \sum_{j=1}^m h\,(f'(\widetilde{x}_j))^2 \approx \frac{h^2}{12} \int_0^1 (f'(x))^2 dx. \end{aligned}$$

Note that this increases as a function of $h$. Now consider the variance. For $h$ small, $1 - p_j \approx 1$, so

$$\begin{aligned} v(x) &\approx \frac{p_j}{nh^2} \\ &= \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right) - x\right)}{nh^2} \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

where we have kept only the dominant term. So,

$$\int_0^1 v(x)dx \approx \frac{1}{nh}.$$

Note that this decreases with $h$. Putting all this together, we get:

**20.4 Theorem.** *Suppose that* $\int (f'(u))^2 du < \infty$. *Then*

$$R(\widehat{f}_n, f) \approx \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}. \tag{20.10}$$

*The value* $h^*$ *that minimizes (20.10) is*

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}. \tag{20.11}$$

*With this choice of binwidth,*

$$R(\widehat{f}_n, f) \approx \frac{C}{n^{2/3}} \tag{20.12}$$

*where* $C = (3/4)^{2/3}\left(\int (f'(u))^2 du\right)^{1/3}$.

Theorem 20.4 is quite revealing. We see that with an optimally chosen binwidth, the MISE decreases to 0 at rate $n^{-2/3}$. By comparison, most parametric estimators converge at rate $n^{-1}$. The slower rate of convergence is the price we pay for being nonparametric. The formula for the optimal binwidth $h^*$ is of theoretical interest but it is not useful in practice since it depends on the unknown function $f$.

A practical way to choose the binwidth is to estimate the risk function and minimize over $h$. Recall that the loss function, which we now write as a function of $h$, is

$$\begin{aligned} L(h) &= \int (\widehat{f}_n(x) - f(x))^2\, dx \\ &= \int \widehat{f}_n^2(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx + \int f^2(x)\, dx. \end{aligned}$$

The last term does not depend on the binwidth $h$ so minimizing the risk is equivalent to minimizing the expected value of

$$J(h) = \int \widehat{f}_n^2(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx.$$

We shall refer to $\mathbb{E}(J(h))$ as the risk, although it differs from the true risk by the constant term $\int f^2(x)\,dx$.

**20.5 Definition.** *The* **cross-validation estimator of risk** *is*

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i) \tag{20.13}$$

*where $\widehat{f}_{(-i)}$ is the histogram estimator obtained after removing the $i^{\text{th}}$ observation. We refer to $\widehat{J}(h)$ as the cross-validation score or estimated risk.*

**20.6 Theorem.** *The cross-validation estimator is nearly unbiased:*

$$\mathbb{E}(\widehat{J}(x)) \approx \mathbb{E}(J(x)).$$

In principle, we need to recompute the histogram $n$ times to compute $\widehat{J}(h)$. Moreover, this has to be done for all values of $h$. Fortunately, there is a shortcut formula.

**20.7 Theorem.** *The following identity holds:*

$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)}\sum_{j=1}^{m} \widehat{p}_j^2. \tag{20.14}$$

**20.8 Example.** We used cross-validation in the astronomy example. The cross-validation function is quite flat near its minimum. Any $m$ in the range of 73 to 310 is an approximate minimizer but the resulting histogram does not change much over this range. The histogram in the top right plot in Figure 20.3 was constructed using $m = 73$ bins. The bottom right plot shows the estimated risk, or more precisely, $\widehat{A}$, plotted versus the number of bins. ■

Next we want a confidence set for $f$. Suppose $\widehat{f}_n$ is a histogram with $m$ bins and binwidth $h = 1/m$. We cannot realistically make confidence statements about the fine details of the true density $f$. Instead, we shall make confidence statements about $f$ at the resolution of the histogram. To this end, define

$$\overline{f}_n(x) = \mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{for } x \in B_j \tag{20.15}$$

where $p_j = \int_{B_j} f(u)du$. Think of $\overline{f}(x)$ as a "histogramized" version of $f$.

**20.9 Definition.** *A pair of functions* $(\ell_n(x), u_n(x))$ *is a* $1-\alpha$ **confidence band (or confidence envelope)** *if*

$$\mathbb{P}\Big(\ell(x) \leq \overline{f}_n(x) \leq u(x) \ \text{ for all } x\Big) \geq 1-\alpha. \tag{20.16}$$

**20.10 Theorem.** *Let* $m = m(n)$ *be the number of bins in the histogram* $\widehat{f}_n$. *Assume that* $m(n) \to \infty$ *and* $m(n)\log n/n \to 0$ *as* $n \to \infty$. *Define*

$$\begin{aligned} \ell_n(x) &= \left(\max\left\{\sqrt{\widehat{f}_n(x)} - c, 0\right\}\right)^2 \\ u_n(x) &= \left(\sqrt{\widehat{f}_n(x)} + c\right)^2 \end{aligned} \tag{20.17}$$

*where*

$$c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}. \tag{20.18}$$

*Then,* $(\ell_n(x), u_n(x))$ *is an approximate* $1-\alpha$ *confidence band.*

PROOF. Here is an outline of the proof. From the central limit theorem, $\widehat{p}_j \approx N(p_j, p_j(1-p_j)/n)$. By the delta method, $\sqrt{\widehat{p}_j} \approx N(\sqrt{p_j}, 1/(4n))$. Moreover, it can be shown that the $\sqrt{\widehat{p}_j}$'s are approximately independent. Therefore,

$$2\sqrt{n}\left(\sqrt{\widehat{p}_j} - \sqrt{p}_j\right) \approx Z_j \tag{20.19}$$

where $Z_1, \ldots, Z_m \sim N(0,1)$. Let

$$A = \left\{\ell_n(x) \leq \overline{f}_n(x) \leq u_n(x) \ \text{ for all } x\right\} = \left\{\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)}\right| \leq c\right\}.$$

Then,

$$\begin{aligned} \mathbb{P}(A^c) &= \mathbb{P}\left(\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)}\right| > c\right) = \mathbb{P}\left(\max_j \left|\sqrt{\frac{\widehat{p}_j}{h}} - \sqrt{\frac{p_j}{h}}\right| > c\right) \\ &= \mathbb{P}\left(\max_j 2\sqrt{n}\left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > z_{\alpha/(2m)}\right) \\ &\approx \mathbb{P}\left(\max_j |Z_j| > z_{\alpha/(2m)}\right) \leq \sum_{j=1}^m \mathbb{P}\left(|Z_j| > z_{\alpha/(2m)}\right) \\ &= \sum_{j=1}^m \frac{\alpha}{m} = \alpha. \quad \blacksquare \end{aligned}$$

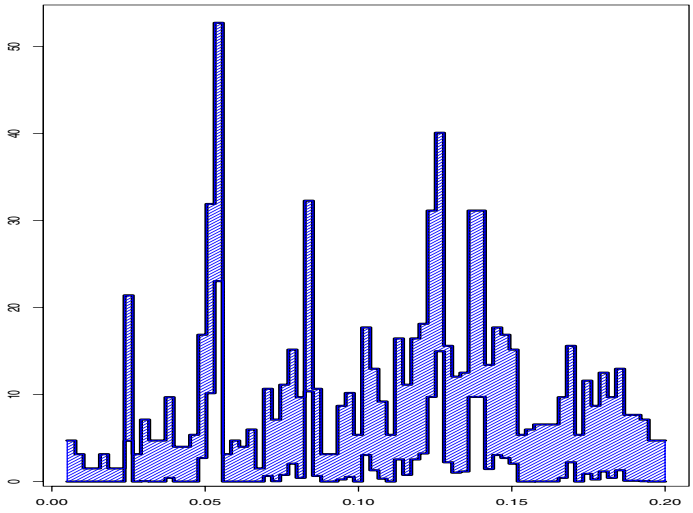



FIGURE 20.4. 95 percent confidence envelope for astronomy data using $m = 73$ bins.

**20.11 Example.** Figure 20.4 shows a 95 percent confidence envelope for the astronomy data. We see that even with over 1,000 data points, there is still substantial uncertainty. ■

## 20.3 Kernel Density Estimation

Histograms are discontinuous. **Kernel density estimators** are smoother and they converge faster to the true density than histograms.

Let $X_1, \ldots, X_n$ denote the observed data, a sample from $f$. In this chapter, a **kernel** is defined to be any smooth function $K$ such that $K(x) \geq 0$, $\int K(x)\, dx = 1$, $\int xK(x)dx = 0$ and $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$. Two examples of kernels are the **Epanechnikov kernel**

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2/5)/\sqrt{5} & |x| < \sqrt{5} \\ 0 & otherwise \end{cases} \tag{20.20}$$

and the Gaussian (Normal) kernel $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

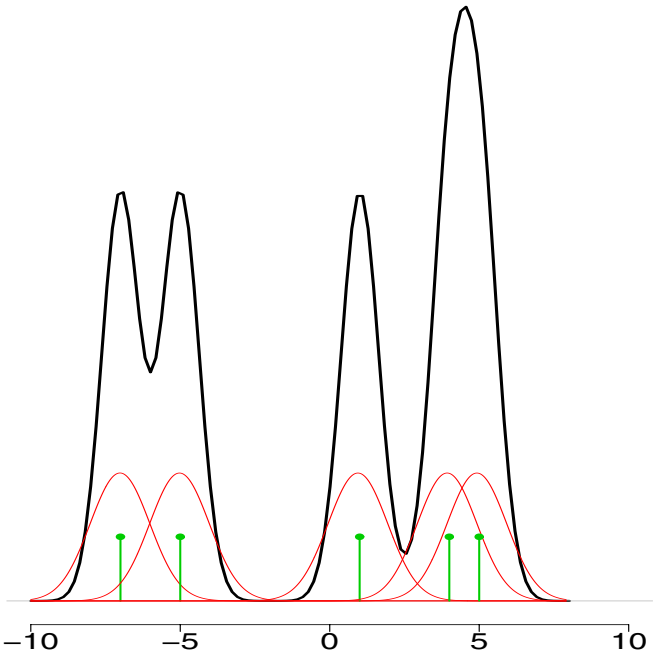



FIGURE 20.5. A kernel density estimator $\widehat{f}$. At each point $x$, $\widehat{f}(x)$ is the average of the kernels centered over the data points $X_i$. The data points are indicated by short vertical bars.

**20.12 Definition.** *Given a kernel $K$ and a positive number $h$, called the* **bandwidth**, *the* **kernel density estimator** *is defined to be*

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{20.21}$$

An example of a kernel density estimator is show in Figure 20.5. The kernel estimator effectively puts a smoothed-out lump of mass of size $1/n$ over each data point $X_i$. The bandwidth $h$ controls the amount of smoothing. When $h$ is close to 0, $\widehat{f}_n$ consists of a set of spikes, one at each data point. The height of the spikes tends to infinity as $h \to 0$. When $h \to \infty$, $\widehat{f}_n$ tends to a uniform density.

**20.13 Example.** Figure 20.6 shows kernel density estimators for the astronomy data using three different bandwidths. In each case we used a Gaussian kernel. The properly smoothed kernel density estimator in the top right panel shows similar structure as the histogram. However, it is easier to see the clusters with the kernel estimator. ■

To construct a kernel density estimator, we need to choose a kernel $K$ and a bandwidth $h$. It can be shown theoretically and empirically that the choice of $K$ is not crucial. [2] However, the choice of bandwidth $h$ is very important. As with the histogram, we can make a theoretical statement about how the risk of the estimator depends on the bandwidth.

**20.14 Theorem.** *Under weak assumptions on $f$ and $K$,*

$$R(f, \widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x)dx}{nh} \tag{20.22}$$

*where $\sigma_K^2 = \int x^2 K(x)dx$. The optimal bandwidth is*

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}} \tag{20.23}$$

*where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $c_3 = \int (f''(x))^2 dx$. With this choice of bandwidth,*

$$R(f, \widehat{f}_n) \approx \frac{c_4}{n^{4/5}}$$

*for some constant $c_4 > 0$.*

PROOF. Write $K_h(x, X) = h^{-1} K\left((x - X)/h\right)$ and $\widehat{f}_n(x) = n^{-1} \sum_i K_h(x, X_i)$. Thus, $\mathbb{E}[\widehat{f}_n(x)] = \mathbb{E}[K_h(x, X)]$ and $\mathbb{V}[\widehat{f}_n(x)] = n^{-1}\mathbb{V}[K_h(x, X)]$. Now,

$$\begin{aligned}
\mathbb{E}[K_h(x, X)] &= \int \frac{1}{h} K\left(\frac{x - t}{h}\right) f(t)\, dt \\
&= \int K(u) f(x - hu)\, du \\
&= \int K(u) \left[f(x) - hf'(x) + \frac{1}{2} f''(x) + \cdots\right] du \\
&= f(x) + \frac{1}{2} h^2 f''(x) \int u^2 K(u)\, du \cdots
\end{aligned}$$

since $\int K(x)\, dx = 1$ and $\int x\, K(x)\, dx = 0$. The bias is

$$\mathbb{E}[K_h(x, X)] - f(x) \approx \frac{1}{2}\sigma_k^2 h^2 f''(x).$$

[2] It can be shown that the Epanechnikov kernel is optimal in the sense of giving smallest asymptotic mean squared error, but it is really the choice of bandwidth which is crucial.

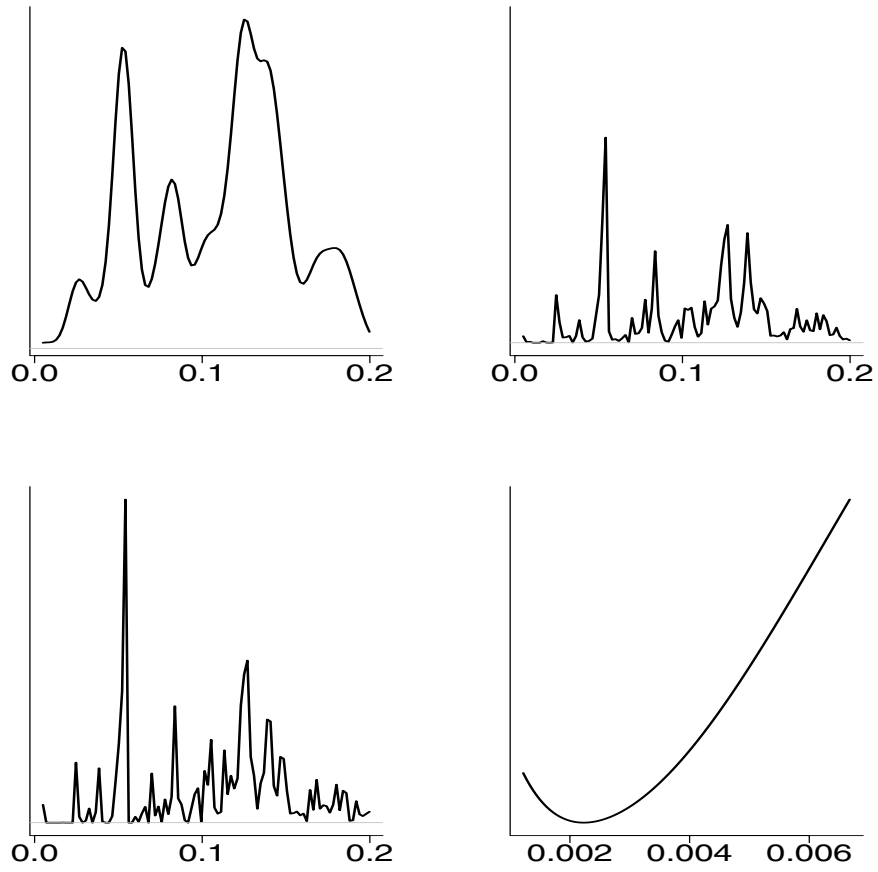



FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth $h$. The bandwidth was chosen to be the value of $h$ where the curve is a minimum.

By a similar calculation,

$$\mathbb{V}[\widehat{f}_n(x)] \approx \frac{f(x)\int K^2(x)\,dx}{n\,h_n}.$$

The result follows from integrating the squared bias plus the variance. ■

We see that kernel estimators converge at rate $n^{-4/5}$ while histograms converge at the slower rate $n^{-2/3}$. It can be shown that, under weak assumptions, there does not exist a nonparametric estimator that converges faster than $n^{-4/5}$.

The expression for $h^*$ depends on the unknown density $f$ which makes the result of little practical use. As with the histograms, we shall use cross-validation to find a bandwidth. Thus, we estimate the risk (up to a constant) by

$$\widehat{J}(h) = \int \widehat{f}^2(x)dz - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i) \tag{20.24}$$

where $\widehat{f}_{-i}$ is the kernel density estimator after omitting the $i^{\text{th}}$ observation.

**20.15 Theorem.** *For any $h > 0$,*

$$\mathbb{E}\left[\widehat{J}(h)\right] = \mathbb{E}\left[J(h)\right].$$

*Also,*

$$\widehat{J}(h) \approx \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) \tag{20.25}$$

*where $K^*(x) = K^{(2)}(x) - 2K(x)$ and $K^{(2)}(z) = \int K(z-y)K(y)dy$. In particular, if $K$ is a N(0,1) Gaussian kernel then $K^{(2)}(z)$ is the $N(0,2)$ density.*

We then choose the bandwidth $h_n$ that minimizes $\widehat{J}(h)$.[3] A justification for this method is given by the following remarkable theorem due to Stone.

**20.16 Theorem** (Stone's Theorem)**.** *Suppose that $f$ is bounded. Let $\widehat{f}_h$ denote the kernel estimator with bandwidth $h$ and let $h_n$ denote the bandwidth chosen by cross-validation. Then,*

$$\frac{\int\left(f(x) - \widehat{f}_{h_n}(x)\right)^2 dx}{\inf_h \int\left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{\text{P}} 1. \tag{20.26}$$

[3]For large data sets, $\widehat{f}$ and (20.25) can be computed quickly using the fast Fourier transform.

**20.17 Example.** The top right panel of Figure 20.6 is based on cross-validation. These data are rounded which problems for cross-validation. Specifically, it causes the minimizer to be $h = 0$. To overcome this problem, we added a small amount of random Normal noise to the data. The result is that $\widehat{J}(h)$ is very smooth with a well defined minimum. ■

**20.18 Remark.** Do not assume that, if the estimator $\widehat{f}$ is wiggly, then cross-validation has let you down. The eye is not a good judge of risk.

To construct confidence bands, we use something similar to histograms. Again, the confidence band is for the smoothed version,

$$\overline{f}_n = \mathbb{E}(\widehat{f}_n(x)) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\, du,$$

of the true density $f$. [4] Assume the density is on an interval $(a, b)$. The band is

$$\ell_n(x) = \widehat{f}_n(x) - q\ \mathsf{se}(x), \quad u_n(x) = \widehat{f}_n(x) + q\ \mathsf{se}(x) \tag{20.27}$$

where

$$\begin{aligned}
\mathsf{se}(x) &= \frac{s(x)}{\sqrt{n}}, \\
s^2(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(x) - \overline{Y}_n(x))^2, \\
Y_i(x) &= \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \\
q &= \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right), \\
m &= \frac{b-a}{\omega}
\end{aligned}$$

where $\omega$ is the width of the kernel. In case the kernel does not have finite width then we take $\omega$ to be the effective width, that is, the range over which the kernel is non-negligible. In particular, we take $\omega = 3h$ for the Normal kernel.

**20.19 Example.** Figure 20.7 shows approximate 95 percent confidence bands for the astronomy data. ■

[4] This is a modified version of the band described in Chaudhuri and Marron (1999).

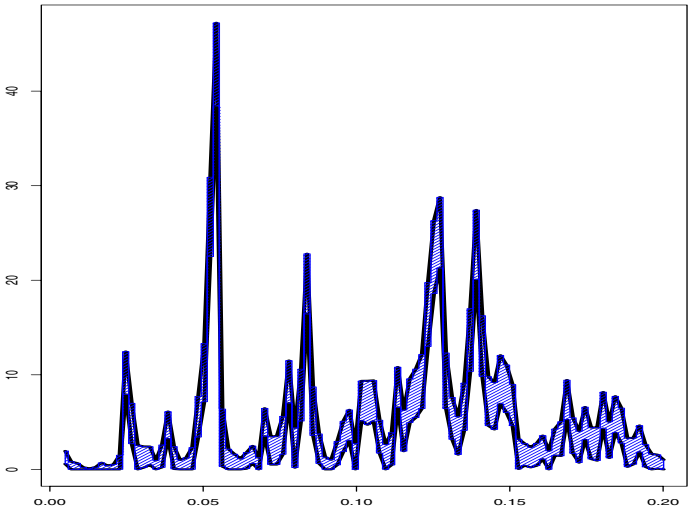



FIGURE 20.7. 95 percent confidence bands for kernel density estimate for the astronomy data.

Suppose now that the data $X_i = (X_{i1}, \ldots, X_{id})$ are $d$-dimensional. The kernel estimator can easily be generalized to $d$ dimensions. Let $h = (h_1, \ldots, h_d)$ be a vector of bandwidths and define

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{20.28}$$

where

$$K_h(x - X_i) = \frac{1}{nh_1 \cdots h_d}\left\{\prod_{j=1}^{d} K\left(\frac{x_i - X_{ij}}{h_j}\right)\right\} \tag{20.29}$$

where $h_1, \ldots, h_d$ are bandwidths. For simplicity, we might take $h_j = s_j h$ where $s_j$ is the standard deviation of the $j^{\text{th}}$ variable. There is now only a single bandwidth $h$ to choose. Using calculations like those in the one-dimensional case, the risk is given by

$$\begin{aligned} R(f, \widehat{f}_n) &\approx \frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx\right] \\ &\quad + \frac{\left(\int K^2(x)dx\right)^d}{nh_1 \cdots h_d} \end{aligned}$$

where $f_{jj}$ is the second partial derivative of $f$. The optimal bandwidth satisfies $h_i \approx c_1 n^{-1/(4+d)}$, leading to a risk of order $n^{-4/(4+d)}$. From this fact, we see

that the risk increases quickly with dimension, a problem usually called the **curse of dimensionality**. To get a sense of how serious this problem is, consider the following table from Silverman (1986) which shows the sample size required to ensure a relative mean squared error less than 0.1 at 0 when the density is multivariate normal and the optimal bandwidth is selected:

| Dimension | Sample Size |
|---|---|
| 1 | 4 |
| 2 | 19 |
| 3 | 67 |
| 4 | 223 |
| 5 | 768 |
| 6 | 2790 |
| 7 | 10,700 |
| 8 | 43,700 |
| 9 | 187,000 |
| 10 | 842,000 |

This is bad news indeed. It says that having 842,000 observations in a ten-dimensional problem is really like having 4 observations in a one-dimensional problem.

## 20.4 Nonparametric Regression

Consider pairs of points $(x_1, Y_1), \ldots, (x_n, Y_n)$ related by

$$Y_i = r(x_i) + \epsilon_i \tag{20.30}$$

where $\mathbb{E}(\epsilon_i) = 0$. We have written the $x_i$'s in lower case since we will treat them as fixed. We can do this since, in regression, it is only the mean of $Y$ conditional on $x$ that we are interested in. We want to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$.

There are many nonparametric regression estimators. Most involve estimating $r(x)$ by taking some sort of weighted average of the $Y_i$'s, giving higher weight to those points near $x$. A popular version is the Nadaraya-Watson kernel estimator.

**20.20 Definition.** *The* **Nadaraya-Watson kernel estimator** *is defined by*

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{20.31}$$

*where $K$ is a kernel and the weights $w_i(x)$ are given by*

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}. \tag{20.32}$$

The form of this estimator comes from first estimating the joint density $f(x,y)$ using kernel density estimation and then inserting the estimate into the formula,

$$r(x) = \mathbb{E}(Y|X=x) = \int y f(y|x)dy = \frac{\int y f(x,y)dy}{\int f(x,y)dy}.$$

**20.21 Theorem.** *Suppose that $\mathbb{V}(\epsilon_i) = \sigma^2$. The risk of the Nadaraya-Watson kernel estimator is*

$$\begin{aligned} R(\widehat{r}_n, r) &\approx \frac{h^4}{4}\left(\int x^2 K^2(x)dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx \\ &\quad + \int \frac{\sigma^2 \int K^2(x)dx}{nhf(x)}dx. \end{aligned} \tag{20.33}$$

*The optimal bandwidth decreases at rate $n^{-1/5}$ and with this choice the risk decreases at rate $n^{-4/5}$.*

In practice, to choose the bandwidth $h$ we minimize the cross validation score

$$\widehat{J}(h) = \sum_{i=1}^n (Y_i - \widehat{r}_{-i}(x_i))^2 \tag{20.34}$$

where $\widehat{r}_{-i}$ is the estimator we get by omitting the $i^{\text{th}}$ variable. Fortunately, there is a shortcut formula for computing $\widehat{J}$.

**20.22 Theorem.** *$\widehat{J}$ can be written as*

$$\widehat{J}(h) = \sum_{i=1}^n (Y_i - \widehat{r}(x_i))^2 \frac{1}{\left(1 - \frac{K(0)}{\sum_{j=1}^n K\left(\frac{x_i - x_j}{h}\right)}\right)^2}. \tag{20.35}$$

**20.23 Example.** Figures 20.8 shows cosmic microwave background (CMB) data from BOOMERaNG (Netterfield et al. (2002)), Maxima (Lee et al. (2001)), and DASI (Halverson et al. (2002))). The data consist of $n$ pairs $(x_1, Y_1)$, ..., $(x_n, Y_n)$ where $x_i$ is called the multipole moment and $Y_i$ is the

estimated power spectrum of the temperature fluctuations. What you are seeing are sound waves in the cosmic microwave background radiation which is the heat, left over from the big bang. If $r(x)$ denotes the true power spectrum, then

$$Y_i = r(x_i) + \epsilon_i$$

where $\epsilon_i$ is a random error with mean 0. The location and size of peaks in $r(x)$ provides valuable clues about the behavior of the early universe. Figure 20.8 shows the fit based on cross-validation as well as an undersmoothed and oversmoothed fit. The cross-validation fit shows the presence of three well-defined peaks, as predicted by the physics of the big bang. ■

The procedure for finding confidence bands is similar to that for density estimation. However, we first need to estimate $\sigma^2$. Suppose that the $x_i$'s are ordered. Assuming $r(x)$ is smooth, we have $r(x_{i+1}) - r(x_i) \approx 0$ and hence

$$Y_{i+1} - Y_i = \Big[r(x_{i+1}) + \epsilon_{i+1}\Big] - \Big[r(x_i) + \epsilon_i\Big] \approx \epsilon_{i+1} - \epsilon_i$$

and hence

$$\mathbb{V}(Y_{i+1} - Y_i) \approx \mathbb{V}(\epsilon_{i+1} - \epsilon_i) = \mathbb{V}(\epsilon_{i+1}) + \mathbb{V}(\epsilon_i) = 2\sigma^2.$$

We can thus use the average of the $n-1$ differences $Y_{i+1} - Y_i$ to estimate $\sigma^2$. Hence, define

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2. \tag{20.36}$$

As with density estimate, the confidence band is for the smoothed version $\overline{r}_n(x) = \mathbb{E}(\widehat{r}_n(x))$ of the true regression function $r$.

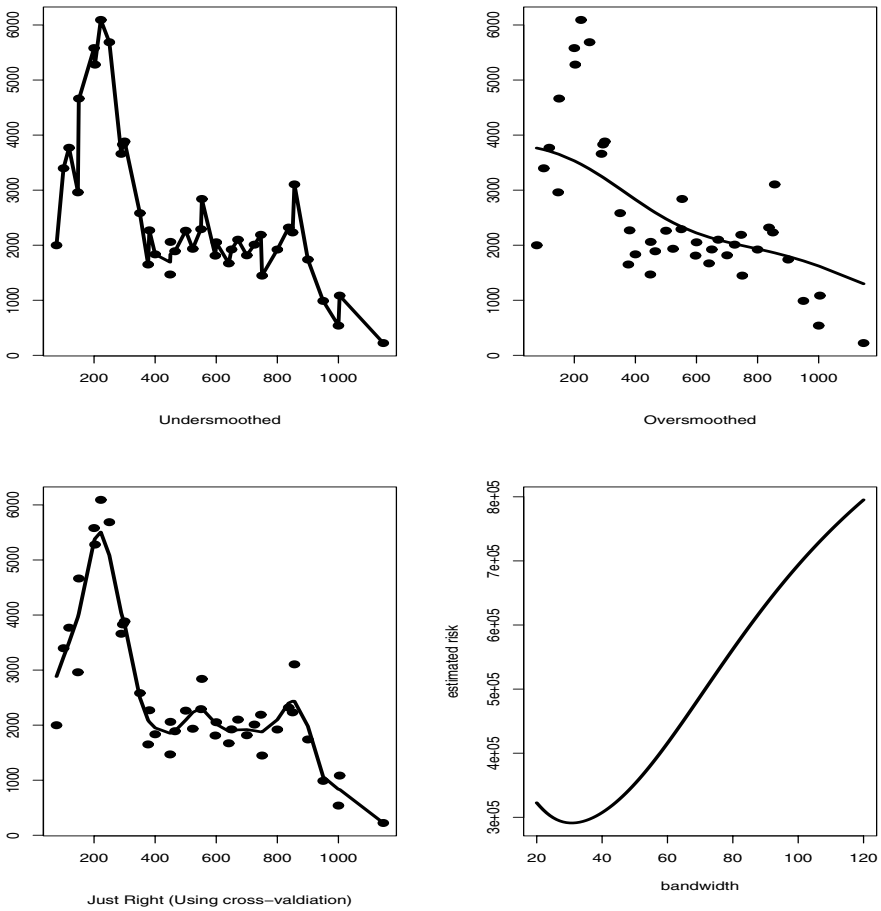



FIGURE 20.8. Regression analysis of the CMB data. The first fit is undersmoothed, the second is oversmoothed, and the third is based on cross-validation. The last panel shows the estimated risk versus the bandwidth of the smoother. The data are from BOOMERaNG, Maxima, and DASI.

**Confidence Bands for Kernel Regression**

An approximate $1-\alpha$ confidence band for $\overline{r}_n(x)$ is

$$\ell_n(x) = \widehat{r}_n(x) - q\ \widehat{\mathsf{se}}(x), \quad u_n(x) = \widehat{r}_n(x) + q\ \widehat{\mathsf{se}}(x) \tag{20.37}$$

where

$$\begin{aligned} \widehat{\mathsf{se}}(x) &= \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w_i^2(x)}, \\ q &= \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right), \\ m &= \frac{b-a}{\omega}, \end{aligned}$$

$\widehat{\sigma}$ is defined in (20.36) and $\omega$ is the width of the kernel. In case the kernel does not have finite width then we take $\omega$ to be the effective width, that is, the range over which the kernel is non-negligible. In particular, we take $\omega = 3h$ for the Normal kernel.

**20.24 Example.** Figure 20.9 shows a 95 percent confidence envelope for the CMB data. We see that we are highly confident of the existence and position of the first peak. We are more uncertain about the second and third peak. At the time of this writing, more accurate data are becoming available that apparently provide sharper estimates of the second and third peak. ■

The extension to multiple regressors $X = (X_1, \ldots, X_p)$ is straightforward. As with kernel density estimation we just replace the kernel with a multivariate kernel. However, the same caveats about the curse of dimensionality apply. In some cases, we might consider putting some restrictions on the regression function which will then reduce the curse of dimensionality. For example, **additive regression** is based on the model

$$Y = \sum_{j=1}^{p} r_j(X_j) + \epsilon. \tag{20.38}$$

Now we only need to fit $p$ one-dimensional functions. The model can be enriched by adding various interactions, for example,

$$Y = \sum_{j=1}^{p} r_j(X_j) + \sum_{j<k} r_{jk}(X_j X_k) + \epsilon. \tag{20.39}$$

Additive models are usually fit by an algorithm called **backfitting**.

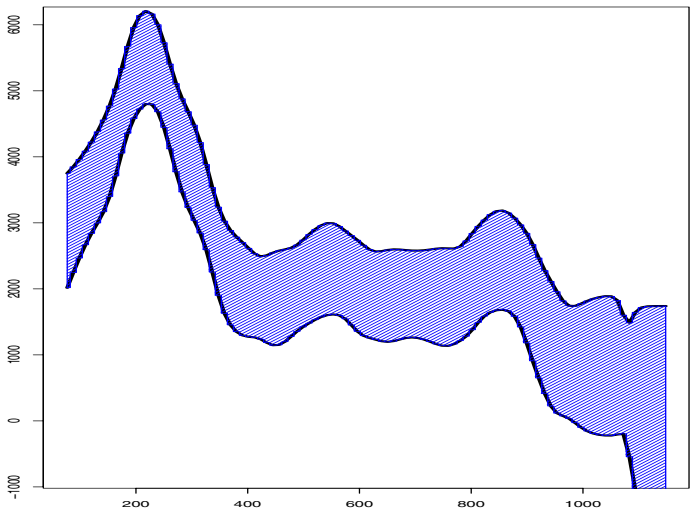



FIGURE 20.9. 95 percent confidence envelope for the CMB data.

**Backfitting**

1. Initialize $r_1(x_1), \ldots, r_p(x_p)$.
2. For $j = 1, \ldots, p$:
   (a) Let $\epsilon_i = Y_i - \sum_{s \neq j} r_s(x_i)$.
   (b) Let $r_j$ be the function estimate obtained by regressing the $\epsilon_i$'s on the $j^{\text{th}}$ covariate.
3. If converged STOP. Else, go back to step 2.

Additive models have the advantage that they avoid the curse of dimensionality and they can be fit quickly, but they have one disadvantage: the model is not fully nonparametric. In other words, the true regression function $r(x)$ may not be of the form (20.38).

## 20.5 Appendix

Confidence Sets and Bias. The confidence bands we computed are not for the density function or regression function but rather for the smoothed

function. For example, the confidence band for a kernel density estimate with bandwidth $h$ is a band for the function one gets by smoothing the true function with a kernel with the same bandwidth. Getting a confidence set for the true function is complicated for reasons we now explain.

Let $\widehat{f}_n(x)$ denote an estimate of the function $f(x)$. Denote the mean and standard deviation of $\widehat{f}_n(x)$ by $\overline{f}_n(x)$ and $s_n(x)$. Then,

$$\frac{\widehat{f}_n(x) - f(x)}{s_n(x)} = \frac{\widehat{f}_n(x) - \overline{f}_n(x)}{s_n(x)} + \frac{\overline{f}_n(x) - f(x)}{s_n(x)}.$$

Typically, the first term converges to a standard Normal from which one derives confidence bands. The second term is the bias divided by the standard deviation. In parametric inference, the bias is usually smaller than the standard deviation of the estimator so this term goes to 0 as the sample size increases. In nonparametric inference, optimal smoothing leads us to balance the bias and the standard deviation. Thus the second term does not vanish even with large sample sizes. This means that the confidence interval will not be centered around the true function $f$.

## 20.6 Bibliographic Remarks

Two very good books on density estimation are Scott (1992) and Silverman (1986). The literature on nonparametric regression is very large. Two good starting points are Hardle (1990) and Loader (1999). The latter emphasizes a class of techniques called local likelihood methods.

## 20.7 Exercises

1. Let $X_1, \ldots, X_n \sim f$ and let $\widehat{f}_n$ be the kernel density estimator using the boxcar kernel:

$$K(x) = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that

$$\mathbb{E}(\widehat{f}(x)) = \frac{1}{h}\int_{x-(h/2)}^{x+(h/2)} f(y)dy$$

   and

$$\mathbb{V}(\widehat{f}(x)) = \frac{1}{nh^2}\left[\int_{x-(h/2)}^{x+(h/2)} f(y)dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y)dy\right)^2\right].$$

(b) Show that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\widehat{f}_n(x) \xrightarrow{\text{P}} f(x)$.

2. Get the data on fragments of glass collected in forensic work from the book website. Estimate the density of the first variable (refractive index) using a histogram and use a kernel density estimator. Use cross-validation to choose the amount of smoothing. Experiment with different binwidths and bandwidths. Comment on the similarities and differences. Construct 95 percent confidence bands for your estimators.

3. Consider the data from question 2. Let $Y$ be refractive index and let $x$ be aluminum content (the fourth variable). Perform a nonparametric regression to fit the model $Y = f(x) + \epsilon$. Use cross-validation to estimate the bandwidth. Construct 95 percent confidence bands for your estimate.

4. Prove Lemma 20.1.

5. Prove Theorem 20.3.

6. Prove Theorem 20.7.

7. Prove Theorem 20.15.

8. Consider regression data $(x_1, Y_1), \ldots, (x_n, Y_n)$. Suppose that $0 \leq x_i \leq 1$ for all $i$. Define bins $B_j$ as in equation (20.7). For $x \in B_j$ define

$$\widehat{r}_n(x) = \overline{Y}_j$$

where $\overline{Y}_j$ is the mean of all the $Y_i$'s corresponding to those $x_i$'s in $B_j$. Find the approximate risk of this estimator. From this expression for the risk, find the optimal bandwidth. At what rate does the risk go to zero?

9. Show that with suitable smoothness assumptions on $r(x)$, $\widehat{\sigma}^2$ in equation (20.36) is a consistent estimator of $\sigma^2$.

10. Prove Theorem 20.22.

# 21
# Smoothing Using Orthogonal Functions

In this chapter we will study an approach to nonparametric curve estimation based on **orthogonal functions**. We begin with a brief introduction to the theory of orthogonal functions, then we turn to density estimation and regression.

## 21.1 Orthogonal Functions and $L_2$ Spaces

Let $v = (v_1, v_2, v_3)$ denote a three-dimensional vector, that is, a list of three real numbers. Let $\mathcal{V}$ denote the set of all such vectors. If $a$ is a scalar (a number) and $v$ is a vector, we define $av = (av_1, av_2, av_3)$. The sum of vectors $v$ and $w$ is defined by $v + w = (v_1 + w_1, v_2 + w_2, v_3 + w_3)$. The **inner product** between two vectors $v$ and $w$ is defined by $\langle v, w \rangle = \sum_{i=1}^{3} v_i w_i$. The **norm (or length)** of a vector $v$ is defined by

$$||v|| = \sqrt{\langle v, v \rangle} = \sqrt{\sum_{i=1}^{3} v_i^2}. \tag{21.1}$$

Two vectors are **orthogonal (or perpendicular)** if $\langle v, w \rangle = 0$. A set of vectors are orthogonal if each pair in the set is orthogonal. A vector is **normal** if $||v|| = 1$.

Let $\phi_1 = (1,0,0)$, $\phi_2 = (0,1,0)$, $\phi_3 = (0,0,1)$. These vectors are said to be an **orthonormal basis** for $\mathcal{V}$ since they have the following properties:
(i) they are orthogonal;
(ii) they are normal;
(iii) they form a basis for $\mathcal{V}$, which means that any $v \in \mathcal{V}$ can be written as a linear combination of $\phi_1$, $\phi_2$, $\phi_3$:

$$v = \sum_{j=1}^{3} \beta_j \phi_j \quad \text{where} \quad \beta_j = \langle \phi_j, v \rangle. \tag{21.2}$$

For example, if $v = (12,3,4)$ then $v = 12\phi_1 + 3\phi_2 + 4\phi_3$. There are other orthonormal bases for $\mathcal{V}$, for example,

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right), \ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

You can check that these three vectors also form an orthonormal basis for $\mathcal{V}$. Again, if $v$ is any vector then we can write

$$v = \sum_{j=1}^{3} \beta_j \psi_j \quad \text{where} \quad \beta_j = \langle \psi_j, v \rangle.$$

For example, if $v = (12,3,4)$ then

$$v = 10.97\psi_1 + 6.36\psi_2 + 2.86\psi_3.$$

Now we make the leap from vectors to functions. Basically, we just replace vectors with functions and sums with integrals. Let $L_2(a,b)$ denote all functions defined on the interval $[a,b]$ such that $\int_a^b f(x)^2 dx < \infty$:

$$L_2(a,b) = \left\{ f : [a,b] \to \mathbb{R}, \quad \int_a^b f(x)^2 dx < \infty \right\}. \tag{21.3}$$

We sometimes write $L_2$ instead of $L_2(a,b)$. The inner product between two functions $f, g \in L_2$ is defined by $\int f(x)g(x)dx$. The norm of $f$ is

$$||f|| = \sqrt{\int f(x)^2 dx}. \tag{21.4}$$

Two functions are orthogonal if $\int f(x)g(x)dx = 0$. A function is normal if $||f|| = 1$.

A sequence of functions $\phi_1, \phi_2, \phi_3, \ldots$ is **orthonormal** if $\int \phi_j^2(x)dx = 1$ for each $j$ and $\int \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$. An orthonormal sequence is **complete** if the only function that is orthogonal to each $\phi_j$ is the zero function.

In this case, the functions $\phi_1, \phi_2, \phi_3, \ldots$ form in basis, meaning that if $f \in L_2$ then $f$ can be written as[1]

$$f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x), \quad \text{where} \quad \beta_j = \int_a^b f(x)\phi_j(x)dx. \tag{21.5}$$

A useful result is **Parseval's relation** which says that

$$||f||^2 \equiv \int f^2(x)\, dx = \sum_{j=1}^{\infty} \beta_j^2 \equiv ||\beta||^2 \tag{21.6}$$

where $\beta = (\beta_1, \beta_2, \ldots)$.

**21.1 Example.** An example of an orthonormal basis for $L_2(0,1)$ is the **cosine basis** defined as follows. Let $\phi_0(x) = 1$ and for $j \geq 1$ define

$$\phi_j(x) = \sqrt{2}\cos(j\pi x). \tag{21.7}$$

The first six functions are plotted in Figure 21.1. ■

**21.2 Example.** Let

$$f(x) = \sqrt{x(1-x)}\, \sin\left(\frac{2.1\pi}{(x+.05)}\right)$$

which is called the "doppler function." Figure 21.2 shows $f$ (top left) and its approximation

$$f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$$

with $J$ equal to 5 (top right), 20 (bottom left), and 200 (bottom right). As $J$ increases we see that $f_J(x)$ gets closer to $f(x)$. The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically. ■

**21.3 Example.** The **Legendre polynomials** on $[-1, 1]$ are defined by

$$P_j(x) = \frac{1}{2^j j!}\frac{d^j}{dx^j}(x^2-1)^j, \quad j = 0, 1, 2, \ldots \tag{21.8}$$

It can be shown that these functions are complete and orthogonal and that

$$\int_{-1}^{1} P_j^2(x)dx = \frac{2}{2j+1}. \tag{21.9}$$

---

[1] The equality in the displayed equation means that $\int (f(x) - f_n(x))^2 dx \to 0$ where $f_n(x) = \sum_{j=1}^{n} \beta_j \phi_j(x)$.

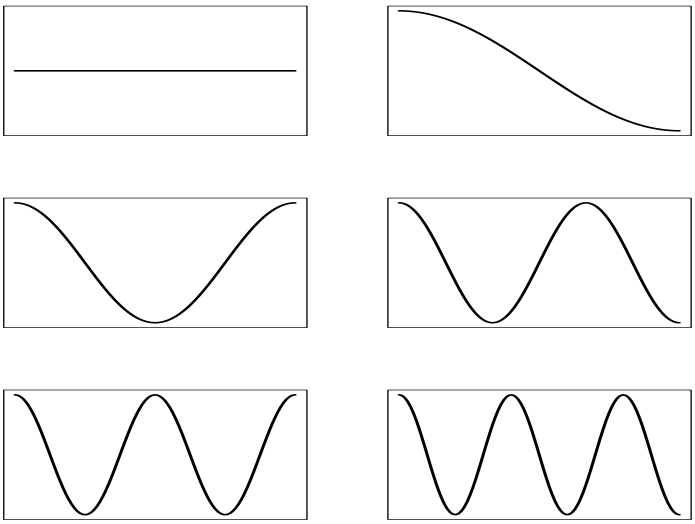


FIGURE 21.1. The first six functions in the cosine basis.

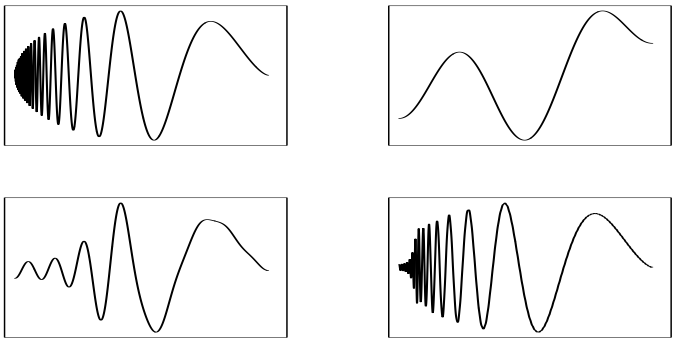


FIGURE 21.2. Approximating the doppler function with its expansion in the cosine basis. The function $f$ (top left) and its approximation $f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ with $J$ equal to 5 (top right), 20 (bottom left), and 200 (bottom right). The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically.

It follows that the functions $\phi_j(x) = \sqrt{(2j+1)/2}P_j(x)$, $j = 0, 1, \ldots$ form an orthonormal basis for $L_2(-1, 1)$. The first few Legendre polynomials are:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{1}{2}\Big(3x^2 - 1\Big), \text{ and} \\ P_3(x) &= \frac{1}{2}\Big(5x^3 - 3x\Big). \end{aligned}$$

These polynomials may be constructed explicitly using the following recursive relation:

$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1}. \quad \blacksquare \tag{21.10}$$

The coefficients $\beta_1, \beta_2, \ldots$ are related to the smoothness of the function $f$. To see why, note that if $f$ is smooth, then its derivatives will be finite. Thus we expect that, for some $k$, $\int_0^1 (f^{(k)}(x))^2 dx < \infty$ where $f^{(k)}$ is the $k^{\text{th}}$ derivative of $f$. Now consider the cosine basis (21.7) and let $f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x)$. Then,

$$\int_0^1 (f^{(k)}(x))^2 dx = 2\sum_{j=1}^{\infty} \beta_j^2 (\pi j)^{2k}.$$

The only way that $\sum_{j=1}^{\infty} \beta_j^2 (\pi j)^{2k}$ can be finite is if the $\beta_j$'s get small when $j$ gets large. To summarize:

> If the function $f$ is smooth, then the coefficients $\beta_j$ will be small when $j$ is large.

For the rest of this chapter, assume we are using the cosine basis unless otherwise specified.

## 21.2 Density Estimation

Let $X_1, \ldots, X_n$ be IID observations from a distribution on $[0, 1]$ with density $f$. Assuming $f \in L_2$ we can write

$$f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x)$$

where $\phi_1, \phi_2, \ldots$ is an orthonormal basis. Define

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i). \tag{21.11}$$

**21.4 Theorem.** *The mean and variance of* $\widehat{\beta}_j$ *are*

$$\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j, \quad \mathbb{V}\left(\widehat{\beta}_j\right) = \frac{\sigma_j^2}{n} \tag{21.12}$$

*where*

$$\sigma_j^2 = \mathbb{V}(\phi_j(X_i)) = \int (\phi_j(x) - \beta_j)^2 f(x) dx. \tag{21.13}$$

PROOF. The mean is

$$\begin{aligned} \mathbb{E}\left(\widehat{\beta}_j\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\phi_j(X_i)\right) \\ &= \mathbb{E}\left(\phi_j(X_1)\right) \\ &= \int \phi_j(x) f(x) dx = \beta_j. \end{aligned}$$

The calculation for the variance is similar. ∎

Hence, $\widehat{\beta}_j$ is an unbiased estimate of $\beta_j$. It is tempting to estimate $f$ by $\sum_{j=1}^{\infty} \widehat{\beta}_j \phi_j(x)$ but this turns out to have a very high variance. Instead, consider the estimator

$$\widehat{f}(x) = \sum_{j=1}^{J} \widehat{\beta}_j \phi_j(x). \tag{21.14}$$

The number of terms $J$ is a smoothing parameter. Increasing $J$ will decrease bias while increasing variance. For technical reasons, we restrict $J$ to lie in the range

$$1 \leq J \leq p$$

where $p = p(n) = \sqrt{n}$. To emphasize the dependence of the risk function on $J$, we write the risk function as $R(J)$.

**21.5 Theorem.** *The risk of* $\widehat{f}$ *is*

$$R(J) = \sum_{j=1}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2. \tag{21.15}$$

An estimate of the risk is

$$\widehat{R}(J) = \sum_{j=1}^{J} \frac{\widehat{\sigma}_j^2}{n} + \sum_{j=J+1}^{p} \left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}_j^2}{n}\right)_+ \tag{21.16}$$

where $a_+ = \max\{a, 0\}$ and

$$\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\phi_j(X_i) - \widehat{\beta}_j\right)^2. \tag{21.17}$$

To motivate this estimator, note that $\widehat{\sigma}_j^2$ is an unbiased estimate of $\sigma_j^2$ and $\widehat{\beta}_j^2 - \widehat{\sigma}_j^2$ is an unbiased estimator of $\beta_j^2$. We take the positive part of the latter term since we know that $\beta_j^2$ cannot be negative. We now choose $1 \leq \widehat{J} \leq p$ to minimize $\widehat{R}(\widehat{f}, f)$. Here is a summary:

**Summary of Orthogonal Function Density Estimation**

1. Let
$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i).$$
2. Choose $\widehat{J}$ to minimize $\widehat{R}(J)$ over $1 \leq J \leq p = \sqrt{n}$ where $\widehat{R}$ is given in equation (21.16).
3. Let
$$\widehat{f}(x) = \sum_{j=1}^{\widehat{J}} \widehat{\beta}_j \phi_j(x).$$

The estimator $\widehat{f}_n$ can be negative. If we are interested in exploring the shape of $f$, this is not a problem. However, if we need our estimate to be a probability density function, we can truncate the estimate and then normalize it. That is, we take $\widehat{f}^* = \max\{\widehat{f}_n(x), 0\}/\int_0^1 \max\{\widehat{f}_n(u), 0\}du$.

Now let us construct a confidence band for $f$. Suppose we estimate $f$ using $J$ orthogonal functions. We are essentially estimating $f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ not the true density $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$. Thus, the confidence band should be regarded as a band for $f_J(x)$.

**21.6 Theorem.** *An approximate $1-\alpha$ confidence band for $f_J$ is $(\ell(x), u(x))$ where*
$$\ell(x) = \widehat{f}_n(x) - c, \quad u(x) = \widehat{f}_n(x) + c \tag{21.18}$$
*where*
$$c = K^2\sqrt{\frac{J\chi^2_{J,\alpha}}{n}} \tag{21.19}$$
*and*
$$K = \max_{1\leq j\leq J} \max_x |\phi_j(x)|.$$
*For the cosine basis, $K = \sqrt{2}$.*

PROOF. Here is an outline of the proof. Let $L = \sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j)^2$. By the central limit theorem, $\widehat{\beta}_j \approx N(\beta_j, \sigma_j^2/n)$. Hence, $\widehat{\beta}_j \approx \beta_j + \sigma_j\epsilon_j/\sqrt{n}$ where

$\epsilon_j \sim N(0,1)$, and therefore

$$L \approx \frac{1}{n}\sum_{j=1}^{J} \sigma_j^2 \epsilon_j^2 \leq \frac{K^2}{n}\sum_{j=1}^{J} \epsilon_j^2 \stackrel{d}{=} \frac{K^2}{n}\chi_J^2. \tag{21.20}$$

Thus we have, approximately, that

$$\mathbb{P}\left(L > \frac{K^2}{n}\chi_{J,\alpha}^2\right) \leq \mathbb{P}\left(\frac{K^2}{n}\chi_J^2 > \frac{K^2}{n}\chi_{J,\alpha}^2\right) = \alpha.$$

Also,

$$\begin{aligned}
\max_x |\widehat{f}_J(x) - f_J(x)| &\leq \max_x \sum_{j=1}^{J} |\phi_j(x)|\ |\widehat{\beta}_j - \beta_j| \\
&\leq K\sum_{j=1}^{J} |\widehat{\beta}_j - \beta_j| \\
&\leq \sqrt{J}\,K\sqrt{\sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j)^2} \\
&= \sqrt{J}\,K\sqrt{L}
\end{aligned}$$

where the third inequality is from the Cauchy-Schwartz inequality (Theorem 4.8). So,

$$\begin{aligned}
\mathbb{P}\left(\max_x |\widehat{f}_J(x) - f_J(x)| > K^2\sqrt{\frac{J\chi_{J,\alpha}^2}{n}}\right) &\leq \mathbb{P}\left(\sqrt{J}\,K\sqrt{L} > K^2\sqrt{\frac{J\chi_{J,\alpha}^2}{n}}\right) \\
&= \mathbb{P}\left(\sqrt{L} > K\sqrt{\frac{\chi_{J,\alpha}^2}{n}}\right) \\
&= \mathbb{P}\left(L > \frac{K^2\chi_{J,\alpha}^2}{n}\right) \\
&\leq \alpha. \quad \blacksquare
\end{aligned}$$

**21.7 Example.** Let

$$f(x) = \frac{5}{6}\phi(x;0,1) + \frac{1}{6}\sum_{j=1}^{5}\phi(x;\mu_j,.1)$$

where $\phi(x;\mu,\sigma)$ denotes a Normal density with mean $\mu$ and standard deviation $\sigma$, and $(\mu_1,\ldots,\mu_5) = (-1,-1/2,0,1/2,1)$. Marron and Wand (1992) call this

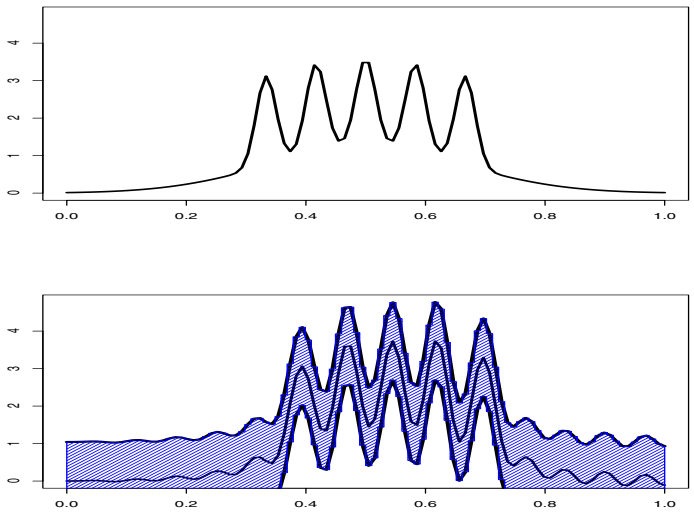



FIGURE 21.3. The top plot is the true density for the Bart Simpson distribution (rescaled to have most of its mass between 0 and 1). The bottom plot is the orthogonal function density estimate and 95 percent confidence band.

"the claw" although the "Bart Simpson" might be more appropriate. Figure 21.3 shows the true density as well as the estimated density based on $n = 5,000$ observations and a 95 percent confidence band. The density has been rescaled to have most of its mass between 0 and 1 using the transformation $y = (x + 3)/6$. ∎

## 21.3 Regression

Consider the regression model

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{21.21}$$

where the $\epsilon_i$ are independent with mean 0 and variance $\sigma^2$. We will initially focus on the special case where $x_i = i/n$. We assume that $r \in L_2(0, 1)$ and hence we can write

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \quad \text{where} \quad \beta_j = \int_0^1 r(x)\phi_j(x)dx \tag{21.22}$$

where $\phi_1, \phi_2, \ldots$ where is an orthonormal basis for $[0, 1]$.

Define

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\,\phi_j(x_i), \quad j = 1, 2, \ldots \tag{21.23}$$

Since $\widehat{\beta}_j$ is an average, the central limit theorem tells us that $\widehat{\beta}_j$ will be approximately Normally distributed.

**21.8 Theorem.**

$$\widehat{\beta}_j \approx N\left(\beta_j, \frac{\sigma^2}{n}\right). \tag{21.24}$$

PROOF. The mean of $\widehat{\beta}_j$ is

$$\begin{aligned}\mathbb{E}(\widehat{\beta}_j) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i)\phi_j(x_i) = \frac{1}{n}\sum_{i=1}^{n} r(x_i)\phi_j(x_i) \\ &\approx \int r(x)\phi_j(x)dx = \beta_j\end{aligned}$$

where the approximate equality follows from the definition of a Riemann integral: $\sum_i \Delta_n h(x_i) \to \int_0^1 h(x)dx$ where $\Delta_n = 1/n$. The variance is

$$\begin{aligned}\mathbb{V}(\widehat{\beta}_j) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(Y_i)\phi_j^2(x_i) \\ &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\phi_j^2(x_i) = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(x_i) \\ &\approx \frac{\sigma^2}{n}\int \phi_j^2(x)dx = \frac{\sigma^2}{n}\end{aligned}$$

since $\int \phi_j^2(x)dx = 1$. ∎

Let

$$\widehat{r}(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x),$$

and let

$$R(J) = \mathbb{E}\int \left(r(x) - \widehat{r}(x)\right)^2 dx$$

be the risk of the estimator.

**21.9 Theorem.** *The risk $R(J)$ of the estimator $\widehat{r}_n(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x)$ is*

$$R(J) = \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty}\beta_j^2. \tag{21.25}$$

To estimate for $\sigma^2 = \mathbb{V}(\epsilon_i)$ we use

$$\widehat{\sigma}^2 = \frac{n}{k} \sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \tag{21.26}$$

where $k = n/4$. To motivate this estimator, recall that if $f$ is smooth, then $\beta_j \approx 0$ for large $j$. So, for $j \geq k$, $\widehat{\beta}_j \approx N(0, \sigma^2/n)$ and thus, $\widehat{\beta}_j \approx \sigma Z_j/\sqrt{n}$ for for $j \geq k$, where $Z_j \sim N(0,1)$. Therefore,

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{n}{k} \sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \approx \frac{n}{k} \sum_{i=n-k+1}^{n} \left( \frac{\sigma}{\sqrt{n}} \widehat{\beta}_j \right)^2 \\ &= \frac{\sigma^2}{k} \sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 = \frac{\sigma^2}{k} \chi_k^2 \end{aligned}$$

since a sum of $k$ Normals has a $\chi_k^2$ distribution. Now $\mathbb{E}(\chi_k^2) = k$ and hence $\mathbb{E}(\widehat{\sigma}^2) \approx \sigma^2$. Also, $\mathbb{V}(\chi_k^2) = 2k$ and hence $\mathbb{V}(\widehat{\sigma}^2) \approx (\sigma^4/k^2)(2k) = (2\sigma^4/k) \to 0$ as $n \to \infty$. Thus we expect $\widehat{\sigma}^2$ to be a consistent estimator of $\sigma^2$. There is nothing special about the choice $k = n/4$. Any $k$ that increases with $n$ at an appropriate rate will suffice.

We estimate the risk with

$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n} \left( \widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n} \right)_+ . \tag{21.27}$$

**21.10 Example.** Figure 21.4 shows the doppler function $f$ and $n = 2,048$ observations generated from the model

$$Y_i = r(x_i) + \epsilon_i$$

where $x_i = i/n$, $\epsilon_i \sim N(0, (.1)^2)$. The figure shows the data and the estimated function. The estimate was based on $\widehat{J} = 234$ terms. ■

We are now ready to give a complete description of the method.

Orthogonal Series Regression Estimator

1. Let
$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(x_i), \quad j = 1, \ldots, n.$$

2. Let
$$\widehat{\sigma}^2 = \frac{n}{k} \sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \tag{21.28}$$

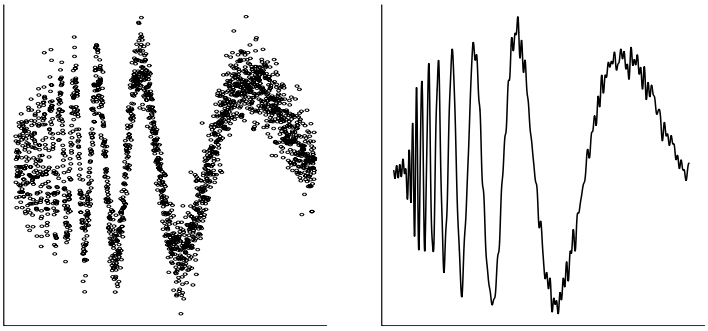


FIGURE 21.4. Data from the doppler test function and the estimated function. See Example 21.10.

where $k \approx n/4$.

3. For $1 \leq J \leq n$, compute the risk estimate

$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n}\left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n}\right)_+.$$

4. Choose $\widehat{J} \in \{1, \ldots n\}$ to minimize $\widehat{R}(J)$.

5. Let

$$\widehat{r}(x) = \sum_{j=1}^{\widehat{J}} \widehat{\beta}_j \phi_j(x).$$

Finally, we turn to confidence bands. As before, these bands are not really for the true function $r(x)$ but rather for the smoothed version of the function $r_J(x) = \sum_{j=1}^{\widehat{J}} \beta_j \phi_j(x)$.

**21.11 Theorem.** *Suppose the estimate $\widehat{r}$ is based on $J$ terms and $\widehat{\sigma}$ is defined as in equation (21.28). Assume that $J < n - k + 1$. An approximate $1 - \alpha$ confidence band for $r_J$ is $(\ell, u)$ where*

$$\ell(x) = \widehat{r}_n(x) - c, \quad u(x) = \widehat{r}_n(x) + c, \tag{21.29}$$

*where*

$$c = \frac{a(x)\ \widehat{\sigma}\ \chi_{J,\alpha}}{\sqrt{n}}, \quad a(x) = \sqrt{\sum_{j=1}^{J}\phi_j^2(x)},$$

*and $\widehat{\sigma}$ is given in equation (21.28).*

PROOF. Let $L = \sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j)^2$. By the central limit theorem, $\widehat{\beta}_j \approx N(\beta_j, \sigma^2/n)$. Hence, $\widehat{\beta}_j \approx \beta_j + \sigma\epsilon_j/\sqrt{n}$ where $\epsilon_j \sim N(0,1)$ and therefore

$$L \approx \frac{\sigma^2}{n}\sum_{j=1}^{J}\epsilon_j^2 \stackrel{d}{=} \frac{\sigma^2}{n}\chi_J^2.$$

Thus,

$$\mathbb{P}\left(L > \frac{\sigma^2}{n}\chi^2_{J,\alpha}\right) = \mathbb{P}\left(\frac{\sigma^2}{n}\chi_J^2 > \frac{\sigma^2}{n}\chi^2_{J,\alpha}\right) = \alpha.$$

Also,

$$\begin{aligned}
|\widehat{r}(x) - r_J(x)| &\leq \sum_{j=1}^{J}|\phi_j(x)|\,|\widehat{\beta}_j - \beta_j| \\
&\leq \sqrt{\sum_{j=1}^{J}\phi_j^2(x)}\sqrt{\sum_{j=1}^{J}(\widehat{\beta}_j - \beta_j^2)} \\
&\leq a(x)\,\sqrt{L}
\end{aligned}$$

by the Cauchy-Schwartz inequality (Theorem 4.8). So,

$$\begin{aligned}
\mathbb{P}\left(\max_x \frac{|\widehat{f}_J(x) - \overline{f}(x)|}{a(x)} > \frac{\widehat{\sigma}\chi_{J,\alpha}}{\sqrt{n}}\right) &\leq \mathbb{P}\left(\sqrt{L} > \frac{\widehat{\sigma}\chi_{J,\alpha}}{\sqrt{n}}\right) \\
&= \alpha
\end{aligned}$$

and the result follows. ∎

**21.12 Example.** Figure 21.5 shows the confidence envelope for the doppler signal. The first plot is based on $J = 234$ (the value of $J$ that minimizes the estimated risk). The second is based on $J = 45 \approx \sqrt{n}$. Larger $J$ yields a higher resolution estimator at the cost of large confidence bands. Smaller $J$ yields a lower resolution estimator but has tighter confidence bands. ∎

So far, we have assumed that the $x_i$'s are of the form $\{1/n, 2/n, \ldots, 1\}$. If the $x_i$'s are on interval $[a, b]$, then we can rescale them so that are in the interval $[0, 1]$. If the $x_i$'s are not equally spaced, the methods we have discussed still apply so long as the $x_i$'s "fill out" the interval [0,1] in such a way so as to not be too clumped together. If we want to treat the $x_i$'s as random instead of fixed, then the method needs significant modifications which we shall not deal with here.

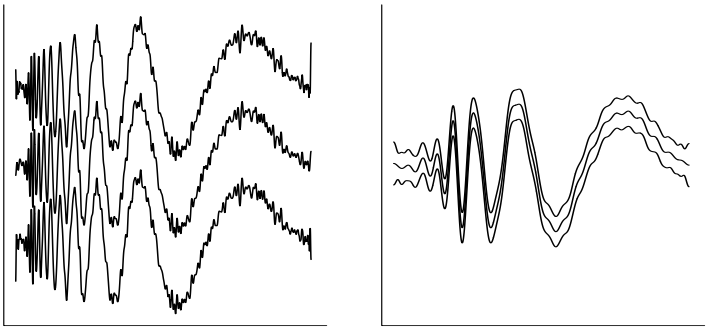


FIGURE 21.5. Estimates and confidence bands for the doppler test function using $n = 2,048$ observations. First plot: $J = 234$ terms. Second plot: $J = 45$ terms.

## 21.4 Wavelets

Suppose there is a sharp jump in a regression function $f$ at some point $x$ but that $f$ is otherwise very smooth. Such a function $f$ is said to be **spatially inhomogeneous**. The doppler function is an example of a spatially inhomogeneous function; it is smooth for large $x$ and unsmooth for small $x$.

It is hard to estimate $f$ using the methods we have discussed so far. If we use a cosine basis and only keep low order terms, we will miss the peak; if we allow higher order terms we will find the peak but we will make the rest of the curve very wiggly. Similar comments apply to kernel regression. If we use a large bandwidth, then we will smooth out the peak; if we use a small bandwidth, then we will find the peak but we will make the rest of the curve very wiggly.

One way to estimate inhomogeneous functions is to use a more carefully chosen basis that allows us to place a "blip" in some small region without adding wiggles elsewhere. In this section, we describe a special class of bases called **wavelets**, that are aimed at fixing this problem. Statistical inference using wavelets is a large and active area. We will just discuss a few of the main ideas to get a flavor of this approach.

We start with a particular wavelet called the **Haar wavelet.** The **Haar father wavelet** or **Haar scaling function** is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{21.30}$$

The **mother Haar wavelet** is defined by

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \le x \le \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < x \le 1. \end{cases} \tag{21.31}$$

For any integers $j$ and $k$ define

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{21.32}$$

The function $\psi_{j,k}$ has the same shape as $\psi$ but it has been rescaled by a factor of $2^{j/2}$ and shifted by a factor of $k$.

See Figure 21.6 for some examples of Haar wavelets. Notice that for large $j$, $\psi_{j,k}$ is a very localized function. This makes it possible to add a blip to a function in one place without adding wiggles elsewhere. Increasing $j$ is like looking in a microscope at increasing degrees of resolution. In technical terms, we say that wavelets provide a **multiresolution analysis** of $L_2(0,1)$.

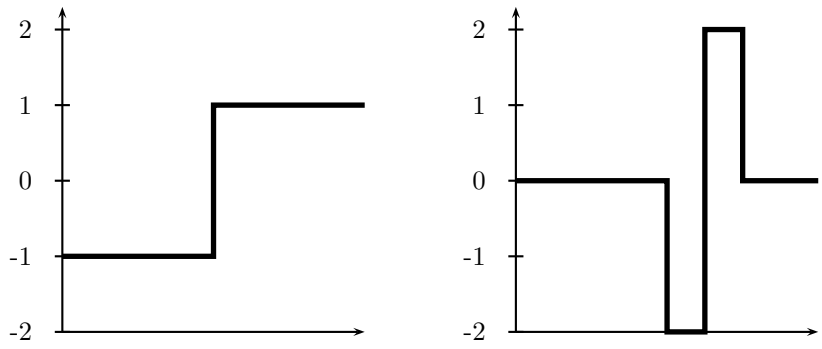



FIGURE 21.6. Some Haar wavelets. Left: the mother wavelet $\psi(x)$; Right: $\psi_{2,2}(x)$.

Let

$$W_j = \{\psi_{jk},\ k = 0, 1, \ldots, 2^j - 1\}$$

be the set of rescaled and shifted mother wavelets at resolution $j$.

**21.13 Theorem.** *The set of functions*

$$\left\{\phi, W_0, W_1, W_2, \ldots, \right\}$$

*is an orthonormal basis for* $L_2(0,1)$.

It follows from this theorem that we can expand any function $f \in L_2(0,1)$ in this basis. Because each $W_j$ is itself a set of functions, we write the expansion as a double sum:

$$f(x) = \alpha\,\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x) \tag{21.33}$$

where

$$\alpha = \int_0^1 f(x)\phi(x)\,dx, \quad \beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)\,dx.$$

We call $\alpha$ the **scaling coefficient** and the $\beta_{j,k}$'s are called the **detail coefficients**. We call the finite sum

$$f_J(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x) \tag{21.34}$$

the **resolution** $J$ approximation to $f$. The total number of terms in this sum is

$$1 + \sum_{j=0}^{J-1} 2^j = 1 + 2^J - 1 = 2^J.$$

**21.14 Example.** Figure 21.7 shows the doppler signal, and its reconstruction using $J = 3, 5$ and $J = 8$. ■

Haar wavelets are localized, meaning that they are zero outside an interval. But they are not smooth. This raises the question of whether there exist smooth, localized wavelets that from an orthonormal basis. In 1988, Ingrid Daubechie showed that such wavelets do exist. These smooth wavelets are difficult to describe. They can be constructed numerically but there is no closed form formula for the smoother wavelets. To keep things simple, we will continue to use Haar wavelets.

Consider the regression model $Y_i = r(x_i) + \sigma\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $x_i = i/n$. To simplify the discussion we assume that $n = 2^J$ for some $J$.

There is one major difference between estimation using wavelets instead of a cosine (or polynomial) basis. With the cosine basis, we used all the terms $1 \le j \le J$ for some $J$. The number of terms $J$ acted as a smoothing parameter. With wavelets, we control smoothing using a method called **thresholding** where we keep a term in the function approximation if its coefficient is large,

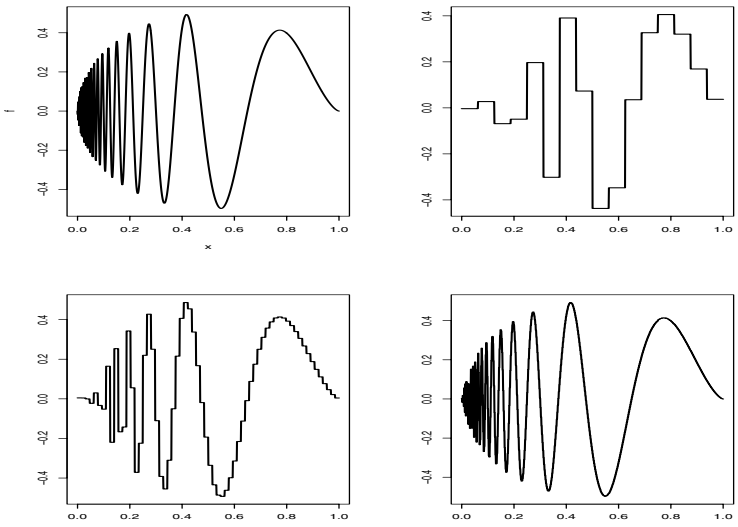



FIGURE 21.7. The doppler signal and its reconstruction $f_J(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_k \beta_{j,k}\psi_{j,k}(x)$ based on $J = 3$, $J = 5$, and $J = 8$.

otherwise, we throw out that term. There are many versions of thresholding. The simplest is called hard, universal thresholding. Let $J = \log_2(n)$ and define

$$\widehat{\alpha} = \frac{1}{n}\sum_i \phi_k(x_i)Y_i \quad \text{and} \quad D_{j,k} = \frac{1}{n}\sum_i \psi_{j,k}(x_i)Y_i \tag{21.35}$$

for $0 \le j \le J-1$.

**Haar Wavelet Regression**

1. Compute $\widehat{\alpha}$ and $D_{j,k}$ as in (21.35), for $0 \le j \le J-1$.
2. Estimate $\sigma$; see (21.37).
3. Apply universal thresholding:

$$\widehat{\beta}_{j,k} = \left\{ \begin{array}{ll} D_{j,k} & \text{if } |D_{j,k}| > \widehat{\sigma}\sqrt{\frac{2\log n}{n}} \\ 0 & \text{otherwise.} \end{array} \right\} \tag{21.36}$$

4. Set $\widehat{f}(x) = \widehat{\alpha}\phi(x) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\widehat{\beta}_{j,k}\psi_{j,k}(x)$.

In practice, we do not compute $S_k$ and $D_{j,k}$ using (21.35). Instead, we use the **discrete wavelet transform (DWT)** which is very fast. The DWT for Haar wavelets is described in the appendix. The estimate of $\sigma$ is

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\left(|D_{J-1,k}| : \ k = 0, \ldots, 2^{J-1} - 1\right)}{0.6745}. \tag{21.37}$$

The estimate for $\sigma$ may look strange. It is similar to the estimate we used for the cosine basis but it is designed to be insensitive to sharp peaks in the function.

To understand the intuition behind universal thresholding, consider what happens when there is no signal, that is, when $\beta_{j,k} = 0$ for all $j$ and $k$.

**21.15 Theorem.** *Suppose that $\beta_{j,k} = 0$ for all $j$ and $k$ and let $\widehat{\beta}_{j,k}$ be the universal threshold estimator. Then*

$$\mathbb{P}(\widehat{\beta}_{j,k} = 0 \text{ for all } j, k) \to 1$$

*as $n \to \infty$.*

PROOF. To simplify the proof, assume that $\sigma$ is known. Now $D_{j,k} \approx N(0, \sigma^2/n)$. We will need Mill's inequality (Theorem 4.7): if $Z \sim N(0,1)$ then $\mathbb{P}(|Z| > t) \leq (c/t)e^{-t^2/2}$ where $c = \sqrt{2/\pi}$ is a constant. Thus,

$$\begin{aligned}
\mathbb{P}(\max |D_{j,k}| > \lambda) &\leq \sum_{j,k} \mathbb{P}(|D_{j,k}| > \lambda) = \sum_{j,k} \mathbb{P}\left(\frac{\sqrt{n}|D_{j,k}|}{\sigma} > \frac{\sqrt{n}\lambda}{\sigma}\right) \\
&\leq \sum_{j,k} \frac{c\sigma}{\lambda\sqrt{n}} \exp\left\{-\frac{1}{2}\frac{n\lambda^2}{\sigma^2}\right\} \\
&= \frac{c}{\sqrt{2\log n}} \to 0. \quad \blacksquare
\end{aligned}$$

**21.16 Example.** Consider $Y_i = r(x_i) + \sigma\epsilon_i$ where $f$ is the doppler signal, $\sigma = .1$ and $n = 2,048$. Figure 21.8 shows the data and the estimated function using universal thresholding. Of course, the estimate is not smooth since Haar wavelets are not smooth. Nonetheless, the estimate is quite accurate. ■

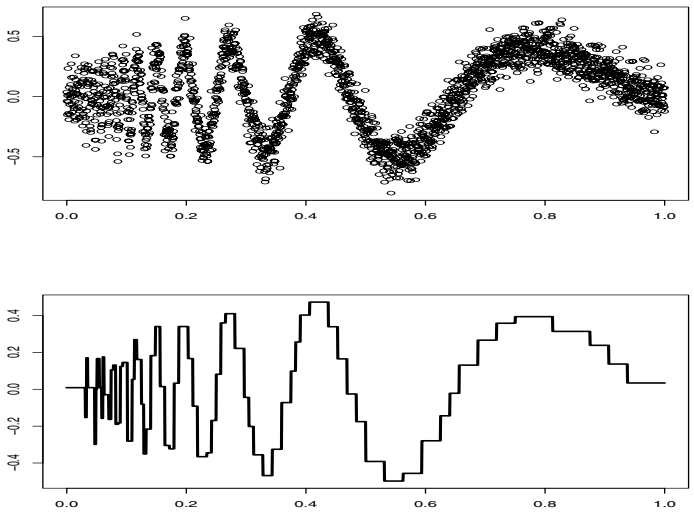



FIGURE 21.8. Estimate of the Doppler function using Haar wavelets and universal thresholding.

## 21.5 Appendix

THE DWT FOR HAAR WAVELETS. Let $y$ be the vector of $Y_i$'s (length $n$) and let $J = \log_2(n)$. Create a list $D$ with elements

$$D[[0]],\ \ldots,\ D[[J-1]].$$

Set:

$$temp \leftarrow y/\sqrt{n}.$$

Then do:

$$\begin{array}{rcl} for(j & in & (J-1):0)\{ \\ m & \leftarrow & 2^j \\ I & \leftarrow & (1:m) \\ D[[j]] & \leftarrow & \Big(temp[2*I] - temp[(2*I)-1]\Big)/\sqrt{2} \\ temp & \leftarrow & \Big(temp[2*I] + temp[(2*I)-1]\Big)/\sqrt{2} \\ & & \} \end{array}$$

## 21.6 Bibliographic Remarks

Efromovich (1999) is a reference for orthogonal function methods. See also Beran (2000) and Beran and Dümbgen (1998). An introduction to wavelets is given in Ogden (1997). A more advanced treatment can be found in Härdle et al. (1998). The theory of statistical estimation using wavelets has been developed by many authors, especially David Donoho and Ian Johnstone. See Donoho and Johnstone (1994), Donoho and Johnstone (1995), Donoho et al. (1995), and Donoho and Johnstone (1998).

## 21.7 Exercises

1. Prove Theorem 21.5.

2. Prove Theorem 21.9.

3. Let

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right),\ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right),\ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

   Show that these vectors have norm 1 and are orthogonal.

4. Prove Parseval's relation equation (21.6).

5. Plot the first five Legendre polynomials. Verify, numerically, that they are orthonormal.

6. Expand the following functions in the cosine basis on $[0, 1]$. For (a) and (b), find the coefficients $\beta_j$ analytically. For (c) and (d), find the coefficients $\beta_j$ numerically, i.e.

$$\beta_j = \int_0^1 f(x)\phi_j(x) \approx \frac{1}{N}\sum_{r=1}^{N} f\left(\frac{r}{N}\right)\phi_j\left(\frac{r}{N}\right)$$

   for some large integer $N$. Then plot the partial sum $\sum_{j=1}^{n}\beta_j\phi_j(x)$ for increasing values of $n$.

   (a) $f(x) = \sqrt{2}\cos(3\pi x)$.

   (b) $f(x) = \sin(\pi x)$.

   (c) $f(x) = \sum_{j=1}^{11} h_j K(x - t_j)$ where $K(t) = (1+\text{sign}(t))/2$, $\text{sign}(x) = -1$ if $x < 0$, $\text{sign}(x) = 0$ if $x = 0$, $\text{sign}(x) = 1$ if $x > 0$,

$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81)$,

$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$.

(d) $f = \sqrt{x(1-x)}\,\sin\left(\frac{2.1\pi}{(x+.05)}\right)$.

7. Consider the glass fragments data from the book's website. Let $Y$ be refractive index and let $X$ be aluminum content (the fourth variable).

   (a) Do a nonparametric regression to fit the model $Y = f(x) + \epsilon$ using the cosine basis method. The data are not on a regular grid. Ignore this when estimating the function. (But do sort the data first according to $x$.) Provide a function estimate, an estimate of the risk, and a confidence band.

   (b) Use the wavelet method to estimate $f$.

8. Show that the Haar wavelets are orthonormal.

9. Consider again the doppler signal:

$$f(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right).$$

   Let $n = 1,024$, $\sigma = 0.1$, and let $(x_1, \ldots, x_n) = (1/n, \ldots, 1)$. Generate data

$$Y_i = f(x_i) + \sigma\epsilon_i$$

   where $\epsilon_i \sim N(0,1)$.

   (a) Fit the curve using the cosine basis method. Plot the function estimate and confidence band for $J = 10, 20, \ldots, 100$.

   (b) Use Haar wavelets to fit the curve.

10. (Haar density Estimation.) Let $X_1, \ldots, X_n \sim f$ for some density $f$ on $[0,1]$. Let's consider constructing a wavelet histogram. Let $\phi$ and $\psi$ be the Haar father and mother wavelet. Write

$$f(x) \approx \phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x)$$

    where $J \approx \log_2(n)$. Let

$$\widehat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(X_i).$$

(a) Show that $\widehat{\beta}_{j,k}$ is an unbiased estimate of $\beta_{j,k}$.

(b) Define the Haar histogram

$$\widehat{f}(x) = \phi(x) + \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \widehat{\beta}_{j,k} \psi_{j,k}(x)$$

for $0 \leq B \leq J-1$.

(c) Find an approximate expression for the MSE as a function of $B$.

(d) Generate $n = 1,000$ observations from a Beta (15,4) density. Estimate the density using the Haar histogram. Use leave-one-out cross validation to choose $B$.

11. In this question, we will explore the motivation for equation (21.37). Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. Let

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\left(|X_1|, \ldots, |X_n|\right)}{0.6745}.$$

(a) Show that $\mathbb{E}(\widehat{\sigma}) = \sigma$.

(b) Simulate $n = 100$ observations from a N(0,1) distribution. Compute $\widehat{\sigma}$ as well as the usual estimate of $\sigma$. Repeat 1,000 times and compare the MSE.

(c) Repeat (b) but add some outliers to the data. To do this, simulate each observation from a N(0,1) with probability .95 and simulate each observation from a N(0,10) with probability .95.

12. Repeat question 6 using the Haar basis.

# 22
# Classification

## 22.1 Introduction

The problem of predicting a discrete random variable $Y$ from another random variable $X$ is called **classification**, **supervised learning**, **discrimination**, or **pattern recognition.**

Consider IID data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$X_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X} \subset \mathbb{R}^d$$

is a $d$-dimensional vector and $Y_i$ takes values in some finite set $\mathcal{Y}$. A **classification rule** is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new $X$, we predict $Y$ to be $h(X)$.

**22.1 Example.** Here is a an example with fake data. Figure 22.1 shows 100 data points. The covariate $X = (X_1, X_2)$ is 2-dimensional and the outcome $Y \in \mathcal{Y} = \{0, 1\}$. The $Y$ values are indicated on the plot with the triangles representing $Y = 1$ and the squares representing $Y = 0$. Also shown is a linear classification rule represented by the solid line. This is a rule of the form

$$h(x) = \begin{cases} 1 & \text{if } a + b_1 x_1 + b_2 x_2 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Everything above the line is classified as a 0 and everything below the line is classified as a 1. ■

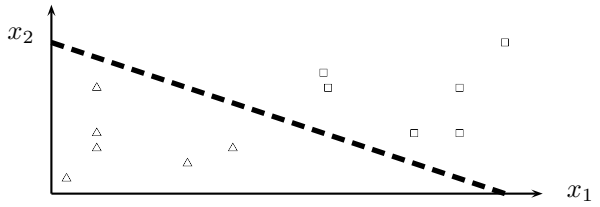



FIGURE 22.1. Two covariates and a linear decision boundary. $\triangle$ means $Y = 1$. $\square$ means $Y = 0$. These two groups are perfectly separated by the linear decision boundary; you probably won't see real data like this.

**22.2 Example.** Recall the the Coronary Risk-Factor Study (CORIS) data from Example 13.17. There are 462 males between the ages of 15 and 64 from three rural areas in South Africa. The outcome $Y$ is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease and there are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. I computed a linear decision boundary using the LDA method based on two of the covariates, systolic blood pressure and tobacco consumption. The LDA method will be explained shortly. In this example, the groups are hard to tell apart. In fact, 141 of the 462 subjects are misclassified using this classification rule. ■

At this point, it is worth revisiting the Statistics/Data Mining dictionary:

| Statistics | Computer Science | Meaning |
|---|---|---|
| classification | supervised learning | predicting a discrete $Y$ from $X$ |
| data | training sample | $(X_1, Y_1), \ldots, (X_n, Y_n)$ |
| covariates | features | the $X_i$'s |
| classifier | hypothesis | map $h : \mathcal{X} \to \mathcal{Y}$ |
| estimation | learning | finding a good classifier |

## 22.2 Error Rates and the Bayes Classifier

Our goal is to find a classification rule $h$ that makes accurate predictions. We start with the following definitions:

**22.3 Definition.** *The* **true error rate**[1] *of a classifier* $h$ *is*

$$L(h) = \mathbb{P}(\{h(X) \neq Y\}) \tag{22.1}$$

*and the* **empirical error rate** *or* **training error rate** *is*

$$\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(X_i) \neq Y_i). \tag{22.2}$$

First we consider the special case where $\mathcal{Y} = \{0, 1\}$. Let

$$r(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$$

denote the **regression function**. From Bayes' theorem we have that

$$\begin{aligned} r(x) &= \mathbb{P}(Y = 1|X = x) \\ &= \frac{f(x|Y = 1)\mathbb{P}(Y = 1)}{f(x|Y = 1)\mathbb{P}(Y = 1) + f(x|Y = 0)\mathbb{P}(Y = 0)} \\ &= \frac{\pi f_1(x)}{\pi f_1(x) + (1 - \pi)f_0(x)} \end{aligned} \tag{22.3}$$

where

$$\begin{aligned} f_0(x) &= f(x|Y = 0) \\ f_1(x) &= f(x|Y = 1) \\ \pi &= \mathbb{P}(Y = 1). \end{aligned}$$

**22.4 Definition.** *The* **Bayes classification rule** $h^*$ *is*

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{22.4}$$

*The set* $\mathcal{D}(h) = \{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 0|X = x)\}$ *is called the* **decision boundary.**

**Warning!** The Bayes rule has nothing to do with Bayesian inference. We could estimate the Bayes rule using either frequentist or Bayesian methods.

The Bayes rule may be written in several equivalent forms:

[1] One can use other loss functions. For simplicity we will use the error rate as our loss function.

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 & \text{otherwise} \end{cases} \tag{22.5}$$

and

$$h^*(x) = \begin{cases} 1 & \text{if } \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{otherwise.} \end{cases} \tag{22.6}$$

**22.5 Theorem.** *The Bayes rule is optimal, that is, if $h$ is any other classification rule then $L(h^*) \leq L(h)$.*

The Bayes rule depends on unknown quantities so we need to use the data to find some approximation to the Bayes rule. At the risk of oversimplifying, there are three main approaches:

1. Empirical Risk Minimization. Choose a set of classifiers $\mathcal{H}$ and find $\widehat{h} \in \mathcal{H}$ that minimizes some estimate of $L(h)$.

2. Regression. Find an estimate $\widehat{r}$ of the regression function $r$ and define

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

3. Density Estimation. Estimate $f_0$ from the $X_i$'s for which $Y_i = 0$, estimate $f_1$ from the $X_i$'s for which $Y_i = 1$ and let $\widehat{\pi} = n^{-1} \sum_{i=1}^n Y_i$. Define

$$\widehat{r}(x) = \widehat{\mathbb{P}}(Y=1|X=x) = \frac{\widehat{\pi}\widehat{f}_1(x)}{\widehat{\pi}\widehat{f}_1(x) + (1-\widehat{\pi})\widehat{f}_0(x)}$$

and

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Now let us generalize to the case where $Y$ takes on more than two values as follows.

**22.6 Theorem.** *Suppose that $Y \in \mathcal{Y} = \{1, \ldots, K\}$. The optimal rule is*

$$h(x) = \text{argmax}_k \, \mathbb{P}(Y=k|X=x) \tag{22.7}$$

$$= \text{argmax}_k \pi_k \, f_k(x) \tag{22.8}$$

*where*

$$\mathbb{P}(Y=k|X=x) = \frac{f_k(x)\pi_k}{\sum_r f_r(x)\pi_r}, \tag{22.9}$$

*$\pi_r = P(Y=r)$, $f_r(x) = f(x|Y=r)$ and* $\text{argmax}_k$ *means "the value of $k$ that maximizes that expression."*

## 22.3 Gaussian and Linear Classifiers

Perhaps the simplest approach to classification is to use the density estimation strategy and assume a parametric model for the densities. Suppose that $\mathcal{Y} = \{0, 1\}$ and that $f_0(x) = f(x|Y = 0)$ and $f_1(x) = f(x|Y = 1)$ are both multivariate Gaussians:

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\}, \quad k = 0, 1.$$

Thus, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$.

**22.7 Theorem.** *If $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$, then the Bayes rule is*

$$h^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 + 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & \text{otherwise} \end{cases} \tag{22.10}$$

*where*

$$r_i^2 = (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i), \quad i = 1, 2 \tag{22.11}$$

*is the* **Manalahobis distance.** *An equivalent way of expressing the Bayes' rule is*

$$h^*(x) = \text{argmax}_k \delta_k(x)$$

*where*

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k \tag{22.12}$$

*and $|A|$ denotes the determinant of a matrix $A$.*

The decision boundary of the above classifier is quadratic so this procedure is called **quadratic discriminant analysis (QDA)**. In practice, we use sample estimates of $\pi, \mu_1, \mu_2, \Sigma_0, \Sigma_1$ in place of the true value, namely:

$$\begin{aligned}
\widehat{\pi}_0 &= \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i), \quad \widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i \\
\widehat{\mu}_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0} X_i, \quad \widehat{\mu}_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1} X_i \\
S_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0} (X_i - \widehat{\mu}_0)(X_i - \widehat{\mu}_0)^T, \quad S_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1} (X_i - \widehat{\mu}_1)(X_i - \widehat{\mu}_1)^T
\end{aligned}$$

where $n_0 = \sum_i (1 - Y_i)$ and $n_1 = \sum_i Y_i$.

A simplification occurs if we assume that $\Sigma_0 = \Sigma_0 = \Sigma$. In that case, the Bayes rule is

$$h^*(x) = \text{argmax}_k \delta_k(x) \tag{22.13}$$

where now

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} + \log \pi_k. \tag{22.14}$$

The parameters are estimated as before, except that the MLE of $\Sigma$ is

$$S = \frac{n_0 S_0 + n_1 S_1}{n_0 + n_1}.$$

The classification rule is

$$h^*(x) = \begin{cases} 1 & \text{if } \delta_1(x) > \delta_0(x) \\ 0 & \text{otherwise} \end{cases} \tag{22.15}$$

where

$$\delta_j(x) = x^T S^{-1} \widehat{\mu}_j - \frac{1}{2} \widehat{\mu}_j^T S^{-1} \widehat{\mu}_j + \log \widehat{\pi}_j$$

is called the **discriminant function**. The decision boundary $\{x : \ \delta_0(x) = \delta_1(x)\}$ is linear so this method is called **linear discrimination analysis (LDA).**

**22.8 Example.** Let us return to the South African heart disease data. The decision rule in in Example 22.2 was obtained by linear discrimination. The outcome was

| | classified as 0 | classified as 1 |
|---|---|---|
| $y = 0$ | 277 | 25 |
| $y = 1$ | 116 | 44 |

The observed misclassification rate is $141/462 = .31$. Including all the covariates reduces the error rate to .27. The results from quadratic discrimination are

| | classified as 0 | classified as 1 |
|---|---|---|
| $y = 0$ | 272 | 30 |
| $y = 1$ | 113 | 47 |

which has about the same error rate $143/462 = .31$. Including all the covariates reduces the error rate to .26. In this example, there is little advantage to QDA over LDA. ■

Now we generalize to the case where $Y$ takes on more than two values.

**22.9 Theorem.** *Suppose that $Y \in \{1, \ldots, K\}$. If $f_k(x) = f(x|Y = k)$ is Gaussian, the Bayes rule is*

$$h(x) = \text{argmax}_k \delta_k(x)$$

*where*

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k. \tag{22.16}$$

*If the variances of the Gaussians are equal, then*

$$\delta_k(x) = x^T \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} + \log \pi_k. \tag{22.17}$$

We estimate $\delta_k(x)$ by by inserting estimates of $\mu_k$, $\Sigma_k$ and $\pi_k$. There is another version of linear discriminant analysis due to Fisher. The idea is to first reduce the dimension of covariates to one dimension by projecting the data onto a line. Algebraically, this means replacing the covariate $X = (X_1, \ldots, X_d)$ with a linear combination $U = w^T X = \sum_{j=1}^d w_j X_j$. The goal is to choose the vector $w = (w_1, \ldots, w_d)$ that "best separates the data." Then we perform classification with the one-dimensional covariate $Z$ instead of $X$.

We need define what we mean by separation of the groups. We would like the two groups to have means that are far apart relative to their spread. Let $\mu_j$ denote the mean of $X$ for $Y_j$ and let $\Sigma$ be the variance matrix of $X$. Then $\mathbb{E}(U|Y = j) = \mathbb{E}(w^T X|Y = j) = w^T \mu_j$ and $\mathbb{V}(U) = w^T \Sigma w$. [2] Define the separation by

$$\begin{aligned} J(w) &= \frac{(\mathbb{E}(U|Y = 0) - \mathbb{E}(U|Y = 1))^2}{w^T \Sigma w} \\ &= \frac{(w^T \mu_0 - w^T \mu_1)^2}{w^T \Sigma w} \\ &= \frac{w^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T \Sigma w}. \end{aligned}$$

We estimate $J$ as follows. Let $n_j = \sum_{i=1}^n I(Y_i = j)$ be the number of observations in group $j$, let $\overline{X}_j$ be the sample mean vector of the $X$'s for group $j$, and let $S_j$ be the sample covariance matrix in group $j$. Define

$$\widehat{J}(w) = \frac{w^T S_B w}{w^T S_W w} \tag{22.18}$$

[2] The quantity $J$ arises in physics, where it is called the Rayleigh coefficient.

where

$$\begin{aligned} S_B &= (\overline{X}_0 - \overline{X}_1)(\overline{X}_0 - \overline{X}_1)^T \\ S_W &= \frac{(n_0-1)S_0 + (n_1-1)S_1}{(n_0-1)+(n_1-1)}. \end{aligned}$$

**22.10 Theorem.** *The vector*

$$w = S_W^{-1}(\overline{X}_0 - \overline{X}_1) \tag{22.19}$$

*is a minimizer of* $\widehat{J}(w)$. *We call*

$$U = w^T X = (\overline{X}_0 - \overline{X}_1)^T S_W^{-1} X \tag{22.20}$$

*the* **Fisher linear discriminant function**. *The midpoint* $m$ *between* $\overline{X}_0$ *and* $\overline{X}_1$ *is*

$$m = \frac{1}{2}(\overline{X}_0 + \overline{X}_1) = \frac{1}{2}(\overline{X}_0 - \overline{X}_1)^T S_B^{-1}(\overline{X}_0 + \overline{X}_1) \tag{22.21}$$

*Fisher's classification rule is*

$$h(x) = \begin{cases} 0 & \text{if } w^T X \geq m \\ 1 & \text{if } w^T X < m. \end{cases}$$

*Fisher's rule is the same as the Bayes linear classifier in equation (22.14) when* $\widehat{\pi} = 1/2$.

## 22.4 Linear Regression and Logistic Regression

A more direct approach to classification is to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$ without bothering to estimate the densities $f_k$. For the rest of this section, we will only consider the case where $\mathcal{Y} = \{0, 1\}$. Thus, $r(x) = \mathbb{P}(Y = 1|X = x)$ and once we have an estimate $\widehat{r}$, we will use the classification rule

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{22.22}$$

The simplest regression model is the linear regression model

$$Y = r(x) + \epsilon = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon \tag{22.23}$$

where $\mathbb{E}(\epsilon) = 0$. This model can't be correct since it does not force $Y = 0$ or 1. Nonetheless, it can sometimes lead to a decent classifier.

Recall that the least squares estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^T$ minimizes the residual sums of squares

$$\text{RSS}(\beta) = \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j \right)^2.$$

Let $\mathbf{X}$ denote the $N \times (d+1)$ matrix of the form

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1d} \\ 1 & X_{21} & \ldots & X_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{nd} \end{bmatrix}.$$

Also let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Then,

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

and the model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. From Theorem 13.13,

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

The predicted values are

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}.$$

Now we use (22.22) to classify, where $\widehat{r}(x) = \widehat{\beta}_0 + \sum_j \widehat{\beta}_j x_j$.

An alternative is to use logistic regression which was also discussed in Chapter 13. The model is

$$r(x) = \mathbb{P}(Y = 1 | X = x) = \frac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}} \tag{22.24}$$

and the MLE $\widehat{\beta}$ is obtained numerically.

**22.11 Example.** Let us return to the heart disease data. The MLE is given in Example 13.17. The error rate, using this model for classification, is .27. The error rate from a linear regression is .26.

We can get a better classifier by fitting a richer model. For example, we could fit

$$\text{logit } \mathbb{P}(Y = 1 | X = x) = \beta_0 + \sum_j \beta_j x_j + \sum_{j,k} \beta_{jk} x_j x_k. \tag{22.25}$$

More generally, we could add terms of up to order $r$ for some integer $r$. Large values of $r$ give a more complicated model which should fit the data better. But there is a bias–variance tradeoff which we'll discuss later.

**22.12 Example.** If we use model (22.25) for the heart disease data with $r = 2$, the error rate is reduced to .22. ■

## 22.5 Relationship Between Logistic Regression and LDA

LDA and logistic regression are almost the same thing. If we assume that each group is Gaussian with the same covariance matrix, then we saw earlier that

$$\begin{aligned}\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) &= \log\left(\frac{\pi_0}{\pi_1}\right) - \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_1-\mu_0) \\ &\quad + x^T\Sigma^{-1}(\mu_1-\mu_0) \\ &\equiv \alpha_0 + \alpha^T x.\end{aligned}$$

On the other hand, the logistic model is, by assumption,

$$\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) = \beta_0 + \beta^T x.$$

These are the same model since they both lead to classification rules that are linear in $x$. The difference is in how we estimate the parameters.

The joint density of a single observation is $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$. In LDA we estimated the whole joint distribution by maximizing the likelihood

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(x_i|y_i)}_{\text{Gaussian}}\underbrace{\prod_i f(y_i)}_{\text{Bernoulli}}. \tag{22.26}$$

In logistic regression we maximized the conditional likelihood $\prod_i f(y_i|x_i)$ but we ignored the second term $f(x_i)$:

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(y_i|x_i)}_{\text{logistic}}\underbrace{\prod_i f(x_i)}_{\text{ignored}}. \tag{22.27}$$

Since classification only requires knowing $f(y|x)$, we don't really need to estimate the whole joint distribution. Logistic regression leaves the marginal

distribution $f(x)$ unspecified so it is more nonparametric than LDA. This is an advantage of the logistic regression approach over LDA.

To summarize: LDA and logistic regression both lead to a linear classification rule. In LDA we estimate the entire joint distribution $f(x, y) = f(x|y)f(y)$. In logistic regression we only estimate $f(y|x)$ and we don't bother estimating $f(x)$.

## 22.6 Density Estimation and Naive Bayes

The Bayes rule is $h(x) = \text{argmax}_k \ \pi_k f_k(x)$. If we can estimate $\pi_k$ and $f_k$ then we can estimate the Bayes classification rule. Estimating $\pi_k$ is easy but what about $f_k$? We did this previously by assuming $f_k$ was Gaussian. Another strategy is to estimate $f_k$ with some nonparametric density estimator $\widehat{f}_k$ such as a kernel estimator. But if $x = (x_1, \ldots, x_d)$ is high-dimensional, nonparametric density estimation is not very reliable. This problem is ameliorated if we assume that $X_1, \ldots, X_d$ are independent, for then, $f_k(x_1, \ldots, x_d) = \prod_{j=1}^d f_{kj}(x_j)$. This reduces the problem to $d$ one-dimensional density estimation problems, within each of the $k$ groups. The resulting classifier is called **the naive Bayes classifier.** The assumption that the components of $X$ are independent is usually wrong yet the resulting classifier might still be accurate. Here is a summary of the steps in the naive Bayes classifier:

The Naive Bayes Classifier

1. For each group $k$, compute an estimate $\widehat{f}_{kj}$ of the density $f_{kj}$ for $X_j$, using the data for which $Y_i = k$.

2. Let
$$\widehat{f}_k(x) = \widehat{f}_k(x_1, \ldots, x_d) = \prod_{j=1}^d \widehat{f}_{kj}(x_j).$$

3. Let
$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^n I(Y_i = k)$$
where $I(Y_i = k) = 1$ if $Y_i = k$ and $I(Y_i = k) = 0$ if $Y_i \neq k$.

4. Let
$$h(x) = \text{argmax}_k \ \widehat{\pi}_k \ \widehat{f}_k(x).$$

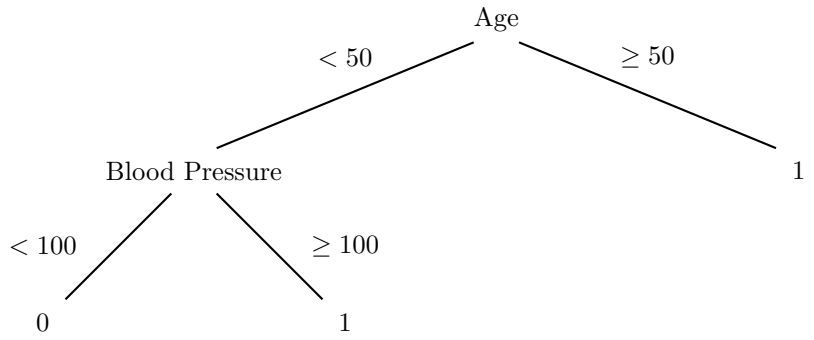



FIGURE 22.2. A simple classification tree.

The naive Bayes classifier is popular when $x$ is high-dimensional and discrete. In that case, $\widehat{f}_{kj}(x_j)$ is especially simple.

## 22.7 Trees

Trees are classification methods that partition the covariate space $\mathcal{X}$ into disjoint pieces and then classify the observations according to which partition element they fall in. As the name implies, the classifier can be represented as a tree.

For illustration, suppose there are two covariates, $X_1 = \text{age}$ and $X_2 = \text{blood pressure}$. Figure 22.2 shows a classification tree using these variables.

The tree is used in the following way. If a subject has Age $\geq 50$ then we classify him as $Y = 1$. If a subject has Age $< 50$ then we check his blood pressure. If systolic blood pressure is $< 100$ then we classify him as $Y = 1$, otherwise we classify him as $Y = 0$. Figure 22.3 shows the same classifier as a partition of the covariate space.

Here is how a tree is constructed. First, suppose that $y \in \mathcal{Y} = \{0, 1\}$ and that there is only a single covariate $X$. We choose a split point $t$ that divides the real line into two sets $A_1 = (-\infty, t]$ and $A_2 = (t, \infty)$. Let $\widehat{p}_s(j)$ be the proportion of observations in $A_s$ such that $Y_i = j$:

$$\widehat{p}_s(j) = \frac{\sum_{i=1}^n I(Y_i = j, X_i \in A_s)}{\sum_{i=1}^n I(X_i \in A_s)} \tag{22.28}$$

for $s = 1, 2$ and $j = 0, 1$. The **impurity** of the split $t$ is defined to be

$$I(t) = \sum_{s=1}^{2} \gamma_s \tag{22.29}$$

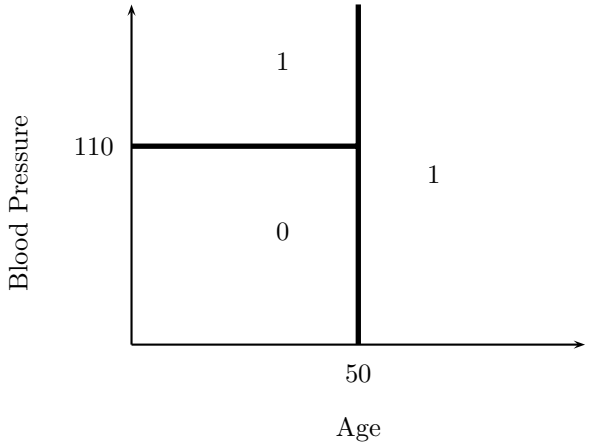



FIGURE 22.3. Partition representation of classification tree.

where

$$\gamma_s = 1 - \sum_{j=0}^{1} \widehat{p}_s(j)^2. \tag{22.30}$$

This particular measure of impurity is known as the **Gini index**. If a partition element $A_s$ contains all 0's or all 1's, then $\gamma_s = 0$. Otherwise, $\gamma_s > 0$. We choose the split point $t$ to minimize the impurity. (Other indices of impurity besides can be used besides the Gini index.)

When there are several covariates, we choose whichever covariate and split that leads to the lowest impurity. This process is continued until some stopping criterion is met. For example, we might stop when every partition element has fewer than $n_0$ data points, where $n_0$ is some fixed number. The bottom nodes of the tree are called the **leaves**. Each leaf is assigned a 0 or 1 depending on whether there are more data points with $Y = 0$ or $Y = 1$ in that partition element.

This procedure is easily generalized to the case where $Y \in \{1, \ldots, K\}$. We simply define the impurity by

$$\gamma_s = 1 - \sum_{j=1}^{k} \widehat{p}_s(j)^2 \tag{22.31}$$

where $\widehat{p}_i(j)$ is the proportion of observations in the partition element for which $Y = j$.

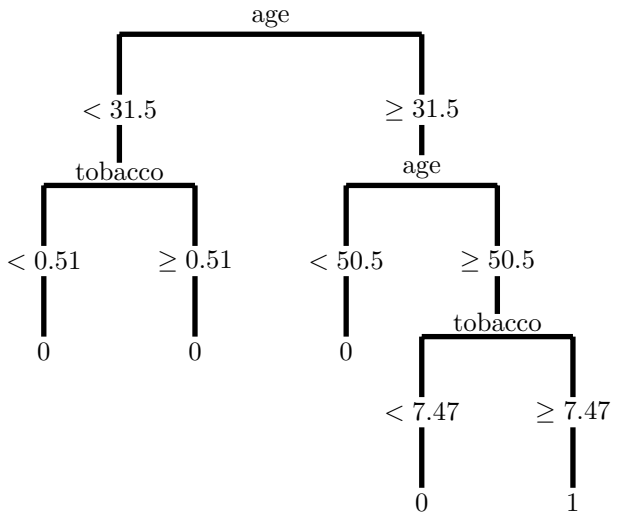



FIGURE 22.4. A classification tree for the heart disease data using two covariates.

**22.13 Example.** A classification tree for the heart disease data yields a misclassification rate of .21. If we build a tree using only tobacco and age, the misclassification rate is then .29. The tree is shown in Figure 22.4. ■

Our description of how to build trees is incomplete. If we keep splitting until there are few cases in each leaf of the tree, we are likely to overfit the data. We should choose the complexity of the tree in such a way that the estimated true error rate is low. In the next section, we discuss estimation of the error rate.

## 22.8 Assessing Error Rates and Choosing a Good Classifier

How do we choose a good classifier? We would like to have a classifier $h$ with a low true error rate $L(h)$. Usually, we can't use the training error rate $\widehat{L}_n(h)$ as an estimate of the true error rate because it is biased downward.

**22.14 Example.** Consider the heart disease data again. Suppose we fit a sequence of logistic regression models. In the first model we include one covariate. In the second model we include two covariates, and so on. The ninth model includes all the covariates. We can go even further. Let's also fit a tenth model that includes all nine covariates plus the first covariate squared. Then

we fit an eleventh model that includes all nine covariates plus the first covariate squared and the second covariate squared. Continuing this way we will get a sequence of 18 classifiers of increasing complexity. The solid line in Figure 22.5 shows the observed classification error which steadily decreases as we make the model more complex. If we keep going, we can make a model with zero observed classification error. The dotted line shows the **10-fold cross-validation estimate** of the error rate (to be explained shortly) which is a better estimate of the true error rate than the observed classification error. The estimated error decreases for a while then increases. This is essentially the bias–variance tradeoff phenomenon we have seen in Chapter 20. ■

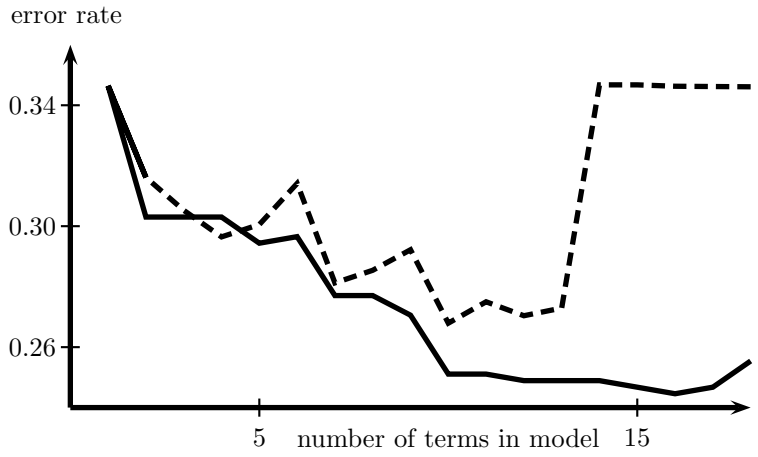



FIGURE 22.5. The solid line is the observed error rate and dashed line is the cross-validation estimate of true error rate.

There are many ways to estimate the error rate. We'll consider two: **cross-validation** and **probability inequalities.**

Cross-Validation. The basic idea of cross-validation, which we have already encountered in curve estimation, is to leave out some of the data when fitting a model. The simplest version of cross-validation involves randomly splitting the data into two pieces: the **training set** $\mathcal{T}$ and the **validation set** $\mathcal{V}$. Often, about 10 per cent of the data might be set aside as the validation set. The classifier $h$ is constructed from the training set. We then estimate

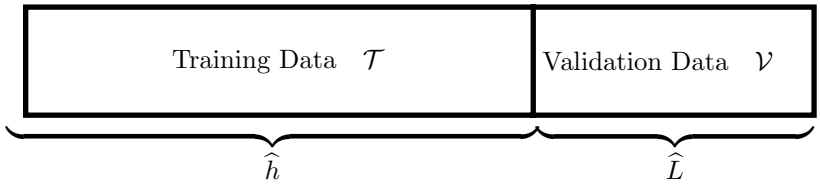



FIGURE 22.6. Cross-validation. The data are divided into two groups: the training data and the validation data. The training data are used to produce an estimated classifier $\widehat{h}$. Then, $\widehat{h}$ is applied to the validation data to obtain an estimate $\widehat{L}$ of the error rate of $\widehat{h}$.

the error by

$$\widehat{L}(h) = \frac{1}{m} \sum_{X_i \in \mathcal{V}} I(h(X_i) \neq Y_I). \tag{22.32}$$

where $m$ is the size of the validation set. See Figure 22.6.

Another approach to cross-validation is **K-fold cross-validation** which is obtained from the following algorithm.

$K$-fold cross-validation.

1. Randomly divide the data into $K$ chunks of approximately equal size. A common choice is $K = 10$.
2. For k = 1 to K, do the following:
   (a) Delete chunk $k$ from the data.
   (b) Compute the classifier $\widehat{h}_{(k)}$ from the rest of the data.
   (c) Use $\widehat{h}_{(k)}$ to the predict the data in chunk $k$. Let $\widehat{L}_{(k)}$ denote the observed error rate.
3. Let

$$\widehat{L}(h) = \frac{1}{K} \sum_{k=1}^{K} \widehat{L}_{(k)}. \tag{22.33}$$

**22.15 Example.** We applied 10-fold cross-validation to the heart disease data. The minimum cross-validation error as a function of the number of leaves occurred at six. Figure 22.7 shows the tree with six leaves. ■

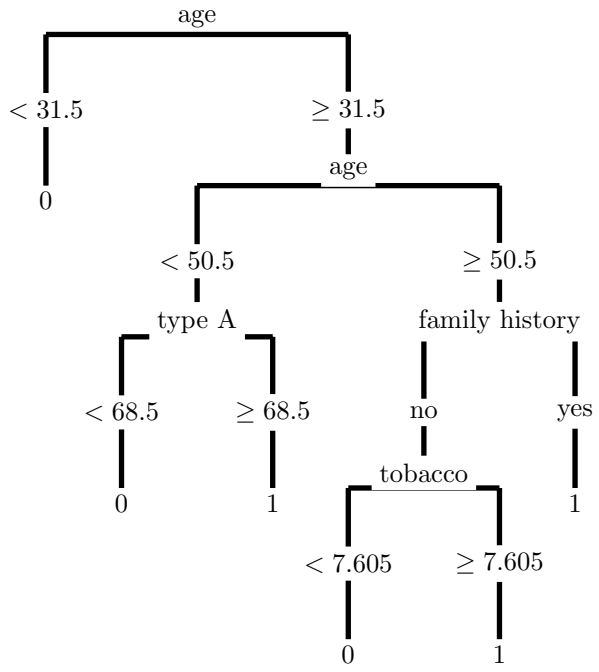



FIGURE 22.7. Smaller classification tree with size chosen by cross-validation.

PROBABILITY INEQUALITIES. Another approach to estimating the error rate is to find a confidence interval for $\widehat{L}_n(h)$ using probability inequalities. This method is useful in the context of **empirical risk minimization**.

Let $\mathcal{H}$ be a set of classifiers, for example, all linear classifiers. Empirical risk minimization means choosing the classifier $\widehat{h} \in \mathcal{H}$ to minimize the training error $\widehat{L}_n(h)$, also called the empirical risk. Thus,

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{L}_n(h) = \text{argmin}_{h \in \mathcal{H}} \left( \frac{1}{n} \sum_i I(h(X_i) \neq Y_i) \right). \tag{22.34}$$

Typically, $\widehat{L}_n(\widehat{h})$ underestimates the true error rate $L(\widehat{h})$ because $\widehat{h}$ was chosen to make $\widehat{L}_n(\widehat{h})$ small. Our goal is to assess how much underestimation is taking place. Our main tool for this analysis is **Hoeffding's inequality** (Theorem 4.5). Recall that if $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, then, for any $\epsilon > 0$,

$$\mathbb{P}\left(|\widehat{p} - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{22.35}$$

where $\widehat{p} = n^{-1} \sum_{i=1}^n X_i$.

First, suppose that $\mathcal{H} = \{h_1, \ldots, h_m\}$ consists of finitely many classifiers. For any fixed $h$, $\widehat{L}_n(h)$ converges in almost surely to $L(h)$ by the law of large numbers. We will now establish a stronger result.

**22.16 Theorem** (Uniform Convergence). *Assume $\mathcal{H}$ is finite and has $m$ elements. Then,*

$$\mathbb{P}\left(\max_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) \leq 2me^{-2n\epsilon^2}.$$

PROOF. We will use Hoeffding's inequality and we will also use the fact that if $A_1, \ldots, A_m$ is a set of events then $\mathbb{P}(\bigcup_{i=1}^m A_i) \leq \sum_{i=1}^m \mathbb{P}(A_i)$. Now,

$$\begin{aligned}
\mathbb{P}\left(\max_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) &= \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) \\
&\leq \sum_{H\in\mathcal{H}} \mathbb{P}\left(|\widehat{L}_n(h) - L(h)| > \epsilon\right) \\
&\leq \sum_{H\in\mathcal{H}} 2e^{-2n\epsilon^2} = 2me^{-2n\epsilon^2}. \quad \blacksquare
\end{aligned}$$

**22.17 Theorem.** *Let*

$$\epsilon = \sqrt{\frac{2}{n}\log\left(\frac{2m}{\alpha}\right)}.$$

*Then $\widehat{L}_n(\widehat{h}) \pm \epsilon$ is a $1-\alpha$ confidence interval for $L(\widehat{h})$.*

PROOF. This follows from the fact that

$$\begin{aligned}
\mathbb{P}(|\widehat{L}_n(\widehat{h}) - L(\widehat{h})| > \epsilon) &\leq \mathbb{P}\left(\max_{h\in\mathcal{H}} |\widehat{L}_n(\widehat{h}) - L(\widehat{h})| > \epsilon\right) \\
&\leq 2me^{-2n\epsilon^2} = \alpha. \quad \blacksquare
\end{aligned}$$

When $\mathcal{H}$ is large the confidence interval for $L(\widehat{h})$ is large. The more functions there are in $\mathcal{H}$ the more likely it is we have "overfit" which we compensate for by having a larger confidence interval.

In practice we usually use sets $\mathcal{H}$ that are infinite, such as the set of linear classifiers. To extend our analysis to these cases we want to be able to say something like

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) \leq \text{something not too big.}$$

One way to develop such a generalization is by way of the **Vapnik-Chervonenkis** or **VC dimension.**

Let $\mathcal{A}$ be a class of sets. Give a finite set $F = \{x_1, \ldots, x_n\}$ let

$$N_{\mathcal{A}}(F) = \#\left\{F \bigcap A : \ A \in \mathcal{A}\right\} \tag{22.36}$$

be the number of subsets of $F$ "picked out" by $\mathcal{A}$. Here $\#(B)$ denotes the number of elements of a set $B$. The **shatter coefficient** is defined by

$$s(\mathcal{A}, n) = \max_{F \in \mathcal{F}_n} N_{\mathcal{A}}(F) \tag{22.37}$$

where $\mathcal{F}_n$ consists of all finite sets of size $n$. Now let $X_1, \ldots, X_n \sim \mathbb{P}$ and let

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_i I(X_i \in A)$$

denote the **empirical probability measure**. The following remarkable theorem bounds the distance between $\mathbb{P}$ and $\mathbb{P}_n$.

**22.18 Theorem** (Vapnik and Chervonenkis (1971)). *For any $\mathbb{P}$, $n$ and $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| > \epsilon\right\} \leq 8 s(\mathcal{A}, n) e^{-n\epsilon^2/32}. \tag{22.38}$$

The proof, though very elegant, is long and we omit it. If $\mathcal{H}$ is a set of classifiers, define $\mathcal{A}$ to be the class of sets of the form $\{x : \ h(x) = 1\}$. We then define $s(\mathcal{H}, n) = s(\mathcal{A}, n)$.

**22.19 Theorem.**

$$\mathbb{P}\left\{\sup_{h \in \mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right\} \leq 8 s(\mathcal{H}, n) e^{-n\epsilon^2/32}.$$

*A $1 - \alpha$ confidence interval for $L(\widehat{h})$ is $\widehat{L}_n(\widehat{h}) \pm \epsilon_n$ where*

$$\epsilon_n^2 = \frac{32}{n} \log\left(\frac{8 s(\mathcal{H}, n)}{\alpha}\right).$$

These theorems are only useful if the shatter coefficients do not grow too quickly with $n$. This is where VC dimension enters.

**22.20 Definition.** *The VC (Vapnik-Chervonenkis) dimension of a class of sets $\mathcal{A}$ is defined as follows. If $s(\mathcal{A}, n) = 2^n$ for all $n$, set $VC(\mathcal{A}) = \infty$. Otherwise, define $VC(\mathcal{A})$ to be the largest $k$ for which $s(\mathcal{A}, n) = 2^k$.*

Thus, the VC-dimension is the size of the largest finite set $F$ that can be **shattered** by $\mathcal{A}$ meaning that $\mathcal{A}$ picks out each subset of $F$. If $\mathcal{H}$ is a set of classifiers we define $VC(\mathcal{H}) = VC(\mathcal{A})$ where $\mathcal{A}$ is the class of sets of the form $\{x : \ h(x) = 1\}$ as $h$ varies in $\mathcal{H}$. The following theorem shows that if $\mathcal{A}$ has finite VC-dimension, then the shatter coefficients grow as a polynomial in $n$.

**22.21 Theorem.** *If $\mathcal{A}$ has finite VC-dimension $v$, then*

$$s(\mathcal{A}, n) \leq n^v + 1.$$

**22.22 Example.** Let $\mathcal{A} = \{(-\infty, a];\ a \in \mathcal{R}\}$. The $\mathcal{A}$ shatters every 1-point set $\{x\}$ but it shatters no set of the form $\{x, y\}$. Therefore, $VC(\mathcal{A}) = 1$. ∎

**22.23 Example.** Let $\mathcal{A}$ be the set of closed intervals on the real line. Then $\mathcal{A}$ shatters $S = \{x, y\}$ but it cannot shatter sets with 3 points. Consider $S = \{x, y, z\}$ where $x < y < z$. One cannot find an interval $A$ such that $A \bigcap S = \{x, z\}$. So, $VC(\mathcal{A}) = 2$. ∎

**22.24 Example.** Let $\mathcal{A}$ be all linear half-spaces on the plane. Any 3-point set (not all on a line) can be shattered. No 4 point set can be shattered. Consider, for example, 4 points forming a diamond. Let $T$ be the left and rightmost points. This can't be picked out. Other configurations can also be seen to be unshatterable. So $VC(\mathcal{A}) = 3$. In general, halfspaces in $\mathcal{R}^d$ have VC dimension $d + 1$. ∎

**22.25 Example.** Let $\mathcal{A}$ be all rectangles on the plane with sides parallel to the axes. Any 4 point set can be shattered. Let $S$ be a 5 point set. There is one point that is not leftmost, rightmost, uppermost, or lowermost. Let $T$ be all points in $S$ except this point. Then $T$ can't be picked out. So $VC(\mathcal{A}) = 4$. ∎

**22.26 Theorem.** *Let $x$ have dimension $d$ and let $\mathcal{H}$ be th set of linear classifiers. The VC-dimension of $\mathcal{H}$ is $d+1$. Hence, a $1-\alpha$ confidence interval for the true error rate is $\widehat{L}(\widehat{h}) \pm \epsilon$ where*

$$\epsilon_n^2 = \frac{32}{n} \log\left(\frac{8(n^{d+1}+1)}{\alpha}\right).$$

## 22.9 Support Vector Machines

In this section we consider a class of linear classifiers called **support vector machines**. Throughout this section, we assume that $Y$ is binary. It will be convenient to label the outcomes as $-1$ and $+1$ instead of 0 and 1. A linear classifier can then be written as

$$h(x) = \text{sign}\Big(H(x)\Big)$$

where $x = (x_1, \ldots, x_d)$,

$$H(x) = a_0 + \sum_{i=1}^{d} a_i x_i$$

and

$$\text{sign}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \\ 1 & \text{if } z > 0. \end{cases}$$

First, suppose that the data are **linearly separable**, that is, there exists a hyperplane that perfectly separates the two classes.

**22.27 Lemma.** *The data can be separated by some hyperplane if and only if there exists a hyperplane $H(x) = a_0 + \sum_{i=1}^{d} a_i x_i$ such that*

$$Y_i H(x_i) \geq 1, \quad i = 1, \ldots, n. \tag{22.39}$$

PROOF. Suppose the data can be separated by a hyperplane $W(x) = b_0 + \sum_{i=1}^{d} b_i x_i$. It follows that there exists some constant $c$ such that $Y_i = 1$ implies $W(X_i) \geq c$ and $Y_i = -1$ implies $W(X_i) \leq -c$. Therefore, $Y_i W(X_i) \geq c$ for all $i$. Let $H(x) = a_0 + \sum_{i=1}^{d} a_i x_i$ where $a_j = b_j/c$. Then $Y_i H(X_i) \geq 1$ for all $i$. The reverse direction is straightforward. ∎

In the separable case, there will be many separating hyperplanes. How should we choose one? Intuitively, it seems reasonable to choose the hyperplane "furthest" from the data in the sense that it separates the +1s and -1s and maximizes the distance to the closest point. This hyperplane is called the **maximum margin hyperplane.** The margin is the distance to from the hyperplane to the nearest point. Points on the boundary of the margin are called **support vectors.** See Figure 22.8.

**22.28 Theorem.** *The hyperplane $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^{d} \widehat{a}_i x_i$ that separates the data and maximizes the margin is given by minimizing $(1/2)\sum_{j=1}^{d} a_j^2$ subject to (22.39).*

It turns out that this problem can be recast as a quadratic programming problem. Let $\langle X_i, X_k \rangle = X_i^T X_k$ denote the inner product of $X_i$ and $X_k$.

**22.29 Theorem.** *Let $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^{d} \widehat{a}_i x_i$ denote the optimal (largest margin) hyperplane. Then, for $j = 1, \ldots, d$,*

$$\widehat{a}_j = \sum_{i=1}^{n} \widehat{\alpha}_i Y_i X_j(i)$$

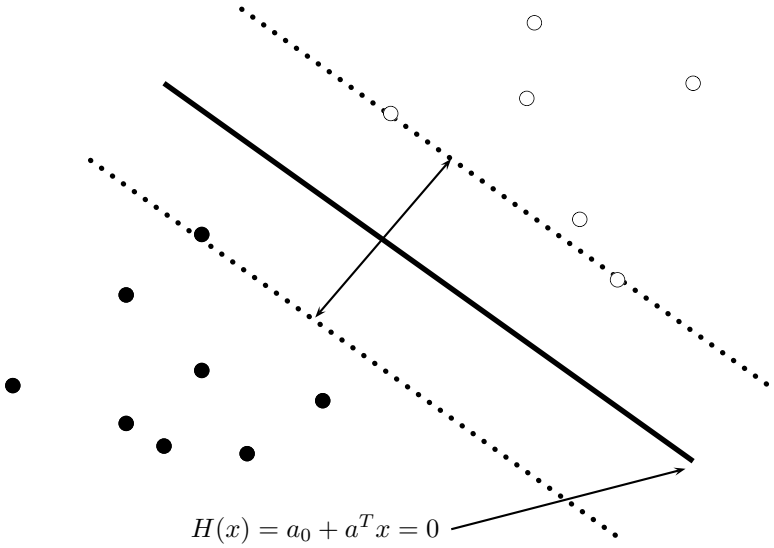



FIGURE 22.8. The hyperplane $H(x)$ has the largest margin of all hyperplanes that separate the two classes.

*where $X_j(i)$ is the value of the covariate $X_j$ for the $i^{\text{th}}$ data point, and $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$ is the vector that maximizes*

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k Y_i Y_k \langle X_i, X_k \rangle \tag{22.40}$$

*subject to*

$$\alpha_i \geq 0$$

*and*

$$0 = \sum_i \alpha_i Y_i.$$

*The points $X_i$ for which $\widehat{\alpha} \neq 0$ are called* **support vectors**. *$\widehat{a}_0$ can be found by solving*

$$\widehat{\alpha}_i \left( Y_i (X_i^T \widehat{a} + \widehat{\beta}_0 \right) = 0$$

*for any support point $X_i$. $\widehat{H}$ may be written as*

$$\widehat{H}(x) = \widehat{\alpha}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i \langle x, X_i \rangle.$$

There are many software packages that will solve this problem quickly. If there is no perfect linear classifier, then one allows overlap between the groups

by replacing the condition (22.39) with

$$Y_i H(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \tag{22.41}$$

The variables $\xi_1, \ldots, \xi_n$ are called **slack variables**.

We now maximize (22.40) subject to

$$0 \leq \xi_i \leq c, \quad i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} \alpha_i Y_i = 0.$$

The constant $c$ is a tuning parameter that controls the amount of overlap.

## 22.10 Kernelization

There is a trick called **kernelization** for improving a computationally simple classifier $h$. The idea is to map the covariate $X$ — which takes values in $\mathcal{X}$ — into a higher dimensional space $\mathcal{Z}$ and apply the classifier in the bigger space $\mathcal{Z}$. This can yield a more flexible classifier while retaining computationally simplicity.

The standard example of this idea is illustrated in Figure 22.9. The covariate $x = (x_1, x_2)$. The $Y_i$s can be separated into two groups using an ellipse. Define a mapping $\phi$ by

$$z = (z_1, z_2, z_3) = \phi(x) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2).$$

Thus, $\phi$ maps $\mathcal{X} = \mathbb{R}^2$ into $\mathcal{Z} = \mathbb{R}^3$. In the higher-dimensional space $\mathcal{Z}$, the $Y_i$'s are separable by a linear decision boundary. In other words,

> a linear classifier in a higher-dimensional space corresponds to a non-linear classifier in the original space.

The point is that to get a richer set of classifiers we do not need to give up the convenience of linear classifiers. We simply map the covariates to a higher-dimensional space. This is akin to making linear regression more flexible by using polynomials.

There is a potential drawback. If we significantly expand the dimension of the problem, we might increase the computational burden. For example, if $x$ has dimension $d = 256$ and we wanted to use all fourth-order terms, then $z = \phi(x)$ has dimension 183,181,376. We are spared this computational

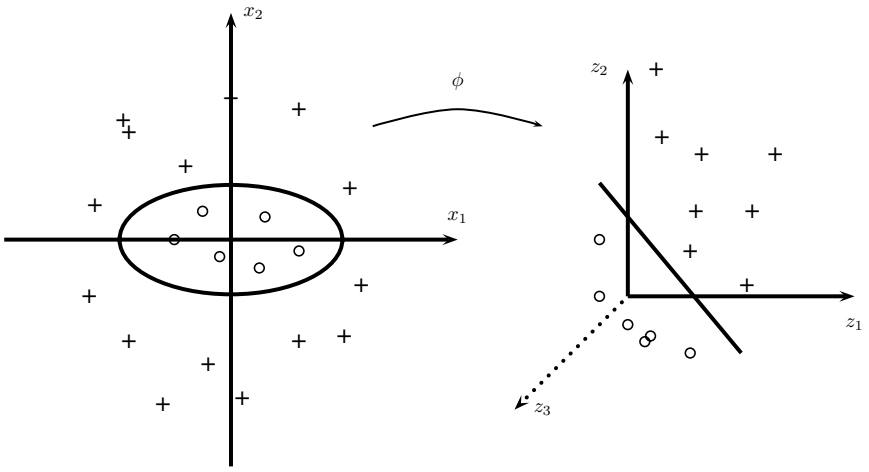



FIGURE 22.9. Kernelization. Mapping the covariates into a higher-dimensional space can make a complicated decision boundary into a simpler decision boundary.

nightmare by the following two facts. First, many classifiers do not require that we know the values of the individual points but, rather, just the inner product between pairs of points. Second, notice in our example that the inner product in $\mathcal{Z}$ can be written

$$\begin{aligned} \langle z, \widetilde{z} \rangle &= \langle \phi(x), \phi(\widetilde{x}) \rangle \\ &= x_1^2 \widetilde{x}_1^2 + 2 x_1 \widetilde{x}_1 x_2 \widetilde{x}_2 + x_2^2 \widetilde{x}_2^2 \\ &= (\langle x, \widetilde{x} \rangle)^2 \equiv K(x, \widetilde{x}). \end{aligned}$$

Thus, we can compute $\langle z, \widetilde{z} \rangle$ without ever computing $Z_i = \phi(X_i)$.

To summarize, kernelization involves finding a mapping $\phi : \mathcal{X} \to \mathcal{Z}$ and a classifier such that:

1. $\mathcal{Z}$ has higher dimension than $\mathcal{X}$ and so leads a richer set of classifiers.
2. The classifier only requires computing inner products.
3. There is a function $K$, called a kernel, such that $\langle \phi(x), \phi(\widetilde{x}) \rangle = K(x, \widetilde{x})$.
4. Everywhere the term $\langle x, \widetilde{x} \rangle$ appears in the algorithm, replace it with $K(x, \widetilde{x})$.

In fact, we never need to construct the mapping $\phi$ at all. We only need to specify a kernel $K(x, \widetilde{x})$ that corresponds to $\langle \phi(x), \phi(\widetilde{x}) \rangle$ for some $\phi$. This raises an interesting question: given a function of two variables $K(x, y)$, does there exist a function $\phi(x)$ such that $K(x, y) = \langle \phi(x), \phi(y) \rangle$? The answer is provided by **Mercer's theorem** which says, roughly, that if $K$ is positive definite — meaning that

$$\int \int K(x, y) f(x) f(y) dx dy \geq 0$$

for square integrable functions $f$ — then such a $\phi$ exists. Examples of commonly used kernels are:

$$\begin{array}{rlcl} \text{polynomial} & K(x, \widetilde{x}) & = & \Big( \langle x, \widetilde{x} \rangle + a \Big)^r \\ \text{sigmoid} & K(x, \widetilde{x}) & = & \tanh(a \langle x, \widetilde{x} \rangle + b) \\ \text{Gaussian} & K(x, \widetilde{x}) & = & \exp\Big( -||x - \widetilde{x}||^2 / (2\sigma^2) \Big) \end{array}$$

Let us now see how we can use this trick in LDA and in support vector machines.

Recall that the Fisher linear discriminant method replaces $X$ with $U = w^T X$ where $w$ is chosen to maximize the Rayleigh coefficient

$$J(w) = \frac{w^T S_B w}{w^T S_W w},$$

$$S_B = (\overline{X}_0 - \overline{X}_1)(\overline{X}_0 - \overline{X}_1)^T$$

and

$$S_W = \left( \frac{(n_0 - 1) S_0}{(n_0 - 1) + (n_1 - 1)} \right) + \left( \frac{(n_1 - 1) S_1}{(n_0 - 1) + (n_1 - 1)} \right).$$

In the kernelized version, we replace $X_i$ with $Z_i = \phi(X_i)$ and we find $w$ to maximize

$$J(w) = \frac{w^T \widetilde{S}_B w}{w^T \widetilde{S}_W w} \tag{22.42}$$

where

$$\widetilde{S}_B = (\overline{Z}_0 - \overline{Z}_1)(\overline{Z}_0 - \overline{Z}_1)^T$$

and

$$S_W = \left( \frac{(n_0 - 1) \widetilde{S}_0}{(n_0 - 1) + (n_1 - 1)} \right) + \left( \frac{(n_1 - 1) \widetilde{S}_1}{(n_0 - 1) + (n_1 - 1)} \right).$$

Here, $\widetilde{S}_j$ is the sample of covariance of the $Z_i$'s for which $Y = j$. However, to take advantage of kernelization, we need to re-express this in terms of inner products and then replace the inner products with kernels.

It can be shown that the maximizing vector $w$ is a linear combination of the $Z_i$'s. Hence we can write

$$w = \sum_{i=1}^{n} \alpha_i Z_i.$$

Also,

$$\overline{Z}_j = \frac{1}{n_j} \sum_{i=1}^{n} \phi(X_i) I(Y_i = j).$$

Therefore,

$$\begin{aligned} w^T \overline{Z}_j &= \left( \sum_{i=1}^{n} \alpha_i Z_i \right)^T \left( \frac{1}{n_j} \sum_{i=1}^{n} \phi(X_i) I(Y_i = j) \right) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \sum_{s=1}^{n} \alpha_i I(Y_s = j) Z_i^T \phi(X_s) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \alpha_i \sum_{s=1}^{n} I(Y_s = j) \phi(X_i)^T \phi(X_s) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \alpha_i \sum_{s=1}^{n} I(Y_s = j) K(X_i, X_s) \\ &= \alpha^T M_j \end{aligned}$$

where $M_j$ is a vector whose $i^{\text{th}}$ component is

$$M_j(i) = \frac{1}{n_j} \sum_{s=1}^{n} K(X_i, X_s) I(Y_i = j).$$

It follows that

$$w^T \widetilde{S}_B w = \alpha^T M \alpha$$

where $M = (M_0 - M_1)(M_0 - M_1)^T$. By similar calculations, we can write

$$w^T \widetilde{S}_W w = \alpha^T N \alpha$$

where

$$N = K_0 \left( I - \frac{1}{n_0} \mathbf{1} \right) K_0^T + K_1 \left( I - \frac{1}{n_1} \mathbf{1} \right) K_1^T,$$

$I$ is the identity matrix, $\mathbf{1}$ is a matrix of all one's, and $K_j$ is the $n \times n_j$ matrix with entries $(K_j)_{rs} = K(x_r, x_s)$ with $x_s$ varying over the observations in group $j$. Hence, we now find $\alpha$ to maximize

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}.$$

All the quantities are expressed in terms of the kernel. Formally, the solution is $\alpha = N^{-1}(M_0 - M_1)$. However, $N$ might be non-invertible. In this case one replaces $N$ by $N + bI$, for some constant $b$. Finally, the projection onto the new subspace can be written as

$$U = w^T \phi(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x).$$

The support vector machine can similarly be kernelized. We simply replace $\langle X_i, X_j \rangle$ with $K(X_i, X_j)$. For example, instead of maximizing (22.40), we now maximize

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k Y_i Y_k K(X_i, X_j). \tag{22.43}$$

The hyperplane can be written as $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i K(X, X_i)$.

## 22.11 Other Classifiers

There are many other classifiers and space precludes a full discussion of all of them. Let us briefly mention a few.

The **k-nearest-neighbors** classifier is very simple. Given a point $x$, find the $k$ data points closest to $x$. Classify $x$ using the majority vote of these $k$ neighbors. Ties can be broken randomly. The parameter $k$ can be chosen by cross-validation.

**Bagging** is a method for reducing the variability of a classifier. It is most helpful for highly nonlinear classifiers such as trees. We draw $B$ bootstrap samples from the data. The $b^{\text{th}}$ bootstrap sample yields a classifier $h_b$. The final classifier is

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \frac{1}{B} \sum_{b=1}^{B} h_b(x) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

**Boosting** is a method for starting with a simple classifier and gradually improving it by refitting the data giving higher weight to misclassified samples. Suppose that $\mathcal{H}$ is a collection of classifiers, for example, trees with only one split. Assume that $Y_i \in \{-1, 1\}$ and that each $h$ is such that $h(x) \in \{-1, 1\}$. We usually give equal weight to all data points in the methods we have discussed. But one can incorporate unequal weights quite easily in most algorithms. For example, in constructing a tree, we could replace the impurity measure with a weighted impurity measure. The original version of boosting, called AdaBoost, is as follows.

1. Set the weights $w_i = 1/n$, $i = 1, \ldots, n$.

2. For $j = 1, \ldots, J$, do the following steps:

   (a) Constructing a classifier $h_j$ from the data using the weights $w_1, \ldots, w_n$.

   (b) Compute the weighted error estimate:

$$\widehat{L}_j = \frac{\sum_{i=1}^n w_i I(Y_i \neq h_j(X_i))}{\sum_{i=1}^n w_i}.$$

   (c) Let $\alpha_j = \log((1 - \widehat{L}_j)/\widehat{L}_j)$.

   (d) Update the weights:

$$w_i \longleftarrow w_i e^{\alpha_j I(Y_i \neq h_j(X_i))}$$

3. The final classifier is

$$\widehat{h}(x) = \text{sign}\left(\sum_{j=1}^J \alpha_j h_j(x)\right).$$

There is now an enormous literature trying to explain and improve on boosting. Whereas bagging is a variance reduction technique, boosting can be thought of as a bias reduction technique. We starting with a simple — and hence highly-biased — classifier, and we gradually reduce the bias. The disadvantage of boosting is that the final classifier is quite complicated.

**Neural Networks** are regression models of the form [3]

$$Y = \beta_0 + \sum_{j=1}^p \beta_j \sigma(\alpha_0 + \alpha^T X)$$

where $\sigma$ is a smooth function, often taken to be $\sigma(v) = e^v/(1 + e^v)$. This is really nothing more than a nonlinear regression model. Neural nets were fashionable for some time but they pose great computational difficulties. In particular, one often encounters multiple minima when trying to find the least squares estimates of the parameters. Also, the number of terms $p$ is essentially a smoothing parameter and there is the usual problem of trying to choose $p$ to find a good balance between bias and variance.

[3] This is the simplest version of a neural net. There are more complex versions of the model.

## 22.12 Bibliographic Remarks

The literature on classification is vast and is growing quickly. An excellent reference is Hastie et al. (2001). For more on the theory, see Devroye et al. (1996) and Vapnik (1998). Two recent books on kernels are Scholkopf and Smola (2002) and Herbich (2002).

## 22.13 Exercises

1. Prove Theorem 22.5.

2. Prove Theorem 22.7.

3. Download the spam data from:

   http://www-stat.stanford.edu/∼tibs/ElemStatLearn/index.html

   The data file can also be found on the course web page. The data contain 57 covariates relating to email messages. Each email message was classified as spam (Y=1) or not spam (Y=0). The outcome Y is the last column in the file. The goal is to predict whether an email is spam or not.

   (a) Construct classification rules using (i) LDA, (ii) QDA, (iii) logistic regression, and (iv) a classification tree. For each, report the observed misclassification error rate and construct a 2-by-2 table of the form

| | $\widehat{h}(x) = 0$ | $\widehat{h}(x) = 1$ |
|---|---|---|
| $Y = 0$ | ?? | ?? |
| $Y = 1$ | ?? | ?? |

   (b) Use 5-fold cross-validation to estimate the prediction accuracy of LDA and logistic regression.

   (c) Sometimes it helps to reduce the number of covariates. One strategy is to compare $X_i$ for the spam and email group. For each of the 57 covariates, test whether the mean of the covariate is the same or different between the two groups. Keep the 10 covariates with the smallest p-values. Try LDA and logistic regression using only these 10 variables.

4. Let $\mathcal{A}$ be the set of two-dimensional spheres. That is, $A \in \mathcal{A}$ if $A = \{(x,y) : (x-a)^2+(y-b)^2 \leq c^2\}$ for some $a, b, c$. Find the VC-dimension of $\mathcal{A}$.

5. Classify the spam data using support vector machines. Free software for the support vector machine is at http://svmlight.joachims.org/

6. Use VC theory to get a confidence interval on the true error rate of the LDA classifier for the iris data (from the book web site).

7. Suppose that $X_i \in \mathbb{R}$ and that $Y_i = 1$ whenever $|X_i| \leq 1$ and $Y_i = 0$ whenever $|X_i| > 1$. Show that no linear classifier can perfectly classify these data. Show that the kernelized data $Z_i = (X_i, X_i^2)$ can be linearly separated.

8. Repeat question 5 using the kernel $K(x, \widetilde{x}) = (1 + x^T \widetilde{x})^p$. Choose $p$ by cross-validation.

9. Apply the $k$ nearest neighbors classifier to the "iris data." Choose $k$ by cross-validation.

10. (Curse of Dimensionality.) Suppose that $X$ has a uniform distribution on the $d$-dimensional cube $[-1/2, 1/2]^d$. Let $R$ be the distance from the origin to the closest neighbor. Show that the median of $R$ is

$$\left( \frac{\left(1 - \left(\frac{1}{2}\right)^{1/n}\right)}{v_d(1)} \right)^{1/d}$$

where

$$v_d(r) = r^d \frac{\pi^{d/2}}{\Gamma((d/2)+1)}$$

is the volume of a sphere of radius $r$. For what dimension $d$ does the median of $R$ exceed the edge of the cube when $n = 100$, $n = 1,000$, $n = 10,000$? (Hastie et al. (2001), p. 22–27.)

11. Fit a tree to the data in question 3. Now apply bagging and report your results.

12. Fit a tree that uses only one split on one variable to the data in question 3. Now apply boosting.

13. Let $r(x) = \mathbb{P}(Y = 1|X = x)$ and let $\widehat{r}(x)$ be an estimate of $r(x)$. Consider the classifier

$$h(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $\widehat{r}(x) \approx N(\overline{r}(x), \sigma^2(x))$ for some functions $\overline{r}(x)$ and $\sigma^2(x)$. Show that, for fixed $x$,

$$\mathbb{P}(Y \neq h(x)) \approx \mathbb{P}(Y \neq h^*(x)) + \Big|2r(x) - 1\Big| \times \left[1 - \Phi\left(\frac{\text{sign}\Big(r(x) - (1/2)\Big)(\overline{r}(x) - (1/2))}{\sigma(x)}\right)\right]$$

where $\Phi$ is the standard Normal CDF and $h^*$ is the Bayes rule. Regard $\text{sign}\Big((r(x) - (1/2))(\overline{r}(x) - (1/2))\Big)$ as a type of bias term. Explain the implications for the bias–variance tradeoff in classification (Friedman (1997)).

Hint: first show that

$$\mathbb{P}(Y \neq h(x)) = |2r(x) - 1|\mathbb{P}(h(x) \neq h^*(x)) + \mathbb{P}(Y \neq h^*(x)).$$

# 23
# Probability Redux: Stochastic Processes

## 23.1 Introduction

Most of this book has focused on IID sequences of random variables. Now we consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on one day is not independent of the temperature on the previous day.

A **stochastic process** $\{X_t : t \in T\}$ is a collection of random variables. We shall sometimes write $X(t)$ instead of $X_t$. The variables $X_t$ take values in some set $\mathcal{X}$ called the **state space**. The set $T$ is called the **index set** and for our purposes can be thought of as time. The index set can be discrete $T = \{0, 1, 2, \ldots\}$ or continuous $T = [0, \infty)$ depending on the application.

**23.1 Example** (IID observations)**.** A sequence of IID random variables can be written as $\{X_t : t \in T\}$ where $T = \{1, 2, 3, \ldots, \}$. Thus, a sequence of IID random variables is an example of a stochastic process. ■

**23.2 Example** (The Weather)**.** Let $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. A typical sequence (depending on where you live) might be

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \text{cloudy}, \cdots$$

This process has a discrete state space and a discrete index set. ■

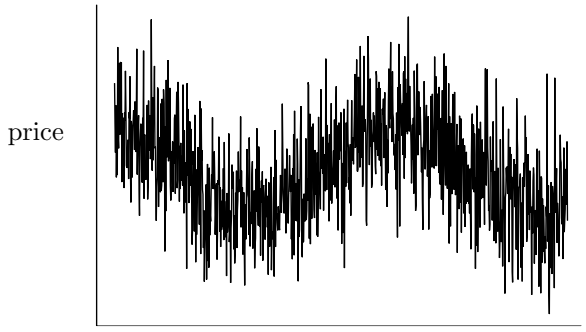



FIGURE 23.1. Stock price over ten week period.

**23.3 Example** (Stock Prices). Figure 23.1 shows the price of a fictitious stock over time. The price is monitored continuously so the index set $T$ is continuous. Price is discrete but for all practical purposes we can treat it as a continuous variable. ■

**23.4 Example** (Empirical Distribution Function). Let $X_1, \ldots, X_n \sim F$ where $F$ is some CDF on [0,1]. Let

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$$

be the empirical CDF. For any fixed value $t$, $\widehat{F}_n(t)$ is a random variable. But the whole empirical CDF

$$\left\{\widehat{F}_n(t) : \ t \in [0,1]\right\}$$

is a stochastic process with a continuous state space and a continuous index set. ■

We end this section by recalling a basic fact. If $X_1, \ldots, X_n$ are random variables, then we can write the joint density as

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} f(x_i|\text{past}_i) \end{aligned} \tag{23.1}$$

where $\text{past}_i = (X_1, \ldots, X_{i-1})$.

## 23.2 Markov Chains

A Markov chain is a stochastic process for which the distribution of $X_t$ depends only on $X_{t-1}$. In this section we assume that the state space is discrete, either $\mathcal{X} = \{1, \ldots, N\}$ or $\mathcal{X} = \{1, 2, \ldots, \}$ and that the index set is $T = \{0, 1, 2, \ldots\}$. Typically, most authors write $X_n$ instead of $X_t$ when discussing Markov chains and I will do so as well.

**23.5 Definition.** *The process* $\{X_n : n \in T\}$ *is a* **Markov chain** *if*

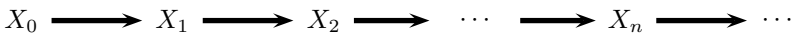
$$\mathbb{P}(X_n = x \mid X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1}) \tag{23.2}$$

*for all* $n$ *and for all* $x \in \mathcal{X}$.

For a Markov chain, equation (23.1) simplifies to

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2) \cdots f(x_n|x_{n-1}).$$

A Markov chain can be represented by the following DAG:




Each variable has a single parent, namely, the previous observation.

The theory of Markov chains is a very rich and complex. We have to get through many definitions before we can do anything interesting. Our goal is to answer the following questions:

1. When does a Markov chain "settle down" into some sort of equilibrium?
2. How do we estimate the parameters of a Markov chain?
3. How can we construct Markov chains that converge to a given equilibrium distribution and why would we want to do that?

We will answer questions 1 and 2 in this chapter. We will answer question 3 in the next chapter. To understand question 1, look at the two chains in Figure 23.2. The first chain oscillates all over the place and will continue to do so forever. The second chain eventually settles into an equilibrium. If we constructed a histogram of the first process, it would keep changing as we got

more and more observations. But a histogram from the second chain would eventually converge to some fixed distribution.

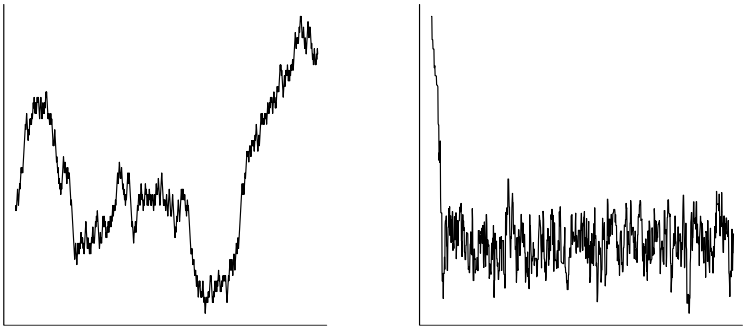



FIGURE 23.2. Two Markov chains. The first chain does not settle down into an equilibrium. The second does.

Transition Probabilities. The key quantities of a Markov chain are the probabilities of jumping from one state into another state. A Markov chain is **homogeneous** if $\mathbb{P}(X_{n+1} = j|X_n = i)$ does not change with time. Thus, for a homogeneous Markov chain, $\mathbb{P}(X_{n+1} = j|X_n = i) = \mathbb{P}(X_1 = j|X_0 = i)$. We shall only deal with homogeneous Markov chains.

**23.6 Definition.** *We call*

$$p_{ij} \equiv \mathbb{P}(X_{n+1} = j|X_n = i) \tag{23.3}$$

*the* **transition probabilities.** *The matrix* $\mathbf{P}$ *whose* $(i, j)$ *element is* $p_{ij}$ *is called the* **transition matrix.**

We will only consider homogeneous chains. Notice that $\mathbf{P}$ has two properties: (i) $p_{ij} \geq 0$ and (ii) $\sum_i p_{ij} = 1$. Each row can be regarded as a probability mass function.

**23.7 Example** (Random Walk With Absorbing Barriers)**.** Let $\mathcal{X} = \{1, \ldots, N\}$. Suppose you are standing at one of these points. Flip a coin with $\mathbb{P}(\text{Heads}) = p$ and $\mathbb{P}(\text{Tails}) = q = 1 - p$. If it is heads, take one step to the right. If it is tails, take one step to the left. If you hit one of the endpoints, stay there. The

transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad \blacksquare$$

**23.8 Example.** Suppose the state space is $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. Then $X_1$, $X_2$, ... represents the weather for a sequence of days. The weather today clearly depends on yesterday's weather. It might also depend on the weather two days ago but as a first approximation we might assume that the dependence is only one day back. In that case the weather is a Markov chain and a typical transition matrix might be

| | Sunny | Cloudy |
|---|---|---|
| Sunny | 0.4 | 0.6 |
| Cloudy | 0.8 | 0.2 |

For example, if it is sunny today, there is a 60 per cent chance it will be cloudy tomorrow. ■

Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i) \tag{23.4}$$

be the probability of of going from state i to state j in $n$ steps. Let $\mathbf{P}_n$ be the matrix whose $(i, j)$ element is $p_{ij}(n)$. These are called the **n-step transition probabilities.**

**23.9 Theorem** (The Chapman-Kolmogorov equations)**.** *The $n$-step probabilities satisfy*

$$p_{ij}(m+n) = \sum_k p_{ik}(m) p_{kj}(n). \tag{23.5}$$

PROOF. Recall that, in general,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y | X = x).$$

This fact is true in the more general form

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z)\mathbb{P}(Y = y | X = x, Z = z).$$

Also, recall the law of total probability:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y).$$

Using these facts and the Markov property we have

$$\begin{aligned} p_{ij}(m+n) &= \mathbb{P}(X_{m+n} = j | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k p_{ik}(m) p_{kj}(n). \quad \blacksquare \end{aligned}$$

Look closely at equation (23.5). This is nothing more than the equation for matrix multiplication. Hence we have shown that

$$\mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n. \tag{23.6}$$

By definition, $\mathbf{P}_1 = \mathbf{P}$. Using the above theorem, $\mathbf{P}_2 = \mathbf{P}_{1+1} = \mathbf{P}_1 \mathbf{P}_1 = \mathbf{P}\mathbf{P} = \mathbf{P}^2$. Continuing this way, we see that

$$\mathbf{P}_n = \mathbf{P}^n \equiv \underbrace{\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}}_{\text{multiply the matrix n times}}. \tag{23.7}$$

Let $\mu_n = (\mu_n(1), \ldots, \mu_n(N))$ be a row vector where

$$\mu_n(i) = \mathbb{P}(X_n = i) \tag{23.8}$$

is the marginal probability that the chain is in state $i$ at time $n$. In particular, $\mu_0$ is called the **initial distribution.** To simulate a Markov chain, all you need to know is $\mu_0$ and $\mathbf{P}$. The simulation would look like this:

**Step 1:** Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = i) = \mu_0(i)$.
**Step 2:** Denote the outcome of step 1 by $i$. Draw $X_1 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_1 = j | X_0 = i) = p_{ij}$.
**Step 3:** Suppose the outcome of step 2 is $j$. Draw $X_2 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_2 = k | X_1 = j) = p_{jk}$.
And so on.

It might be difficult to understand the meaning of $\mu_n$. Imagine simulating the chain many times. Collect all the outcomes at time $n$ from all the chains. This histogram would look approximately like $\mu_n$. A consequence of theorem 23.9 is the following:

**23.10 Lemma.** *The marginal probabilities are given by*

$$\mu_n = \mu_0 \mathbf{P}^n.$$

PROOF.

$$\begin{aligned} \mu_n(j) &= \mathbb{P}(X_n = j) \\ &= \sum_i \mathbb{P}(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_i \mu_0(i) p_{ij}(n) = \mu_0 \mathbf{P}^n. \quad \blacksquare \end{aligned}$$

Summary of Terminology

1. Transition matrix: $\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$.
2. $n$-step matrix: $\mathbf{P}_n(i,j) = \mathbb{P}(X_{n+m} = j | X_m = i)$.
3. $\mathbf{P}_n = \mathbf{P}^n$.
4. Marginal: $\mu_n(i) = \mathbb{P}(X_n = i)$.
5. $\mu_n = \mu_0 \mathbf{P}^n$.

STATES. The states of a Markov chain can be classified according to various properties.

**23.11 Definition.** *We say that $i$* **reaches** *$j$ (or $j$ is* **accessible** *from $i$) if $p_{ij}(n) > 0$ for some $n$, and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that $i$ and $j$* **communicate.**

**23.12 Theorem.** *The communication relation satisfies the following properties:*

1. *$i \leftrightarrow i$.*
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of* **classes** *$\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \cdots$ where two states $i$ and $j$ communicate with each other if and only if they are in the same class.*

If all states communicate with each other, then the chain is called **irreducible**. A set of states is **closed** if, once you enter that set of states you never leave. A closed set consisting of a single state is called an **absorbing state.**

**23.13 Example.** Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}, \{3\}$ and $\{4\}$. State 4 is an absorbing state. ∎

Suppose we start a chain in state $i$. Will the chain ever return to state $i$? If so, that state is called persistent or recurrent.

**23.14 Definition.** *State $i$ is* **recurrent** *or* **persistent** *if*

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1.$$

*Otherwise, state $i$ is* **transient**.

**23.15 Theorem.** *A state $i$ is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty. \tag{23.9}$$

*A state $i$ is transient if and only if*

$$\sum_n p_{ii}(n) < \infty. \tag{23.10}$$

PROOF. Define

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases}$$

The number of times that the chain is in state $i$ is $Y = \sum_{n=0}^{\infty} I_n$. The mean of $Y$, given that the chain starts in state $i$, is

$$\mathbb{E}(Y|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{E}(I_n|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i|X_0 = i) = \sum_{n=0}^{\infty} p_{ii}(n).$$

Define $a_i = \mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i)$. If $i$ is recurrent, $a_i = 1$. Thus, the chain will eventually return to $i$. Once it does return to $i$, we argue again

that since $a_i = 1$, the chain will return to state $i$ again. By repeating this argument, we conclude that $\mathbb{E}(Y|X_0 = i) = \infty$. If $i$ is transient, then $a_i < 1$. When the chain is in state $i$, there is a probability $1 - a_i > 0$ that it will never return to state $i$. Thus, the probability that the chain is in state $i$ exactly $n$ times is $a_i^{n-1}(1 - a_i)$. This is a geometric distribution which has finite mean. ∎

**23.16 Theorem.** *Facts about recurrence.*

1. *If state $i$ is recurrent and $i \leftrightarrow j$, then $j$ is recurrent.*
2. *If state $i$ is transient and $i \leftrightarrow j$, then $j$ is transient.*
3. *A finite Markov chain must have at least one recurrent state.*
4. *The states of a finite, irreducible Markov chain are all recurrent.*

**23.17 Theorem** (Decomposition Theorem)**.** *The state space $\mathcal{X}$ can be written as the disjoint union*

$$\mathcal{X} = \mathcal{X}_T \bigcup \mathcal{X}_1 \bigcup \mathcal{X}_2 \cdots$$

*where $\mathcal{X}_T$ are the transient states and each $\mathcal{X}_i$ is a closed, irreducible set of recurrent states.*

**23.18 Example** (Random Walk)**.** Let $\mathcal{X} = \{\ldots, -2, -1, 0, 1, 2, \ldots, \}$ and suppose that $p_{i,i+1} = p$, $p_{i,i-1} = q = 1 - p$. All states communicate, hence either all the states are recurrent or all are transient. To see which, suppose we start at $X_0 = 0$. Note that

$$p_{00}(2n) = \binom{2n}{n} p^n q^n \tag{23.11}$$

since the only way to get back to 0 is to have $n$ heads (steps to the right) and $n$ tails (steps to the left). We can approximate this expression using Stirling's formula which says that

$$n! \sim n^n \sqrt{n} e^{-n} \sqrt{2\pi}.$$

Inserting this approximation into (23.11) shows that

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{n\pi}}.$$

It is easy to check that $\sum_n p_{00}(n) < \infty$ if and only if $\sum_n p_{00}(2n) < \infty$. Moreover, $\sum_n p_{00}(2n) = \infty$ if and only if $p = q = 1/2$. By Theorem (23.15), the chain is recurrent if $p = 1/2$ otherwise it is transient. ∎

CONVERGENCE OF MARKOV CHAINS. To discuss the convergence of chains, we need a few more definitions. Suppose that $X_0 = i$. Define the **recurrence time**

$$T_{ij} = \min\{n > 0 : \ X_n = j\} \tag{23.12}$$

assuming $X_n$ ever returns to state $i$, otherwise define $T_{ij} = \infty$. The **mean recurrence time** of a recurrent state $i$ is

$$m_i = \mathbb{E}(T_{ii}) = \sum_n n f_{ii}(n) \tag{23.13}$$

where

$$f_{ij}(n) = \mathbb{P}(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i).$$

A recurrent state is **null** if $m_i = \infty$ otherwise it is called **non-null** or **positive.**

**23.19 Lemma.** *If a state is null and recurrent, then* $p_{ii}^n \to 0$.

**23.20 Lemma.** *In a finite state Markov chain, all recurrent states are positive.*

Consider a three-state chain with transition matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Suppose we start the chain in state 1. Then we will be in state 3 at times 3, 6, 9, .... This is an example of a periodic chain. Formally, the **period** of state $i$ is $d$ if $p_{ii}(n) = 0$ whenever $n$ is not divisible by $d$ and $d$ is the largest integer with this property. Thus, $d = \gcd\{n : \ p_{ii}(n) > 0\}$ where gcd means "greater common divisor." State $i$ is **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$. A state with period 1 is called **aperiodic**.

**23.21 Lemma.** *If state $i$ has period $d$ and $i \leftrightarrow j$ then $j$ has period $d$.*

**23.22 Definition.** *A state is* **ergodic** *if it is recurrent, non-null and aperiodic. A chain is ergodic if all its states are ergodic.*

Let $\pi = (\pi_i : \ i \in \mathcal{X})$ be a vector of non-negative numbers that sum to one. Thus $\pi$ can be thought of as a probability mass function.

**23.23 Definition.** *We say that $\pi$ is a* **stationary** *(or* **invariant***) distribution if $\pi = \pi \mathbf{P}$.*

Here is the intuition. Draw $X_0$ from distribution $\pi$ and suppose that $\pi$ is a stationary distribution. Now draw $X_1$ according to the transition probability of the chain. The distribution of $X_1$ is then $\mu_1 = \mu_0\mathbf{P} = \pi\mathbf{P} = \pi$. The distribution of $X_2$ is $\pi\mathbf{P}^2 = (\pi\mathbf{P})\mathbf{P} = \pi\mathbf{P} = \pi$. Continuing this way, we see that the distribution of $X_n$ is $\pi\mathbf{P}^n = \pi$. In other words:

> If at any time the chain has distribution $\pi$, then it will continue to have distribution $\pi$ forever.

**23.24 Definition.** *We say that a chain has* **limiting distribution** *if*

$$\mathbf{P}^n \to \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}$$

*for some $\pi$, that is, $\pi_j = \lim_{n\to\infty} \mathbf{P}^n_{ij}$ exists and is independent of $i$.*

Here is the main theorem about convergence. The theorem says that an ergodic chain converges to its stationary distribution. Also, sample averages converge to their theoretical expectations under the stationary distribution.

**23.25 Theorem.** *An irreducible, ergodic Markov chain has a unique stationary distribution $\pi$. The limiting distribution exists and is equal to $\pi$. If $g$ is any bounded function, then, with probability 1,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \sum_j g(j)\pi_j. \tag{23.14}$$

Finally, there is another definition that will be useful later. We say that $\pi$ satisfies **detailed balance** if

$$\pi_i p_{ij} = p_{ji}\pi_j. \tag{23.15}$$

Detailed balance guarantees that $\pi$ is a stationary distribution.

**23.26 Theorem.** *If $\pi$ satisfies detailed balance, then $\pi$ is a stationary distribution.*

PROOF. We need to show that $\pi\mathbf{P} = \pi$. The $j^{\text{th}}$ element of $\pi\mathbf{P}$ is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$. ∎

The importance of detailed balance will become clear when we discuss Markov chain Monte Carlo methods in Chapter 24.

**Warning!** Just because a chain has a stationary distribution does not mean it converges.

**23.27 Example.** Let

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Let $\pi = (1/3, 1/3, 1/3)$. Then $\pi P = \pi$ so $\pi$ is a stationary distribution. If the chain is started with the distribution $\pi$ it will stay in that distribution. Imagine simulating many chains and checking the marginal distribution at each time $n$. It will always be the uniform distribution $\pi$. But this chain does not have a limit. It continues to cycle around forever. ■

Examples of Markov Chains.

**23.28 Example.** Let $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Then $C_1 = \{1, 2\}$ and $C_2 = \{5, 6\}$ are irreducible closed sets. States 3 and 4 are transient because of the path $3 \to 4 \to 6$ and once you hit state 6 you cannot return to 3 or 4. Since $p_{ii}(1) > 0$, all the states are aperiodic. In summary, 3 and 4 are transient while 1, 2, 5, and 6 are ergodic. ■

**23.29 Example** (Hardy-Weinberg)**.** Here is a famous example from genetics. Suppose a gene can be type $A$ or type $a$. There are three types of people (called genotypes): AA, Aa, and aa. Let $(p, q, r)$ denote the fraction of people of each genotype. We assume that everyone contributes one of their two copies of the gene at random to their children. We also assume that mates are selected at random. The latter is not realistic however, it is often reasonable to assume that you do not choose your mate based on whether they are AA, Aa, or aa. (This would be false if the gene was for eye color and if people chose mates based on eye color.) Imagine if we pooled everyone's genes together. The proportion of $A$ genes is $P = p + (q/2)$ and the proportion of $a$ genes is

$Q = r + (q/2)$. A child is AA with probability $P^2$, aA with probability $2PQ$, and aa with probability $Q^2$. Thus, the fraction of $A$ genes in this generation is

$$P^2 + PQ = \left(p + \frac{q}{2}\right)^2 + \left(p + \frac{q}{2}\right)\left(r + \frac{q}{2}\right).$$

However, $r = 1 - p - q$. Substitute this in the above equation and you get $P^2 + PQ = P$. A similar calculation shows that the fraction of "a" genes is $Q$. We have shown that the proportion of type A and type a is $P$ and $Q$ and this remains stable after the first generation. The proportion of people of type AA, Aa, aa is thus $(P^2, 2PQ, Q^2)$ from the second generation and on. This is called the Hardy-Weinberg law.

Assume everyone has exactly one child. Now consider a fixed person and let $X_n$ be the genotype of their $n^{th}$ descendant. This is a Markov chain with state space $\mathcal{X} = \{AA, Aa, aa\}$. Some basic calculations will show you that the transition matrix is

$$\begin{bmatrix} P & Q & 0 \\ \frac{P}{2} & \frac{P+Q}{2} & \frac{Q}{2} \\ 0 & P & Q \end{bmatrix}.$$

The stationary distribution is $\pi = (P^2, 2PQ, Q^2)$. ■

**23.30 Example** (Markov chain Monte Carlo)**.** In Chapter 24 we will present a simulation method called Markov chain Monte Carlo (MCMC). Here is a brief description of the idea. Let $f(x)$ be a probability density on the real line and suppose that $f(x) = cg(x)$ where $g(x)$ is a known function and $c > 0$ is unknown. In principle, we can compute $c$ since $\int f(x)dx = 1$ implies that $c = 1/\int g(x)dx$. However, it may not be feasible to perform this integral, nor is it necessary to know $c$ in the following algorithm. Let $X_0$ be an arbitrary starting value. Given $X_0, \ldots, X_i$, draw $X_{i+1}$ as follows. First, draw $W \sim N(X_i, b^2)$ where $b > 0$ is some fixed constant. Let

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

Draw $U \sim \text{Uniform}(0, 1)$ and set

$$X_{i+1} = \begin{cases} W & \text{if } U < r \\ X_i & \text{if } U \geq r. \end{cases}$$

We will see in Chapter 24 that, under weak conditions, $X_0, X_1, \ldots,$ is an ergodic Markov chain with stationary distribution $f$. Hence, we can regard the draws as a sample from $f$. ■

INFERENCE FOR MARKOV CHAINS. Consider a chain with finite state space $\mathcal{X} = \{1, 2, \ldots, N\}$. Suppose we observe $n$ observations $X_1, \ldots, X_n$ from this chain. The unknown parameters of a Markov chain are the initial probabilities $\mu_0 = (\mu_0(1), \mu_0(2), \ldots, )$ and the elements of the transition matrix $\mathbf{P}$. Each row of $\mathbf{P}$ is a multinomial distribution. So we are essentially estimating $N$ distributions (plus the initial probabilities). Let $n_{ij}$ be the observed number of transitions from state $i$ to state $j$. The likelihood function is

$$\mathcal{L}(\mu_0, \mathbf{P}) = \mu_0(x_0) \prod_{r=1}^{n} p_{X_{r-1}, X_r} = \mu_0(x_0) \prod_{i=1}^{N} \prod_{j=1}^{N} p_{ij}^{n_{ij}}.$$

There is only one observation on $\mu_0$ so we can't estimate that. Rather, we focus on estimating $\mathbf{P}$. The MLE is obtained by maximizing $\mathcal{L}(\mu_0, \mathbf{P})$ subject to the constraint that the elements are non-negative and the rows sum to 1. The solution is

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i}$$

where $n_i = \sum_{j=1}^{N} n_{ij}$. Here we are assuming that $n_i > 0$. If not, then we set $\widehat{p}_{ij} = 0$ by convention.

**23.31 Theorem** (Consistency and Asymptotic Normality of the MLE)**.** *Assume that the chain is ergodic. Let $\widehat{p}_{ij}(n)$ denote the* MLE *after $n$ observations. Then $\widehat{p}_{ij}(n) \xrightarrow{\text{P}} p_{ij}$. Also,*

$$\left[\sqrt{N_i(n)}(\widehat{p}_{ij} - p_{ij})\right] \rightsquigarrow N(0, \Sigma)$$

*where the left-hand side is a matrix, $N_i(n) = \sum_{r=1}^{n} I(X_r = i)$ and*

$$\Sigma_{ij,k\ell} = \begin{cases} p_{ij}(1 - p_{ij}) & (i,j) = (k,\ell) \\ -p_{ij}p_{i\ell} & i = k, j \neq \ell \\ 0 & \text{otherwise.} \end{cases}$$

## 23.3 Poisson Processes

The Poisson process arises when we count occurrences of events over time, for example, traffic accidents, radioactive decay, arrival of email messages, etc. As the name suggests, the Poisson process is intimately related to the Poisson distribution. Let's first review the Poisson distribution.

Recall that $X$ has a Poisson distribution with parameter $\lambda$ — written $X \sim \text{Poisson}(\lambda)$ — if

$$\mathbb{P}(X = x) \equiv p(x;\ \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Also recall that $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$. If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\nu)$ and $X \amalg Y$, then $X+Y \sim \text{Poisson}(\lambda+\nu)$. Finally, if $N \sim \text{Poisson}(\lambda)$ and $Y|N = n \sim \text{Binomial}(n,p)$, then the marginal distribution of $Y$ is $Y \sim \text{Poisson}(\lambda p)$.

Now we describe the Poisson process. Imagine that you are at your computer. Each time a new email message arrives you record the time. Let $X_t$ be the number of messages you have received up to and including time $t$. Then, $\{X_t : t \in [0,\infty)\}$ is a stochastic process with state space $\mathcal{X} = \{0,1,2,\ldots\}$. A process of this form is called a **counting process**. A Poisson process is a counting process that satisfies certain conditions. In what follows, we will sometimes write $X(t)$ instead of $X_t$. Also, we need the following notation. Write $f(h) = o(h)$ if $f(h)/h \to 0$ as $h \to 0$. This means that $f(h)$ is smaller than $h$ when $h$ is close to 0. For example, $h^2 = o(h)$.

**23.32 Definition.** *A* **Poisson process** *is a stochastic process* $\{X_t : t \in [0,\infty)\}$ *with state space* $\mathcal{X} = \{0,1,2,\ldots\}$ *such that*

1. $X(0) = 0$.
2. *For any* $0 = t_0 < t_1 < t_2 < \cdots < t_n$, *the increments*

$$X(t_1) - X(t_0),\ X(t_2) - X(t_1),\ \cdots, X(t_n) - X(t_{n-1})$$

*are independent.*

3. *There is a function* $\lambda(t)$ *such that*

$$\mathbb{P}(X(t+h) - X(t) = 1) = \lambda(t)h + o(h) \tag{23.16}$$
$$\mathbb{P}(X(t+h) - X(t) \geq 2) = o(h). \tag{23.17}$$

*We call* $\lambda(t)$ *the* **intensity function**.

The last condition means that the probability of an event in $[t, t+h]$ is approximately $h\lambda(t)$ while the probability of more than one event is small.

**23.33 Theorem.** *If* $X_t$ *is a Poisson process with intensity function* $\lambda(t)$, *then*

$$X(s+t) - X(s) \sim \text{Poisson}(m(s+t) - m(s))$$

*where*

$$m(t) = \int_0^t \lambda(s)\, ds.$$

*In particular,* $X(t) \sim \text{Poisson}(m(t))$. *Hence,* $\mathbb{E}(X(t)) = m(t)$ *and* $\mathbb{V}(X(t)) = m(t)$.

**23.34 Definition.** *A Poisson process with intensity function $\lambda(t) \equiv \lambda$ for some $\lambda > 0$ is called a* **homogeneous Poisson process** *with rate $\lambda$. In this case,*

$$X(t) \sim \text{Poisson}(\lambda t).$$

Let $X(t)$ be a homogeneous Poisson process with rate $\lambda$. Let $W_n$ be the time at which the $n^{th}$ event occurs and set $W_0 = 0$. The random variables $W_0, W_1, \ldots,$ are called **waiting times**. Let $S_n = W_{n+1} - W_n$. Then $S_0, S_1, \ldots,$ are called **sojourn times** or **interarrival times**.

**23.35 Theorem.** *The sojourn times $S_0, S_1, \ldots$ are* IID *random variables. Their distribution is exponential with mean $1/\lambda$, that is, they have density*

$$f(s) = \lambda e^{-\lambda s}, \quad s \geq 0.$$

*The waiting time $W_n \sim \text{Gamma}(n, 1/\lambda)$ i.e., it has density*

$$f(w) = \frac{1}{\Gamma(n)} \lambda^n w^{n-1} e^{-\lambda t}.$$

*Hence, $\mathbb{E}(W_n) = n/\lambda$ and $\mathbb{V}(W_n) = n/\lambda^2$.*

PROOF. First, we have

$$\mathbb{P}(S_1 > t) = \mathbb{P}(X(t) = 0) = e^{-\lambda t}$$

with shows that the CDF for $S_1$ is $1 - e^{-\lambda t}$. This shows the result for $S_1$. Now,

$$\begin{aligned} \mathbb{P}(S_2 > t | S_1 = s) &= \mathbb{P}(\text{no events in } (s, s+t] | S_1 = s) \\ &= \mathbb{P}(\text{no events in } (s, s+t]) \ \text{ (increments are independent)} \\ &= e^{-\lambda t}. \end{aligned}$$

Hence, $S_2$ has an exponential distribution and is independent of $S_1$. The result follows by repeating the argument. The result for $W_n$ follows since a sum of exponentials has a Gamma distribution. ■

**23.36 Example.** Figure 23.3 shows requests to a WWW server in Calgary.[1] Assuming that this is a homogeneous Poisson process, $N \equiv X(T) \sim \text{Poisson}(\lambda T)$. The likelihood is

$$\mathcal{L}(\lambda) \propto e^{-\lambda T} (\lambda T)^N$$

[1] See http://ita.ee.lbl.gov/html/contrib/Calgary-HTTP.html for more information.

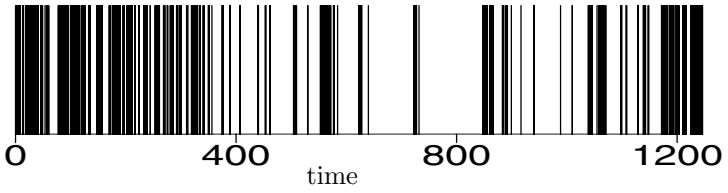



FIGURE 23.3. Hits on a web server. Each vertical line represents one event.

which is maximized at

$$\widehat{\lambda} = \frac{N}{T} = 48.0077$$

in units per minute. Let's now test the assumption that the data follow a homogeneous Poisson process using a goodness-of-fit test. We divide the interval $[0, T]$ into 4 equal length intervals $I_1, I_2, I_3, I_4$. If the process is a homogeneous Poisson process then, given the total number of events, the probability that an event falls into any of these intervals must be equal. Let $p_i$ be the probability of a point being in $I_i$. The null hypothesis is that $p_1 = p_2 = p_3 = p_4 = 1/4$. We can test this hypothesis using either a likelihood ratio test or a $\chi^2$ test. The latter is

$$\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

where $O_i$ is the number of observations in $I_i$ and $E_i = n/4$ is the expected number under the null. This yields $\chi^2 = 252$ with a p-value near 0. This is strong evidence against the null so we reject the hypothesis that the data are from a homogeneous Poisson process. This is hardly surprising since we would expect the intensity to vary as a function of time. ■

## 23.4 Bibliographic Remarks

This is standard material and there are many good references including Grimmett and Stirzaker (1982), Taylor and Karlin (1994), Guttorp (1995), and Ross (2002). The following exercises are from those texts.

## 23.5 Exercises

1. Let $X_0, X_1, \ldots$ be a Markov chain with states $\{0, 1, 2\}$ and transition matrix
$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.9 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.1 \end{bmatrix}$$
Assume that $\mu_0 = (0.3, 0.4, 0.3)$. Find $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 2)$ and $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 1)$.

2. Let $Y_1, Y_2, \ldots$ be a sequence of iid observations such that $\mathbb{P}(Y = 0) = 0.1$, $\mathbb{P}(Y = 1) = 0.3$, $\mathbb{P}(Y = 2) = 0.2$, $\mathbb{P}(Y = 3) = 0.4$. Let $X_0 = 0$ and let
$$X_n = \max\{Y_1, \ldots, Y_n\}.$$
Show that $X_0, X_1, \ldots$ is a Markov chain and find the transition matrix.

3. Consider a two-state Markov chain with states $\mathcal{X} = \{1, 2\}$ and transition matrix
$$\mathbf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$
where $0 < a < 1$ and $0 < b < 1$. Prove that
$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}.$$

4. Consider the chain from question 3 and set $a = .1$ and $b = .3$. Simulate the chain. Let
$$\begin{aligned} \widehat{p}_n(1) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 1) \\ \widehat{p}_n(2) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 2) \end{aligned}$$
be the proportion of times the chain is in state 1 and state 2. Plot $\widehat{p}_n(1)$ and $\widehat{p}_n(2)$ versus $n$ and verify that they converge to the values predicted from the answer in the previous question.

5. An important Markov chain is the **branching process** which is used in biology, genetics, nuclear physics, and many other fields. Suppose that an animal has $Y$ children. Let $p_k = \mathbb{P}(Y = k)$. Hence, $p_k \geq 0$ for all $k$ and $\sum_{k=0}^{\infty} p_k = 1$. Assume each animal has the same lifespan and

that they produce offspring according to the distribution $p_k$. Let $X_n$ be the number of animals in the $n^{th}$ generation. Let $Y_1^{(n)}, \ldots, Y_{X_n}^{(n)}$ be the offspring produced in the $n^{th}$ generation. Note that

$$X_{n+1} = Y_1^{(n)} + \cdots + Y_{X_n}^{(n)}.$$

Let $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathbb{V}(Y)$. Assume throughout this question that $X_0 = 1$. Let $M(n) = \mathbb{E}(X_n)$ and $V(n) = \mathbb{V}(X_n)$.

(a) Show that $M(n+1) = \mu M(n)$ and $V(n+1) = \sigma^2 M(n) + \mu^2 V(n)$.

(b) Show that $M(n) = \mu^n$ and that $V(n) = \sigma^2 \mu^{n-1}(1 + \mu + \cdots + \mu^{n-1})$.

(c) What happens to the variance if $\mu > 1$? What happens to the variance if $\mu = 1$? What happens to the variance if $\mu < 1$?

(d) The population goes extinct if $X_n = 0$ for some $n$. Let us thus define the extinction time $N$ by

$$N = \min\{n : \ X_n = 0\}.$$

Let $F(n) = \mathbb{P}(N \leq n)$ be the CDF of the random variable $N$. Show that

$$F(n) = \sum_{k=0}^{\infty} p_k (F(n-1))^k, \quad n = 1, 2, \ldots$$

Hint: Note that the event $\{N \leq n\}$ is the same as event $\{X_n = 0\}$. Thus, $\mathbb{P}(\{N \leq n\}) = \mathbb{P}(\{X_n = 0\})$. Let $k$ be the number of offspring of the original parent. The population becomes extinct at time $n$ if and only if each of the $k$ sub-populations generated from the $k$ offspring goes extinct in $n-1$ generations.

(e) Suppose that $p_0 = 1/4$, $p_1 = 1/2$, $p_2 = 1/4$. Use the formula from (5d) to compute the CDF $F(n)$.

6. Let

$$\mathbf{P} = \begin{bmatrix} 0.40 & 0.50 & 0.10 \\ 0.05 & 0.70 & 0.25 \\ 0.05 & 0.50 & 0.45 \end{bmatrix}$$

Find the stationary distribution $\pi$.

7. Show that if $i$ is a recurrent state and $i \leftrightarrow j$, then $j$ is a recurrent state.

8. Let
$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$
Which states are transient? Which states are recurrent?

9. Let
$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
Show that $\pi = (1/2, 1/2)$ is a stationary distribution. Does this chain converge? Why/why not?

10. Let $0 < p < 1$ and $q = 1 - p$. Let
$$\mathbf{P} = \begin{bmatrix} q & p & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ q & 0 & 0 & p & 0 \\ q & 0 & 0 & 0 & p \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
Find the limiting distribution of the chain.

11. Let $X(t)$ be an inhomogeneous Poisson process with intensity function $\lambda(t) > 0$. Let $\Lambda(t) = \int_0^t \lambda(u)du$. Define $Y(s) = X(t)$ where $s = \Lambda(t)$. Show that $Y(s)$ is a homogeneous Poisson process with intensity $\lambda = 1$.

12. Let $X(t)$ be a Poisson process with intensity $\lambda$. Find the conditional distribution of $X(t)$ given that $X(t + s) = n$.

13. Let $X(t)$ be a Poisson process with intensity $\lambda$. Find the probability that $X(t)$ is odd, i.e. $\mathbb{P}(X(t) = 1, 3, 5, \ldots)$.

14. Suppose that people logging in to the University computer system is described by a Poisson process $X(t)$ with intensity $\lambda$. Assume that a person stays logged in for some random time with CDF $G$. Assume these times are all independent. Let $Y(t)$ be the number of people on the system at time $t$. Find the distribution of $Y(t)$.

15. Let $X(t)$ be a Poisson process with intensity $\lambda$. Let $W_1, W_2, \ldots,$ be the waiting times. Let $f$ be an arbitrary function. Show that
$$\mathbb{E}\left(\sum_{i=1}^{X(t)} f(W_i)\right) = \lambda \int_0^t f(w)dw.$$

16. A two-dimensional Poisson point process is a process of random points on the plane such that (i) for any set $A$, the number of points falling in $A$ is Poisson with mean $\lambda\mu(A)$ where $\mu(A)$ is the area of $A$, (ii) the number of events in non-overlapping regions is independent. Consider an arbitrary point $x_0$ in the plane. Let $X$ denote the distance from $x_0$ to the nearest random point. Show that

$$\mathbb{P}(X > t) = e^{-\lambda\pi t^2}$$

and

$$\mathbb{E}(X) = \frac{1}{2\sqrt{\lambda}}.$$

# 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

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

If $\theta = (\theta_1, \ldots, \theta_k)$ is multidimensional, then we might be interested in the posterior for one of the components, $\theta_1$, 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 $\theta$ 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 $\delta$ 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 $\delta$ 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 $\delta$ 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 $n = m = 10$,

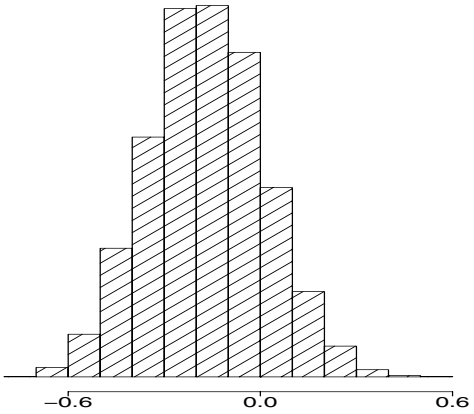



FIGURE 24.1. Posterior of $\delta$ from simulation.

$X = 8$ 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 $Y_i \sim$ 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 $\delta$ 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 $\delta$. The posterior mean of $\delta$ 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 $\delta$ 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

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

$\delta$ 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 $\theta$. 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 $\epsilon_i$ with a thicker tail, for example, a $t$-distribution with $\nu$ 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 $\nu$ 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 $\epsilon_i$ has

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

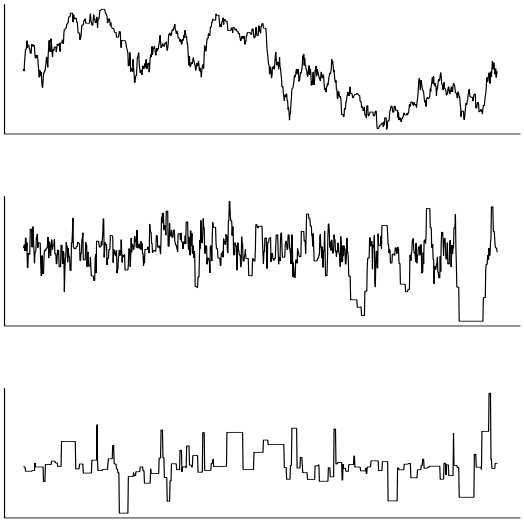


FIGURE 24.2. Three Metropolis chains corresponding to $b = .1$, $b = 1$, $b = 10$.

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 $\pi$ satisfies **detailed balance** for a Markov chain if

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

We showed that if $\pi$ 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 $\pi$ 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 $\epsilon_i$ 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 $\psi$ is more accurate than the Normal approximation for $p$ so we shall work with $\psi$. We shall treat $\sigma_i$ 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 $\mu$. 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 $\psi_i$ 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 $\psi_i$ 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 $\mu$. Figure 24.4 shows the posterior for $\mu$ 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 $\tau$ controls the amount of shrinkage. We set $\tau = 1$ but, in practice, we should treat $\tau$ as another unknown parameter and let the data determine how much shrinkage is needed. ■

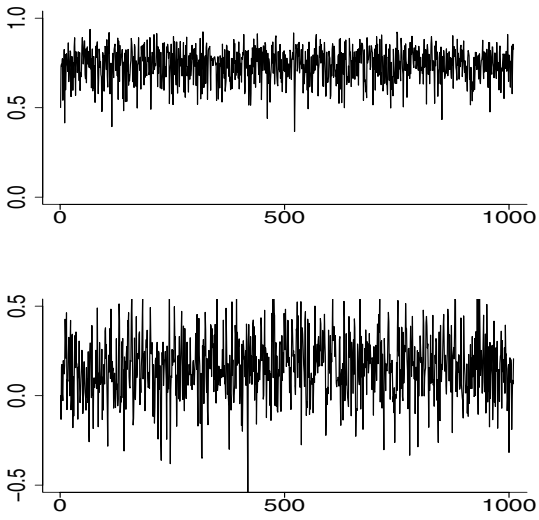



FIGURE 24.3. Posterior simulation for Example 24.11. The top panel shows simulated values of $p_1$. The top panel shows simulated values of $\mu$.

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.

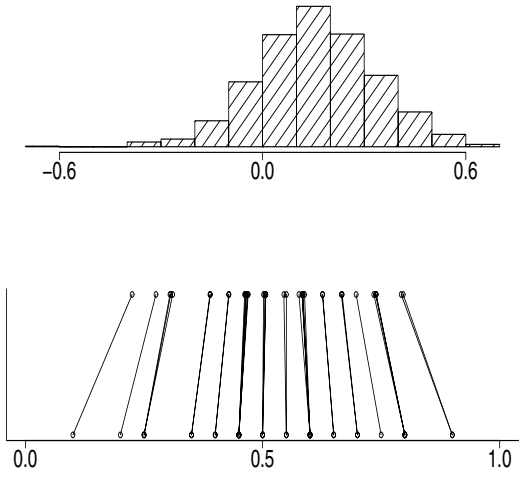



FIGURE 24.4. Example 24.11. Top panel: posterior histogram of $\mu$. Lower panel: raw proportions and the Bayes posterior estimates. The Bayes estimates have been shrunk closer together than the raw proportions.

## 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 $\beta_j$'s. Get the posterior mean and a 95 percent posterior interval for each $\beta_j$.

(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 $\int_0^\infty y^{\alpha-1} e^{-y} dy$ |
| $\Omega$ | sample space (set of outcomes) |
| $\omega$ | outcome, element, point |
| $A$ | event (subset of $\Omega$) |
| $I_A(\omega)$ | 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 |
| $A \wasylozenge\!\!\!\!\!\!\sim B$ | $A$ 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 $F$ |
| $\phi$ | standard Normal probability density |
| $\Phi$ | standard Normal distribution function |
| $z_\alpha$ | upper $\alpha$ 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 |
| $\theta$ | 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($p$) | $p(1-p)^{x-1}I(x\geq 1)$ | $1/p$ | $\frac{1-p}{p^2}$ | $\frac{pe^t}{1-(1-p)e^t}\ \big(t<-\log(1-p)\big)$ |
| Poisson($\lambda$) | $\frac{\lambda^x e^{-\lambda}}{x!}$ | $\lambda$ | $\lambda$ | $e^{\lambda(e^t-1)}$ |
| 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($\mu,\sigma^2$) | $\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ | $\mu$ | $\sigma^2$ | $\exp\left\{\mu t+\frac{\sigma^2t^2}{2}\right\}$ |
| Exponential($\beta$) | $\frac{e^{-x/\beta}}{\beta}$ | $\beta$ | $\beta^2$ | $\frac{1}{1-\beta t}\ \big(t<1/\beta\big)$ |
| Gamma($\alpha,\beta$) | $\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$ | $\alpha\beta$ | $\alpha\beta^2$ | $\left(\frac{1}{1-\beta t}\right)^\alpha\ (t<1/\beta)$ |
| Beta($\alpha,\beta$) | $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ | $1+\sum_{k=1}^\infty\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$ |
| $t_\nu$ | $\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$ | $0$ (if $\nu>1$) | $\frac{\nu}{\nu-2}$ (if $\nu>2$) | does not exist |
| $\chi^2_p$ | $\frac{1}{\Gamma(p/2)2^{p/2}}x^{(p/2)-1}e^{-x/2}$ | $p$ | $2p$ | $\left(\frac{1}{1-2t}\right)^{p/2}\ \big(t<1/2\big)$ |

# Index