Statistics and Computing

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Statistics and Computing

Gentle: Numerical Linear Algebra for Applications in Statistics.
Gentle: Random Number Generation and Monte Carlo Methods.
Härdle/Klinke/Turlach: XploRe: An Interactive Statistical Computing Environment.
Krause/Olson: The Basics of S and S-PLUS, 2nd Edition.
Lange: Numerical Analysis for Statisticians.
Loader: Local Regression and Likelihood.
Ó Ruanaidh/Fitzgerald: Numerical Bayesian Methods Applied to Signal Processing.
Pannatier: VARIOWIN: Software for Spatial Data Analysis in 2D.
Pinheiro/Bates: Mixed-Effects Models in S and S-PLUS.
Venables/Ripley: Modern Applied Statistics with S-PLUS, 3rd Edition.
Venables/Ripley: S Programming.
Wilkinson: The Grammar of Graphics.

Kenneth Lange

Numerical Analysis for Statisticians

Springer

Kenneth Lange
Departments of Biomathematics and Human Genetics
UCLA School of Medicine
Los Angeles, CA 90095
USA
klange@ucla.edu

With 9 figures.

Library of Congress Cataloging-in-Publication Data
Lange, Kenneth.
Numerical analysis for statisticians / Kenneth Lange.
p. cm. — (Statistics and computing)
Includes bibliographical references and index.
ISBN 0-387-94979-8 (hardcover: alk. paper)
1. Numerical analysis. 2. Mathematical statistics. I. Title.
II. Series.
QA297.L34 1998
519.4—dc21 98-16688

ISBN 0-387-94979-8 SPIN 10782832

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface

This book, like many books, was born in frustration. When in the fall of 1994 I set out to teach a second course in computational statistics to doctoral students at the University of Michigan, none of the existing texts seemed exactly right. On the one hand, the many decent, even inspiring, books on elementary computational statistics stress the nuts and bolts of using packaged programs and emphasize model interpretation more than numerical analysis. On the other hand, the many theoretical texts in numerical analysis almost entirely neglect the issues of most importance to statisticians. The closest book to my ideal was the classical text of Kennedy and Gentle [2]. More than a decade and a half after its publication, this book still has many valuable lessons to teach statisticians. However, upon reflecting on the rapid evolution of computational statistics, I decided that the time was ripe for an update.

The book you see before you represents a biased selection of those topics in theoretical numerical analysis most relevant to statistics. By intent this book is not a compendium of tried and trusted algorithms, is not a consumer's guide to existing statistical software, and is not an exposition of computer graphics or exploratory data analysis. My focus on principles of numerical analysis is intended to equip students to craft their own software and to understand the advantages and disadvantages of different numerical methods. Issues of numerical stability, accurate approximation, computational complexity, and mathematical modeling share the limelight and take precedence over philosophical questions of statistical inference. Accordingly, you must look elsewhere for a discussion of the merits of frequentist versus Bayesian inference. My attitude is that good data deserve inspec-

tion from a variety of perspectives. More often than not, these different perspectives reinforce and clarify rather than contradict one another.

Having declared a truce on issues of inference, let me add that I have little patience with the view that mathematics is irrelevant to statistics. While it is demeaning to statistics to view it simply as a branch of mathematics, it is also ridiculous to contend that statistics can prosper without the continued influx of new mathematical ideas. Nowhere is this more evident than in computational statistics. Statisticians need to realize that the tensions existing between statistics and mathematics mirror the tensions between other disciplines and mathematics. If physicists and economists can learn to live with mathematics, then so can statisticians. Theoreticians in any science will be attracted to mathematics and practitioners repelled. In the end, it really is just a matter of choosing the relevant parts of mathematics and ignoring the rest. Of course, the hard part is deciding what is irrelevant.

Each of the chapters of this book weaves a little mathematical tale with a statistical moral. My hope is to acquaint students with the main principles behind a numerical method without overwhelming them with detail. On first reading, this assertion may seem debatable, but you only have to delve a little more deeply to learn that many chapters have blossomed into full books written by better informed authors. In the process of writing, I have had to educate myself about many topics. I am sure my ignorance shows, and to the experts I apologize. If there is anything fresh here, it is because my own struggles have made me more sensitive to the struggles of my classroom students. Students deserve to have logical answers to logical questions. I do not believe in pulling formulas out of thin air and expecting students to be impressed. Of course, this attitude reflects my mathematical bent and my willingness to slow the statistical discussion to attend to the mathematics.

The mathematics in this book is a mix of old and new. One of the charms of applying mathematics is that there is little guilt attached to resurrecting venerable subjects such as continued fractions. If you feel that I pay too much attention to these museum pieces, just move on to the next chapter. Note that although there is a logical progression tying certain chapters together—for instance, the chapters on optimization theory and the chapters on numerical integration—many chapters can be read as independent essays. At the opposite extreme of continued fractions, several chapters highlight recent statistical developments such as wavelets, the bootstrap, and Markov chain Monte Carlo methods. These modern topics were unthinkable to previous generations unacquainted with today's computers.

Any instructor contemplating a one-semester course based on this book will have to decide which chapters to cover and which to omit. It is difficult for me to provide sound advice because the task of writing is still so fresh in my mind. In reading the prepublication reviews of my second draft, I

was struck by the reviewers' emphasis on the contents of Chapters 5, 7, 10, 11, 21, and 24. Instructors may want to cover material from Chapters 20 and 23 as a prelude to Chapters 21 and 24. Another option is to devote the entire semester to a single topic such as optimization theory. Finally, given the growing importance of computational statistics, a good case can be made for a two-semester course. This book contains adequate material for a rapidly paced yearlong course.

As with any textbook, the problems are nearly as important as the main text. Most problems merely serve to strengthen intellectual muscles strained by the introduction of new theory; some problems extend the theory in significant ways. The majority of any theoretical and typographical errors are apt to be found in the problems. I will be profoundly grateful to readers who draw to my attention errors anywhere in the book, no matter how small.

I have several people to thank for their generous help. Robert Jennrich taught me the rudiments of computational statistics many years ago. His influence pervades the book. Let me also thank the students in my graduate course at Michigan for enduring a mistake-ridden first draft. Ruzong Fan, in particular, checked and corrected many of the exercises. Michael Newton of the University of Wisconsin and Yingnian Wu of the University of Michigan taught from a corrected second draft. Their comments have been helpful in further revision. Robert Strawderman kindly brought to my attention Example 18.4.2, shared his notes on the bootstrap, and critically read Chapter 22. David Hunter prepared the index, drew several figures, and contributed substantially to the content of Chapter 20. Last of all, I thank John Kimmel of Springer for his patient encouragement and editorial advice.

This book is dedicated to the memory of my brother Charles. His close friend and colleague at UCLA, Nick Grossman, dedicated his recent book on celestial mechanics to Charles with the following farewell comments:

> His own work was notable for its devotion to real problems arising from the real world, for the beauty of the mathematics he invoked, and for the elegance of its exposition. Chuck died in summer, 1993, at the age of 51, leaving much undone. Many times since his death I have missed his counsel, and I know that this text would be far less imperfect if I could have asked him about a host of questions that vexed me. Reader, I hope that you have such a friend. [1]

It is impossible for me to express my own regrets more poetically.

References

[1] Grossman N (1996) *The Sheer Joy of Celestial Mechanics.* Birkhäuser, Boston

[2] Kennedy WJ Jr, Gentle JE (1980) *Statistical Computing*. Marcel Dekker, New York

Los Angeles, California Kenneth Lange

Contents

1
Recurrence Relations

1.1 Introduction

Recurrence relations are ubiquitous in computational statistics and probability. Devising good recurrence relations is both an art and a science. One general theme is the alpha and omega principle; namely, most recurrences are derived by considering either the first or last event in a chain of events. The following examples illustrate this principle and some other commonly employed techniques.

1.2 Binomial Coefficients

Let $\binom{n}{k}$ be the number of subsets of size k from a set of size n. Pascal's triangle is the recurrence scheme specified by

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \tag{1}$$

together with the boundary conditions $\binom{n}{0} = \binom{n}{n} = 1$. To derive (1) we take a set of size $n+1$ and divide it into a set of size n and a set of size 1. We can either choose $k-1$ elements from the n-set and combine them with the single element from the 1-set or choose all k elements from the n-set. The first choice can be made in $\binom{n}{k-1}$ ways and the second in $\binom{n}{k}$ ways.

As indicated by its name, we visualize Pascal's triangle as an infinite lower triangular matrix with n as row index and k as column index. The

boundary values specify the first column and the diagonal as the constant 1. The recurrence proceeds row by row. If one desires only the binomial coefficients for a single final row, it is advantageous in coding Pascal's triangle to proceed from right to left along the current row. This minimizes computer storage by making it possible to overwrite safely the contents of the previous row with the contents of the current row. Pascal's triangle also avoids the danger of computer overflows caused by computing binomial coefficients via factorials.

1.3 Number of Partitions of a Set

Let B_n be the number of partitions of a set with n elements. By a partition we mean a division of the set into disjoint blocks. A partition induces an equivalence relation on the set in the sense that two elements are equivalent if and only if they belong to the same block. Two partitions are the same if and only if they induce the same equivalence relation.

Starting with $B_0 = 1$, the B_n satisfy the recurrence relation

$$\begin{aligned} B_{n+1} &= \sum_{k=0}^{n} \binom{n}{k} B_{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} B_k. \end{aligned} \tag{2}$$

The reasoning leading to (2) is basically the same as in our last example. We divide our set with $n+1$ elements into an n-set and a 1-set. The 1-set can form a block by itself, and the n-set can be partitioned in B_n ways. Or we can choose $k \geq 1$ elements from the n-set in $\binom{n}{k}$ ways and form a block consisting of these elements and the single element from the 1-set. The remaining $n-k$ elements of the n-set can be partitioned in B_{n-k} ways.

1.4 Horner's Method

Suppose we desire to evaluate the polynomial

$$p(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n$$

for a particular value of x. If we proceed naively, then it takes $n-1$ multiplications to form the powers $x^k = x \cdot x^{k-1}$ for $2 \leq k \leq n$, n multiplications to multiply each power x^k by its coefficient a_{n-k}, and n additions to sum the resulting terms. This amounts to $3n-1$ operations in all. Horner's method exploits the fact that $p(x)$ can be expressed as

$$\begin{aligned} p(x) &= x(a_0 x^{n-1} + a_1 x^{n-2} + \cdots + a_{n-1}) + a_n \\ &= x b_{n-1}(x) + a_n. \end{aligned}$$

Since the polynomial $b_{n-1}(x)$ of degree $n-1$ can be similarly reduced, a complete recursive scheme for evaluating $p(x)$ is given by

$$\begin{aligned} b_0(x) &= a_0 \\ b_k(x) &= xb_{k-1}(x) + a_k, \quad k = 1, \ldots, n. \end{aligned} \tag{3}$$

This scheme, known as Horner's method, requires only n multiplications and n additions to compute $p(x) = b_n(x)$.

Interestingly enough, Horner's method can be modified to produce the derivative $p'(x)$ as well as $p(x)$. This modification is useful, for instance, in searching for a root of $p(x)$ by Newton's method. To discover the algorithm for evaluating $p'(x)$, we differentiate (3). This gives the amended Horner scheme

$$\begin{aligned} b_1'(x) &= b_0(x) \\ b_k'(x) &= xb_{k-1}'(x) + b_{k-1}(x), \quad k = 2, \ldots, n, \end{aligned}$$

requiring an additional $n-1$ multiplications and $n-1$ additions to compute $p'(x) = b_n'(x)$.

1.5 Sample Means and Variances

Consider a sequence $x_1, \ldots, x_n$ of n real numbers. After you have computed the sample mean and variance

$$\begin{aligned} \mu_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_n)^2, \end{aligned}$$

suppose you are presented with a new observation x_{n+1}. It is possible to adjust the sample mean and variance without revisiting all of the previous observations. For example, it is obvious that

$$\mu_{n+1} = \frac{1}{n+1}(n\mu_n + x_{n+1}).$$

Because

$$\begin{aligned} (n+1)\sigma_{n+1}^2 &= \sum_{i=1}^{n+1}(x_i - \mu_{n+1})^2 \\ &= \sum_{i=1}^{n}(x_i - \mu_{n+1})^2 + (x_{n+1} - \mu_{n+1})^2 \\ &= \sum_{i=1}^{n}(x_i - \mu_n)^2 + n(\mu_{n+1} - \mu_n)^2 + (x_{n+1} - \mu_{n+1})^2 \end{aligned}$$

and

$$\begin{aligned} n(\mu_{n+1} - \mu_n)^2 &= n(\mu_{n+1} - \frac{n+1}{n}\mu_{n+1} + \frac{1}{n}x_{n+1})^2 \\ &= \frac{1}{n}(x_{n+1} - \mu_{n+1})^2, \end{aligned}$$

it follows that

$$\sigma_{n+1}^2 = \frac{n}{n+1}\sigma_n^2 + \frac{1}{n}(x_{n+1} - \mu_{n+1})^2.$$

1.6 Expected Family Size

A married couple desires a family consisting of at least s sons and d daughters. At each birth the mother independently bears a son with probability p and a daughter with probability $q = 1 - p$. They will quit having children when their objective is reached. Let N_{sd} be the random number of children born to them. Suppose we wish to compute the expected value $\mathrm{E}(N_{sd})$. Two cases are trivial. If either $s = 0$ or $d = 0$, then N_{sd} follows a negative binomial distribution. It follows that $\mathrm{E}(N_{0d}) = d/q$ and $\mathrm{E}(N_{s0}) = s/p$. When both s and d are positive, the distribution of N_{sd} is not so obvious. However, in this case we can condition on the outcome of the first birth and compute

$$\begin{aligned} \mathrm{E}(N_{sd}) &= p[1 + \mathrm{E}(N_{s-1,d})] + q[1 + \mathrm{E}(N_{s,d-1})] \\ &= 1 + p\,\mathrm{E}(N_{s-1,d}) + q\,\mathrm{E}(N_{s,d-1}). \end{aligned}$$

There are many variations on this idea. For instance, suppose we wish to compute the probability R_{sd} that the couple reaches its quota of s sons before its quota of d daughters. Then the R_{sd} satisfy the boundary conditions $R_{0d} = 1$ for $d > 0$ and $R_{s0} = 0$ for $s > 0$. When s and d are both positive, we have the recurrence relation

$$R_{sd} = pR_{s-1,d} + qR_{s,d-1}.$$

1.7 Poisson-Binomial Distribution

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with a possibly different success probability p_k for each X_k. The sum $S_n = \sum_{k=1}^n X_k$ is said to have a Poisson-binomial distribution. If all $p_k = p$, then S_n has a binomial distribution with n trials and success probability p. If each p_i is small, but the sum $\mu = \sum_{k=1}^n p_k$ is moderate in size, then S_n is approximately Poisson with mean μ. In many applications it is unnecessary to invoke this approximation because the exact distribution $p_n(i) = \Pr(S_n = i)$ can be calculated recursively by incrementing the number of summands n. Note

first that $p_1(0) = 1 - p_1$ and $p_1(1) = p_1$. With these initial values, one can proceed inductively via

$$\begin{aligned} p_j(0) &= (1-p_j)p_{j-1}(0) \\ p_j(i) &= p_j p_{j-1}(i-1) + (1-p_j)p_{j-1}(i), \quad 1 \le i \le j-1 \\ p_j(j) &= p_j p_{j-1}(j-1) \end{aligned} \tag{4}$$

until reaching $j = n$. For the binomial distribution, this method can be superior to calculating each term directly by the standard formula

$$p_n(i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

Just as with Pascal's triangle, it is preferable to proceed from right to left along a row. Furthermore, if the values $p_n(i)$ are only needed for a limited range $0 \le i \le k$, then the recurrence (4) can be carried out with this proviso.

1.8 A Multinomial Test Statistic

For relatively sparse multinomial data with known but unequal probabilities per category, it is useful to have alternatives to the classical chi-square test. For instance, the number of categories W_d with d or more observations can be a sensitive indicator of clustering. This statistic has mean $\lambda = \sum_{i=1}^m \mu_i$, where

$$\mu_i = \sum_{k=d}^n \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

is the probability that the count N_i of category i satisfies $N_i \ge d$. Here we assume n trials, m categories, and a probability p_i attached to category i. If the variance of W_d is close to λ, then W_d follows an approximate Poisson distribution with mean λ [1, 3].

As a supplement to this approximation, it is possible to compute the distribution function $\Pr(W_d \le j)$ recursively by adapting a technique of Sandell [5]. Once this is done, the p-value of an experimental result w_d can be recovered via $\Pr(W_d \ge w_d) = 1 - \Pr(W_d \le w_d - 1)$. The recursive scheme can be organized by defining $t_{j,k,l}$ to be the probability that $W_d \le j$ given k trials and l categories. The indices j, k, and l are confined to the ranges $0 \le j \le w_d - 1$, $0 \le k \le n$, and $1 \le l \le m$. The l categories implicit in $t_{j,k,l}$ refer to the first l of the overall m categories; the ith of these l categories is assigned the conditional probability $p_i/(p_1 + \cdots + p_l)$.

With these definitions in mind, note first the obvious initial values (a) $t_{0,k,1} = 1$ for $k < d$, (b) $t_{0,k,1} = 0$ for $k \ge d$, and (c) $t_{j,k,1} = 1$ for $j > 0$. Now beginning with $l = 1$, compute $t_{j,k,l}$ recursively by conditioning on how many observations fall in category l. Since at most $d - 1$ observations

can fall in category l without increasing W_d by 1, the recurrence relation for $j = 0$ is

$$t_{0,k,l} = \sum_{i=0}^{\min\{d-1,k\}} \binom{k}{i} \left(\frac{p_l}{p_1 + \cdots + p_l}\right)^i \left(1 - \frac{p_l}{p_1 + \cdots + p_l}\right)^{k-i} t_{0,k-i,l-1},$$

and the recurrence relation for $j > 0$ is

$$\begin{aligned} t_{j,k,l} = &\sum_{i=0}^{\min\{d-1,k\}} \binom{k}{i} \left(\frac{p_l}{p_1 + \cdots + p_l}\right)^i \left(1 - \frac{p_l}{p_1 + \cdots + p_l}\right)^{k-i} t_{j,k-i,l-1} \\ &+ \sum_{i=d}^{k} \binom{k}{i} \left(\frac{p_l}{p_1 + \cdots + p_l}\right)^i \left(1 - \frac{p_l}{p_1 + \cdots + p_l}\right)^{k-i} t_{j-1,k-i,l-1}. \end{aligned}$$

These recurrence relations jointly permit replacing the matrix $(t_{j,k,l-1})$ by the matrix $(t_{j,k,l})$. At the end of this recursive scheme on $l = 2, \ldots, m$, we extract the desired probability $t_{w_d-1,n,m}$.

The binomial probabilities occurring in these formulas can be computed by our previous algorithm for the Poisson-binomial distribution. It is noteworthy that the Poisson-binomial recurrence increments the number of trials whereas the recurrence for the distribution function of W_d increments the number of categories in the multinomial distribution.

1.9 An Unstable Recurrence

Not all recurrence relations are numerically stable. Henrici [2] dramatically illustrates this point using the integrals

$$y_n = \int_0^1 \frac{x^n}{x + a} dx. \tag{5}$$

The recurrence $y_n = 1/n - ay_{n-1}$ follows directly from the identity

$$\begin{aligned} \int_0^1 \frac{x^{n-1}(x + a - a)}{x + a} dx &= \int_0^1 x^{n-1} dx - a \int_0^1 \frac{x^{n-1}}{x + a} dx \\ &= \frac{1}{n} - a \int_0^1 \frac{x^{n-1}}{x + a} dx. \end{aligned}$$

In theory this recurrence furnishes a convenient method for calculating the y_n starting with the initial value $y_0 = \ln \frac{1+a}{a}$. Table 1.1 records the results of our computations in single precision when $a = 10$. It is clear that something has gone amiss. Computing in double precision only delays the onset of the instability.

We can diagnose the source of the problem by noting that for n moderately large most of the mass of the integral occurs near $x = 1$. Thus, to a

good approximation

$$\begin{aligned} y_{n-1} &\approx \frac{1}{1+a}\int_0^1 x^{n-1}dx \\ &= \frac{1}{(1+a)n}. \end{aligned}$$

When a is large, the fraction $a/(1+a)$ in the difference

$$\begin{aligned} y_n &\approx \frac{1}{n} - a\frac{1}{(1+a)n} \\ &= \frac{1}{n}\left(1 - \frac{a}{1+a}\right) \end{aligned}$$

is close to 1. We lose precision whenever we subtract two numbers of the same sign and comparable magnitude. The moral here is that we must exercise caution in using recurrence relations involving subtraction. Fortunately, many recurrences in probability theory arise by conditioning arguments and consequently entail only addition and multiplication of nonnegative numbers.

1.10 Quick Sort

Statisticians sort lists of numbers to compute sample quantiles and plot empirical distribution functions. It is a pleasant fact that the fastest sorting algorithm can be explained by a probabilistic argument [6]. At the heart of this argument is a recurrence relation specifying the average number of operations encountered in sorting n numbers. In this problem, we can explicitly solve the recurrence relation and estimate the rate of growth of its solution as a function of n. The recurrence relation is not so much an end in itself as a means to understanding the behavior of the sorting algorithm.

The quick sort algorithm is based on the idea of finding a splitting entry x_i of a sequence $x_1, \ldots, x_n$ of n distinct numbers in the sense that $x_j < x_i$ for $j < i$ and $x_j > x_i$ for $j > i$. In other words, a splitter x_i is already correctly ordered relative to the rest of the entries of the sequence. Finding

TABLE 1.1. Computed Values of the Integral y_n

n	y_n	n	y_n
0	0.095310	5	0.012960
1	0.046898	6	0.037064
2	0.031020	7	−0.227781
3	0.023130	8	2.402806
4	0.018704	9	−23.916945

a splitter reduces the computational complexity of sorting because it is easier to sort both of the subsequences $x_1, \ldots, x_{i-1}$ and $x_{i+1}, \ldots, x_n$ than it is to sort the original sequence. At this juncture, one can reasonably object that no splitter need exist, and even if one does, it may be difficult to locate. The quick sort algorithm avoids these difficulties by randomly selecting a splitting value and then slightly rearranging the sequence so that this splitting value occupies the correct splitting location.

In the background of quick sort is the probabilistic assumption that all $n!$ permutations of the n values are equally likely. The algorithm begins by randomly selecting one of the n values and moving it to the leftmost or first position of the sequence. Through a sequence of exchanges, this value is then promoted to its correct location. In the probabilistic setting adopted, the correct location of the splitter is uniformly distributed over the n positions of the sequence.

The promotion process works by exchanging or swapping entries to the right of the randomly chosen splitter x_1, which is kept in position 1 until a final swap. Let j be the current position of the sequence as we examine it from left to right. In the sequence up to position j, a candidate position i for the insertion of x_1 must satisfy the conditions $x_k < x_1$ for $1 < k \leq i$ and $x_k > x_1$ for $i < k \leq j$. Clearly, the choice $i = j$ works when $j = 1$ because then the set $\{k : 1 < k \leq i \text{ or } i < k \leq j\}$ is empty. Now suppose we examine position $j + 1$. If $x_{j+1} > x_1$, then we keep the current candidate position i. If $x_{j+1} < x_1$, then we swap x_{i+1} and x_{j+1} and replace i by $i+1$. In either case, the two required conditions imposed on i continue to obtain. Thus, we can inductively march from the left end to the right end of the sequence, carrying out a few swaps in the process, so that when $j = n$, the value i marks the correct position to insert x_1. Once this insertion is made, the subsequences $x_1, \ldots, x_{i-1}$ and $x_{i+1}, \ldots, x_n$ can be sorted separately by the same splitting procedure.

Now let e_n be the expected number of operations involved in quick sorting a sequence of n numbers. By convention $e_0 = 0$. If we base our analysis only on how many positions j must be examined at each stage and not on how many swaps are involved, then we can write the recurrence relation

$$\begin{aligned} e_n &= n - 1 + \frac{1}{n} \sum_{i=1}^{n} (e_{i-1} + e_{n-i}) \\ &= n - 1 + \frac{2}{n} \sum_{i=1}^{n} e_{i-1} \end{aligned} \tag{6}$$

by conditioning on the correct position i of the first splitter.

The recurrence relation (6) looks formidable, but a few algebraic maneuvers render it solvable. Multiplying equation (6) by n produces

$$ne_n = n(n-1) + 2 \sum_{i=1}^{n} e_{i-1}.$$

If we subtract from this the corresponding expression for $(n-1)e_{n-1}$, then we get

$$ne_n - (n-1)e_{n-1} = 2n - 2 + 2e_{n-1},$$

which can be rearranged to give

$$\frac{e_n}{n+1} = \frac{2(n-1)}{n(n+1)} + \frac{e_{n-1}}{n}. \tag{7}$$

Equation (7) can be iterated to yield

$$\begin{aligned}\frac{e_n}{n+1} &= 2\sum_{k=1}^{n} \frac{(k-1)}{k(k+1)} \\ &= 2\sum_{k=1}^{n} \left(\frac{2}{k+1} - \frac{1}{k}\right) \\ &= 2\sum_{k=1}^{n} \frac{1}{k} - \frac{4n}{n+1}.\end{aligned}$$

Because $\sum_{k=1}^{n} \frac{1}{k}$ approximates $\int_1^n \frac{1}{x}dx = \ln n$, it follows that

$$\lim_{n\to\infty} \frac{e_n}{2n\ln n} = 1.$$

Quick sort is indeed a very efficient algorithm on average. Press et al. [4] provide good computer code implementing it.

1.11 Problems

1. Let f_n be the number of subsets of $\{1, \ldots, n\}$ that do not contain two consecutive integers. Show that $f_1 = 2$, $f_2 = 3$, and $f_n = f_{n-1} + f_{n-2}$ for $n > 2$.

2. Suppose n, j, and $r_1, \ldots, r_j$ are positive integers with $n = r_1 + \cdots + r_j$ and with $r_1 \geq r_2 \geq \cdots \geq r_j \geq 1$. Such a decomposition is called a partition of n with largest part r_1. For example, $6 = 4 + 1 + 1$ is a partition of 6 into three parts with largest part 4. Let q_{nk} be the number of partitions of n with largest part k. Show that

$$q_{nk} = q_{n-1,k-1} + q_{n-k,k}.$$

3. In Horner's method suppose x_0 is a root of $p(x)$. Show that the numbers $b_k(x_0)$ produced by (3) yield the deflated polynomial

$$b_0(x_0)x^{n-1} + b_1(x_0)x^{n-2} + \cdots + b_{n-1}(x_0) = \frac{p(x)}{x - x_0}.$$

4. Give a recursive method for computing the second moments $\mathrm{E}(N_{sd}^2)$ in the family-planning model.

5. In the family-planning model, suppose the couple has an upper limit m on the number of children they can afford. Hence, they stop whenever they reach their goal of s sons and d daughters or m total children, whichever comes first. Let N_{sdm} now be their random number of children. Give a recursive method for computing $\mathrm{E}(N_{sdm})$.

6. In the family-planning model, suppose the husband and wife are both carriers of a recessive genetic disease. On average one quarter of their children will be afflicted. If the parents want at least s normal sons and at least d normal daughters, let T_{sd} be their random number of children. Give a recursive method for computing $\mathrm{E}(T_{sd})$.

7. Consider the multinomial model with m categories, n trials, and probability p_i attached to the ith category. Express the distribution function of the maximum number of counts $\max_i N_i$ observed in any category in terms of the distribution functions of the W_d. How can the algorithm for computing the distribution function of W_d be simplified to give an algorithm for computing a p-value of $\max_i N_i$?

8. Define the statistic U_d to be the number of categories i with $N_i < d$. Express the right-tail probability $\Pr(U_d \geq j)$ in terms of the distribution function of W_d. This gives a method for computing p-values of the statistic U_d. In some circumstances U_d has an approximate Poisson distribution. What do you conjecture about these circumstances?

9. Demonstrate that the integral y_n defined by equation (5) can be expanded in the infinite series

$$y_n = \sum_{k=0}^{\infty} \frac{(-1)^k}{(n+k+1)a^{k+1}}$$

when $a > 1$. This does provide a reasonably stable method of computing y_n for large a.

10. Show that the worst case of quick sort takes on the order of n^2 operations.

11. Let p be the probability that a randomly chosen permutation of n distinct numbers contains at least one pre-existing splitter. Show by an inclusion-exclusion argument that

$$p = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i!} \approx 1 - e^{-1}.$$

12. Continuing Problem 11, demonstrate that both the mean and variance of the number of pre-existing splitters equal 1.

References

[1] Barbour AD, Holst L, Janson S (1992) *Poisson Approximation.* Oxford University Press, Oxford

[2] Henrici P (1982) *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*. Wiley, New York
[3] Kolchin VF, Sevast'yanov BA, Chistyakov VP (1978) *Random Allocations*. Winston, Washington DC
[4] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[5] Sandell D (1991) Computing probabilities in a generalized birthday problem. *Math Scientist* 16:78–82
[6] Wilf HS (1986) *Algorithms and Complexity*. Prentice-Hall, New York

2
Power Series Expansions

2.1 Introduction

Power series expansions are old friends of all workers in the mathematical sciences [4, 5, 10]. This chapter emphasizes special techniques for handling and generating the power series encountered in computational statistics. Most expansions can be phrased in terms of recurrence relations. Logarithmic differentiation is one powerful device for developing recurrences. Our applications of logarithmic differentiation to problems such as the conversion between moments and cumulants illustrate some of the interesting possibilities.

Power series expansions are also available for many of the well-known distribution functions of statistics. Although such expansions are usually guaranteed to converge, roundoff error for an alternating series can be troublesome. Thus, either high-precision arithmetic should be used in expanding a distribution function, or the distribution function should be modified so that only positive terms are encountered in the series defining the modified function. Our expansions are coordinated with the discussion of special functions in *Numerical Recipes* [9]. We particularly stress connections among the various distribution functions.

2.2 Expansion of $P(s)^n$

Suppose $P(s) = \sum_{k=0}^{\infty} p_k s^k$ is a power series with $p_0 \neq 0$. If n is a positive integer, then the recurrence relation of J.C.P. Miller [5] permits one to compute the coefficients of $Q(s) = \sum_{k=0}^{\infty} q_k s^k = P(s)^n$ from those of $P(s)$. This clever formula is derived by differentiating $Q(s)$ and then multiplying the result by $P(s)$. By definition of $Q(s)$, this yields

$$P(s)Q'(s) = nP'(s)Q(s). \tag{1}$$

If we equate the coefficients of s^{k-1} on both sides of (1), then it follows that

$$\sum_{j=1}^{k} p_{k-j} j q_j = n \sum_{j=0}^{k-1} (k-j) p_{k-j} q_j,$$

which can be solved for q_k in the form

$$q_k = \frac{1}{kp_0} \sum_{j=0}^{k-1} [n(k-j) - j] p_{k-j} q_j. \tag{2}$$

The obvious initial condition is $q_0 = p_0^n$. Sometimes it is more natural to compute q_k^*, where $q_k^*/k! = q_k$ and $p_k^*/k! = p_k$. Then the recurrence relation (2) can be rewritten as

$$q_k^* = \frac{1}{kp_0^*} \sum_{j=0}^{k-1} \binom{k}{j} [n(k-j) - j] p_{k-j}^* q_j^*. \tag{3}$$

2.2.1 Application to Moments

Suppose $X_1, \ldots, X_n$ are independent, identically distributed random variables. Let μ_k be the kth moment of X_1, and let ω_k be the kth moment of $S_n = \sum_{i=1}^{n} X_i$. Applying the recurrence (3) to the moment generating functions of X_1 and S_n gives

$$\omega_k = \frac{1}{k} \sum_{j=0}^{k-1} \binom{k}{j} [n(k-j) - j] \mu_{k-j} \omega_j.$$

As a concrete example, suppose $n = 10$ and X_1 has a uniform distribution on $[0, 1]$. Then $\mu_k = 1/(k+1)$. Table 2.1 records the first 10 moments ω_k of S_{10}.

TABLE 2.1. The Moments ω_k of the Sum of Ten Uniform Deviates

k	ω_k	k	ω_k
1	$.50000 \times 10^1$	6	$.24195 \times 10^5$
2	$.25833 \times 10^2$	7	$.14183 \times 10^6$
3	$.13750 \times 10^3$	8	$.84812 \times 10^6$
4	$.75100 \times 10^3$	9	$.51668 \times 10^7$
5	$.42167 \times 10^4$	10	$.32029 \times 10^8$

2.3 Expansion of $e^{P(s)}$

Again let $P(s)$ be a power series, and put $Q(s) = e^{P(s)}$ [8]. If one equates the coefficients of s^{k-1} in the obvious identity

$$Q'(s) = P'(s)Q(s),$$

then it follows that

$$q_k = \frac{1}{k}\sum_{j=0}^{k-1}(k-j)p_{k-j}q_j. \tag{4}$$

Clearly, $q_0 = e^{p_0}$. When $q_k^*/k! = q_k$ and $p_k^*/k! = p_k$, equation (4) becomes

$$q_k^* = \sum_{j=0}^{k-1}\binom{k-1}{j}p_{k-j}^* q_j^*. \tag{5}$$

2.3.1 Moments to Cumulants and Vice Versa

If $Q(s) = \sum_{k=0}^{\infty}\frac{m_k}{k!}s^k = e^{P(s)}$ is the moment generating function of a random variable, then $P(s) = \sum_{k=0}^{\infty}\frac{c_k}{k!}s^k$ is the corresponding cumulant generating function. Clearly, $m_0 = 1$ and $c_0 = 0$. The recurrence (5) can be rewritten as

$$m_k = \sum_{j=0}^{k-1}\binom{k-1}{j}c_{k-j}m_j.$$

From this one can deduce the equally useful recurrence

$$c_k = m_k - \sum_{j=1}^{k-1}\binom{k-1}{j}c_{k-j}m_j$$

converting moments to cumulants.

2.3.2 Compound Poisson Distributions

Consider a random sum $S_N = X_1 + \cdots + X_N$ of a random number N of independent, identically distributed random variables X_k. If N is indepen-

dent of the X_k and has a Poisson distribution with mean λ, then S_N is said to have a compound Poisson distribution. If $R(s)$ is the common moment generating function of the X_k, then $Q(s) = e^{-\lambda+\lambda R(s)}$ is the moment generating function of S_N. Likewise, if the X_k assume only nonnegative integer values, and if $R(s)$ is their common probability generating function, then $Q(s) = e^{-\lambda+\lambda R(s)}$ is the probability generating function of S_N. Thus, the moments $\mathrm{E}(S_N^i)$ and probabilities $\Pr(S_N = i)$ can be recursively computed from the corresponding quantities for the X_k.

2.3.3 Evaluation of Hermite Polynomials

The Hermite polynomials $H_k(x)$ can be defined by the generating function

$$\sum_{k=0}^{\infty} \frac{H_k(x)}{k!} s^k = e^{xs-\frac{1}{2}s^2}.$$

From this definition it is clear that $H_0(x) = 1$ and $H_1(x) = x$. The recurrence (5) takes the form

$$H_k(x) = xH_{k-1}(x) - (k-1)H_{k-2}(x)$$

for $k \geq 2$. In general, any sequence of orthogonal polynomials can be generated by a linear, two-term recurrence relation. We will meet the Hermite polynomials later when we consider Gaussian quadrature and Edgeworth expansions.

2.4 Standard Normal Distribution Function

Consider the standard normal distribution

$$F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{y^2}{2}} dy.$$

If we expand

$$e^{-\frac{y^2}{2}} = \sum_{n=0}^{\infty} \frac{(-1)^n y^{2n}}{2^n n!}$$

and integrate term by term, then it is clear that

$$F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2^n (2n+1) n!}.$$

This is an alternating series that entails severe roundoff error even for x as small as 4.

To derive a more stable expansion, let

$$g(x) = e^{\frac{x^2}{2}} \int_0^x e^{-\frac{y^2}{2}} dy$$

$$= \sum_{n=0}^{\infty} c_n x^{2n+1}.$$

By inspection, $g(x)$ satisfies the differential equation

$$g'(x) = xg(x) + 1. \tag{6}$$

Now $c_0 = 1$ because $g'(0) = 0g(0) + 1$. All subsequent coefficients are also positive. Indeed, equating coefficients of x^{2n} in (6) gives the recurrence relation

$$c_n = \frac{1}{2n+1} c_{n-1}.$$

Thus, the series for $g(x)$ converges stably for all $x > 0$. Since $g(x)$ is an odd function, only positive x need be considered. In evaluating

$$\begin{aligned} F(x) - \frac{1}{2} &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} g(x) \\ &= \sum_{n=0}^{\infty} a_n, \end{aligned}$$

we put $a_0 = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x$ and $a_n = a_{n-1} \frac{x^2}{2n+1}$. Then the partial sums $\sum_{i=0}^{n} a_i$ are well scaled, and $a_n = 0$ gives a machine independent test for the convergence of the series at its nth term.

2.5 Incomplete Gamma Function

The distribution function of a gamma random variable with parameters a and b is defined by

$$\begin{aligned} P(a, bx) &= \frac{1}{\Gamma(a)} \int_0^x b^a y^{a-1} e^{-by} dy \\ &= \frac{1}{\Gamma(a)} \int_0^{bx} z^{a-1} e^{-z} dz. \end{aligned} \tag{7}$$

We can expand $P(a, x)$ in a power series by repeated integration by parts. In fact,

$$\begin{aligned} P(a, x) &= \frac{x^a}{a\Gamma(a)} e^{-x} + \frac{1}{a\Gamma(a)} \int_0^x z^a e^{-z} dz \\ &= \frac{e^{-x} x^a}{\Gamma(a+1)} + P(a+1, x) \end{aligned}$$

leads to the stable series

$$P(a,x) = e^{-x}x^a \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(a+n+1)}. \tag{8}$$

For the expansion (8) to be practical, we must have some method for evaluating the ordinary gamma function. One option is to iterate the functional identity

$$\ln \Gamma(a) = \ln \Gamma(a+1) - \ln a$$

until k is large enough so that $\ln \Gamma(a+k)$ is well approximated by Stirling's formula.

2.6 Incomplete Beta Function

For a and b positive, the incomplete beta function is defined by

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1}(1-y)^{b-1} dy.$$

Suppose we attempt to expand this distribution function in the form

$$\begin{aligned} I_x(a,b) &= x^a(1-x)^b \sum_{n=0}^{\infty} c_n x^n \\ &= \sum_{n=0}^{\infty} c_n x^{n+a}(1-x)^b. \end{aligned} \tag{9}$$

If we divide the derivative

$$\begin{aligned} \frac{d}{dx} I_x(a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \\ &= \sum_{n=0}^{\infty} c_n[(n+a)(1-x) - bx]x^{n+a-1}(1-x)^{b-1} \end{aligned}$$

by $x^{a-1}(1-x)^{b-1}$, then it follows that

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \sum_{n=0}^{\infty} c_n[(n+a)(1-x) - bx]x^n. \tag{10}$$

Equating the coefficients of x^n on both sides of (10) gives for $n = 0$

$$c_0 = \frac{\Gamma(a+b)}{a\Gamma(a)\Gamma(b)} = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)}$$

and for $n > 0$

$$c_n(n+a) - c_{n-1}(n-1+a+b) = 0,$$

which collapses to the recurrence relation

$$c_n = \frac{n-1+a+b}{n+a} c_{n-1}.$$

Therefore, all coefficients c_n are positive. The ratio test indicates that the power series (9) converges for $0 \leq x < 1$. For x near 1, the symmetry relation $I_x(a,b) = 1 - I_{1-x}(b,a)$ can be employed to get a more quickly converging series.

2.7 Connections to Other Distributions

Evaluation of many classical distribution functions reduces to the cases already studied. Here are some examples.

2.7.1 Chi-Square and Standard Normal

A chi-square random variable χ^2_n with n degrees of freedom has a gamma distribution with parameters $a = n/2$ and $b = 1/2$. Hence, in terms of definition (7), we have $\Pr(\chi^2_n \leq x) = P(\frac{n}{2}, \frac{x}{2})$. If X has a standard normal distribution, then X^2 has a chi-square distribution with one degree of freedom. Obvious symmetry arguments therefore imply for $x \geq 0$ that $\Pr(X \leq x) = \frac{1}{2} + \frac{1}{2}P(\frac{1}{2}, \frac{x^2}{2})$.

2.7.2 Poisson

The distribution function of a Poisson random variable X with mean λ can be expressed in terms of the incomplete gamma function (7) as

$$\Pr(X \leq k-1) = 1 - P(k, \lambda).$$

The most illuminating proof of this result relies on constructing a Poisson process of unit intensity on $[0, \infty)$. In this framework $\Pr(X \leq k-1)$ is the probability of $k-1$ or fewer random points on $[0, \lambda]$. Since the waiting time until the kth random point in the process follows a gamma distribution with parameters $a = k$ and $b = 1$, the probability of $k-1$ or fewer random points on $[0, \lambda]$ coincides with the probability $1 - P(k, \lambda)$ that the kth random point falls beyond λ.

2.7.3 Binomial and Negative Binomial

Let X be a binomially distributed random variable with n trials and success probability p. We can express the distribution function of X in terms of the incomplete beta function (9) as

$$\Pr(X \leq k-1) = 1 - I_p(k, n-k+1). \tag{11}$$

To validate this expression, imagine distributing n points randomly on $[0,1]$. The probability $\Pr(X \le k-1)$ is just the probability that $k-1$ or fewer of the random points occur on $[0,p]$. This latter probability is also the probability that the kth random point to the right of 0 falls on $[p,1]$. But standard arguments from the theory of order statistics show that the kth random point to the right of 0 has beta density $n\binom{n-1}{k-1}y^{k-1}(1-y)^{n-k}$.

Alternatively, if we drop random points indefinitely on $[0,1]$ and record the trial Y at which the kth point falls to the left of p, then Y follows a negative binomial distribution. By the above argument,

$$\Pr(Y > n) = \Pr(X \le k-1),$$

which clearly entails $\Pr(Y \le n) = I_p(k, n-k+1)$. If we focus on failures rather than total trials in the definition of the negative binomial, then the random variable $Z = Y - k$ is representative of this point of view. In this case, $\Pr(Z \le m) = I_p(k, m+1)$.

2.7.4 F and Student's t

An $F_{m,n}$ random variable can be written as the ratio

$$F_{m,n} = \frac{n\chi^2_m}{m\chi^2_n}$$

of two independent chi-square random variables scaled to have unit means. Straightforward algebra gives

$$\begin{aligned}\Pr(F_{m,n} \le x) &= \Pr\left(\frac{\chi^2_m}{\chi^2_n} \le \frac{mx}{n}\right) \\ &= \Pr\left(\frac{\chi^2_n}{\chi^2_m + \chi^2_n} \ge \frac{n}{mx+n}\right).\end{aligned}$$

If $p = n/2$ is an integer, then $\chi^2_n/2 = W_p$ is a gamma distributed random variable that can be interpreted as the waiting time until the pth random point in a Poisson process on $[0,\infty)$. Similarly, if $q = m/2$ is an integer, then $\chi^2_m/2 = W_q$ can be interpreted as the waiting time from the pth random point until the $(p+q)$th random point of the same Poisson process. In this setting, the ratio

$$\begin{aligned}\frac{\chi^2_n}{\chi^2_m + \chi^2_n} &= \frac{W_p}{W_q + W_p} \\ &\ge u\end{aligned}$$

if and only if the waiting time until the pth point is a fraction u or greater of the waiting time until the $(p+q)$th point. Now conditional on the waiting time $W_p + W_q$ until random point $p+q$, the $p+q-1$ previous random points are uniformly and independently distributed on the interval $[0, W_p + W_q]$.

It follows from equation (11) that

$$\begin{aligned}\Pr\left(\frac{W_p}{W_q+W_p}\geq u\right) &= \sum_{j=0}^{p-1}\binom{p+q-1}{j}u^j(1-u)^{p+q-1-j}\\ &= 1-I_u(p,p+q-1-p+1)\\ &= I_{1-u}(q,p).\end{aligned}$$

In general, regardless of whether n or m is even, the identity

$$\Pr(F_{m,n}\leq x) = I_{\frac{mx}{mx+n}}\left(\frac{m}{2},\frac{n}{2}\right) \tag{12}$$

holds, relating the F distribution to the incomplete beta function [1].

By definition a random variable t_n follows Student's t distribution with n degrees of freedom if it is symmetric around 0 and its square t_n^2 has an $F_{1,n}$ distribution. Therefore according to equation (12),

$$\begin{aligned}\Pr(t_n\leq x) &= \frac{1}{2}+\frac{1}{2}\Pr(t_n^2\leq x^2)\\ &= \frac{1}{2}+\frac{1}{2}\Pr(F_{1,n}\leq x^2)\\ &= \frac{1}{2}+\frac{1}{2}I_{\frac{x^2}{x^2+n}}\left(\frac{1}{2},\frac{n}{2}\right)\end{aligned}$$

for $x\geq 0$.

2.7.5 Monotonic Transformations

Suppose X is a random variable with known distribution function $F(x)$ and $h(x)$ is a strictly increasing, continuous function. Then the random variable $h(X)$ has distribution function

$$\Pr[h(X)\leq x] = F[h^{-1}(x)],$$

where $h^{-1}(x)$ is the functional inverse of $h(x)$. If $h(x)$ is strictly decreasing and continuous, then

$$\begin{aligned}\Pr[h(X)<x] &= \Pr[X>h^{-1}(x)]\\ &= 1-F[h^{-1}(x)].\end{aligned}$$

Many common distributions fit this paradigm. For instance, if X is normal, then e^X is lognormal. If X is chi-square, then $1/X$, $1/\sqrt{X}$, and $\ln X$ are inverse chi-square, inverse chi, and log chi-square, respectively. If X has an $F_{m,n}$ distribution, then $\frac{1}{2}\ln X$ has Fisher's z distribution. Calculating any of these distributions therefore reduces to evaluating either an incomplete beta or an incomplete gamma function.

2.8 Problems

1. A symmetric random walk on the integer lattice points of R^k starts at the origin and at each epoch randomly chooses one of the $2k$ possible coordinate directions and takes a unit step in that direction [3]. If u_{2n} is the probability that the walk returns to the origin at epoch $2n$, then one can show that

$$\sum_{n=0}^{\infty} \frac{u_{2n}}{(2n)!} x^{2n} = \Big[\sum_{n=0}^{\infty} \frac{1}{(2k)^{2n}(n!)^2} x^{2n} \Big]^k.$$

Derive a recurrence relation for computing u_{2n}, and implement it when $k = 2$. Check your numerical results against the exact formula

$$u_{2n} = \Big[\frac{1}{2^{2n}} \binom{2n}{n} \Big]^2.$$

Discuss possible sources of numerical error in using the recurrence relation.

2. Write recurrence relations for the Taylor coefficients of the functions $\left(\frac{1+s}{1-s}\right)^n$ and $\exp(\frac{1+s}{1-s})$.

3. Show that the coefficients of the exponential generating function

$$\sum_{n=0}^{\infty} \frac{B_n}{n!} s^n = e^{e^s - 1}$$

satisfy the recurrence relation (2) of Chapter 1. Check the initial condition $B_0 = 1$, and conclude that the coefficient B_n determines the number of partitions of a set with n elements.

4. Suppose the coefficients of a power series $\sum_{n=0}^{\infty} b_n x^n$ satisfy $b_n = p(n)$ for some polynomial p. Find a power series $\sum_{n=0}^{\infty} a_n x^n$ such that

$$p\Big(x \frac{d}{dx}\Big) \sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n.$$

5. Show that $\sum_{n=1}^{m} n^2 = m(m+1)(2m+1)/6$ by evaluating

$$\Big(x \frac{d}{dx}\Big)^2 \sum_{n=0}^{m} x^n = \Big(x \frac{d}{dx}\Big)^2 \frac{x^{m+1} - 1}{x - 1}$$

at $x = 1$.

6. A family of discrete density functions $p_n(\theta)$ defined on $\{0, 1, \ldots\}$ and indexed by a parameter $\theta > 0$ is said to be a power series family if for all n

$$p_n(\theta) = \frac{c_n \theta^n}{g(\theta)}, \tag{13}$$

where $c_n \geq 0$, and where $g(\theta) = \sum_{k=0}^{\infty} c_k \theta^k$ is the appropriate normalizing constant. Show that the mean $\mu(\theta)$ and variance $\sigma^2(\theta)$ of the

$p_n(\theta)$ reduce to

$$\mu(\theta) = \frac{\theta g'(\theta)}{g(\theta)}$$
$$\sigma^2(\theta) = \theta\mu'(\theta).$$

7. Continuing Problem 6, suppose $X_1, \ldots, X_m$ is a random sample from the power series distribution (13). Show that $S_m = X_1 + \cdots + X_m$ follows a power series distribution with

$$\Pr(S_m = n) = \frac{a_{mn}\theta^n}{g(\theta)^m},$$

where a_{mn} is the coefficient of θ^n in $g(\theta)^m$. If $a_{mn} = 0$ for $n < 0$, then also prove that $a_{m,S_m-r}/a_{m,S_m}$ is an unbiased estimator of θ^r. This estimator is, in fact, the uniformly minimum variance, unbiased estimator of θ^r [6].

8. Suppose $f_n(x)$ and $F_n(x)$ represent, respectively, the density and distribution functions of a chi-square random variable with n degrees of freedom. The noncentral chi-square density [1] with noncentrality parameter 2λ and degrees of freedom n can be written as the Poisson mixture

$$f_{\lambda,n}(x) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} f_{n+2k}(x).$$

Show that

$$F_{n+2(k+1)}(x) = F_{n+2k}(x) - \frac{e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}+k}}{\Gamma\left(\frac{n}{2}+k+1\right)}.$$

Hence, in evaluating the distribution function $F_{\lambda,n}(x)$ of $f_{\lambda,n}(x)$, it suffices to compute only the single incomplete gamma function $F_n(x)$. Prove the error estimate

$$0 \le F_{\lambda,n}(x) - \sum_{k=0}^{m} \frac{\lambda^k}{k!} e^{-\lambda} F_{n+2k}(x)$$
$$\le \left(1 - \sum_{k=0}^{m} \frac{\lambda^k}{k!} e^{-\lambda}\right) F_{n+2(m+1)}(x).$$

9. To generalize Problem 8, consider the sum $S_N = \sum_{i=1}^{N} X_i$, where the summands X_i are independent, exponentially distributed random variables with common mean $1/\nu$, and the number of summands N is a nonnegative, integer-valued random variable independent of the X_i. By definition, $S_N = 0$ when $N = 0$. If $\Pr(N = n) = p_n$, then show that

$$\Pr(S_N \le x) = \sum_{n=0}^{\infty} p_n P(n, \nu x)$$

$$\mathrm{E}(e^{-\theta S_N}) = \sum_{n=0}^{\infty} p_n \left(\frac{\nu}{\nu+\theta}\right)^n,$$

where $P(n,x)$ is the incomplete gamma function. In view of Problem 8, how would you proceed in evaluating the distribution function $\Pr(S_N \leq x)$? Finally, demonstrate that $\mathrm{E}(S_N) = \mathrm{E}(N)/\nu$ and $\mathrm{Var}(S_N) = [\mathrm{E}(N) + \mathrm{Var}(N)]/\nu^2$.

10. Prove the incomplete beta function identities

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^{b-1} + I_x(a+1,b-1), \quad b > 1$$
$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^b + I_x(a+1,b).$$

These two relations form the basis of a widely used algorithm [7] for computing $I_x(a,b)$. (Hints: For the first, integrate by parts, and for the second, show that both sides have the same derivative.)

11. Suppose that Z has discrete density

$$\Pr(Z = j) = \binom{k+j-1}{j} p^k (1-p)^j,$$

where $k > 0$ and $0 < p < 1$. In other words, Z follows a negative binomial distribution counting failures, not total trials. Show that $\Pr(Z \leq m) = I_p(k, m+1)$ regardless of whether k is an integer. (Hint: Use one of the identities of the previous problem.)

12. Let $X_{(k)}$ be the kth order statistic from a finite sequence $X_1, \ldots, X_n$ of independent, identically distributed random variables with common distribution function $F(x)$. Show that $X_{(k)}$ has distribution function $\Pr(X_{(k)} \leq x) = I_{F(x)}(k, n-k+1)$.

13. Suppose the bivariate normal random vector $(X_1, X_2)^t$ has means $\mathrm{E}(X_i) = \mu_i$, variances $\mathrm{Var}(X_i) = \sigma_i^2$, and correlation ρ. Verify the decomposition

$$X_1 = \sigma_1 |\rho|^{\frac{1}{2}} Y + \sigma_1 (1-|\rho|)^{\frac{1}{2}} Z_1 + \mu_1$$
$$X_2 = \sigma_2 \operatorname{sgn}(\rho) |\rho|^{\frac{1}{2}} Y + \sigma_2 (1-|\rho|)^{\frac{1}{2}} Z_2 + \mu_2,$$

where Y, Z_1, and Z_2 are independent, standard normal random variables. Use this decomposition to deduce that

$$\Pr(X_1 \leq x_1, X_2 \leq x_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Phi\left[\frac{x_1 - \mu_1 - \sigma_1 |\rho|^{\frac{1}{2}} y}{\sigma_1 (1-|\rho|)^{\frac{1}{2}}}\right] \times \Phi\left[\frac{x_2 - \mu_2 - \sigma_2 \operatorname{sgn}(\rho) |\rho|^{\frac{1}{2}} y}{\sigma_2 (1-|\rho|)^{\frac{1}{2}}}\right] e^{-\frac{y^2}{2}} dy,$$

where $\Phi(x)$ is the standard normal distribution [2].

References

[1] Bickel PJ, Doksum KA (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland, CA
[2] Curnow RN, Dunnett CW (1962) The numerical evaluation of certain multivariate normal integrals. *Ann Math Stat* 33:571–579
[3] Feller W (1968) *An Introduction to Probability Theory and Its Applications, Vol 1*, 3rd ed. Wiley, New York
[4] Graham RL, Knuth DE, Patashnik O (1988) *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA
[5] Henrici P (1974) *Applied and Computational Complex Analysis, Vol 1*. Wiley, New York
[6] Lehmann EL (1991) *Theory of Point Estimation*. Wadsworth, Belmont, CA
[7] Majumder KL, Bhattacharjee GP (1973) Algorithm AS 63. The incomplete beta integral. *Appl Stat* 22:409–411
[8] Pourhamadi M (1984) Taylor expansion of $\exp(\sum_{k=0}^{\infty} a_k z^k)$ and some applications. *Amer Math Monthly* 91:303–307
[9] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[10] Wilf HS (1990) *generatingfunctionology*. Academic Press, New York

3
Continued Fraction Expansions

3.1 Introduction

A continued fraction [2, 3, 4, 5] is a sequence of fractions

$$f_n = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdots + \cfrac{a_n}{b_n}}}} \tag{1}$$

formed from two sequences $a_1, a_2, \ldots$ and $b_0, b_1, \ldots$ of numbers. For typographical convenience, definition (1) is usually recast as

$$f_n = b_0 + \frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots\frac{a_n}{b_n}.$$

In many practical examples, the approximant f_n converges to a limit, which is typically written as

$$\lim_{n\to\infty} f_n = b_0 + \frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots.$$

Because the elements a_n and b_n of the two defining sequences can depend on a variable x, continued fractions offer an alternative to power series in expanding functions such as distribution functions. In fact, continued fractions can converge where power series diverge, and where both types of expansions converge, continued fractions often converge faster.

A lovely little example of a continued fraction is furnished by

$$\begin{aligned}\sqrt{2}-1 &= \frac{1}{2+(\sqrt{2}-1)}\\ &= \frac{1}{2+\dfrac{1}{2+(\sqrt{2}-1)}}\\ &= \frac{1}{2+\dfrac{1}{2+\dfrac{1}{2+(\sqrt{2}-1)}}}.\end{aligned}$$

One can easily check numerically that the limit

$$\sqrt{2} = 1 + \frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\cdots$$

is correct. It is harder to prove this analytically. For the sake of brevity, we will largely avoid questions of convergence. Readers interested in a full treatment of continued fractions can consult the references [2, 3, 5]. Problems 7 through 10 prove convergence when the sequences a_n and b_n are positive.

Before giving more examples, it is helpful to consider how we might go about evaluating the approximant f_n. One obvious possibility is to work from the bottom of the continued fraction (1) to the top. This obvious approach can be formalized by defining fractional linear transformations $t_0(x) = b_0 + x$ and $t_n(x) = a_n/(b_n + x)$ for $n > 0$. If the circle symbol $\circ$ denotes functional composition, then we take $x = 0$ and compute

$$\begin{aligned}t_n(0) &= \frac{a_n}{b_n}\\ t_{n-1} \circ t_n(0) &= \frac{a_{n-1}}{b_{n-1} + t_n(0)}\\ &\vdots\\ t_0 \circ t_1 \circ \cdots \circ t_n(0) &= f_n.\end{aligned}$$

This turns out to be a rather inflexible way to proceed because if we want the next approximant f_{n+1}, we are forced to start all over again. In 1655 J. Wallis [6] suggested an alternative strategy. (This is a venerable but often neglected subject.)

3.2 Wallis's Algorithm

According to Wallis,

$$t_0 \circ t_1 \circ \cdots \circ t_n(x) = \frac{A_{n-1}x + A_n}{B_{n-1}x + B_n} \tag{2}$$

for a certain pair of auxiliary sequences A_n and B_n. Taking $x = 0$ gives the approximant $f_n = A_n/B_n$. The sequences A_n and B_n satisfy the initial conditions

$$\begin{pmatrix} A_{-1} \\ B_{-1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} = \begin{pmatrix} b_0 \\ 1 \end{pmatrix} \tag{3}$$

and for $n > 0$ the recurrence relation

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = b_n \begin{pmatrix} A_{n-1} \\ B_{n-1} \end{pmatrix} + a_n \begin{pmatrix} A_{n-2} \\ B_{n-2} \end{pmatrix}. \tag{4}$$

From the initial conditions (3), it is clear that

$$t_0(x) = b_0 + x = \frac{A_{-1}x + A_0}{B_{-1}x + B_0}.$$

The general case of (2) is proved by induction. Suppose the formula is true for an arbitrary nonnegative integer n. Then the induction hypothesis and the recurrence relation (4) together imply

$$\begin{aligned} t_0 \circ t_1 \circ \cdots \circ t_{n+1}(x) &= t_0 \circ t_1 \circ \cdots \circ t_n\Big(\frac{a_{n+1}}{b_{n+1} + x}\Big) \\ &= \frac{A_{n-1}\frac{a_{n+1}}{b_{n+1}+x} + A_n}{B_{n-1}\frac{a_{n+1}}{b_{n+1}+x} + B_n} \\ &= \frac{A_n x + (b_{n+1}A_n + a_{n+1}A_{n-1})}{B_n x + (b_{n+1}B_n + a_{n+1}B_{n-1})} \\ &= \frac{A_n x + A_{n+1}}{B_n x + B_{n+1}}. \end{aligned}$$

3.3 Equivalence Transformations

The same continued fraction can be defined by more than one pair of sequences a_n and b_n. For instance, if $b_n \neq 0$ for all $n > 0$, then it is possible to concoct an equivalent continued fraction given by a pair of sequences a'_n and b'_n with $b'_0 = b_0$ and $b'_n = 1$ for all $n > 0$. This can be demonstrated most easily by defining the transformed auxiliary sequences

$$\begin{pmatrix} A'_n \\ B'_n \end{pmatrix} = \frac{1}{\prod_{k=1}^{n} b_k} \begin{pmatrix} A_n \\ B_n \end{pmatrix},$$

with the understanding that $\prod_{k=1}^{n} b_k = 1$ when $n = -1$ or 0. ¿From this definition it follows that A'_n and B'_n satisfy the same initial conditions (3) as A_n and B_n. Furthermore, the recurrence relation (4) becomes

$$\begin{pmatrix} A'_1 \\ B'_1 \end{pmatrix} = \begin{pmatrix} A'_0 \\ B'_0 \end{pmatrix} + \frac{a_1}{b_1} \begin{pmatrix} A'_{-1} \\ B'_{-1} \end{pmatrix}$$
$$\begin{pmatrix} A'_n \\ B'_n \end{pmatrix} = \begin{pmatrix} A'_{n-1} \\ B'_{n-1} \end{pmatrix} + \frac{a_n}{b_{n-1}b_n} \begin{pmatrix} A'_{n-2} \\ B'_{n-2} \end{pmatrix}, \qquad n > 1$$

after division by $\prod_{k=1}^{n} b_k$. Thus, the transformed auxiliary sequences correspond to the choice $a'_1 = a_1/b_1$ and $a'_n = a_n/(b_{n-1}b_n)$ for $n > 1$. By definition, the approximants $f'_n = A'_n/B'_n$ and $f_n = A_n/B_n$ coincide.

Faster convergence can often be achieved by taking the even part of a continued fraction. This is a new continued fraction whose approximant f'_n equals the approximant f_{2n} of the original continued fraction. For the sake of simplicity, suppose that we start with a transformed continued fraction with all $b_n = 1$ for $n > 0$. We can then view the approximant f_n as the value at $x = 1$ of the iterated composition

$$s_1 \circ \cdots \circ s_n(x) = b_0 + \cfrac{a_1}{1 + \cfrac{a_2}{1 + \cfrac{a_3}{1 + \cdots + \cfrac{a_{n-1}}{1 + a_n x}}}}$$

of the n functions $s_1(x) = b_0 + a_1 x$ and $s_k(x) = 1/(1 + a_k x)$ for $2 \le k \le n$. If we compose these functions $r_n(x) = s_{2n-1} \circ s_{2n}(x)$ two by two, we get

$$\begin{aligned} r_1(x) &= b_0 + \frac{a_1}{1 + a_2 x}, \\ r_n(x) &= \cfrac{1}{1 + \cfrac{a_{2n-1}}{1 + a_{2n}x}} \\ &= 1 - \frac{a_{2n-1}}{1 + a_{2n-1} + a_{2n}x}, \qquad n > 1. \end{aligned}$$

The approximant f'_n of the new continued fraction is just

$$\begin{aligned} f'_n &= r_1 \circ \cdots \circ r_n(1) \\ &= b_0 + \cfrac{a_1}{1 + a_2 - \cfrac{a_2 a_3}{1 + a_3 + a_4 - \cdots - \cfrac{a_{2n-2}a_{2n-1}}{1 + a_{2n-1} + a_{2n}}}}. \end{aligned}$$

From this expansion of the even part, we read off the sequences $a'_1 = a_1$, $a'_n = -a_{2n-2}a_{2n-1}$ for $n > 1$, $b'_0 = b_0$, $b'_1 = 1 + a_2$, and $b'_n = 1 + a_{2n-1} + a_{2n}$ for $n > 1$.

3.4 Gauss's Expansion of Hypergeometric Functions

The hypergeometric function ${}_2F_1(a,b,c;x)$ is given by the power series

$$ {}_2F_1(a,b,c;x) = \sum_{n=0}^{\infty} \frac{a^{\overline{n}} b^{\overline{n}}}{c^{\overline{n}}} \frac{x^n}{n!}, \tag{5} $$

where $a^{\overline{n}}$, $b^{\overline{n}}$, and $c^{\overline{n}}$ are rising factorial powers defined by

$$ a^{\overline{n}} = \begin{cases} 1 & n = 0 \\ a(a+1)\cdots(a+n-1) & n > 0, \end{cases} $$

and so forth. To avoid division by 0, the constant c in (5) should be neither 0 nor a negative integer. If either a or b is 0 or a negative integer, then the power series reduces to a polynomial. The binomial series

$$ \begin{aligned} \frac{1}{(1-x)^a} &= \sum_{n=0}^{\infty} \binom{a+n-1}{n} x^n \\ &= \sum_{n=0}^{\infty} a^{\overline{n}} \frac{x^n}{n!} \\ &= {}_2F_1(a,1,1;x) \end{aligned} $$

and the incomplete beta function

$$ \begin{aligned} I_x(a,b) &= \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^b \sum_{n=0}^{\infty} \frac{(a+b)^{\overline{n}}}{(a+1)^{\overline{n}}} x^n \\ &= \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} x^a (1-x)^b \, {}_2F_1(a+b,1,a+1;x) \end{aligned} \tag{6} $$

involve typical hypergeometric expansions. Straightforward application of the ratio test shows that the hypergeometric series (5) converges for all $|x| < 1$. As the binomial series makes clear, convergence can easily fail for $|x| \geq 1$.

In 1812 Gauss [1] described a method of converting ratios of hypergeometric functions into continued fractions. His point of departure was the simple identity

$$ \frac{a^{\overline{n}}(b+1)^{\overline{n}}}{(c+1)^{\overline{n}} n!} - \frac{a^{\overline{n}} b^{\overline{n}}}{c^{\overline{n}} n!} = \frac{a(c-b)}{c(c+1)} \frac{(a+1)^{\overline{n-1}} (b+1)^{\overline{n-1}}}{(c+2)^{\overline{n-1}} (n-1)!}. $$

Multiplying this by x^n and summing on n yields the hypergeometric function identity

$$ \begin{aligned} &{}_2F_1(a,b+1,c+1;x) - {}_2F_1(a,b,c;x) \\ &\quad = \frac{a(c-b)x}{c(c+1)} {}_2F_1(a+1,b+1,c+2;x), \end{aligned} $$

which can be rewritten as

$$\frac{{}_2F_1(a,b+1,c+1;x)}{{}_2F_1(a,b,c;x)} = \frac{1}{1-\dfrac{a(c-b)x}{c(c+1)}\cdot\dfrac{{}_2F_1(a+1,b+1,c+2;x)}{{}_2F_1(a,b+1,c+1;x)}}. \tag{7}$$

If in this derivation we interchange the roles of a and b and then replace b by $b+1$ and c by $c+1$, then we arrive at the similar identity

$$\frac{{}_2F_1(a+1,b+1,c+2;x)}{{}_2F_1(a,b+1,c+1;x)} = \frac{1}{1-\dfrac{(b+1)(c+1-a)x}{(c+1)(c+2)}\cdot\dfrac{{}_2F_1(a+1,b+2,c+3;x)}{{}_2F_1(a+1,b+1,c+2;x)}}. \tag{8}$$

Now the ratio (8) can be substituted for the ratio appearing in the denominator on the right of equation (7). Likewise, the ratio (7) with a, b, and c replaced by $a+1$, $b+1$, and $c+2$, respectively, can be substituted for the ratio appearing in the denominator on the right of equation (8). Alternating these successive substitutions produces Gauss's continued fraction expansion

$$\frac{{}_2F_1(a,b+1,c+1;x)}{{}_2F_1(a,b,c;x)} = \frac{1}{1+}\frac{d_1x}{1+}\frac{d_2x}{1+}\cdots \tag{9}$$

with

$$d_{2n+1} = -\frac{(a+n)(c-b+n)}{(c+2n)(c+2n+1)}$$
$$d_{2n+2} = -\frac{(b+n+1)(c-a+n+1)}{(c+2n+1)(c+2n+2)}$$

for $n \geq 0$.

Gauss's expansion (9) is most useful when $b = 0$, for then

$${}_2F_1(a,b,c;x) = 1.$$

For instance, the hypergeometric expansion of $(1-x)^{-a}$ has coefficients

$$d_1 = -a$$
$$d_{2n+1} = -\frac{(a+n)}{2(2n+1)}, \qquad n \geq 1$$
$$d_{2n+2} = -\frac{(n+1-a)}{2(2n+1)}, \qquad n \geq 0.$$

In this example, note that the identity (7) continues to hold for $b = c = 0$, provided ${}_2F_1(a,0,0;x)$ and the ratio $(c-b)/c$ are both interpreted as 1.

The hypergeometric function ${}_2F_1(a+b,1,a+1;x)$ determining the incomplete beta function (6) can be expanded with coefficients

$$d_{2n+1} = -\frac{(a+b+n)(a+n)}{(a+2n)(a+2n+1)}$$
$$d_{2n+2} = -\frac{(n+1)(n+1-b)}{(a+2n+1)(a+2n+2)}$$

for $n \geq 0$. Press et al. [4] claim that this continued fraction expansion for the incomplete beta function is superior to the power series expansion (6) for all values of the argument x, provided one switches to the expansion of $I_{1-x}(b,a) = 1 - I_x(a,b)$ when $x > (a+1)/(a+b+2)$.

3.5 Expansion of the Incomplete Gamma Function

To expand the incomplete gamma function as a continued fraction, we take a detour and first examine the integral

$$J_x(a,b) = \frac{1}{\Gamma(a)} \int_0^\infty \frac{e^{-y} y^{a-1}}{(1+xy)^b} dy$$

for $a > 0$ and $x \geq 0$. This integral exhibits the surprising symmetry $J_x(a,b) = J_x(b,a)$. In fact, when both a and $b > 0$,

$$\begin{aligned} J_x(a,b) &= \frac{1}{\Gamma(a)} \int_0^\infty e^{-y} y^{a-1} \frac{1}{\Gamma(b)} \int_0^\infty e^{-z(1+xy)} z^{b-1} dz dy \\ &= \frac{1}{\Gamma(b)} \int_0^\infty e^{-z} z^{b-1} \frac{1}{\Gamma(a)} \int_0^\infty e^{-y(1+xz)} y^{a-1} dy dz \\ &= J_x(b,a). \end{aligned}$$

Because $J_x(a,0) = 1$ by definition of the gamma function, this symmetry relation yields $\lim_{a \to 0} J_x(a,b) = \lim_{a \to 0} J_x(b,a) = 1$. Thus, it is reasonable to define $J_x(0,b) = 1$ for $b > 0$.

To forge a connection to the incomplete gamma function, we consider $J_{x^{-1}}(1, 1-a)$. An obvious change of variables then implies

$$\begin{aligned} J_{x^{-1}}(1,1-a) &= \int_0^\infty e^{-y} \Big(1 + \frac{y}{x}\Big)^{a-1} dy \\ &= x^{1-a} \int_0^\infty e^{-y} (x+y)^{a-1} dy \\ &= x^{1-a} e^x \int_x^\infty e^{-z} z^{a-1} dz. \end{aligned}$$

A final simple rearrangement gives

$$\frac{1}{\Gamma(a)}\int_0^x e^{-z}z^{a-1}dz = 1 - \frac{e^{-x}x^{a-1}}{\Gamma(a)}J_{x^{-1}}(1, 1-a). \tag{10}$$

The integral $J_x(a,b)$ also satisfies identities similar to equations (7) and (8) for the hypergeometric function. For instance,

$$\begin{aligned} J_x(a,b) &= \frac{1}{\Gamma(a)}\int_0^\infty \frac{e^{-y}y^{a-1}(1+xy)}{(1+xy)^{b+1}}dy \\ &= J_x(a,b+1) + \frac{ax}{a\Gamma(a)}\int_0^\infty \frac{e^{-y}y^a}{(1+xy)^{b+1}}dy \\ &= J_x(a,b+1) + axJ_x(a+1,b+1) \end{aligned} \tag{11}$$

can be rearranged to give

$$\frac{J_x(a,b+1)}{J_x(a,b)} = \frac{1}{1+ax\dfrac{J_x(a+1,b+1)}{J_x(a,b+1)}}. \tag{12}$$

Exploiting the symmetry $J_x(a,b) = J_x(b,a)$ when $b > 0$ or integrating by parts in general, we find

$$J_x(a,b+1) = J_x(a+1,b+1) + (b+1)xJ_x(a+1,b+2).$$

This in turn yields

$$\frac{J_x(a+1,b+1)}{J_x(a,b+1)} = \frac{1}{1+(b+1)x\dfrac{J_x(a+1,b+2)}{J_x(a+1,b+1)}}. \tag{13}$$

Substituting equation (13) into equation (12) and vice versa in an alternating fashion leads to a continued fraction expansion of the form (9) for the ratio $J_x(a,b+1)/J_x(a,b)$. The coefficients of this expansion can be expressed as

$$d_{2n+1} = a+n$$

$$d_{2n+2} = b+n+1$$

for $n \geq 0$. The special case $b = 0$ is important because $J_x(a,0) = 1$.

If $a = 0$, then $J_x(0,b) = 1$, and it is advantageous to expand the continued fraction starting with identity (13) rather than identity (12). For example, the function $J_{x^{-1}}(1, 1-a)$ appearing in expression (10) for the incomplete gamma function can be expanded with coefficients

$$d_{2n+1} = 1-a+n$$

$$d_{2n+2} = n+1,$$

provided we replace x in (9) by $1/x$, commence the continued fraction with identity (13), and take $a = 0$ and $b+1 = 1-a$. (See Problem 5.) Press

et al. [4] recommend this continued fraction expansion for the incomplete gamma function on $x > a + 1$ and the previously discussed power series expansion on $x < a + 1$.

One subtle point in dealing with the case $a = 0$ is that we want the limiting value $J_x(0, b) = 1$ to hold for all b, not just for $b > 0$. To prove this slight extension, observe that iterating recurrence (11) leads to the representation

$$J_x(a, b) = J_x(a, b + n) + ax \sum_{k=1}^{n} p_k(x) J_x(a + k, b + n), \tag{14}$$

where the $p_k(x)$ are polynomials. If n is so large that $b+n > 0$, then taking limits in (14) again yields $\lim_{a\to 0} J_x(a, b) = 1$.

3.6 Problems

1. Suppose a continued fraction has all $a_k = x$, $b_0 = 0$, and all remaining $b_k = 1 - x$ for $|x| \neq 1$. Show that the nth approximant $f_n(x)$ satisfies

$$f_n(x) = \frac{x[1 - (-x)^n]}{1 - (-x)^{n+1}}.$$

Conclude that

$$\lim_{n\to\infty} f_n(x) = \begin{cases} x & |x| < 1 \\ -1 & |x| > 1. \end{cases}$$

Thus, the same continued fraction converges to two different analytic functions on two different domains.

2. Verify the identities

$$\ln(1 - x) = -x \, {}_2F_1(1, 1, 2; x)$$
$$\arctan(x) = x \, {}_2F_1\Big(\frac{1}{2}, 1, \frac{3}{2}; -x^2\Big)$$
$$\int_1^\infty \frac{e^{-xy}}{y^n} dy = \frac{e^{-x}}{x} \int_0^\infty \frac{e^{-u}}{\big(1 + \frac{u}{x}\big)^n} du.$$

3. Find continued fraction expansions for each of the functions in the previous problem.

4. If ${}_1F_1(b, c; x) = \sum_{n=0}^{\infty} \frac{\overline{b^n}}{\overline{c^n}} \frac{x^n}{n!}$, then prove that

$${}_1F_1(b, c; x) = \lim_{a\to\infty} {}_2F_1(a, b, c; \frac{x}{a}). \tag{15}$$

Noting that $e^x = {}_1F_1(1, 1; x)$, demonstrate that e^x has a continued fraction expansion given by the right-hand side of equation (9) with $d_{2n+1} = -(4n + 2)^{-1}$ and $d_{2n+2} = -(4n + 2)^{-1}$. (Hint: For the expansion of e^x, derive two recurrence relations by taking appropriate limits in equations (7) and (8).)

5. Check that the function $J_{x^{-1}}(1, 1-a)$ appearing in our discussion of the incomplete gamma function has the explicit expansion

$$J_{x^{-1}}(1,1-a) = x\Big(\frac{1}{x+}\frac{1-a}{1+}\frac{1}{x+}\frac{2-a}{1+}\frac{2}{x+}\cdots\Big).$$

Show that the even part of this continued fraction expansion amounts to

$$J_{x^{-1}}(1,1-a) = x\Big(\frac{1}{x+1-a-}\ \frac{1\cdot(1-a)}{x+3-a-}\ \frac{2\cdot(2-a)}{x+5-a-}\cdots\Big).$$

6. Lentz's method of evaluating the continued fraction (1) is based on using the ratios $C_n = A_n/A_{n-1}$ and $D_n = B_{n-1}/B_n$ and calculating f_n by $f_n = f_{n-1}C_nD_n$. This avoids underflows and overflows when the A_n or B_n tend to very small or very large values. Show that the ratios satisfy the recurrence relations

$$C_n = b_n + \frac{a_n}{C_{n-1}}$$
$$D_n = \frac{1}{b_n + a_nD_{n-1}}.$$

7. Prove the determinant formulas

$$\det\begin{pmatrix} A_n & A_{n-1} \\ B_n & B_{n-1}\end{pmatrix} = (-1)^{n-1}\prod_{k=1}^{n} a_k \tag{16}$$

$$\det\begin{pmatrix} A_{n+1} & A_{n-1} \\ B_{n+1} & B_{n-1}\end{pmatrix} = (-1)^{n-1}b_{n+1}\prod_{k=1}^{n} a_k \tag{17}$$

in the notation of Wallis' algorithm.

8. Suppose the two sequences a_n and b_n generating a continued fraction have all elements nonnegative. Show that the approximants f_n satisfy $f_1 \geq f_3 \geq \cdots \geq f_{2n+1} \geq f_{2n} \geq \cdots \geq f_2 \geq f_0$. It follows that $\lim_{n\to\infty} f_{2n}$ and $\lim_{n\to\infty} f_{2n+1}$ exist, but unless further assumptions are made, there can be a gap between these two limits. (Hint: Use equation (17) from the previous problem to prove $f_{2n} \geq f_{2n-2}$ and $f_{2n+1} \leq f_{2n-1}$. Use equation (16) to prove $f_{2n+1} \geq f_{2n}$.)

9. Provided all $a_n \neq 0$, prove that the continued fraction (1) is also generated by the sequences $a'_n = 1$ and $b'_n = b_n\prod_{k=1}^{n} a_k^{(-1)^{n-k+1}}$.

10. Suppose that the sequences a_n and b_n are positive. The Stern-Stolz theorem [2] says that

$$\sum_{n=0}^{\infty} b_n \prod_{k=1}^{n} a_k^{(-1)^{n-k+1}} = \infty$$

is a necessary and sufficient condition for the convergence of the approximants f_n to the continued fraction (1). To prove the sufficiency of this condition, verify that:

(a) It is enough by the previous problem to take all $a_n = 1$.
(b) The approximants then satisfy

$$\begin{aligned} f_{2n+1} - f_{2n} &= \frac{A_{2n+1}}{B_{2n+1}} - \frac{A_{2n}}{B_{2n}} \\ &= \frac{A_{2n+1}B_{2n} - A_{2n}B_{2n+1}}{B_{2n}B_{2n+1}} \\ &= \frac{1}{B_{2n}B_{2n+1}} \end{aligned}$$

by the determinant formula (16).
(c) Because $B_n = b_n B_{n-1} + B_{n-2}$, the sequence B_n satisfies

$$\begin{aligned} B_{2n} &\geq B_0 \\ &= 1 \\ B_{2n+1} &\geq b_1. \end{aligned}$$

(d) The recurrence $B_n = b_n B_{n-1} + B_{n-2}$ and part (c) together imply

$$\begin{aligned} B_{2n} &\geq (b_{2n} + b_{2n-2} + \cdots + b_2)b_1 + 1 \\ B_{2n+1} &\geq b_{2n+1} + b_{2n-1} + \cdots + b_1; \end{aligned}$$

consequently, either $\lim_{n\to\infty} B_{2n} = \infty$, or $\lim_{n\to\infty} B_{2n+1} = \infty$.
(e) The sufficiency part of the theorem now follows from parts (b) and (d).

11. The Stieltjes function $F(x) = F(0)\int_0^\infty \frac{1}{1+xy} dG(y)$ plays an important role in the theoretical development of continued fractions [5]. Here $G(y)$ is an arbitrary probability distribution function concentrated on $[0, \infty)$ and $F(0) > 0$. In the region $\{x : x \neq 0,\ |\arg(x)| < \pi\}$ of the complex plane C excluding the negative real axis and 0, show that $F(x)$ has the following properties:

(a) $\frac{1}{x}F(\frac{1}{x}) = F(0)\int_0^\infty \frac{1}{x+y} dG(y)$,
(b) $F(x)$ is an analytic function,
(c) $\lim_{x\to 0} F(x) = F(0)$,
(d) The imaginary part of $F(x)$ satisfies

$$\mathrm{Im}F(x) = \begin{cases} < 0 & \mathrm{Im}(x) > 0 \\ = 0 & \mathrm{Im}(x) = 0 \\ > 0 & \mathrm{Im}(x) < 0 \end{cases}.$$

References

[1] Gauss CF (1812) Disquisitiones Generales circa Seriem Infinitam $1 + \frac{\alpha\beta}{1\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot 2\gamma(\gamma+1)}x^2 + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot 2\cdot 3\gamma(\gamma+1)(\gamma+2)}x^3 +$ etc. Pars prior, *Commentationes Societatis Regiae Scientiarium Gottingensis Recentiores* 2:1–46

[2] Jones WB, Thron WJ (1980) *Continued Fractions: Analytic Theory and Applications*. Volume 11 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley, Reading, MA
[3] Lorentzen L, Waadeland H (1992) *Continued Fractions with Applications*. North-Holland, Amsterdam
[4] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[5] Wall HS (1948) *Analytic Theory of Continued Fractions*. Van Nostrand, New York
[6] Wallis J (1695) in *Opera Mathematica Vol 1*. Oxoniae e Theatro Shedoniano, reprinted by Georg Olms Verlag, Hildeshein, New York, 1972, p 355

4
Asymptotic Expansions

4.1 Introduction

Asymptotic analysis is a branch of mathematics dealing with the order of magnitude and limiting behavior of functions, particularly at boundary points of their domains of definition [1, 2, 4, 5, 7]. Consider, for instance, the function

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

It is obvious that $f(x)$ resembles the function x as $x \to \infty$. However, one can be more precise. The expansion

$$\begin{aligned} f(x) &= \frac{x^2+1}{x(1+\frac{1}{x})} \\ &= \Big(x+\frac{1}{x}\Big)\sum_{k=0}^{\infty}\Big(\frac{-1}{x}\Big)^k \\ &= x-1-2\sum_{k=1}^{\infty}\Big(\frac{-1}{x}\Big)^k \end{aligned}$$

indicates that $f(x)$ more closely resembles $x-1$ for large x. Furthermore, $f(x)-x+1$ behaves like $2/x$ for large x. We can refine the precision of the approximation by taking more terms in the infinite series. How far we

continue in this and other problems is usually dictated by the application at hand.

4.2 Order Relations

Order relations are central to the development of asymptotic analysis. Suppose we have two functions $f(x)$ and $g(x)$ defined on a common interval I, which may extend to ∞ on the right or to $-\infty$ on the left. Let x_0 be either an internal point or a boundary point of I with $g(x) \neq 0$ for x close, but not equal, to x_0. Then the function $f(x)$ is said to be $O(g(x))$ if there exists a constant M such that $|f(x)| \leq M|g(x)|$ as $x \to x_0$. If $\lim_{x\to x_0} f(x)/g(x) = 0$, then $f(x)$ is said to be $o(g(x))$. Obviously, the relation $f(x) = o(g(x))$ implies the weaker relation $f(x) = O(g(x))$. Finally, if $\lim_{x\to x_0} f(x)/g(x) = 1$, then $f(x)$ is said to be asymptotic to $g(x)$. This is usually written $f(x) \asymp g(x)$. In many problems, the functions $f(x)$ and $g(x)$ are defined on the integers $\{1, 2, \ldots\}$ instead of on an interval I, and x_0 is taken as ∞.

For example, on $I = (1, \infty)$ one has $e^x = O(\sinh x)$ as $x \to \infty$ because

$$\frac{e^x}{\dfrac{e^x - e^{-x}}{2}} = \frac{2}{1 - e^{-2x}} \leq \frac{2}{1 - e^{-2}}.$$

On $(0, \infty)$ one has $\sin^2 x = o(x)$ as $x \to 0$ because

$$\begin{aligned} \lim_{x\to 0} \frac{\sin^2 x}{x} &= \lim_{x\to 0} \sin x \lim_{x\to 0} \frac{\sin x}{x} \\ &= 0 \times 1. \end{aligned}$$

On $I = (0, \infty)$, our initial example can be rephrased as $(x^2+1)/(x+1) \asymp x$ as $x \to \infty$.

If $f(x)$ is bounded in a neighborhood of x_0, then we write $f(x) = O(1)$ as $x \to x_0$, and if $\lim_{x\to x_0} f(x) = 0$, we write $f(x) = o(1)$ as $x \to x_0$. The notation $f(x) = g(x) + O(h(x))$ means $f(x) - g(x) = O(h(x))$ and similarly for the o notation. For example,

$$\frac{x^2 + 1}{x + 1} = x - 1 + O\Big(\frac{1}{x}\Big).$$

If $f(x)$ is differentiable at point x_0, then

$$f(x_0 + h) - f(x_0) = f'(x_0)h + o(h).$$

There is a host of miniature theorems dealing with order relations. Among these are

$$\begin{aligned} O(g) + O(g) &= O(g) \\ o(g) + o(g) &= o(g) \end{aligned}$$

$$\begin{aligned} O(g_1)O(g_2) &= O(g_1 g_2) \\ o(g_1)O(g_2) &= o(g_1 g_2) \\ |O(g)|^\lambda &= O(|g|^\lambda), \quad \lambda > 0 \\ |o(g)|^\lambda &= o(|g|^\lambda), \quad \lambda > 0. \end{aligned}$$

4.3 Finite Taylor Expansions

One easy way of generating approximations to a function is via finite Taylor expansions. Suppose $f(x)$ has $n+1$ continuous derivatives near $x_0 = 0$. Then

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k + O(x^{n+1})$$

as $x \to 0$. This order relation is validated by l'Hopital's rule applied $n+1$ times to the quotient

$$\frac{f(x) - \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k}{x^{n+1}}.$$

Of course, it is more informative to write the Taylor expansion with an explicit error term; for instance,

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k + \frac{x^{n+1}}{n!} \int_0^1 f^{(n+1)}(tx)(1-t)^n dt. \tag{1}$$

This integral $\frac{x^{n+1}}{n!} \int_0^1 f^{(n+1)}(tx)(1-t)^n dt$ form of the remainder $R_n(x)$ after n terms can be derived by noting the recurrence relation

$$R_n(x) = -\frac{x^n}{n!} f^{(n)}(0) + R_{n-1}(x)$$

and the initial condition

$$R_0(x) = f(x) - f(0),$$

both of which follow from integration by parts. One virtue of formula (1) emerges when the derivatives of $f(x)$ satisfy $(-1)^k f^{(k)}(x) \geq 0$ for all $k > 0$. If this condition holds, then

$$\begin{aligned} 0 &\leq (-1)^{n+1} R_n(x) \\ &= \frac{x^{n+1}}{n!} \int_0^1 (-1)^{n+1} f^{(n+1)}(tx)(1-t)^n dt \\ &\leq \frac{x^{n+1}}{n!} (-1)^{n+1} f^{(n+1)}(0) \int_0^1 (1-t)^n dt \\ &= \frac{x^{n+1}}{(n+1)!} (-1)^{n+1} f^{(n+1)}(0) \end{aligned}$$

for any $x > 0$. In other words, the remainders $R_n(x)$ alternate in sign and are bounded in absolute value by the next term of the expansion. As an example, the function $f(x) = -\ln(1+x)$ satisfies the inequalities $(-1)^k f^{(k)}(x) \geq 0$ and consequently also an infinity of Taylor expansion inequalities beginning with $0 \leq -\ln(1+x) + x \leq x^2/2$.

In large sample theory, finite Taylor expansions are invoked to justify asymptotic moment formulas for complicated random variables. The next proposition [6] is one species of a genus of results.

Proposition 4.3.1. *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with common mean $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Suppose that I is some interval with $\Pr(X_i \in I) = 1$ and that $m \geq 4$ is an even integer such that the first m moments of X_i exist. If $h(x)$ is any function whose mth derivative $h^{(m)}(x)$ is bounded on I, then the sample mean $A_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies*

$$\mathrm{E}[h(A_n)] = h(\mu) + \frac{\sigma^2}{2n}h''(\mu) + O\Big(\frac{1}{n^2}\Big) \tag{2}$$

as $n \to \infty$. If $h(x)^2$ satisfies the same hypothesis as $h(x)$ with a possibly different m, then

$$\mathrm{Var}[h(A_n)] = \frac{\sigma^2}{n}h'(\mu)^2 + O\Big(\frac{1}{n^2}\Big) \tag{3}$$

as $n \to \infty$.

Proof. Let us begin by finding the order of magnitude of the kth moment μ_{nk} of the centered sum $S_n = \sum_{i=1}^n (X_i - \mu)$. We claim that μ_{nk} is a polynomial in n of degree $\lfloor k/2 \rfloor$ or less, where $\lfloor \cdot \rfloor$ is the least integer function. This assertion is certainly true for $k \leq 2$ because $\mu_{n0} = 1$, $\mu_{n1} = 0$, and $\mu_{n2} = n\sigma^2$. The general case can be verified by letting c_j be the jth cumulant of $X_i - \mu$. Because a cumulant of a sum of independent random variables is the sum of the cumulants, nc_j is the jth cumulant of S_n. According to our analysis in Chapter 2, we can convert cumulants to moments via

$$\begin{aligned}\mu_{nk} &= \sum_{j=0}^{k-1} \binom{k-1}{j} nc_{k-j}\mu_{nj} \\ &= \sum_{j=0}^{k-2} \binom{k-1}{j} nc_{k-j}\mu_{nj},\end{aligned}$$

where the fact $c_1 = 0$ permits us to omit the last term in the sum. This formula and mathematical induction evidently imply that μ_{nk} is a polynomial in n whose degree satisfies

$$\deg \mu_{nk} \leq 1 + \max_{0 \leq j \leq k-2} \deg \mu_{nj}$$

$$\begin{aligned} &\leq 1 + \left\lfloor \frac{k-2}{2} \right\rfloor \\ &= \left\lfloor \frac{k}{2} \right\rfloor. \end{aligned}$$

This calculation validates the claim.

Now consider the Taylor expansion

$$h(A_n) - \sum_{k=0}^{m-1} \frac{h^{(k)}(\mu)}{k!}(A_n - \mu)^k = \frac{h^{(m)}(\eta)}{m!}(A_n - \mu)^m \tag{4}$$

for η between A_n and μ. In view of the fact that $|h^{(m)}(A_n)| \leq b$ for some constant b and all possible values of A_n, taking expectations in equation (4) yields

$$\left| \mathrm{E}[h(A_n)] - \sum_{k=0}^{m-1} \frac{h^{(k)}(\mu)}{k!} \frac{\mu_{nk}}{n^k} \right| \leq \frac{b}{m!} \frac{\mu_{nm}}{n^m}. \tag{5}$$

Because μ_{nk} is a polynomial of degree at most $\lfloor k/2 \rfloor$ in n, the factor μ_{nk}/n^k is $O\Big(n^{-k+\lfloor k/2 \rfloor}\Big)$. This fact in conjunction with inequality (5) clearly gives the expansion (2).

If $h(x)^2$ satisfies the same hypothesis as $h(x)$, then

$$\mathrm{E}[h(A_n)^2] = h(\mu)^2 + \frac{\sigma^2}{2n} 2[h(\mu)h''(\mu) + h'(\mu)^2] + O\Big(\frac{1}{n^2}\Big).$$

Straightforward algebra now indicates that the difference

$$\mathrm{Var}[h(A_n)] = \mathrm{E}[h(A_n)^2] - \mathrm{E}[h(A_n)]^2$$

takes the form (3). □

The proposition is most easily applied if the X_i are bounded or $h(x)$ is a polynomial. For example, if the X_i are Bernoulli random variables with success probability p, then $h(A_n) = A_n(1 - A_n)$ is the maximum likelihood estimate of the Bernoulli variance $\sigma^2 = p(1-p)$. Proposition 4.3.1 implies

$$\begin{aligned} \mathrm{E}[A_n(1 - A_n)] &= p(1-p) - \frac{p(1-p)2}{2n} \\ &= (1 - \frac{1}{n})p(1-p) \\ \mathrm{Var}[A_n(1 - A_n)] &= \frac{p(1-p)(1-2p)^2}{n} + O\Big(\frac{1}{n^2}\Big). \end{aligned}$$

The expression for the mean $\mathrm{E}[A_n(1 - A_n)]$ is exact since the third and higher derivatives of $h(x) = x(1-x)$ vanish.

4.4 Expansions via Integration by Parts

Integration by parts often works well as a formal device for generating asymptotic expansions. Here are three examples.

4.4.1 Exponential Integral

Suppose Y has exponential density e^{-y} with unit mean. Given Y, let a point X be chosen uniformly from the interval $[0, Y]$. Then it is easy to show that X has density $E_1(x) = \int_x^\infty e^{-y}y^{-1}dy$ and distribution function $1 - e^{-x} + xE_1(x)$. To generate an asymptotic expansion of the exponential integral $E_1(x)$ as $x \to \infty$, one can repeatedly integrate by parts. This gives

$$\begin{aligned} E_1(x) &= -\frac{e^{-y}}{y}\Big|_x^\infty - \int_x^\infty \frac{e^{-y}}{y^2}dy \\ &= \frac{e^{-x}}{x} + \frac{e^{-y}}{y^2}\Big|_x^\infty + 2\int_x^\infty \frac{e^{-y}}{y^3}dy \\ &\vdots \\ &= e^{-x}\sum_{k=1}^{n}(-1)^{k-1}\frac{(k-1)!}{x^k} + (-1)^n n!\int_x^\infty \frac{e^{-y}}{y^{n+1}}dy. \end{aligned}$$

This is emphatically not a convergent series in powers of $1/x$. In fact, for any fixed x, we have $\lim_{k\to\infty} |(-1)^{(k-1)}(k-1)!/x^k| = \infty$.

Fortunately, the remainders $R_n(x) = (-1)^n n! \int_x^\infty e^{-y}y^{-n-1}dy$ alternate in sign and are bounded in absolute value by

$$\begin{aligned} |R_n(x)| &\le \frac{n!}{x^{n+1}}\int_x^\infty e^{-y}dy \\ &= \frac{n!}{x^{n+1}}e^{-x}, \end{aligned}$$

the absolute value of the next term of the expansion. This suggests that we truncate the expansion when n is the largest integer with

$$\frac{\frac{n!}{x^{n+1}}e^{-x}}{\frac{(n-1)!}{x^n}e^{-x}} \le 1.$$

In other words, we should choose $n \approx x$. If we include more terms, then the approximation degrades. This is in striking contrast to what happens with a convergent series.

Table 4.1 illustrates these remarks by tabulating a few representative values of the functions

$$I(x) = xe^x E_1(x)$$

TABLE 4.1. Asymptotic Approximation of the Exponential Integral

x	$I(x)$	$S_1(x)$	$S_2(x)$	$S_3(x)$	$S_4(x)$	$S_5(x)$	$S_6(x)$
1	0.59634	1.0	0.0	2.0	−4.0		
2	0.72266	1.0	0.5	1.0	0.25	1.75	
3	0.78625	1.0	0.667	0.8999	0.6667	0.9626	0.4688
5	0.85212	1.0	0.8	0.88	0.8352	0.8736	0.8352

$$S_n(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{(k-1)!}{x^{k-1}}.$$

For larger values of x, the approximation noticeably improves. For instance, $I(10) = 0.91563$ while $S_{10}(10) = 0.91544$ and $I(100) = 0.99019 = S_4(100)$.

4.4.2 Incomplete Gamma Function

Repeated integration by parts of the right-tail probability of a gamma distributed random variable produces in the same manner

$$\frac{1}{\Gamma(a)} \int_x^\infty y^{a-1} e^{-y} dy$$
$$= x^a e^{-x} \sum_{k=1}^{n} \frac{1}{x^k \Gamma(a-k+1)} + \frac{1}{\Gamma(a-n)} \int_x^\infty y^{a-n-1} e^{-y} dy.$$

If a is a positive integer, then the expansion stops at $n = a$ with remainder 0. Otherwise, if n is so large that $a - n - 1$ is negative, then the remainder satisfies

$$\left| \frac{1}{\Gamma(a-n)} \int_x^\infty y^{a-n-1} e^{-y} dy \right| \le \left| \frac{1}{\Gamma(a-n)} \right| x^{a-n-1} e^{-x}.$$

Reasoning as above, we deduce that it is optimal to truncate the expansion when $|a - n|/x \approx 1$. The right-tail probability

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{y^2}{2}} dy = \frac{1}{2\Gamma(\frac{1}{2})} \int_{\frac{x^2}{2}}^\infty z^{\frac{1}{2}-1} e^{-z} dz$$

of the standard normal random variable is covered by the special case $a = 1/2$ for $x > 0$; namely,

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{y^2}{2}} dy = \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}} \Big(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{3 \cdot 5}{x^6} + \cdots \Big).$$

4.4.3 Laplace Transforms

The Laplace transform of a function $f(x)$ is defined by

$$\hat{f}(\lambda) = \int_0^\infty e^{-\lambda x} f(x) dx.$$

Repeated integration by parts yields

$$\begin{aligned}\hat{f}(\lambda) &= \frac{f(0)}{\lambda} + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x} f'(x) dx \\ &\vdots \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(0)}{\lambda^{k+1}} + \frac{1}{\lambda^{n+1}}\int_0^\infty e^{-\lambda x} f^{(n+1)}(x) dx,\end{aligned}$$

provided $f(x)$ is sufficiently well behaved that the required derivatives $f^{(k)}(0)$ and integrals $\int_0^\infty e^{-\lambda x}|f^{(k)}(x)|dx$ exist. The remainder satisfies

$$\lambda^{-n-1}\int_0^\infty e^{-\lambda x} f^{(n+1)}(x) dx = o(\lambda^{-n-1})$$

as $\lambda \to \infty$. Watson's lemma significantly generalizes this result [7].

4.5 General Definition of an Asymptotic Expansion

The previous examples suggest Poincaré's definition of an asymptotic expansion. Let $\phi_n(x)$ be a sequence of functions such that $\phi_{n+1} = o(\phi_n(x))$ as $x \to x_0$. Then $\sum_{k=1}^{\infty} c_k \phi_k(x)$ is an asymptotic expansion for $f(x)$ if $f(x) = \sum_{k=1}^{n} c_k \phi_k(x) + o(\phi_n(x))$ holds as $x \to x_0$ for every $n \geq 1$. The constants c_n are uniquely determined by the limits

$$c_n = \lim_{x \to x_0} \frac{f(x) - \sum_{k=1}^{n-1} c_k \phi_k(x)}{\phi_n(x)}$$

taken recursively starting with $c_1 = \lim_{x \to x_0} f(x)/\phi_1(x)$. Implicit in this definition is the assumption that $\phi_n(x) \neq 0$ for x close, but not equal, to x_0.

4.6 Laplace's Method

Laplace's method gives asymptotic approximations for integrals

$$\int_c^d f(y) e^{-xg(y)} dy \tag{6}$$

depending on a parameter x as $x \to \infty$. Here the boundary points c and d can be finite or infinite. There are two cases of primary interest. If c is finite, and the minimum of $g(y)$ occurs at c, then the contributions to the integral around c dominate as $x \to \infty$. Without loss of generality, let us take $c = 0$ and $d = \infty$. (If d is finite, then we can extend the range of integration by defining $f(x) = 0$ to the right of d.) Now the supposition that the dominant contributions occur around 0 suggests that we can replace $f(y)$ by $f(0)$ and $g(y)$ by its first-order Taylor expansion $g(y) \approx g(0) + g'(0)y$. Making these substitutions leads us to conjecture that

$$\begin{aligned}\int_0^\infty f(y)e^{-xg(y)}dy &\asymp f(0)e^{-xg(0)}\int_0^\infty e^{-xyg'(0)}dy \\ &= \frac{f(0)e^{-xg(0)}}{xg'(0)}. \end{aligned} \tag{7}$$

In essence, we have reduced the integral to integration against the exponential density with mean $[xg'(0)]^{-1}$. As this mean approaches 0, the approximation becomes better and better. Under the weaker assumption that $f(y) \asymp ay^{b-1}$ as $y \to 0$ for $b > 0$, the integral (6) can be replaced by an integral involving a gamma density. In this situation,

$$\int_0^\infty f(y)e^{-xg(y)}dy \asymp \frac{a\Gamma(b)e^{-xg(0)}}{[xg'(0)]^b} \tag{8}$$

as $x \to \infty$.

The other case occurs when $g(y)$ assumes its minimum at an interior point, say 0, between, say, $c = -\infty$ and $d = \infty$. Now we replace $g(y)$ by its second-order Taylor expansion $g(y) = g(0) + \frac{1}{2}g''(0)y^2 + o(y^2)$. Assuming that the region around 0 dominates, we conjecture that

$$\begin{aligned}\int_{-\infty}^\infty f(y)e^{-xg(y)}dy &\asymp f(0)e^{-xg(0)}\int_{-\infty}^\infty e^{-\frac{xg''(0)y^2}{2}}dy \\ &= f(0)e^{-xg(0)}\sqrt{\frac{2\pi}{xg''(0)}}. \end{aligned} \tag{9}$$

In other words, we reduce the integral to integration against the normal density with mean 0 and variance $[xg''(0)]^{-1}$. As this variance approaches 0, the approximation improves.

The asymptotic equivalences (8) and (9) and their generalizations constitute Laplace's method. Before rigorously stating and proving the second of these conjectures, let us briefly consider some applications.

4.6.1 Moments of an Order Statistic

Our first application of Laplace's method involves a problem in order statistics. Let $X_1, \ldots, X_n$ be i.i.d. positive, random variables with common distribution function $F(x)$. We assume that $F(x) \asymp ax^b$ as $x \to 0$.

Now consider the first order statistic $X_{(1)} = \min_{1 \le i \le n} X_i$. One can express the kth moment of $X_{(1)}$ in terms of its right-tail probability

$$\Pr(X_{(1)} > x) = [1 - F(x)]^n$$

as

$$\begin{aligned} \mathrm{E}(X_{(1)}^k) &= k \int_0^\infty x^{k-1} [1 - F(x)]^n dx \\ &= k \int_0^\infty x^{k-1} e^{n \ln[1 - F(x)]} dx \\ &= \frac{k}{b} \int_0^\infty u^{\frac{k}{b}-1} e^{n \ln[1 - F(u^{\frac{1}{b}})]} du, \end{aligned}$$

where the last integral arises from the change of variable $u = x^b$. Now the function $g(u) = -\ln[1 - F(u^{\frac{1}{b}})]$ has its minimum at $u = 0$, and an easy calculation invoking the assumption $F(x) \asymp ax^b$ yields $g'(0) = a$. Hence, the first form (8) of Laplace's method implies

$$\mathrm{E}(X_{(1)}^k) \asymp \frac{k\Gamma(\frac{k}{b})}{b(na)^{\frac{k}{b}}}. \tag{10}$$

This asymptotic equivalence has an amusing consequence for a birthday problem. Suppose that people are selected one by one from a large crowd until two of the chosen people share a birthday. We would like to know how many people are selected on average before a match occurs. One way of conceptualizing this problem is to imagine drawing people at random times dictated by a Poisson process with unit intensity. The expected time until the first match then coincides with the expected number of people drawn [3]. Since the choice of a birthday from the available $n = 365$ days of the year is made independently for each random draw, we are in effect watching the evolution of n independent Poisson processes, each with intensity $1/n$.

Let X_i be the time when the second random point happens in the ith process. The time when the first birthday match occurs in the overall process is $X_{(1)} = \min_{1 \le i \le n} X_i$. Now X_i has right-tail probability

$$\Pr(X_i > x) = (1 + \frac{x}{n}) e^{-\frac{x}{n}}$$

because 0 or 1 random points must occur on $[0, x]$ in order for $X_i > x$. It follows that X_i has distribution function

$$\begin{aligned} \Pr(X_i \le x) &= 1 - (1 + \frac{x}{n}) e^{-\frac{x}{n}} \\ &\asymp \frac{x^2}{2n^2}, \end{aligned}$$

and according to our calculation (10) with $a = 1/(2n^2)$, $b = 2$, and $k = 1$,

$$\mathrm{E}(X_{(1)}) \asymp \frac{\Gamma(\frac{1}{2})}{2(n\frac{1}{2n^2})^{\frac{1}{2}}} = \frac{1}{2}\sqrt{2\pi n}.$$

For $n = 365$ we get $\mathrm{E}(X_{(1)}) \approx 23.9$, a reasonably close approximation to the true value of 24.6.

4.6.2 Stirling's Formula

The behavior of the gamma function

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

as $\lambda \to \infty$ can be ascertained by Laplace's method. If we define $z = y/\lambda$, then

$$\Gamma(\lambda + 1) = \lambda^{\lambda+1} \int_0^\infty e^{-\lambda g(z)} dz$$

for the function $g(z) = z - \ln z$, which has its minimum at $z = 1$. Applying Laplace's second approximation (9) at $z = 1$ gives Stirling's asymptotic formula

$$\Gamma(\lambda + 1) \asymp \sqrt{2\pi} \lambda^{\lambda+\frac{1}{2}} e^{-\lambda}$$

as $\lambda \to \infty$.

4.6.3 Posterior Expectations

In Bayesian calculations one is often confronted with the need to evaluate the posterior expectation

$$\frac{\int e^{h(\theta)} e^{l_n(\theta)+\pi(\theta)} d\theta}{\int e^{l_n(\theta)+\pi(\theta)} d\theta} \tag{11}$$

of some function $e^{h(\theta)}$ of the parameter θ. In formula (11), $\pi(\theta)$ is the logprior and $l_n(\theta)$ is the loglikelihood of n observations. If n is large and the observations are independent, then usually the logposterior $l_n(\theta)+\pi(\theta)$ is sharply peaked in the vicinity of the posterior mode $\hat{\theta}$.

In the spirit of Laplace's method, this suggests that the denominator in (11) can be approximated by

$$\begin{aligned} \int e^{l_n(\theta)+\pi(\theta)} d\theta &\approx e^{l_n(\hat{\theta})+\pi(\hat{\theta})} \int e^{\frac{1}{2}[l_n''(\hat{\theta})+\pi''(\hat{\theta})](\theta-\hat{\theta})^2} d\theta \\ &= e^{l_n(\hat{\theta})+\pi(\hat{\theta})} \sqrt{\frac{2\pi}{-[l_n''(\hat{\theta}) + \pi''(\hat{\theta})]}}. \end{aligned}$$

If we also approximate the numerator of (11) by expanding the sum $h(\theta) + l_n(\theta) + \pi(\theta)$ around its maximum point $\tilde{\theta}$, then the ratio (11) can be approximated by

$$\frac{\int e^{h(\theta)} e^{l_n(\theta)+\pi(\theta)} d\theta}{\int e^{l_n(\theta)+\pi(\theta)} d\theta}$$

$$\approx e^{[h(\tilde{\theta})+l_n(\tilde{\theta})+\pi(\tilde{\theta})-l_n(\hat{\theta})-\pi(\hat{\theta})]}\sqrt{\frac{l_n''(\hat{\theta})+\pi''(\hat{\theta})}{h''(\tilde{\theta})+l_n''(\tilde{\theta})+\pi''(\tilde{\theta})}}. \tag{12}$$

The major virtue of this approximation is that it substitutes optimization for integration. The approximation extends naturally to multidimensional settings, where the difficulty of integration is especially acute. Tierney and Kadane [8] provide a detailed analysis of the order of magnitude of the errors committed in using formula (12).

4.7 Validation of Laplace's Method

Here we undertake a formal proof of the second Laplace asymptotic formula (9). Proof of the first formula (7) is similar.

Proposition 4.7.1. *If the conditions*

(a) for every $\delta > 0$ there exists a $\rho > 0$ with $g(y) - g(0) \geq \rho$ for $|y| \geq \delta$,
(b) $g(y)$ is twice continuously differentiable in a neighborhood of 0 and $g''(0) > 0$,
(c) $f(y)$ is continuous in a neighborhood of 0 and $f(0) > 0$,
(d) the integral $\int_{-\infty}^{\infty} f(y)e^{-xg(y)}dy$ is absolutely convergent for $x \geq x_1$,

are satisfied, then the asymptotic relation (9) obtains.

Proof. By multiplying both sides of the asymptotic relation (9) by $e^{xg(0)}$, we can assume without loss of generality that $g(0) = 0$. Because $g(y)$ has its minimum at $y = 0$, l'Hopital's rule implies $g(y) - \frac{1}{2}g''(0)y^2 = o(y^2)$ as $y \to 0$. Now let a small $\epsilon > 0$ be given, and choose $\delta > 0$ sufficiently small so that the inequalities

$$\begin{aligned}(1-\epsilon)f(0) &\leq f(y)\\ &\leq (1+\epsilon)f(0)\\ |g(y) - \frac{1}{2}g''(0)y^2| &\leq \epsilon y^2\end{aligned}$$

hold for $|y| \leq \delta$. Assumption (a) guarantees the existence of a $\rho > 0$ with $g(y) \geq \rho$ for $|y| \geq \delta$.

We next show that the contributions to the Laplace integral from the region $|y| \geq \delta$ are negligible as $x \to \infty$. Indeed, for $x \geq x_1$,

$$\begin{aligned}\left|\int_{\delta}^{\infty} f(y)e^{-xg(y)}dy\right| &\leq \int_{\delta}^{\infty} |f(y)|e^{-(x-x_1)g(y)}e^{-x_1g(y)}dy\\ &\leq e^{-(x-x_1)\rho}\int_{\delta}^{\infty} |f(y)|e^{-x_1g(y)}dy\\ &= O(e^{-\rho x}).\end{aligned}$$

Likewise, $\int_{-\infty}^{-\delta} f(y)e^{-xg(y)}dy = O(e^{-\rho x})$.

Owing to our choice of δ, the central portion of the integral satisfies

$$\int_{-\delta}^{\delta} f(y)e^{-xg(y)}dy \leq (1+\epsilon)f(0)\int_{-\delta}^{\delta} e^{-\frac{x}{2}[g''(0)-2\epsilon]y^2}dy.$$

Duplicating the above reasoning,

$$\int_{-\infty}^{-\delta} e^{-\frac{x}{2}[g''(0)-2\epsilon]y^2}dy + \int_{\delta}^{\infty} e^{-\frac{x}{2}[g''(0)-2\epsilon]y^2}dy = O(e^{-\omega x}),$$

where $\omega = \frac{1}{2}[g''(0)-2\epsilon]\delta^2$. Thus,

$$\begin{aligned}
&(1+\epsilon)f(0)\int_{-\delta}^{\delta} e^{-\frac{x}{2}[g''(0)-2\epsilon]y^2}dy \\
&\quad = (1+\epsilon)f(0)\int_{-\infty}^{\infty} e^{-\frac{x}{2}[g''(0)-2\epsilon]y^2}dy + O(e^{-\omega x}) \\
&\quad = (1+\epsilon)f(0)\sqrt{\frac{2\pi}{x[g''(0)-2\epsilon]}} + O(e^{-\omega x}).
\end{aligned}$$

Assembling all of the relevant pieces, we now conclude that

$$\begin{aligned}
\int_{-\infty}^{\infty} f(y)e^{-xg(y)}dy \leq (1+\epsilon)f(0)&\sqrt{\frac{2\pi}{x[g''(0)-2\epsilon]}} \\
&+ O(e^{-\rho x}) + O(e^{-\omega x}).
\end{aligned}$$

Hence,

$$\limsup_{x\to\infty} \sqrt{x}\int_{-\infty}^{\infty} f(y)e^{-xg(y)}dy \leq (1+\epsilon)f(0)\sqrt{\frac{2\pi}{[g''(0)-2\epsilon]}},$$

and sending $\epsilon \to 0$ produces

$$\limsup_{x\to\infty} \sqrt{x}\int_{-\infty}^{\infty} f(y)e^{-xg(y)}dy \leq f(0)\sqrt{\frac{2\pi}{g''(0)}}.$$

A similar argument gives

$$\liminf_{x\to\infty} \sqrt{x}\int_{-\infty}^{\infty} f(y)e^{-xg(y)}dy \geq f(0)\sqrt{\frac{2\pi}{g''(0)}}$$

and proves the proposition. □

4.8 Problems

1. Prove the following order relations:

(a) $1-\cos^2 x = O(x^2)$ as $x \to 0$,

(b) $\ln x = o(x^\alpha)$ as $x \to \infty$ for any $\alpha > 0$,
(c) $\frac{x^2}{1+x^3} + \ln(1+x^2) = O(x^2)$ as $x \to 0$,
(d) $\frac{x^2}{1+x^3} + \ln(1+x^2) = O(\ln x)$ as $x \to \infty$.

2. Show that $f(x) \asymp g(x)$ as $x \to x_0$ does not entail the stronger relation $e^{f(x)} \asymp e^{g(x)}$ as $x \to x_0$. Argue that the condition $f(x) = g(x) + o(1)$ is sufficient to imply $e^{f(x)} \asymp e^{g(x)}$.

3. For two positive functions $f(x)$ and $g(x)$, prove that $f(x) \asymp g(x)$ as $x \to x_0$ implies $\ln f(x) = \ln g(x) + o(1)$ as $x \to x_0$. Hence, $\lim_{x\to x_0} \ln f(x) \neq 0$ entails $\ln f(x) \asymp \ln g(x)$ as $x \to x_0$.

4. Suppose in Proposition 4.3.1 we replace $h(A_n)$ by $h(c_n A_n)$, where the sequence of constants $c_n = 1 + an^{-1} + O(n^{-2})$. How does this change the right hand sides of the asymptotic expressions (2) and (3)?

5. Continuing Problem 4, derive asymptotic expressions for the mean and variance of $\Phi\Big[(u - A_n)\sqrt{n/(n-1)}\Big]$, where u is a constant, A_n is the sample mean of a sequence $X_1, \ldots, X_n$ of i.i.d. normal random variables with mean μ and variance 1, and $\Phi(x)$ is the standard normal distribution function. The statistic $\Phi\Big[(u - A_n)\sqrt{n/(n-1)}\Big]$ is the uniformly minimum variance unbiased estimator of the percentile $p = \Phi(X_i \leq u)$ [6].

6. Find an asymptotic expansion for $\int_x^\infty e^{-y^4} dy$ as $x \to \infty$.

7. Suppose that $0 < c < \infty$ and that $f(x)$ is bounded and continuous on $[0, c]$. If $f(c) \neq 0$, then show that

$$\int_0^c x^n f(x) dx \asymp \frac{c^{n+1}}{n} f(c)$$

as $n \to \infty$.

8. Let $F(x)$ be a distribution function concentrated on $[0, \infty)$ with moments $m_k = \int_0^\infty y^k dF(y)$. For $x \geq 0$ define the Stieltjes function $f(x) = \int_0^\infty \frac{1}{1+xy} dF(y)$. Show that $\sum_{k=0}^\infty (-1)^k m_k x^k$ is an asymptotic expansion for $f(x)$ satisfying

$$f(x) - \sum_{k=0}^{n} (-1)^k m_k x^k = (-x)^{n+1} \int_0^\infty \frac{y^{n+1}}{1+xy} dF(y).$$

Argue, therefore, that the remainders of the expansion alternate in sign and are bounded in absolute value by the first omitted term.

9. Show that $\int_0^\infty \frac{e^{-y}}{1+xy} dy \asymp \frac{\ln x}{x}$ as $x \to \infty$. (Hints: Write

$$\int_0^\infty \frac{e^{-y}}{1+xy} dy = \frac{1}{x} \int_0^\infty \frac{d}{dy} \ln(1+xy) e^{-y} dy,$$

and use integration by parts and the dominated convergence theorem.)

10. Prove that

$$\int_0^{\frac{\pi}{2}} e^{-x\tan y}dy \asymp \frac{1}{x}$$
$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (y+2)e^{-x\cos y}dy \asymp \frac{4}{x}$$

as $x \to \infty$.

11. For $0 < \lambda < 1$, demonstrate the asymptotic equivalence

$$\sum_{k=0}^{n} \binom{n}{k} k! n^{-k} \lambda^k \asymp \frac{1}{1-\lambda}$$

as $n \to \infty$. (Hint: Use the identity $k!n^{-k-1} = \int_0^\infty y^k e^{-ny}dy$.)

12. Demonstrate the asymptotic equivalence

$$\sum_{k=0}^{n} \binom{n}{k} k! n^{-k} \asymp \sqrt{\frac{\pi n}{2}}$$

as $n \to \infty$. (Hint: See Problem (11).)

13. The von Mises density

$$\frac{e^{\kappa\cos(y-\alpha)}}{2\pi I_0(\kappa)}, \qquad -\pi < y \leq \pi,$$

is used to model random variation on a circle. Here α is a location parameter, $\kappa > 0$ is a concentration parameter, and the modified Bessel function $I_0(\kappa)$ is the normalizing constant

$$I_0(\kappa) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\kappa\cos y}dy.$$

Verify that Laplace's method yields

$$I_0(\kappa) \asymp \frac{e^\kappa}{\sqrt{2\pi\kappa}}$$

as $\kappa \to \infty$. For large κ it is clear that the von Mises distribution is approximately normal.

References

[1] Barndorff-Nielsen OE, Cox DR (1989) *Asymptotic Techniques for Use in Statistics*. Chapman & Hall, London

[2] Bender CM, Orszag SA (1978) *Advanced Mathematical Methods for Scientists and Engineers*. McGraw-Hill, New York

[3] Bloom G, Holst L, Sandell D (1994) *Problems and Snapshots from the World of Probability*. Springer-Verlag, New York

[4] de Bruijn NG (1981) *Asymptotic Methods in Analysis*. Dover, New York

[5] Graham RL, Knuth DE, Patashnik O (1988) *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading MA
[6] Lehmann EL (1991) *Theory of Point Estimation*. Wadsworth, Belmont, CA
[7] Murray JD (1984) *Asymptotic Analysis*. Springer-Verlag, New York
[8] Tierney L, Kadane J (1986) Accurate approximations for posterior moments and marginal densities. *J Amer Stat Soc* 81:82–86

5
Solution of Nonlinear Equations

5.1 Introduction

Solving linear and nonlinear equations is a major preoccupation of applied mathematics and statistics. For nonlinear equations, closed-form solutions are the exception rather than the rule. Here we will concentrate on three simple techniques—bisection, functional iteration, and Newton's method—for solving equations in one variable. Insight into how these methods operate can be gained by a combination of theory and examples. Since functional iteration and Newton's method generalize to higher-dimensional problems, it is particularly important to develop intuition about their strengths and weaknesses. Equipped with this intuition, we can tackle harder problems with more confidence and understanding.

5.2 Bisection

Bisection is a simple, robust method of finding solutions to the equation $g(x) = 0$. In contrast to faster techniques such as Newton's method, no derivatives of $g(x)$ are required. Furthermore, under minimal assumptions on $g(x)$, bisection is guaranteed to converge to some root. Suppose that $g(x)$ is continuous, and an interval $[a, b]$ has been identified such that $g(a)$ and $g(b)$ are of opposite sign. If $g(x)$ is continuous, then the intermediate value theorem implies that $g(x)$ vanishes somewhere on $[a, b]$. Consider the midpoint $c = (a + b)/2$ of $[a, b]$. If $g(c) = 0$, then we are done. Otherwise,

either $g(a)$ and $g(c)$ are of opposite sign, or $g(b)$ and $g(c)$ are of opposite sign. In the former case, the interval $[a, c]$ brackets a root; in the latter case, the interval $[c, b]$ does. In either case, we replace $[a, b]$ by the corresponding subinterval and continue. If we bisect $[a, b]$ a total of n times, then the final bracketing interval has length $2^{-n}(b - a)$. For n large enough, we can stop and approximate the bracketed root by the midpoint of the final bracketing interval. If we want to locate nearly all of the roots of $g(x)$ on $[a, b]$, then we can subdivide $[a, b]$ into many small adjacent intervals and apply bisection to each small interval in turn.

5.2.1 Computation of Quantiles by Bisection

Suppose we are given a continuous distribution function $F(x)$ and desire to find the α-quantile of $F(x)$. This amounts to solving the equation $g(x) = 0$ for $g(x) = F(x) - \alpha$. Bisection is applicable if we can find a bracketing interval to start the process. One strategy exploiting the monotonicity of $g(x)$ is to take an arbitrary initial point a and examine $g(a)$. If $g(a) < 0$, then we look for the first positive integer k with $g(a + k) > 0$. When this integer is found, the interval $[a + k - 1, a + k]$ brackets the α-quantile. If $g(a) > 0$, then we look for the first negative integer k such that $g(a+k) < 0$. In this case $[a + k, a + k + 1]$ brackets the α-quantile. Once a bracketing interval is found, bisection can begin. An obvious candidate for a is the mean. Instead of incrementing or decrementing by 1 in finding the initial bracketing interval, it usually is preferable to increment or decrement by the standard deviation of $F(x)$.

As a numerical example, consider the problem of calculating the .95–quantile of a t distribution with $n = 5$ degrees of freedom. A random variable with this distribution has mean 0 and standard deviation $\sqrt{n/(n-2)}$. Using the search tactic indicated above, we find an initial bracketing interval of $[1.291, 2.582]$. This and the subsequent bracketing intervals produced by bisection are noted in Table 5.1.

TABLE 5.1. Bracketing Intervals Given by Bisection

Iteration n	**Interval**	**Iteration** n	**Interval**
0	[1.291,2.582]	6	[1.997,2.017]
1	[1.936,2.582]	7	[2.007,2.017]
2	[1.936,2.259]	8	[2.012,2.017]
3	[1.936,2.098]	9	[2.015,2.017]
4	[1.936,2.017]	10	[2.015,2.016]
5	[1.977,2.017]	11	[2.015,2.015]

5.2.2 Shortest Confidence Interval

In forming a confidence interval or a Bayesian credible interval, it is natural to ask for the shortest interval $[a, b]$ with fixed content $H(b) - H(a) = \alpha$ for some probability distribution $H(x)$. This problem is not always well posed. To avoid logical difficulties, let us assume that $H(x)$ possesses a density $h(x)$ and ask for the region S_α of smallest Lebesgue measure $\mu(S_\alpha)$ satisfying $\int_{S_\alpha} h(x)dx = \alpha$. If $h(x)$ is unimodal, strictly increasing to the left of its mode, and strictly decreasing to the right of its mode, then S_α is a well-defined interval, and its Lebesgue measure is just its length.

In general, the reformulated problem makes sense in m-dimensional space R^m [1]. Its solution is given by

$$S_\alpha = \{x : h(x) \geq \lambda(\alpha)\}$$

for some number $\lambda(\alpha)$ depending on α. Such a number $\lambda(\alpha)$ exists if $\int_{\{x:h(x)=\lambda\}} h(x)\,dx = 0$ for all values of λ. In fact, if X is a random vector with density $h(x)$, then this condition guarantees that the right-tail probability

$$\Pr(h(X) \geq \lambda) = \int_{\{x:h(x)\geq\lambda\}} h(x)dx$$

is continuous and decreasing as a function of λ. In view of the intermediate value theorem, at least one λ must then qualify for each $\alpha \in (0, 1)$.

The solution set S_α is unique, but only up to a set of Lebesgue measure 0. This can be checked by supposing that T also satisfies $\int_T h(x)dx = \alpha$. Subtracting this equation from the same equation for S_α yields

$$\int_{S_\alpha \setminus T} h(x)dx - \int_{T \setminus S_\alpha} h(x)dx = 0. \tag{1}$$

Because $h(x) \geq \lambda(\alpha)$ on $S_\alpha \setminus T$, it follows that

$$\int_{S_\alpha \setminus T} h(x)dx \geq \lambda(\alpha)\mu(S_\alpha \setminus T). \tag{2}$$

If $\mu(T \setminus S_\alpha) > 0$, it likewise follows that

$$\int_{T \setminus S_\alpha} h(x)dx < \lambda(\alpha)\mu(T \setminus S_\alpha). \tag{3}$$

Now if $\mu(T) \leq \mu(S_\alpha)$, then

$$\mu(T \setminus S_\alpha) \leq \mu(S_\alpha \setminus T). \tag{4}$$

The three inequalities (2), (3), and (4) are inconsistent with equality (1) unless $\mu(T \setminus S_\alpha) = 0$. But if $\mu(T \setminus S_\alpha) = 0$, then

$$\alpha = \int_T h(x)\,dx = \int_{T \cap S_\alpha} h(x)\,dx < \int_{S_\alpha} h(x)\,dx$$

TABLE 5.2. Bisection Iterates for the Shortest .95 Confidence Interval

Iteration n	c_λ	d_λ	**Iteration** n	c_λ	d_λ
1	0.2290	2.6943	9	0.0423	4.7669
2	0.1007	3.7064	10	0.0427	4.7559
3	0.0478	4.6198	11	0.0425	4.7614
4	0.0233	5.4844	12	0.0424	4.7642
5	0.0354	4.9831	13	0.0423	4.7656
6	0.0415	4.7893	14	0.0424	4.7649
7	0.0446	4.7019	15	0.0424	4.7652
8	0.0431	4.7449			

unless $\mu(S_\alpha \setminus T) = 0$. Therefore, both $\mu(T \setminus S_\alpha)$ and $\mu(S_\alpha \setminus T)$ equal 0, and S_α and T differ by at most a set of measure 0.

As a concrete illustration of this principle, consider the problem of finding the shortest interval $[c, d]$ with a fixed probability $\int_c^d h(x)dx = \alpha$ for the gamma density $h(x) = \Gamma(a)^{-1}b^a x^{a-1}e^{-bx}$. Because

$$\frac{b^a}{\Gamma(a)} \int_c^d x^{a-1}e^{-bx}dx = \frac{1}{\Gamma(a)} \int_{bc}^{bd} z^{a-1}e^{-z}dz,$$

it suffices to take the scale constant $b = 1$. If $a \leq 1$, then the gamma density $h(x)$ is strictly decreasing in x, and the left endpoint of the shortest interval is given by $c = 0$. The right endpoint d can be found by bisection using our previously devised methods of evaluating the incomplete gamma function $P(a, x)$.

If the constant $a > 1$, $h(x)$ first increases and then decreases. Its modal value $\Gamma(a)^{-1}(a-1)^{a-1}e^{-(a-1)}$ occurs at $x = a-1$. One strategy for finding the shortest interval is to consider for each λ satisfying

$$0 < \lambda < \frac{1}{\Gamma(a)}(a-1)^{a-1}e^{-(a-1)}$$

the interval $[c_\lambda, d_\lambda]$ where $h(x) \geq \lambda$. The endpoints c_λ and d_λ are implicitly defined by $h(c_\lambda) = h(d_\lambda) = \lambda$ and can be found by bisection or Newton's method. Once $[c_\lambda, d_\lambda]$ is determined, the corresponding probability

$$\frac{1}{\Gamma(a)} \int_{c_\lambda}^{d_\lambda} x^{a-1}e^{-x}dx = P(a, d_\lambda) - P(a, c_\lambda)$$

can be expressed in terms of the incomplete gamma function. Thus, the original problem reduces to finding the particular λ satisfying

$$P(a, d_\lambda) - P(a, c_\lambda) = \alpha. \tag{5}$$

This λ can be straightforwardly computed by bisection. Note that this iterative process involves inner iterations to find c_λ and d_λ within each outer bisection iteration on λ.

Table 5.2 displays the endpoints c_λ and d_λ generated by the successive midpoints λ in a bisection scheme to find the particular λ satisfying equation (5) for $a = 2$ and $\alpha = 0.95$.

5.3 Functional Iteration

Suppose we are interested in finding a root of the equation $g(x) = 0$. If we let $f(x) = g(x) + x$, then this equation is trivially equivalent to the equation $x = f(x)$. In many examples, the iterates $x_n = f(x_{n-1})$ converge to a root of $g(x)$ starting from any point x_0 nearby. For obvious reasons, a root of $g(x)$ is said to be a fixed point of $f(x)$. Precise sufficient conditions for the existence of a unique fixed point of $f(x)$ and convergence to it are offered by the following proposition.

Proposition 5.3.1. *Suppose the function $f(x)$ defined on a closed interval I satisfies the conditions*

(a) $f(x) \in I$ whenever $x \in I$,
(b) $|f(y) - f(x)| \leq \lambda|y - x|$ for any two points x and y in I.

Then provided the Lipschitz constant λ is in $[0, 1)$, $f(x)$ has a unique fixed point $x_\infty \in I$, and the functional iterates $x_n = f(x_{n-1})$ converge to x_∞ regardless of their starting point $x_0 \in I$. Furthermore, we have the precise error estimate

$$|x_n - x_\infty| \leq \frac{\lambda^n}{1-\lambda}|x_1 - x_0|. \tag{6}$$

Proof. The inequality

$$\begin{aligned} |x_{k+1} - x_k| &= |f(x_k) - f(x_{k-1})| \\ &\leq \lambda|x_k - x_{k-1}| \\ &\vdots \\ &\leq \lambda^k|x_1 - x_0| \end{aligned}$$

implies for $m > n$ the further inequality

$$\begin{aligned} |x_n - x_m| &\leq \sum_{k=n}^{m-1} |x_k - x_{k+1}| \\ &\leq \sum_{k=n}^{m-1} \lambda^k|x_1 - x_0| \\ &\leq \frac{\lambda^n}{1-\lambda}|x_1 - x_0|. \end{aligned} \tag{7}$$

It follows from inequality (7) that x_n is a Cauchy sequence. Because the interval I is closed, the limit x_∞ of the sequence x_n exists in I. Invoking

the continuity of $f(x)$ in the defining relation $x_n = f(x_{n-1})$ shows that x_∞ is a fixed point. Existence of a fixed point $y_\infty \neq x_\infty$ in I is incompatible with the inequality

$$\begin{aligned} |x_\infty - y_\infty| &= |f(x_\infty) - f(y_\infty)| \\ &\leq \lambda |x_\infty - y_\infty|. \end{aligned}$$

Finally, the explicit bound (6) follows from inequality (7) by sending m to ∞. □

A function $f(x)$ having a Lipschitz constant $\lambda < 1$ is said to be contractive. In practice λ is taken to be any convenient upper bound of $|f'(x)|$ on the interval I. Such a choice is valid because of the mean value equality $f(x) - f(y) = f'(z)(x - y)$, where z is some number between x and y. In the vicinity of a fixed point x_∞ with $|f'(x_\infty)| < 1$, we can usually find a closed interval $I_d = [x_\infty - d, x_\infty + d]$ pertinent to the proposition. For instance, if $f(x)$ is continuously differentiable, then all sufficiently small, positive constants d yield $\lambda = \sup_{z \in I_d} |f'(z)| < 1$. Furthermore, $f(x)$ maps I_d into itself because

$$\begin{aligned} |f(x) - x_\infty| &= |f(x) - f(x_\infty)| \\ &\leq \lambda |x - x_\infty| \\ &\leq d \end{aligned}$$

for $x \in I_d$.

A fixed point x_∞ with $|f'(x_\infty)| < 1$ is said to be attractive. If x_∞ satisfies $f'(x_\infty) \in (-1, 0)$, then iterates $x_n = f(x_{n-1})$ converging to x_∞ eventually oscillate from side to side of x_∞. Convergence is eventually monotone if $f'(x_\infty) \in (0, 1)$. If the inequality $|f'(x_\infty)| > 1$ holds, then the fixed point x_∞ is said to be repelling. Indeed, the mean value theorem implies in this situation that

$$\begin{aligned} |f(x) - x_\infty| &= |f'(z)(x - x_\infty)| \\ &> |x - x_\infty| \end{aligned}$$

for all x sufficiency close to x_∞. The case $|f'(x_\infty)| = 1$ is indeterminate and requires further investigation.

5.3.1 Fractional Linear Transformations

A fractional linear transformation $f(x) = (ax + b)/(cx + d)$ maps the complex plane into itself. For the sake of simplicity, let us assume that the complex constants a, b, c, and d are real and satisfy $ad - bc \neq 0$. Now consider the possibility of finding a real root of $x = f(x)$ by functional iteration. The solutions, if any, coincide with the two roots of the quadratic equation $cx^2 + (d - a)x - b = 0$. These roots can be expressed by the

standard quadratic formula as

$$\begin{aligned} r_\pm &= \frac{-(d-a) \pm \sqrt{(d-a)^2 + 4bc}}{2c} \\ &= \frac{-(d-a) \pm \sqrt{(d+a)^2 - 4(ad-bc)}}{2c}. \end{aligned}$$

Both roots are purely real if and only if $(d+a)^2 \geq 4(ad-bc)$. Let us assume that this discriminant condition holds. Either root r is then locally attractive to the functional iterates, provided the derivative

$$\begin{aligned} f'(r) &= \frac{a}{cr+d} - \frac{(ar+b)c}{(cr+d)^2} \\ &= \frac{ad-bc}{(cr+d)^2} \end{aligned}$$

satisfies $|f'(r)| < 1$. It is locally repelling when $|f'(r)| > 1$.

Consider the product $(cr_+ + d)(cr_- + d) = ad - bc$. One of three things can happen. Either (a)

$$\begin{aligned} |cr_+ + d| &= |cr_- + d| \\ &= \sqrt{|ad-bc|}, \end{aligned}$$

or (b)

$$\begin{aligned} |cr_+ + d| &> \sqrt{|ad-bc|} \\ |cr_- + d| &< \sqrt{|ad-bc|}, \end{aligned}$$

or (c)

$$\begin{aligned} |cr_+ + d| &< \sqrt{|ad-bc|} \\ |cr_- + d| &> \sqrt{|ad-bc|}. \end{aligned}$$

Case (a) is indeterminate because $|f'(r_+)| = |f'(r_-)| = 1$. It turns out that functional iteration converges to the common root $r_+ = r_-$ when it exists [8]. Otherwise, case (a) leads to divergence unless the initial point is a root to begin with. In case (b) functional iteration converges locally to r_+ and diverges locally from r_-. This local behavior, in fact, holds globally [8]. In case (c), the opposite behavior relative to the two roots is observed. This analysis explains, for instance, why the continued fraction generated by the fractional linear transformation $f(x) = 1/(2+x)$ converges to $\sqrt{2}-1$ rather than to $-\sqrt{2}-1$.

5.3.2 Extinction Probabilities by Functional Iteration

In a branching process [2], particles reproduce independently at the end of each generation according to the same probabilistic law. Let p_k be the probability that a particle present at the current generation is replaced by k

daughter particles at the next generation. Starting with a single particle at generation 0, we can ask for the probability s_∞ that the process eventually goes extinct. To characterize s_∞, we condition on the number of daughter particles k born to the initial particle. If extinction is to occur, then each line of descent emanating from a daughter particle must die out. If there are k daughter particles and consequently k lines of descent, then by independence of reproduction, all k lines of descent go extinct with probability s_∞^k. It follows that s_∞ satisfies the functional equation $s = \sum_{k=0}^\infty p_k s^k = P(s)$, where $P(s)$ is the generating function of the progeny distribution.

One can find the extinction probability by functional iteration starting at $s = 0$. Let s_n be the probability that extinction occurs in the branching process at or before generation n. Then $s_0 = 0$, $s_1 = p_0 = P(s_0)$, and, in general, $s_{n+1} = P(s_n)$. This recurrence relation can be deduced by again conditioning on the number of daughter particles in the first generation. If extinction is to occur at or before generation $n + 1$, then extinction must occur in n additional generations or sooner for each line of descent emanating from a daughter particle of the original particle.

On probabilistic grounds it is clear that the sequence s_n increases monotonely to the extinction probability s_∞. To understand what is happening numerically, we need to know the number of fixed points of $s = P(s)$ and which of these fixed points is s_∞. Since $P''(s) = \sum_{k=2}^\infty k(k-1)p_k s^{k-2} \geq 0$, the curve $P(s)$ is convex. It starts at $P(0) = p_0 > 0$ above the diagonal line $t = s$. (Note that if $p_0 = 0$, then the process can never go extinct.) On the interval $[0, 1]$, the curve $P(s)$ and the diagonal line $t = s$ intersect in either one or two points. Figure 5.1 depicts the situation of two inter-

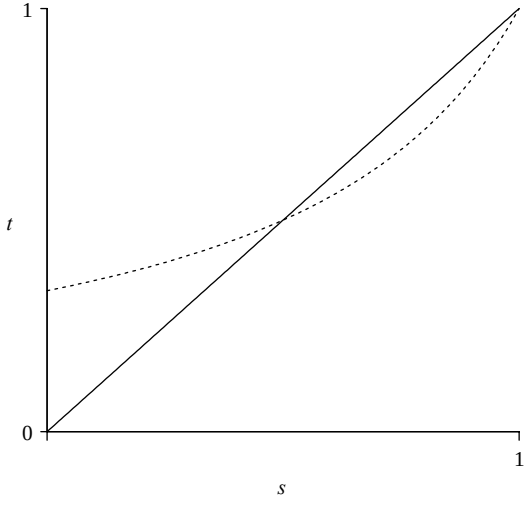

FIGURE 5.1. Intersection points for a supercritical branching process.

section points. The point $s = 1$ is certainly one intersection point because $P(1) = \sum_{k=0}^{\infty} p_k = 1$. There is a second intersection point to the left of $s = 1$ if and only if the slope of $P(s)$ at $s = 1$ is strictly greater than 1. The curve $P(s)$ then intersects $t = s$ at $s = 1$ from below. The slope $P'(1) = \sum_{k=0} kp_k$ equals the mean number of particles of the progeny distribution. Extinction is certain when the mean $P'(1) \leq 1$. When $P'(1) > 1$, the point $s = 1$ repels the iterates $s_n = P(s_{n-1})$. Hence, in this case the extinction probability is the smaller of the two fixed points of $s = P(s)$ on $[0, 1]$, and extinction is not certain.

As a numerical example, consider the data of Lotka [4, 5] on the extinction of surnames among white males in the United States. Using 1920 census data, he computed the progeny generating function

$$P(s) = .4982 + .2103s + .1270s^2 + .0730s^3 + .0418s^4 + .0241s^5 + .0132s^6 + .0069s^7 + .0035s^8 + .0015s^9 + .0005s^{10}.$$

Table 5.3 lists some representative functional iterates. Convergence to the correct extinction probability 0.880 is relatively slow.

5.4 Newton's Method

Newton's method can be motivated by the mean value theorem. Let x_{n-1} approximate the root x_∞ of the equation $g(x) = 0$. According to the mean value theorem,

$$\begin{aligned} g(x_{n-1}) &= g(x_{n-1}) - g(x_\infty) \\ &= g'(z)(x_{n-1} - x_\infty) \end{aligned}$$

for some z on the interval between x_{n-1} and x_∞. If we substitute x_{n-1} for z and the next approximant x_n for x_∞, then this equality can be rearranged to provide the definition

$$x_n = x_{n-1} - \frac{g(x_{n-1})}{g'(x_{n-1})} \tag{8}$$

TABLE 5.3. Functional Iteration for an Extinction Probability

Iteration n	**Iterate** s_n	**Iteration** n	**Iterate** s_n
0	0.000	10	0.847
1	0.498	20	0.873
2	0.647	30	0.878
3	0.719	40	0.879
4	0.761	50	0.880
5	0.788		

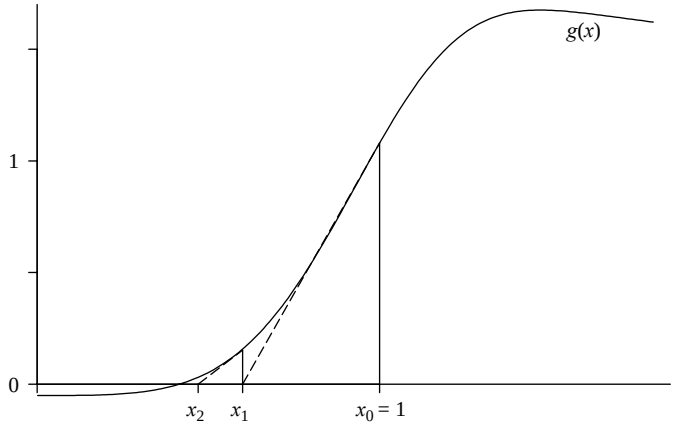

FIGURE 5.2. Two steps of Newton's method starting from $x_0 = 1$ and moving toward the unique zero of $g(x) = 1.95 - e^{-2/x} - 2e^{-x^4}$ on $(0, \infty)$. The iterate x_n is taken as the point of intersection of the x-axis and the tangent drawn through $(x_{n-1}, g(x_{n-1}))$. Newton's method fails to converge if x_0 is chosen too far to the left or right.

of Newton's method. From the perspective of functional iteration, Newton's method can be rephrased as $x_n = f(x_{n-1})$, where $f(x) = x - g(x)/g'(x)$. See Figure 5.2 for a geometric interpretation of Newton's method.

The local convergence properties of Newton's method are determined by

$$\begin{aligned} f'(x_\infty) &= 1 - \frac{g'(x_\infty)}{g'(x_\infty)} + \frac{g(x_\infty)g''(x_\infty)}{g'(x_\infty)^2} \\ &= 0. \end{aligned}$$

If we let $e_n = x_n - x_\infty$ be the current error in approximating x_∞, then executing a second-order Taylor expansion around x_∞ yields

$$\begin{aligned} e_n &= f(x_{n-1}) - f(x_\infty) \\ &= f'(x_\infty)e_{n-1} + \frac{1}{2}f''(z)e_{n-1}^2 \\ &= \frac{1}{2}f''(z)e_{n-1}^2, \end{aligned} \tag{9}$$

where z again lies between x_{n-1} and x_∞. Provided $f''(z)$ is continuous and x_0 is close enough to x_∞, the error representation (9) makes it clear that Newton's method converges and that

$$\lim_{n\to\infty} \frac{e_n}{e_{n-1}^2} = \frac{1}{2}f''(x_\infty).$$

This property is referred to as quadratic convergence. If an iteration function $f(x)$ satisfies $0 < |f'(x_\infty)| < 1$, then a first-order Taylor ex-

pansion implies $\lim_{n\to\infty} e_n/e_{n-1} = f'(x_\infty)$, which is referred to as linear convergence.

All else being equal, quadratic convergence is preferred to linear convergence. In practice, Newton's method can fail miserably if started too far from a desired root x_∞. Furthermore, it can be expensive to evaluate the derivative $g'(x)$. For these reasons, simpler, more robust methods such as bisection are often employed instead of Newton's method. The following two examples highlight favorable circumstances ensuring global convergence of Newton's method on a properly defined domain.

5.4.1 Division Without Dividing

Forming the reciprocal of a number a is equivalent to solving for a root of the equation $g(x) = a - x^{-1}$. Newton's method (8) iterates according to

$$x_n = x_{n-1} - \frac{a - x_{n-1}^{-1}}{x_{n-1}^{-2}} = x_{n-1}(2 - ax_{n-1}),$$

which involves multiplication and subtraction but no division. If x_n is to be positive, then x_{n-1} must lie on the interval $(0, 2/a)$. If x_{n-1} does indeed reside there, then x_n will reside on the shorter interval $(0, 1/a)$ because the quadratic $x(2 - ax)$ attains its maximum of $1/a$ at $x = 1/a$. Furthermore, $x_n > x_{n-1}$ if and only if $2 - ax_{n-1} > 1$, and this latter inequality holds if and only if $x_{n-1} < 1/a$. Thus, starting on $(0, 1/a)$, the iterates x_n monotonely increase to their limit $1/a$. Starting on $[1/a, 2/a)$, the first iterate satisfies $x_1 \leq 1/a$, and subsequent iterates monotonely increase to $1/a$.

5.4.2 Extinction Probabilities by Newton's Method

Newton's method offers an alternative to functional iteration in computing the extinction probability s_∞ of a branching process. If $P(s)$ is the progeny generating function, then Newton's method starts with $x_0 = 0$ and iterates according to

$$x_n = x_{n-1} + \frac{P(x_{n-1}) - x_{n-1}}{1 - P'(x_{n-1})}. \tag{10}$$

Because extinction is certain when $P'(1) \leq 1$, we will make the contrary assumption $P'(1) > 1$. For such a supercritical process, $s_\infty < 1$. Because the curve $P(s)$ intersects the diagonal line $h(s) = s$ from above at s_∞, we infer that $P'(s_\infty) < 1$ in the supercritical case. This fact is important in avoiding division by 0 in the Newton's iterates (10).

It is useful to compare the sequence (10) to the sequence $s_n = P(s_{n-1})$ generated by functional iteration. Both schemes start at 0. We will show by induction that (a) $x_n \leq s_\infty$, (b) $x_{n-1} \leq x_n$, and (c) $s_n \leq x_n$ hold for all $n \geq 0$. Conditions (a) and (c) are true by definition when $n = 0$,

while condition (b) is vacuous. In our inductive proof, we use the fact that condition (b) is logically equivalent to the condition $x_{n-1} \le P(x_{n-1})$. With this in mind, suppose all three conditions hold for an arbitrary $n \ge 0$.

Because (a) is true for n and $P'(s)$ is increasing in s, the mean value theorem implies

$$\begin{aligned} s_\infty - P(x_n) &= P(s_\infty) - P(x_n) \\ &\ge P'(x_n)(s_\infty - x_n). \end{aligned}$$

Adding $x_n - s_\infty$ to this inequality leads to

$$x_n - P(x_n) \ge [1 - P'(x_n)](x_n - s_\infty),$$

which can be divided by $1 - P'(x_n)$ to yield

$$\frac{x_n - P(x_n)}{1 - P'(x_n)} \ge x_n - s_\infty.$$

Simple rearrangement gives the desired inequality (a) for $n + 1$.

Because $x_{n-1} \le P(x_{n-1})$ and $x_{n-1} \le x_n$ both hold by the induction hypothesis, it follows that the mean value theorem and definition (10) imply

$$\begin{aligned} P(x_n) - x_n &\ge P(x_{n-1}) + P'(x_{n-1})(x_n - x_{n-1}) - x_n \\ &= P(x_{n-1}) - x_{n-1} - [1 - P'(x_{n-1})](x_n - x_{n-1}) \\ &= P(x_{n-1}) - x_{n-1} - [1 - P'(x_{n-1})]\frac{P(x_{n-1}) - x_{n-1}}{1 - P'(x_{n-1})} \\ &= 0. \end{aligned}$$

This proves the alternate form $P(x_n) \ge x_n$ of (b) for $n + 1$.

To prove condition (c), we note that condition (b) implies

$$-P(x_n) \le -x_n.$$

Multiplying this inequality by $P'(x_n)$ and then adding $P(x_n)$ yield

$$P(x_n)[1 - P'(x_n)] \le P(x_n) - x_n P'(x_n).$$

Finally, dividing by $1 - P'(x_n)$ gives

$$\begin{aligned} P(x_n) &\le \frac{P(x_n) - x_n P'(x_n)}{1 - P'(x_n)} \\ &= x_n + \frac{P(x_n) - x_n}{1 - P'(x_n)} = x_{n+1}. \end{aligned}$$

TABLE 5.4. Newton's Method for an Extinction Probability

Iteration n	**Iterate** x_n	**Iteration** n	**Iterate** x_n
0	0.000	3	0.860
1	0.631	4	0.878
2	0.800	5	0.880

Since $s_{n+1} = P(s_n) \leq P(x_n) \leq x_{n+1}$, this completes the proof of (c).

Application of Newton's method to the Lotka branching process data produces the iterates displayed in Table 5.4. Comparison of this table with Table 5.3 illustrates the much faster convergence of Newton's method. Properties (a), (b), and (c) are evident in these two tables. For those readers acquainted with multitype branching process, it is noteworthy that all aspects of our comparison generalize if the differential $dP(\mathbf{1})$ of the vector of progeny generating functions is primitive and possesses a dominant eigenvalue strictly greater than 1.

5.5 Problems

1. Consider the quadratic function $x^2 - 2Ax + B$ whose coefficients A and B are independent, exponentially distributed random variables with common mean $1/\alpha$. The probability $p(\alpha)$ that both roots of this quadratic are real is given by the quantity

$$p(\alpha) = 1 - \sqrt{\pi\alpha}e^{\frac{\alpha}{4}}\left[1 - \Phi\left(\sqrt{\frac{\alpha}{2}}\right)\right],$$

where $\Phi(z)$ is the standard normal distribution function. Plot $p(\alpha)$ as a function of α. Find via bisection the minimum point and minimum value of $p(\alpha)$.

2. Let $f(x)$ be a probability density and $g(x)$ a positive, measurable function. To minimize $\int_{S_\alpha} g(x)dx$ subject to $\int_{S_\alpha} f(x)dx = \alpha$, show that one should choose $\lambda(\alpha)$ and $S_\alpha = \{x : f(x)/g(x) > \lambda(\alpha)\}$ so that the constraint $\int_{S_\alpha} f(x)dx = \alpha$ is satisfied. If $f(x)$ and $g(x)$ are defined on an interval of the real line, and the ratio $f(x)/g(x)$ is increasing to the left and decreasing to the right of its mode, then S_α will be an interval.

3. To apply the Neyman–Pearson lemma of Problem 2, let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. The statistic

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is a pivot that follows a chi-squared distribution with $n-1$ degrees of freedom. This pivot can be inverted to give a confidence interval for σ^2 of the form $(S^2/b, S^2/a)$. Design and implement an algorithm for computing the shortest confidence interval with a given confidence level. (Hint: As suggested in [3], use Problem 2 with $g(x) = x^{-2}$. You can check your results against the tables in [7].)

4. Show that the map $f(x) = \sqrt{2 + x}$ is contractive on $[0, \infty)$. What is the smallest value of the Lipschitz constant? Identify the limit of the functional iterates $x_n = f(x_{n-1})$ from any starting point x_0.

5. Kepler's problem of celestial mechanics involves finding the eccentric anomaly E in terms of the mean anomaly M and the eccentricity $0 \leq \epsilon < 1$ of an elliptical orbit. These three quantities are related by the equation $E = M + \epsilon \sin E$. Demonstrate that the corresponding function $f(E) = M + \epsilon \sin E$ is contractive on $(-\infty, \infty)$ with Lipschitz constant ϵ. Hence, the solution can be found by functional iteration.

6. Suppose $f(x) = -x^2 + x + \frac{1}{4}$. Prove that the iterates $x_n = f(x_{n-1})$ diverge if $x_0 < -\frac{1}{2}$ or $x_0 > \frac{3}{2}$, converge to $\frac{1}{2}$ if $-\frac{1}{2} < x_0 < \frac{3}{2}$, and converge to $-\frac{1}{2}$ if $x_0 = -\frac{1}{2}$ or $x_0 = \frac{3}{2}$.

7. For $0 \leq a \leq 4$, the function $f(x) = ax(1-x)$ maps the unit interval $[0, 1]$ onto itself. Show that:

(a) The point 0 is a fixed point that is globally attractive when $a \leq 1$ and locally repelling when $a > 1$. Note that the rate of convergence to 0 is less than geometric when $a = 1$.
(b) The point $1 - a^{-1}$ is a fixed point for $a > 1$. It is locally attractive when $1 < a < 3$ and locally repelling when $3 < a \leq 4$.
(c) For $1 < a \leq 2$, the fixed point $r = 1 - a^{-1}$ is globally attractive on $(0, 1)$. (Hint: Write $f(x) - r = (x - r)(1 - ax)$.)

For $2 < a \leq 3$, the fixed point $1 - a^{-1}$ continues to be globally attractive on $(0, 1)$, but the proof of this fact is harder. For $a > 3$, the iterates $x_n = f(x_{n-1})$ no longer reliably converge. They periodically oscillate between several limit points until at $a = 4$ they behave completely chaotically. See [6] for a nice intuitive discussion.

8. Functional iteration can often be accelerated. In searching for a fixed point of $x = f(x)$, consider the iteration scheme $x_n = f_\alpha(x_{n-1})$, where α is some constant and $f_\alpha(x) = (1-\alpha)x + \alpha f(x)$. Prove that any fixed point x_∞ of $f(x)$ is also a fixed point of $f_\alpha(x)$ and vice versa. Since $|f'_\alpha(x_\infty)|$ determines the rate of convergence of $x_n = f_\alpha(x_{n-1})$ to x_∞, find the α that minimizes $|f'_\alpha(x_\infty)|$ when $|f'(x_\infty)| < 1$. Unfortunately, neither x_∞ nor $f'(x_\infty)$ is typically known in advance.

9. The last problem is relevant to the branching process example of the text. Investigate numerically the behavior of the iterates

$$x_n = (1 - \alpha)x_{n-1} + \alpha P(x_{n-1})$$

for the choice $\alpha = 1/[1 - P'(0)]$ in the Lotka data. Is convergence to the extinction probability s_∞ faster than in ordinary functional iteration?

10. In the context of Problems 8 and 9, assume that $P'(1) > 1$. Show that the choice $\alpha = 1/[1 - P'(0)]$ guarantees that the iterates increase monotonely from $x_0 = 0$ to the extinction probability $s_\infty < 1$.

11. Suppose the function $g(x)$ mapping a closed interval I into itself has a k-fold composition $f(x) = g \circ \cdots \circ g(x)$ satisfying the assumptions of Proposition 5.3.1. Prove that $g(x)$ has a unique fixed point.

12. What happens when you apply Newton's method to the functions

$$f(x) = \begin{cases} \sqrt{x} & x \geq 0 \\ -\sqrt{-x} & x < 0 \end{cases}$$

and $g(x) = \sqrt[3]{x}$?

13. Newton's method can be used to extract roots. Consider the function $g(x) = x^m - c$ for some integer $m > 1$ and $c > 0$. Show that Newton's method is defined by

$$x_n = x_{n-1}\left(1 - \frac{1}{m} + \frac{c}{mx_{n-1}^m}\right).$$

Prove that $x_n \geq c^{\frac{1}{m}}$ for all $x_{n-1} > 0$ and that $x_n \leq x_{n-1}$ whenever $x_{n-1} \geq c^{\frac{1}{m}}$. Thus, if $x_0 \geq c^{\frac{1}{m}}$, then x_n monotonely decreases to $c^{\frac{1}{m}}$. If $0 < x_0 < c^{\frac{1}{m}}$, then $x_1 > c^{\frac{1}{m}}$, but thereafter, x_n monotonely decreases to $c^{\frac{1}{m}}$.

References

[1] Box GEP, Tiao G (1973) *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA

[2] Feller W (1968) *An Introduction to Probability Theory and its Applications, Vol 1*, 3rd ed. Wiley, New York

[3] Juola RC (1993) More on shortest confidence intervals. *Amer Statistician* 47:117–119

[4] Lotka AJ (1931) Population analysis—the extinction of families I. *J Wash Acad Sci* 21:377–380

[5] Lotka AJ (1931) Population analysis—the extinction of families II. *J Wash Acad Sci* 21:453–459

[6] Strang G (1986) *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, Wellesley, MA

[7] Tate RF, Klett GW (1969) Optimal confidence intervals for the variance of a normal distribution. *J Amer Stat Assoc* 54:674–682

[8] Wall HS (1948) *Analytic Theory of Continued Fractions*. Van Nostrand, New York

6
Vector and Matrix Norms

6.1 Introduction

In multidimensional calculus, vector and matrix norms quantify notions of topology and convergence [2, 3, 4, 5, 7, 9]. Because norms are also devices for deriving explicit bounds, theoretical developments in numerical analysis rely heavily on norms. They are particularly useful in establishing convergence and in estimating rates of convergence of iterative methods for solving linear and nonlinear equations. Norms also arise in almost every other branch of theoretical numerical analysis. Functional analysis, which deals with infinite-dimensional vector spaces, uses norms on functions.

6.2 Elementary Properties of Vector Norms

In our exposition of norms, we will assume a nodding familiarity with the Euclidean vector norm $||x||_2 = \sqrt{\sum_{i=1}^{m} x_i^2}$ in m-dimensional space R^m. This norm and others generalize the absolute value of a number on the real line. A norm on R^m is formally defined by four properties:

(a) $||x|| \geq 0$,
(b) $||x|| = 0$ if and only if $x = \mathbf{0}$,
(c) $||cx|| = |c| \cdot ||x||$ for every real number c,
(d) $||x + y|| \leq ||x|| + ||y||$.

In property (b), $\mathbf{0}$ is the vector with all m components 0. Property (d) is known as the triangle inequality. One immediate consequence of the triangle inequality is the further inequality $|\,||x|| - ||y||\,| \leq ||x - y||$.

Two other simple but helpful norms are

$$||x||_1 = \sum_{i=1}^{m} |x_i|$$
$$||x||_\infty = \max_{1 \leq i \leq m} |x_i|.$$

For each of the norms $||x||_p$, $p = 1$, 2, and ∞, a sequence of vectors x_n converges to a vector y if and only if each component sequence x_{ni} converges to y_i. Thus, all three norms give the same topology on R^m. The next proposition clarifies and generalizes this property.

Proposition 6.2.1. *Let $||x||$ be any norm on R^m. Then there exist positive constants k_l and k_u such that $k_l||x||_1 \leq ||x|| \leq k_u||x||_1$ holds for all $x \in R^m$.*

Proof. Let $e_1, \ldots, e_m$ be the standard basis vectors for R^m. Then the conditions (c) and (d) defining a norm indicate that $x = \sum_i x_i e_i$ satisfies

$$\begin{aligned} ||x|| &\leq \sum_i |x_i| \cdot ||e_i|| \\ &\leq \Big(\max_i ||e_i|| \Big) ||x||_1. \end{aligned}$$

This proves the upper bound with $k_u = \max_i ||e_i||$.

To establish the lower bound, we note that property (c) of a norm allows us to restrict attention to the set $S = \{x : ||x||_1 = 1\}$. Now the function $x \to ||x||$ is uniformly continuous on R^m because

$$\begin{aligned} |\,||x|| - ||y||\,| &\leq ||x - y|| \\ &\leq k_u||x - y||_1 \end{aligned}$$

follows from the upper bound just demonstrated. Since the set S is compact (closed and bounded), the function $x \to ||x||$ attains its lower bound k_l on S. Because of property (b), $k_l > 0$. □

Proposition 6.2.1 immediately implies that $\sup_{x \neq \mathbf{0}} ||x||/||x||^\dagger$ is finite for every pair of norms $||x||$ and $||x||^\dagger$. For instance, it is straightforward to verify that

$$||x||_q \leq ||x||_p \tag{1}$$
$$||x||_p \leq m^{\frac{1}{p} - \frac{1}{q}} ||x||_q \tag{2}$$

when p and q are chosen from $\{1, 2, \infty\}$ and $p < q$. These inequalities are sharp. Equality holds in (1) when $x = (1, 0, \ldots, 0)^t$, and equality holds in (2) when $x = (1, 1, \ldots, 1)^t$.

6.3 Elementary Properties of Matrix Norms

From one perspective an $m \times m$ matrix $A = (a_{ij})$ is simply a vector in R^{m^2}. Accordingly, we can define many norms involving A. However, it is profitable for a matrix norm also to be compatible with matrix multiplication. Thus, to the list of properties (a) through (d) for a vector norm, we add the requirement

(e) $||AB|| \leq ||A|| \cdot ||B||$

for any product of $m \times m$ matrices A and B. With this addition the Euclidean norm $||A||_E = \sqrt{\sum_{i=1}^m \sum_{j=1}^m a_{ij}^2} = \sqrt{\text{tr}(AA^t)} = \sqrt{\text{tr}(A^tA)}$ qualifies as a matrix norm. (Our reasons for writing $||A||_E$ rather than $||A||_2$ will soon be apparent.) Conditions (a) through (d) need no checking. Property (e) is verified by invoking the Cauchy–Schwarz inequality in

$$\begin{aligned} ||AB||_E^2 &= \sum_{i,j} \Big| \sum_k a_{ik}b_{kj} \Big|^2 \\ &\leq \sum_{i,j} \Big(\sum_k a_{ik}^2 \Big) \Big(\sum_l b_{lj}^2 \Big) \\ &= \Big(\sum_{i,k} a_{ik}^2 \Big) \Big(\sum_{l,j} b_{lj}^2 \Big) \\ &= ||A||_E^2 ||B||_E^2. \end{aligned}$$

Corresponding to any vector norm $||x||$ on R^m, there is an induced matrix norm $||A||$ on $m \times m$ matrices defined by

$$\begin{aligned} ||A|| &= \sup_{x \neq \mathbf{0}} \frac{||Ax||}{||x||} \\ &= \sup_{||x||=1} ||Ax||. \end{aligned} \tag{3}$$

Using the same symbol for both the vector and inherited matrix norm ordinarily causes no confusion. All of the defining properties of a matrix norm are trivial to check for definition (3). For instance, consider property (e):

$$\begin{aligned} ||AB|| &= \sup_{||x||=1} ||ABx|| \\ &\leq ||A|| \sup_{||x||=1} ||Bx|| \\ &= ||A|| \cdot ||B||. \end{aligned}$$

Definition (3) also entails the equality $||I|| = 1$, where I is the $m \times m$ identity matrix. Because $||I||_E = \sqrt{m}$, the Euclidean norm $||A||_E$ and the induced matrix norm $||A||_2$ are definitely different.

In the following proposition, $\rho(C)$ denotes the absolute value of the dominant eigenvalue of the matrix C. This quantity is called the spectral radius of C.

Proposition 6.3.1. *If $A = (a_{ij})$ is an $m \times m$ matrix, then*

(a) $||A||_1 = \max_j \sum_i |a_{ij}|$,
(b) $||A||_2 = \sqrt{\rho(A^t A)}$, *which reduces to* $\rho(A)$ *if A is symmetric,*
(c) $||A||_\infty = \max_i \sum_j |a_{ij}|$.

Proof. To prove (a) note that

$$\begin{aligned} ||A||_1 &= \sup_{||x||_1=1} \sum_i \Big| \sum_j a_{ij} x_j \Big| \\ &\le \sup_{||x||_1=1} \sum_i \sum_j |a_{ij}| \cdot |x_j| \\ &= \sup_{||x||_1=1} \sum_j |x_j| \sum_i |a_{ij}| \\ &\le \sup_{||x||_1=1} \Big(\sum_j |x_j| \Big) \Big(\max_k \sum_i |a_{ik}| \Big) \\ &= \max_k \sum_i |a_{ik}|. \end{aligned}$$

Equality holds throughout for the standard basis vector $x = e_k$ whose index k maximizes $\sum_i |a_{ik}|$.

To prove (b) choose an orthonormal basis of eigenvectors $u_1, \ldots, u_m$ for the symmetric matrix $A^t A$ with corresponding eigenvalues arranged so that $0 \le \lambda_1 \le \cdots \le \lambda_m$. If $x = \sum_i c_i u_i$ is a unit vector, then $\sum_i c_i^2 = 1$, and

$$||A||_2^2 = \sup_{||x||_2=1} x^t A^t A x = \sum_i \lambda_i c_i^2 \le \lambda_m.$$

Equality is achieved when $c_m = 1$ and all other $c_i = 0$. If A is symmetric with eigenvalues μ_i arranged so that $|\mu_1| \le \cdots \le |\mu_m|$, then the u_i can be chosen to be the corresponding eigenvectors. In this case, clearly $\lambda_i = \mu_i^2$.

To prove (c) note that

$$\begin{aligned} ||A||_\infty &= \sup_{||x||_\infty=1} \max_i \Big| \sum_j a_{ij} x_j \Big| \\ &\le \sup_{||x||_\infty=1} \max_i \sum_j |a_{ij}| \Big(\max_k |x_k| \Big) \\ &= \max_i \sum_j |a_{ij}|. \end{aligned}$$

Equality holds throughout for

$$x_j = \begin{cases} \dfrac{a_{kj}}{|a_{kj}|} & a_{kj} \neq 0 \\ 0 & a_{kj} = 0 \end{cases}$$

if k is an index with maximum row sum $\sum_j |a_{kj}|$. □

For theoretical purposes, it is convenient to consider vector and matrix norms defined over the complex vector space C^m. All of the properties studied so far generalize naturally to this setting. One needs to exercise a little care for the norm $||x||_2 = \sqrt{\sum_{i=1}^m |x_i|^2}$, where $|x_i|^2$ replaces x_i^2. This norm is induced by the complex inner product

$$\langle x, y \rangle = \sum_{i=1}^m x_i y_i^*,$$

with y_i^* denoting the complex conjugate of y_i. Proposition 6.3.1 (b) now refers to Hermitian matrices $A = (a_{ij}) = (a_{ji}^*) = A^*$ rather than to symmetric matrices. One of the advantages of extending norms to C^m is the following generalization of Proposition 6.3.1 (b) to arbitrary matrices.

Proposition 6.3.2. *The spectral radius $\rho(A)$ of a matrix A satisfies*

$$\rho(A) \leq ||A||$$

for any induced matrix norm. Furthermore, for any A and $\epsilon > 0$, there exists some induced matrix norm with $||A|| \leq \rho(A) + \epsilon$.

Proof. If λ is an eigenvalue of A with nontrivial eigenvector u, then the equality $||Au|| = |\lambda| \cdot ||u||$ for a vector norm entails the corresponding inequality $|\lambda| \leq ||A||$ for the induced matrix norm.

Suppose A and $\epsilon > 0$ are given. There exists an invertible matrix S and an upper triangular matrix $T = (t_{ij})$ such that $A = STS^{-1}$. This fact follows directly from the Jordan canonical form or the Schur decomposition of A [6, 8, 7]. For $\delta > 0$ consider the diagonal matrix $D(\delta)$ whose ith diagonal entry is δ^{i-1}. It is straightforward to check that $[SD(\delta)]^{-1}ASD(\delta) = T(\delta)$ is upper triangular with entries $(t_{ij}\delta^{j-i})$ and consequently that the upper off-diagonal entries of $T(\delta)$ tend to 0 as $\delta \to 0$. It is also easy to check that $||x||_\delta = ||[SD(\delta)]^{-1}x||_\infty$ defines a vector norm whose induced matrix norm is $||A||_\delta = ||[SD(\delta)]^{-1}ASD(\delta)||_\infty = ||T(\delta)||_\infty$. (See Problem 7.) According to Proposition 6.3.1 (c),

$$||A||_\delta = \max_i \sum_j |t_{ij}|\delta^{j-i}.$$

Because the eigenvalues of A coincide with the diagonal entries of T, we can take $\delta > 0$ so small that

$$\max_i \sum_j |t_{ij}|\delta^{j-i} \leq \rho(A) + \epsilon.$$

Such a choice implies $||A||_\delta \leq \rho(A) + \epsilon$. □

6.4 Iterative Solution of Linear Equations

Many numerical problems involve iterative schemes of the form

$$x_n = Bx_{n-1} + w \tag{4}$$

for solving the vector-matrix equation $(I - B)x = w$. Clearly, the map $f(x) = Bx + w$ satisfies

$$\begin{aligned} ||f(y) - f(x)|| &= ||B(y - x)|| \\ &\leq ||B|| \cdot ||y - x|| \end{aligned}$$

and therefore is contractive for a vector norm $||x||$ if $||B|| < 1$ holds for the induced matrix norm. If we substitute norms for absolute values in our convergence proof for one-dimensional functional iteration, then that proof generalizes to this vector setting, and we find that the iterates x_n converge to the unique solution x of $(I - B)x = w$. In light of the fact that w is arbitrary, it follows that $I - B$ is invertible. These facts are incorporated in the next proposition.

Proposition 6.4.1. *Let B be an arbitrary matrix with spectral radius $\rho(B)$. Then $\rho(B) < 1$ if and only if $||B|| < 1$ for some induced matrix norm. The inequality $||B|| < 1$ implies*

(a) $\lim_{n\to\infty} ||B^n|| = 0$,
(b) $(I - B)^{-1} = \sum_{n=0}^{\infty} B^n$,
(c) $\frac{1}{1+||B||} \leq ||(I - B)^{-1}|| \leq \frac{1}{1-||B||}$.

Proof. The first claim is an immediate consequence of Proposition 6.3.2. Assertion (a) follows from $||B^n|| \leq ||B||^n$. Assertion (b) follows if we let $x_0 = 0$ in the iteration scheme (4). Then $x_n = \sum_{i=0}^{n-1} B^i w$, and

$$\begin{aligned} (I - B)^{-1} w &= \lim_{n\to\infty} x_n \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} B^i w. \end{aligned}$$

To prove the first inequality of assertion (c), note that taking norms in $I = (I - B)(I - B)^{-1}$ implies

$$\begin{aligned} 1 &\leq ||I - B|| \cdot ||(I - B)^{-1}|| \\ &\leq (1 + ||B||)||(I - B)^{-1}||. \end{aligned}$$

For the second inequality, use the identity $(I - B)^{-1} = I + B(I - B)^{-1}$. Taking norms now produces

$$||(I - B)^{-1}|| \leq 1 + ||B|| \cdot ||(I - B)^{-1}||,$$

which can be rearranged to give the desired result. □

Linear iteration is especially useful in solving the equation $Ax = b$ for x when an approximation C to A^{-1} is known. If this is the case, then one

can set $B = I - CA$ and $w = Cb$ and iterate via (4). Provided $||B|| < 1$, the unique fixed point of the scheme (4) satisfies $x = (I - CA)x + Cb$. If C^{-1} exists, then this is equivalent to $Ax = b$.

6.4.1 Jacobi's Method

Jacobi's method offers a typical example of this strategy. Suppose for the sake of simplicity that $A = (a_{ij})$ is strictly diagonally dominant in the sense that $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ holds for all rows i. Let D be the diagonal matrix with ith diagonal entry a_{ii}. Then the matrix $C = D^{-1}$ can be considered an approximate inverse A. The matrix $B = I - CA$ has diagonal elements $b_{ii} = 0$ and off-diagonal elements $b_{ij} = a_{ij}/a_{ii}$. By definition

$$||B||_\infty = \max_i \sum_{j \neq i} |b_{ij}| < 1.$$

This analysis has the side effect of showing that every strictly diagonally dominant matrix A is invertible.

6.4.2 Pan and Reif's Iteration Scheme

In practice, the approximate inverse C can be rather crude. For instance, Pan and Reif [10] suggest the surprising choice $C = \epsilon A^t$ for ϵ small and positive. Because $A^t A$ is positive definite, its eigenvalues can be arranged as $0 < \lambda_1 \leq \cdots \leq \lambda_m$. The eigenvalues of the symmetric matrix $I - \epsilon A^t A$ are then $1 - \epsilon\lambda_1, \ldots 1 - \epsilon\lambda_m$. As long as $1 - \epsilon\lambda_m > -1$, all eigenvalues of $I - \epsilon A^t A$ will occur on the interval $(-1, 1)$, which according to part (b) of Proposition 6.3.1 implies $||I - \epsilon A^t A||_2 < 1$. In other words, if $\epsilon < 2/\lambda_m = 2/||A||_2^2$, then linear iteration can be employed to solve $Ax = b$. Since finding the norm $||A||_2$ is cumbersome, one can replace it in bounding ϵ with more simply computed upper bounds. For instance, the inequalities $||A||_2 \leq ||A||_E$ and $||A||_2 \leq \sqrt{||A||_\infty ||A||_1}$ discussed in Problems 5 and 6 often serve well.

6.4.3 Equilibrium Distribution of a Markov Chain

A slightly different problem is to determine the equilibrium distribution of a finite state Markov chain. Recall that movement among the m states of a Markov chain is governed by its $m \times m$ transition matrix $P = (p_{ij})$, whose entries are nonnegative and satisfy $\sum_j p_{ij} = 1$ for all i. A column vector x with nonnegative entries and norm $||x||_1 = \sum_i x_i = 1$ is said to be an equilibrium distribution for P provided $x^t P = x^t$, or equivalently $Qx = x$ for $Q = P^t$. Because the norm $||Q||_1 = 1$, one cannot immediately invoke the contraction mapping principle. However, if we restrict attention to the closed set $S = \{x : x_i \geq 0,\ i = 1, \ldots, m,\ \sum_i x_i = 1\}$, then we do get a contraction map under the hypothesis that some power Q^k has all entries

positive [1]. Let $c > 0$ be the minimum entry of Q^k and $\mathbf{1}$ be the column vector of all 1's. The matrix $R = Q^k - c\mathbf{1}\mathbf{1}^t$ has all entries nonnegative and norm $||R||_1 < 1$.

Consider two vectors x and y from S. Owing to the fact $\mathbf{1}^t(x-y) = 0$, we get

$$\begin{aligned} ||Q^k x - Q^k y||_1 &= ||R(x-y)||_1 \\ &< ||R||_1 ||x-y||_1, \end{aligned}$$

and it follows that the map $x \to Q^k x$ is contractive on S with unique fixed point x_∞. Now for any $x \in S$,

$$Qx_\infty = Q \lim_{n\to\infty} Q^{nk}x = \lim_{n\to\infty} Q^{nk}Qx = x_\infty.$$

Thus, x_∞ is a fixed point of $x \to Qx$ as well. Because any integer n can be represented uniquely as $kl + r$ with $0 \le r < k$, the inequality

$$\begin{aligned} ||Q^n x - x_\infty||_1 &= ||Q^{kl}(Q^r x - Q^r x_\infty)||_1 \\ &< ||R||_1^l ||Q^r x - Q^r x_\infty||_1 \end{aligned}$$

can be invoked to show that $\lim_{n\to\infty} Q^n x = x_\infty$ for all $x \in S$.

6.5 Condition Number of a Matrix

Consider the apparently innocuous matrix

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \tag{5}$$

concocted by R. S. Wilson [2]. This matrix is symmetric and positive definite. The unique solution to the linear equation $Ax = b$ can be expressed as $x = A^{-1}b$. For the choice $b = (32, 23, 33, 31)^t$, we find $x = (1, 1, 1, 1)^t$. The slightly perturbed vector $b + \Delta b = (32.1, 22.9, 33.1, 30.9)^t$ leads to the violently perturbed solution $x + \Delta x = (9.2, -12.6, 4.5, -1.1)^t$. When we start with $b = (4, 3, 3, 1)^t$, then the solution of $Ax = b$ is $x = (1, -1, 1, -1)^t$. If we perturb A to $A + .01I$, then the solution of $(A + .01I)(x + \Delta x) = b$ is $x + \Delta x = (.59, -.32, .82, -.89)^t$. Thus, a relatively small change in A propagates to a large change in the solution of the linear equation.

One can gain insight into these disturbing patterns by defining the condition number of an invertible matrix. Consider a vector norm $||x||$ and its induced matrix norm $||A||$. If $Ax = b$ and $A(x + \Delta x) = b + \Delta b$, then by definition of the induced matrix norm,

$$\begin{aligned} ||b|| &\le ||A|| \cdot ||x|| \\ ||\Delta x|| &\le ||A^{-1}|| \cdot ||\Delta b||. \end{aligned}$$

Dividing the second of these inequalities by the first produces

$$\frac{||\Delta x||}{||x||} \leq \text{cond}(A)\frac{||\Delta b||}{||b||}, \tag{6}$$

where $\text{cond}(A) = ||A|| \cdot ||A^{-1}||$ is termed the condition number of the matrix A relative to the given norm. Inequality (6) is sharp. To achieve equality, one merely needs to choose x so that $||Ax|| = ||A|| \cdot ||x||$ and Δb so that $||A^{-1}\Delta b|| = ||A^{-1}|| \cdot ||\Delta b||$.

Now suppose $Ax = b$ and $(A + \Delta A)(x + \Delta x) = b$. It then follows from $\Delta x = -A^{-1}\Delta A(x + \Delta x)$ that

$$||\Delta x|| \leq ||A^{-1}|| \cdot ||\Delta A|| \cdot ||x + \Delta x||, \tag{7}$$

or equivalently

$$\frac{||\Delta x||}{||x + \Delta x||} \leq \text{cond}(A)\frac{||\Delta A||}{||A||}. \tag{8}$$

Inequality (8) is also sharp; see Problem 12.

A bound on the change $||\Delta x||/||x||$ is, perhaps, preferable to the bound (8). For $||\Delta A||$ small, one can argue that $x+\Delta x = (I+A^{-1}\Delta A)^{-1}x$ because

$$\begin{aligned} x &= A^{-1}b \\ &= A^{-1}(A + \Delta A)(x + \Delta x) \\ &= (I + A^{-1}\Delta A)(x + \Delta x). \end{aligned}$$

The identity $x + \Delta x = (I + A^{-1}\Delta A)^{-1}x$ in turn implies

$$\begin{aligned} ||x + \Delta x|| &\leq ||(I + A^{-1}\Delta A)^{-1}|| \cdot ||x|| \\ &\leq \frac{||x||}{1 - ||A^{-1}\Delta A||} \\ &\leq \frac{||x||}{1 - ||A^{-1}|| \cdot ||\Delta A||} \end{aligned}$$

in view of part (c) of Proposition 6.4.1. Substituting this bound for $||x+\Delta x||$ in inequality (7) yields

$$\begin{aligned} \frac{||\Delta x||}{||x||} &\leq \frac{||A^{-1}|| \cdot ||\Delta A||}{1 - ||A^{-1}|| \cdot ||\Delta A||} \\ &= \text{cond}(A)\frac{||\Delta A||}{(||A|| - \text{cond}(A)||\Delta A||)}. \end{aligned}$$

The mysteries of the matrix (5) disappear when we compute its condition number $\text{cond}_2(A)$ relative to the matrix norm $||A||_2$. Recalling part (b) of Proposition 6.3.1, it is clear that $\text{cond}_2(A)$ is the ratio of the largest and smallest eigenvalues λ_4 and λ_1 of A. For the matrix (5), it turns out that $\lambda_1 = 0.01015$, $\lambda_4 = 30.2887$, and $\text{cond}_2(A) = 2984$. We will learn later how to compute the dominant eigenvalues of A and A^{-1}. If A^{-1} is available,

we can elect another more easily computed norm and calculate cond(A) relative to it.

6.6 Problems

1. Verify the vector norm inequalities (1) and (2) for p and q chosen from $\{1, 2, \infty\}$.

2. Show that $||x||_2^2 \leq ||x||_\infty ||x||_1 \leq \sqrt{m}||x||_2^2$ for any vector $x \in R^m$.

3. Suppose T is a symmetric matrix. What further conditions on T guarantee that $||x|| = \sqrt{|x^t T x|}$ is a vector norm?

4. Prove that $1 \leq ||I||$ and $||A||^{-1} \leq ||A^{-1}||$ for any matrix norm.

5. For an $m \times m$ matrix A, show that

$$\frac{1}{\sqrt{m}}||A||_1 \leq ||A||_2 \leq \sqrt{m}||A||_1$$

$$\frac{1}{\sqrt{m}}||A||_\infty \leq ||A||_2 \leq \sqrt{m}||A||_\infty$$

$$\frac{1}{\sqrt{m}}||A||_E \leq ||A||_2 \leq ||A||_E.$$

(Hint: Use the vector norm inequalities (1) and (2) and Proposition 6.3.1.)

6. Prove the inequality $||A||_2 \leq \sqrt{||A||_\infty ||A||_1}$. (Hint: If the dominant eigenvalue $\lambda \geq 0$ of $A^t A$ has eigenvector u, then bound $\lambda||u||_1$.)

7. Suppose $||x||$ is a vector norm and T is an invertible matrix. Show that $||x||^\dagger = ||Tx||$ defines a vector norm whose induced matrix norm is $||A||^\dagger = ||TAT^{-1}||$.

8. Define $||A|| = \max_{i,j} |a_{ij}|$ for $A = (a_{ij})$. Show that this defines a vector norm but not a matrix norm on $m \times m$ matrices A.

9. Let O_n be a sequence of orthogonal matrices. Show that there exists a subsequence O_{n_k} that converges to an orthogonal matrix. (Hint: Compute the norm $||O_n||_2$.)

10. Demonstrate that $\rho(A) = \lim_{n\to\infty} ||A^n||^{1/n}$ for any induced matrix norm. (Hints: $\rho(A^n)^{1/n} = \rho(A)$ and $[(\rho(A) + \epsilon)^{-1}A]^n \to 0$.)

11. Prove that the series $B_n = \sum_{k=0}^{n} \frac{A^k}{k!}$ converges. Its limit is the matrix exponential e^A.

12. Show that inequality (8) is sharp by choosing $w \neq \mathbf{0}$ so that

$$||A^{-1}w|| = ||A^{-1}|| \cdot ||w||.$$

Then take successively $\Delta x = -\beta A^{-1}w$, $x + \Delta x = w$, $\Delta A = \beta I$, and $b = (A + \Delta A)w$, where β is any nonzero number such that $A + \beta I$ invertible.

13. Relative to any induced matrix norm, show that $\text{cond}(A) \geq 1$ and that $\text{cond}(A^{-1}) = \text{cond}(A)$ and $\text{cond}(cA) = \text{cond}(A)$ for any scalar $c \neq 0$. Also verify that if U is orthogonal, then $\text{cond}_2(U) = 1$ and

$$\text{cond}_2(A) = \text{cond}_2(AU) = \text{cond}_2(UA).$$

14. If $A + \Delta A$ is invertible, prove that

$$\frac{||(A + \Delta A)^{-1} - A^{-1}||}{||(A + \Delta A)^{-1}||} \leq \text{cond}(A) \frac{||\Delta A||}{||A||}.$$

References

[1] Baldwin JT (1989) On Markov processes in elementary mathematics courses. *Amer Math Monthly* 96:147–153

[2] Ciarlet PG (1989) *Introduction to Numerical Linear Algebra and Optimization*. Cambridge University Press, Cambridge

[3] Gill PE, Murray W, Wright MH (1991) *Numerical Linear Algebra and Optimization, Vol 1*. Addison-Wesley, Reading, MA

[4] Golub GH, Van Loan CF (1989) *Matrix Computations*, 2nd ed. Johns Hopkins University Press, Baltimore, MD

[5] Hämmerlin G, Hoffmann K-H (1991) *Numerical Mathematics*. Springer-Verlag, New York

[6] Hoffman K, Kunze R (1971) *Linear Algebra*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ

[7] Isaacson E, Keller HB (1966) *Analysis of Numerical Methods*. Wiley, New York

[8] Lang S (1971) *Linear Algebra*, 2nd ed. Addison-Wesley, Reading, MA

[9] Ortega JM (1990) *Numerical Analysis: A Second Course*. Society for Industrial and Applied Mathematics, Philadelphia

[10] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge

7

Linear Regression and Matrix Inversion

7.1 Introduction

The sweep operator [1, 2, 4, 5, 6] is the workhorse of computational statistics. The matrices that appear in linear regression and multivariate analysis are almost invariably symmetric. Sweeping exploits this symmetry. Although there are faster and numerically more stable algorithms for inverting a matrix or solving a least-squares problem, no algorithm matches the conceptual simplicity and utility of sweeping. To highlight some of the typical quantities that sweeping calculates with surprising ease, we briefly review a few key ideas from linear regression and multivariate analysis.

Gram–Schmidt orthogonalization, particularly in its modified form, offers a numerically more stable method of computing linear regression estimates than sweeping [9, 10]. Although this is reason enough for introducing one of the major algorithms of matrix orthogonalization, we will meet further motivation in Chapter 14, where we discuss the computation of asymptotic standard errors of maximum likelihood estimates subject to linear constraints.

Woodbury's formula occupies a somewhat different niche than sweeping or orthogonalization [7, 8]. Many statistical models involve the inversion of matrices that are low-rank perturbations of matrices with known inverses. For instance, if D is an invertible diagonal matrix and u is a compatible column vector, the Sherman–Morrison formula [7, 8]

$$(D + uu^t)^{-1} = D^{-1} - \frac{1}{1 + u^t D^{-1} u} D^{-1} u u^t D^{-1}$$

provides the inverse of the symmetric, rank-one perturbation $D+uu^t$ of D. Woodbury's formula generalizes the Sherman–Morrison formula. Both the original Sherman–Morrison formula and Woodbury's generalization permit straightforward computation of the determinant of the perturbed matrix from the determinant of the original matrix.

7.2 Motivation from Linear Regression

As motivation for the sweep operator, we briefly review linear regression. The basic setup involves p independent observations that individually take the form

$$y_i = \sum_{j=1}^{q} x_{ij}\beta_j + u_i. \tag{1}$$

Here y_i depends linearly on the unknown parameters β_j through the known constants x_{ij}. The error u_i is assumed to be normally distributed with mean 0 and variance σ^2. If we collect the y_i into a $p \times 1$ observation vector y, the x_{ij} into a $p \times q$ design matrix X, the β_j into a $q \times 1$ parameter vector β, and the u_j into a $p \times 1$ error vector u, then the linear regression model can be rewritten in vector notation as $y = X\beta + u$. A maximum likelihood estimator $\hat{\beta}$ of β solves the normal equations $X^tX\beta = X^ty$. Provided the matrix X is of full rank, $\hat{\beta} = (X^tX)^{-1}X^ty$. This is also the least-squares estimator of β even when the error vector u is nonnormal. In general, if u has uncorrelated components with common mean 0 and common variance σ^2, then the estimator $\hat{\beta}$ has mean and variance

$$\begin{aligned} \mathrm{E}(\hat{\beta}) &= \beta \\ \mathrm{Var}(\hat{\beta}) &= \sigma^2(X^tX)^{-1}. \end{aligned}$$

The difference $y-\hat{y} = y-X\hat{\beta}$ between the actual and predicted observations is termed the residual vector. Its Euclidean norm $||y-\hat{y}||_2^2$ squared, known as the residual sum of squares, is fundamentally important in inference. For example, σ^2 is usually estimated by $||y-\hat{y}||_2^2/(p-q)$. A single application of the sweep operator permits simultaneous computation of $\hat{\beta}$, $\mathrm{Var}(\hat{\beta})$, and $||y-\hat{y}||_2^2$.

7.3 Motivation from Multivariate Analysis

A random vector $X \in R^p$ with mean vector μ, covariance matrix Ω, and density

$$(2\pi)^{-\frac{p}{2}} \det(\Omega)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^t\Omega^{-1}(x-\mu)}$$

is said to follow a multivariate normal distribution. The sweep operator permits straightforward calculation of the quadratic form $(x-\mu)^t\Omega^{-1}(x-\mu)$ and the determinant of Ω. If we partition X and its mean and covariance so that

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \qquad \Omega = \begin{pmatrix} \Omega_Y & \Omega_{YZ} \\ \Omega_{ZY} & \Omega_Z \end{pmatrix},$$

then conditional on the event $Y = y$ the subvector Z follows a multivariate normal density with conditional mean and variance

$$\begin{aligned} \mathrm{E}(Z \mid Y = y) &= \mu_Z + \Omega_{ZY}\Omega_Y^{-1}(y - \mu_Y) \\ \mathrm{Var}(Z \mid Y = y) &= \Omega_Z - \Omega_{ZY}\Omega_Y^{-1}\Omega_{YZ}. \end{aligned}$$

These quantities and the conditional density of Z given $Y = y$ can all be easily evaluated via the sweep operator.

7.4 Definition of the Sweep Operator

Suppose $A = (a_{ij})$ is an $m \times m$ symmetric matrix. Sweeping on the kth diagonal entry $a_{kk} \neq 0$ of A yields a new symmetric matrix $\hat{A} = (\hat{a}_{ij})$ with entries

$$\begin{aligned} \hat{a}_{kk} &= -\frac{1}{a_{kk}} \\ \hat{a}_{ik} &= \frac{a_{ik}}{a_{kk}} \\ \hat{a}_{kj} &= \frac{a_{kj}}{a_{kk}} \\ \hat{a}_{ij} &= a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \end{aligned}$$

for $i \neq k$ and $j \neq k$. Sweeping on the kth diagonal entry can be undone by inverse sweeping on the kth diagonal entry. Inverse sweeping sends the matrix $A = (a_{ij})$ into $\check{A} = (\check{a}_{ij})$ with entries

$$\begin{aligned} \check{a}_{kk} &= -\frac{1}{a_{kk}} \\ \check{a}_{ik} &= -\frac{a_{ik}}{a_{kk}} \\ \check{a}_{kj} &= -\frac{a_{kj}}{a_{kk}} \\ \check{a}_{ij} &= a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \end{aligned}$$

for $i \neq k$ and $j \neq k$. Because sweeping and inverse sweeping preserve symmetry, all operations can be carried out on either the lower or upper-triangular part of A alone. This saves both computation and storage. In practice, it is wise to carry out sweeping in double precision.

7.5 Properties of the Sweep Operator

We now develop the basic properties of the sweep operator following the exposition of Jennrich [4]. Readers familiar with Gaussian elimination or the simplex algorithm in linear programming have already been exposed to the major themes of pivoting and exchange [3]. Sweeping is a symmetrized version of Gauss–Jordan pivoting.

Proposition 7.5.1. *Let A be an $m \times m$ matrix and U and V be $p \times m$ matrices with columns $u_1, \ldots, u_m$ and $v_1, \ldots, v_m$, respectively. If $V = UA$ before sweeping on the kth diagonal entry of A, then $\hat{V} = \hat{U}\hat{A}$ after sweeping on the kth diagonal entry of A. Here A is sent into $\hat{A}$, the matrix $\hat{U}$ coincides with U except for the exchange of column u_k for column v_k, and the matrix $\hat{V}$ coincides with V except for the exchange of column v_k for $-u_k$. The inverse sweep produces the same result except that u_k is exchanged for $-v_k$ and v_k is exchanged for u_k. Consequently, an inverse sweep undoes a sweep on the same diagonal entry and vice versa. An inverse sweep also coincides with a sweep cubed.*

Proof. By definition $v_{jl} = \sum_i u_{ji}a_{il}$ for all pairs j and l. After sweeping on a_{kk},

$$\begin{aligned}
\hat{v}_{jk} &= -u_{jk} \\
&= -\frac{1}{a_{kk}}(v_{jk} - \sum_{i \neq k} u_{ji}a_{ik}) \\
&= \hat{u}_{jk}\hat{a}_{kk} + \sum_{i \neq k} \hat{u}_{ji}\hat{a}_{ik} \\
&= \sum_i \hat{u}_{ji}\hat{a}_{ik},
\end{aligned}$$

and for $l \neq k$,

$$\begin{aligned}
\hat{v}_{jl} &= v_{jl} \\
&= \sum_{i \neq k} u_{ji}a_{il} + u_{jk}a_{kl} \\
&= \sum_{i \neq k} u_{ji}a_{il} + (v_{jk} - \sum_{i \neq k} u_{ji}a_{ik})\frac{a_{kl}}{a_{kk}} \\
&= \sum_{i \neq k} \hat{u}_{ji}\hat{a}_{il} + \hat{u}_{jk}\hat{a}_{kl} \\
&= \sum_i \hat{u}_{ji}\hat{a}_{il}.
\end{aligned}$$

Thus, $\hat{V} = \hat{U}\hat{A}$. Similar reasoning applies to an inverse sweep.

If a sweep is followed by an inverse sweep on the same diagonal entry, then the doubly transformed matrix $\check{\hat{A}}$ satisfies the equation $V = U\check{\hat{A}}$.

Choosing square matrices U and V such that U is invertible allows one to write both A and $\hat{\hat{A}}$ as $U^{-1}V$. Likewise, it is easy to check that the inverse and cube of a sweep transform U and V into exactly the same matrices $\check{U}$ and $\check{V}$. □

Performing a sequence of sweeps leads to the results stated in the next proposition.

Proposition 7.5.2. *Let the symmetric matrix A be partitioned as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

If possible, sweeping on the diagonal entries of A_{11} yields

$$\hat{A} = \begin{pmatrix} -A_{11}^{-1} & A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}. \tag{2}$$

In other words, sweeping on a matrix in block form conforms to the same rules as sweeping on the matrix entry by entry. Furthermore, if it is possible to sweep on a set of diagonal elements in more than one order, then the result is independent of the order chosen.

Proof. Applying Proposition 7.5.1 repeatedly in the equality

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I_{11} & \mathbf{0}_{12} \\ \mathbf{0}_{21} & I_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

leads to

$$\begin{pmatrix} -I_{11} & A_{12} \\ \mathbf{0}_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & \mathbf{0}_{12} \\ A_{21} & I_{22} \end{pmatrix} \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{pmatrix},$$

where I_{11} and I_{22} are identity matrices and $\mathbf{0}_{12}$ and $\mathbf{0}_{21}$ are zero matrices. This implies

$$\begin{aligned} -I_{11} &= A_{11}\hat{A}_{11} \\ A_{12} &= A_{11}\hat{A}_{12} \\ \mathbf{0}_{21} &= A_{21}\hat{A}_{11} + \hat{A}_{21} \\ A_{22} &= A_{21}\hat{A}_{12} + \hat{A}_{22}. \end{aligned}$$

Solving these equations for the blocks of $\hat{A}$ yields the claimed results. □

Sweeping is also a device for monitoring the positive definiteness of a matrix.

Proposition 7.5.3. *A symmetric matrix A is positive definite if and only if each diagonal entry can be swept in succession and is positive until it is swept. When a diagonal entry of a positive definite matrix A is swept, it becomes negative and remains negative thereafter. Furthermore, taking the product of the diagonal entries just before each is swept yields the determinant of A.*

Proof. The equivalence of the two conditions characterizing A is obvious if A is a 1×1 matrix. If A is $m \times m$, then suppose it has the form given in Proposition 7.5.2. Now the matrix identity

$$\begin{pmatrix} A_{11} & \mathbf{0}_{12} \\ \mathbf{0}_{21} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix} = \begin{pmatrix} I_{11} & \mathbf{0}_{12} \\ -A_{21}A_{11}^{-1} & I_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I_{11} & -A_{11}^{-1}A_{12} \\ \mathbf{0}_{21} & I_{22} \end{pmatrix} \tag{3}$$

shows that A is positive definite if and only if A_{11} and $A_{22} - A_{21}A_{11}^{-1}A_{12}$ are both positive definite. In view of equation (2) of Proposition 7.5.2, the equivalence of the sweeping condition and positive definiteness of A follows inductively from the same equivalence applied to the smaller matrices A_{11} and $A_{22} - A_{21}A_{11}^{-1}A_{12}$.

Once a diagonal entry of A has been swept, the diagonal entry forms part of the matrix $-A_{11}^{-1}$, which is negative definite. Hence, the swept diagonal entries must be negative. Finally, formula (3) shows that

$$\det A = \det(A_{11}) \det(A_{22} - A_{21}A_{11}^{-1}A_{12}). \tag{4}$$

The validity of the asserted procedure for calculating $\det A$ now follows inductively since it is obviously true for a 1×1 matrix. □

The determinant formula (4) is of independent interest. It does not depend on A being symmetric. Obviously, the analogous formula

$$\det A = \det(A_{22}) \det(A_{11} - A_{12}A_{22}^{-1}A_{21}) \tag{5}$$

also holds.

7.6 Applications of Sweeping

The representation (2) is of paramount importance. For instance in linear regression, suppose we construct the matrix

$$\begin{pmatrix} X^tX & X^ty \\ y^tX & y^ty \end{pmatrix} \tag{6}$$

and sweep on the diagonal entries of X^tX. Then the basic theoretical ingredients

$$\begin{pmatrix} -(X^tX)^{-1} & (X^tX)^{-1}X^ty \\ y^tX(X^tX)^{-1} & y^ty - y^tX(X^tX)^{-1}X^ty \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}\operatorname{Var}(\hat{\beta}) & \hat{\beta} \\ \hat{\beta}^t & ||y - \hat{y}||_2^2 \end{pmatrix}$$

magically emerge.

When we construct the matrix

$$\begin{pmatrix} \Omega & x-\mu \\ x^t-\mu^t & 0 \end{pmatrix}$$

for the multivariate normal distribution and sweep on the diagonal entries of Ω, we get the quadratic form $-(x-\mu)^t\Omega^{-1}(x-\mu)$ in the lower right block of the swept matrix. In the process we can also accumulate $\det\Omega$. To avoid underflows and overflows, it is better to compute $\ln\det\Omega$ by summing the logarithms of the diagonal entries as we sweep them. If we partition X as $(Y^t, Z^t)^t$ and sweep on the upper left block of

$$\begin{pmatrix} \Omega_Y & \Omega_{YZ} & \mu_Y-y \\ \Omega_{ZY} & \Omega_Z & \mu_Z \\ \mu_Y^t-y^t & \mu_Z^t & 0 \end{pmatrix},$$

then the conditional mean $\mathrm{E}(Z \mid Y=y) = \mu_Z + \Omega_{ZY}\Omega_Y^{-1}(y-\mu_Y)$ and conditional variance $\mathrm{Var}(Z \mid Y=y) = \Omega_Z - \Omega_{ZY}\Omega_Y^{-1}\Omega_{YZ}$ are immediately available.

7.7 Gram–Schmidt Orthogonalization

Consider again the regression model (1). In the Gram–Schmidt approach to regression, we decompose X into the product $X = UR$, where U is a $p \times q$ matrix with orthonormal columns $u_1, \ldots, u_q$ and R is a $q \times q$ invertible upper-triangular matrix. Before we explain how to construct this decomposition, let us rephrase the normal equations $X^tX\beta = X^ty$ as

$$R^tU^tUR\beta = R^tU^ty.$$

Because R^t is invertible and U^tU equals the $q \times q$ identity matrix I_q, it follows that the normal equations reduce to $R\beta = U^ty = z$. This system of equations is trivial to solve. Beginning with $\hat{\beta}_q = z_q/r_{qq}$, we simply backsolve for all of the entries of $\hat{\beta}$ via the recurrence

$$\hat{\beta}_j = \frac{1}{r_{jj}}\left(z_j - \sum_{k=j+1}^{q} r_{jk}\hat{\beta}_k\right).$$

Gram–Schmidt orthogonalization takes a collection of vectors such as the columns $x_1, \ldots, x_q$ of the design matrix X into an orthonormal collection of vectors $u_1, \ldots, u_q$ spanning the same column space. The algorithm begins by defining

$$u_1 = \frac{1}{\|x_1\|_2}x_1.$$

This definition compels the upper left entry of R to be $r_{11} = \|x_1\|_2$. Given $u_1, \ldots, u_{k-1}$, the next unit vector u_k in the sequence is defined by dividing

the column vector

$$v_k = x_k - \sum_{j=1}^{k-1} u_j^t x_k \, u_j \tag{7}$$

by its Euclidean norm. In other words, we subtract from x_k its projections onto each of the previously created u_j and normalize the result. A simple induction argument shows that the vectors $u_1, \ldots, u_k$ form an orthonormal basis of the subspace spanned by $x_1, \ldots, x_k$, assuming of course that these latter vectors are independent. The upper-triangular entries of the matrix R are given by the formulas $r_{jk} = u_j^t x_k$ for $1 \leq j < k$ and $r_{kk} = \|v_k\|_2$.

Computational experience has shown that the numerical stability of Gram–Schmidt orthogonalization can be improved by a simple device. In equation (7) we subtract from x_k all of its projections simultaneously. If the columns of X are nearly collinear, it is better to subtract off the projections sequentially. Thus, we let $v_k^{(1)} = x_k$ and sequentially compute

$$v_k^{(j+1)} = v_k^{(j)} - u_j^t v_k^{(j)} \, u_j, \qquad r_{jk} = u_j^t v_k^{(j)}$$

until we reach $v_k = v_k^{(k)}$. As before, $r_{kk} = \|v_k\|_2$. With perfect arithmetic, the modified algorithm arrives at the same outcome as the previous algorithm. However, with imperfect arithmetic, the vectors $u_1, \ldots, u_q$ computed under the modified algorithm are more nearly orthogonal.

7.8 Woodbury's Formula

Suppose A is a $p \times p$ matrix with known inverse A^{-1} and known determinant $\det A$. If U and V are $p \times q$ matrices of rank q, then $A + UV^t$ is a rank q perturbation of A. In many applications q is much smaller than p. If U has columns $u_1, \ldots, u_q$ and V has columns $v_1, \ldots, v_q$, then $A + UV^t$ can also be expressed as

$$A + UV^t = A + \sum_{i=1}^{q} u_i v_i^t.$$

Woodbury's formula amounts to

$$(A + UV^t)^{-1} = A^{-1} - A^{-1}U(I_q + V^t A^{-1} U)^{-1} V^t A^{-1}, \tag{8}$$

where I_q is the $q \times q$ identity matrix [7, 8]. Equation (8) is valuable because the $q \times q$ matrix $I_q + V^t A^{-1} U$ is typically much easier to invert than the $p \times p$ matrix $A + UV^t$. When $V = U$, Woodbury's formula is a consequence of sweeping the matrix

$$\begin{pmatrix} -A & U \\ U^t & I_q \end{pmatrix}$$

first on its upper left block and then on its lower right block and comparing the results to sweeping on these blocks in reverse order.

In solving the linear equation $(A + UV^t)x = b$, computing the whole inverse $(A+UV^t)^{-1}$ is unnecessary. Press et al. [8] recommend the following procedure: First compute the column vectors $z_1, \ldots, z_q$ of $Z = A^{-1}U$ by solving each linear equation $Az_i = u_i$. Then calculate $H = (I_q + V^tZ)^{-1}$. Finally, solve the linear equation $Ay = b$ for y. The solution to the linear equation $(A + UV^t)x = b$ can then be written as $x = y - ZHV^ty$.

If $A + UU^t$ is the covariance matrix of a multivariate normal random vector X, then to evaluate the density of X it is necessary to compute $\det(A + UU^t)$. (Observe that choosing $V = U$ preserves the symmetry of A.) The identity

$$\det(A + UV^t) = \det A \det(I_q + V^tA^{-1}U) \tag{9}$$

permits easy computation of $\det(A + UV^t)$. This identity also evidently implies that $A + UV^t$ is invertible if and only if $I_q + V^tA^{-1}U$ is invertible. To prove (9), we note that

$$\begin{aligned}\det(A + UV^t) &= \det A \det(I_p + A^{-1}UV^t) \\ &= \det A \det \begin{pmatrix} I_q & V^t \\ -A^{-1}U & I_p \end{pmatrix} \\ &= \det A \det(I_q + V^tA^{-1}U)\end{aligned}$$

follows directly from equations (4) and (5).

7.9 Problems

1. Consider the matrix

$$A = \frac{1}{3}\begin{pmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{pmatrix}.$$

Compute its inverse by sweeping. Determine whether A is positive definite based on the intermediate results of sweeping.

2. Calculate how many arithmetic operations it takes to compute one sweep of an $m \times m$ symmetric matrix A. If you calculate only the upper-triangular part of the result, how many operations do you save? Note that the revised sweeping scheme

$$\begin{aligned}\hat{a}_{kk} &= -\frac{1}{a_{kk}} \\ \hat{a}_{ik} &= -\hat{a}_{kk}a_{ik} \\ \hat{a}_{kj} &= -\hat{a}_{kk}a_{kj} \\ \hat{a}_{ij} &= a_{ij} - \hat{a}_{ik}a_{kj}\end{aligned}$$

for $i \neq k$ and $j \neq k$ is more efficient than the original sweeping scheme. How many operations does it take to compute A^{-1} assuming all of the required diagonal sweeps are possible?

3. Suppose the positive definite matrix $A = (a_{ij})$ has inverse $B = (b_{ij})$. Show that $a_{ii}^{-1} \leq b_{ii}$ with equality if and only if $a_{ij} = a_{ji} = 0$ for all $j \neq i$. If A is an expected information matrix, what implications does this result have for maximum likelihood estimation in large samples?

4. The jackknife method of regression analysis can be implemented by replacing the linear regression matrix (6) by the matrix

$$\begin{pmatrix} X^t \\ I_p \\ y^t \end{pmatrix} \begin{pmatrix} X & I_p & y \end{pmatrix} = \begin{pmatrix} X^tX & X^t & X^ty \\ X & I_p & y \\ y^tX & y^t & y^ty \end{pmatrix},$$

sweeping on its upper left block X^tX, and then sweeping on its $(q+k)$th diagonal entry for some k between 1 and p. Prove that this action yields the necessary ingredients for regression analysis omitting the kth observation y_k and the corresponding kth row of the $p \times q$ design matrix X [1]. (Hint: The additional sweep is equivalent to replacing the kth regression equation $y_k = \sum_{l=1}^q x_{kl}\beta_l + e_k$ by the regression equation $y_k = \sum_{l=1}^q x_{kl}\beta_l + \beta_{q+k} + e_k$ involving an additional parameter; the other regression equations are untouched. The parameter β_{q+k} can be adjusted to give a perfect fit to y_k. Hence, the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_q$ after the additional sweep depend only on the observations y_i for $i \neq k$.)

5. Continuing Problem 4, let h_{kk} be the kth diagonal entry of the projection matrix $X(X^tX)^{-1}X^t$. If $\hat{y}_k$ is the predicted value of y_k and $\hat{y}_k^{-k}$ is the predicted value of y_k omitting this observation, then demonstrate that

$$y_k - \hat{y}_k^{-k} = \frac{y_k - \hat{y}_k}{1 - h_{kk}}.$$

(Hint: Apply the Sherman–Morrison–Woodbury formula.)

6. Let A be an $m \times m$ positive definite matrix. The Cholesky decomposition B of A is a lower-triangular matrix with positive diagonal entries such that $A = BB^t$. To prove that such a decomposition exists we can argue by induction. Why is the case of a 1×1 matrix trivial? Now suppose A is partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Applying the induction hypothesis, there exist matrices C_{11} and D_{22} such that

$$\begin{aligned} C_{11}C_{11}^t &= A_{11} \\ D_{22}D_{22}^t &= A_{22} - A_{21}A_{11}^{-1}A_{12}. \end{aligned}$$

Prove that

$$B = \begin{pmatrix} C_{11} & 0_{12} \\ A_{21}(C_{11}^t)^{-1} & D_{22} \end{pmatrix}$$

gives the desired decomposition. Extend this argument to show that B is uniquely determined.

7. Continuing Problem 6, show that one can compute the Cholesky decomposition $B = (b_{ij})$ of $A = (a_{ij})$ by the recurrence relations

$$b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2}$$
$$b_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}}{b_{jj}}, \quad i > j$$

for columns $j = 1$, $j = 2$, and so forth until column $j = m$. How can you compute $\det A$ in terms of the entries of B?

8. Based on Problems 6 and 7, suppose the matrix $A = (a_{ij})$ is banded in the sense that $a_{ij} = 0$ when $|i - j| > d$. Prove that the Cholesky decomposition $B = (b_{ij})$ also satisfies the band condition $b_{ij} = 0$ when $|i - j| > d$.

9. In view of Problem 7, how many arithmetic operations does it take to compute the Cholesky decomposition of an $m \times m$ positive definite matrix A? You may count a square root as a single operation. If the solution of the linear system $Au = b$ is desired, describe how the Cholesky decomposition comes into play and estimate the number of extra operations needed to find $u = A^{-1}b$ once the Cholesky decomposition is obtained.

10. Describe how the Cholesky decomposition of Problem 6 can be used to generate a random sample from a multivariate normal random vector with mean μ and covariance Ω from a random sample of univariate, standard normal deviates.

11. Find the Gram–Schmidt decomposition UR of the matrix

$$X = \begin{pmatrix} 1 & 3 & 3 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \\ 1 & 1 & 3 \end{pmatrix}.$$

12. If $X = UR$ is the Gram–Schmidt decomposition of X, then show that the projection matrix $X(X^tX)^{-1}X^t = UU^t$.

13. Prove that the inverse of an upper-triangular matrix is upper triangular.

14. Consider the $p \times p$ matrix

$$M = \begin{pmatrix} a & b & \cdots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \vdots & \vdots \\ b & b & \cdots & a \end{pmatrix}.$$

If $\mathbf{1}$ is the column vector with all entries 1, then show that M has inverse $\frac{1}{a-b}(I_p - \frac{b}{a+(p-1)b}\mathbf{1}\mathbf{1}^t)$ and determinant $(a-b)^{p-1}[a+(p-1)b]$.

15. Let u be a vector with norm $||u||_2 = 1$. Compute the inverse and determinant of the Householder matrix $I - 2uu^t$.

16. Prove the slight generalization

$$(A + UDV^t)^{-1} = A^{-1} - A^{-1}U(D^{-1} + V^tA^{-1}U)^{-1}V^tA^{-1}$$

of the Woodbury formula (8) for compatible matrices A, U, D, and V.

17. Let A be an $m \times n$ matrix of full rank. One can easily show that

$$P = I_n - A^t(AA^t)^{-1}A$$

is the unique $n \times n$ orthogonal matrix projecting onto the null space of A; in other words, $P = P^t$, $P^2 = P$, and $Px = x$ if and only if $Ax = \mathbf{0}$. If b is a vector such that $b^tPb \neq 0$, then verify that the rank-one perturbation

$$Q = P - \frac{1}{b^tPb}Pbb^tP$$

is the unique orthogonal projection onto the null space of $\begin{pmatrix} A \\ b^t \end{pmatrix}$. If $b^tPb = 0$, then verify that $Q = P$ serves as the orthogonal projector.

References

[1] Dempster AP (1969) *Continuous Multivariate Analysis*. Addison-Wesley, Reading, MA

[2] Goodnight JH (1979) A tutorial on the SWEEP operator. *Amer Statistician* 33:149–158

[3] Henrici P (1982) *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*. Wiley, New York

[4] Jennrich RI (1977) Stepwise regression. *Statistical Methods for Digital Computers*. Enslein K, Ralston A, Wilf HS, editors, Wiley-Interscience, New York, pp 58–75

[5] Kennedy WJ Jr, Gentle JE (1980) *Statistical Computing*. Marcel Dekker, New York

[6] Little RJA, Rubin DB (1987) *Statistical Analysis with Missing Data*. Wiley, New York

[7] Miller KS (1987) *Some Eclectic Matrix Theory*. Robert E Krieger Publishing, Malabar, FL

[8] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[9] Strang G (1986) *Introduction to Applied Mathematics.* Wellesley-Cambridge Press, Wellesley, MA
[10] Thisted RA (1988) *Elements of Statistical Computing.* Chapman & Hall, New York

8

Eigenvalues and Eigenvectors

8.1 Introduction

Finding the eigenvalues and eigenvectors of a symmetric matrix is one of the basic tasks of computational statistics. For instance, in principal components analysis [9], a random m-vector X with covariance matrix Ω is postulated. As a symmetric matrix, Ω can be decomposed as

$$\Omega = UDU^t,$$

where D is the diagonal matrix of eigenvalues of Ω and U is the corresponding orthogonal matrix of eigenvectors. If the eigenvalues are distinct and ordered $0 \leq \lambda_1 < \lambda_2 < \ldots < \lambda_m$, then the columns $U_1, \ldots, U_m$ of U are unique up to sign. The random variables $V_j = U_j^t X$, $j = 1, \ldots, m$, have covariance matrix $U^t \Omega U = D$. These random variables are termed principal components. They are uncorrelated and increase in variance from the first, V_1, to the last, V_m.

Besides this classical application, there are other reasons for being interested in eigenvalues and eigenvectors. We have already seen how the dominant eigenvalue of a symmetric matrix Ω determines its norm $||\Omega||_2$. If Ω is the covariance matrix of a normally distributed random vector X with mean $\mathrm{E}(X) = \mu$, then the quadratic form and the determinant

$$(x-\mu)^t \Omega^{-1} (x-\mu) = [U^t(x-\mu)]^t D^{-1} U^t (x-\mu)$$
$$\det \Omega = \prod_i \lambda_i$$

appearing in the density of X are trivial to calculate if Ω can be diagonalized explicitly. Understanding the eigenstructure of matrices also is crucial in proving convergence for maximum likelihood algorithms. This leads us to consider the Rayleigh quotient later in this chapter.

8.2 Jacobi's Method

Rather than survey the variety of methods for computing the eigenvalues and eigenvectors of a symmetric matrix Ω, we will focus on just one, the classical Jacobi method [1, 2, 3, 5, 10]. This is not necessarily the fastest method, but it does illustrate some useful ideas for proving convergence of iterative methods in general. One attractive feature of Jacobi's method is the ease with which it can be implemented on parallel computers. This fact suggests that it may regain its competitiveness on large-scale problems.

The idea of the Jacobi method is to gradually transform Ω to a diagonal matrix by a sequence of similarity transformations. Each similarity transformation involves a rotation designed to increase the sum of squares of the diagonal entries of the matrix currently similar to Ω. In two dimensions, a rotation is a matrix

$$R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{1}$$

that rotates vectors in the plane clockwise by an angle θ [6]. Besides satisfying the orthogonality property $RR^t = I$, a rotation R also has $\det R = 1$. This distinguishes a rotation from a reflection, which is an orthogonal transformation R with $\det R = -1$. These are the only two choices for $\det R$ because $\det R = \det R^t$ and $\det I = 1$.

Suppose two $m \times m$ matrices A and B are related by the orthogonal similarity transformation $B = W^t AW$. Of crucial importance in Jacobi's method are the identities $\mathrm{tr}(B) = \mathrm{tr}(A)$ and $||B||_E^2 = ||A||_E^2$. These are verified by appealing to the circular permutation property of the trace function in the following computations

$$\begin{aligned} \mathrm{tr}(B) &= \mathrm{tr}(W^t AW) \\ &= \mathrm{tr}(AWW^t) \\ &= \mathrm{tr}(A) \\ ||B||_E^2 &= \mathrm{tr}(B^t B) \\ &= \mathrm{tr}(W^t A^t WW^t AW) \\ &= \mathrm{tr}(A^t AWW^t) \\ &= \mathrm{tr}(A^t A) \\ &= ||A||_E^2. \end{aligned}$$

Now consider the effect of a rotation involving row k and column l of the $m \times m$ matrix $A = (a_{ij})$. Without loss of generality, we take $k = 1$ and $l = 2$ and form the orthogonal matrix

$$W = \begin{pmatrix} R & \mathbf{0} \\ \mathbf{0}^t & I_{m-2} \end{pmatrix}.$$

The diagonal entry b_{ii} of $B = W^t AW$ equals a_{ii} when $i > 2$. The entries of the upper left block of B are given by

$$\begin{aligned} b_{11} &= a_{11}\cos^2\theta - 2a_{12}\cos\theta\sin\theta + a_{22}\sin^2\theta \\ b_{12} &= (a_{11} - a_{22})\cos\theta\sin\theta + a_{12}(\cos^2\theta - \sin^2\theta) \\ b_{22} &= a_{11}\sin^2\theta + 2a_{12}\cos\theta\sin\theta + a_{22}\cos^2\theta. \end{aligned} \tag{2}$$

By virtue of the trigonometric identities

$$\begin{aligned} \cos^2\theta - \sin^2\theta &= \cos(2\theta) \\ \cos\theta\sin\theta &= \frac{1}{2}\sin(2\theta), \end{aligned}$$

it follows that

$$b_{12} = \frac{a_{11} - a_{22}}{2}\sin(2\theta) + a_{12}\cos(2\theta).$$

When $a_{22} - a_{11} \neq 0$, there is a unique $|\theta| < \pi/4$ such that

$$\tan(2\theta) = \frac{2a_{12}}{a_{22} - a_{11}}.$$

Making this choice of θ forces $b_{12} = 0$. When $a_{22} - a_{11} = 0$, the choice $\theta = \pi/4$ also gives $b_{12} = 0$.

Given $b_{12} = 0$, we infer from the first two formulas of (2) that

$$\begin{aligned} b_{11} &= a_{11}\cos^2\theta - 2a_{12}\cos\theta\sin\theta + a_{22}\sin^2\theta + b_{12}\tan\theta \\ &= a_{11} - a_{12}\tan\theta. \end{aligned} \tag{3}$$

The trace identity $b_{11} + b_{22} = a_{11} + a_{22}$ then yields the corresponding equality

$$b_{22} = a_{22} + a_{12}\tan\theta.$$

The two-dimensional version of the identity $||B||_E^2 = ||A||_E^2$ applies to the upper left blocks of B and A. In other words, if $b_{12} = 0$, then

$$b_{11}^2 + b_{22}^2 = a_{11}^2 + 2a_{12}^2 + a_{22}^2.$$

In terms of the sums of squares of the diagonal entries of the matrices B and A, this translates into

$$S(B) = \sum_{i=1}^{m} b_{ii}^2$$

$$\begin{aligned} &= \sum_{i=1}^{m} a_{ii}^2 + 2a_{12}^2 \\ &= S(A) + 2a_{12}^2. \end{aligned}$$

Thus, choosing $b_{12} = 0$ forces $S(B) > S(A)$ whenever $a_{12} \neq 0$.

Beginning with a symmetric matrix Ω, Jacobi's method employs a sequence of rotations U_n as designed above to steadily decrease the sum of squares

$$||\Omega_n||_E^2 - S(\Omega_n) = ||\Omega||_E^2 - S(\Omega_n)$$

of the off-diagonal entries of the transformed matrices

$$\Omega_n = U_n^t \cdots U_1^t \Omega U_1 \cdots U_n.$$

For large n, approximate eigenvalues of Ω can be extracted from the diagonal of the nearly diagonal matrix Ω_n, and approximate eigenvectors can be extracted from the columns of the orthogonal matrix $O_n = U_1 \cdots U_n$.

In fact, there are several competing versions of Jacobi's method. The classical method selects the row i and column j giving the largest increase $2a_{ij}^2$ in the sum of squares $S(A)$ of the diagonal entries. The disadvantage of this strategy is that it necessitates searching through all off-diagonal entries of the current matrix A. A simpler strategy is to cycle through the off-diagonal entries according to some fixed schedule. In the threshold Jacobi method, this cyclic strategy is modified so that a rotation is undertaken only when the current off-diagonal entry a_{ij} is sufficiently large in absolute value. For purposes of theoretical exposition, it is simplest to adopt the classical strategy.

With this decision in mind, consider the sum of squares of the off-diagonal entries

$$\begin{aligned} L(A) &= \sum_k \sum_{l \neq k} a_{kl}^2 \\ &= ||A||_E^2 - S(A), \end{aligned}$$

and suppose that $B = W^t A W$, where W is a Jacobi rotation in the plane of i and j. Because a_{ij} is the largest off-diagonal entry in absolute value, $L(A) \leq m(m-1)a_{ij}^2$. Together with the relation $L(B) = L(A) - 2a_{ij}^2$, this implies

$$L(B) \leq \left[1 - \frac{2}{m(m-1)}\right] L(A).$$

Thus, the function $L(A)$ is driven to 0 at least as fast as the successive powers of $1 - 2/[m(m-1)]$. This clearly suggests the convergence of Jacobi's method to the diagonal matrix D.

Rigorously proving the convergence of Jacobi's method requires the next technical result.

Proposition 8.2.1. *Let x_n be a bounded sequence in R^p, and suppose*

$$\lim_{n\to\infty} ||x_{n+1} - x_n|| = 0$$

for some norm $||x||$. Then the set T of limit points of x_n is connected. If T is finite, it follows that T reduces to a single point and that $\lim_{n\to\infty} x_n = x_\infty$ exists.

Proof. It is straightforward to prove that T is a compact set. If it is disconnected, then it is contained in the union of two disjoint, open subsets S_1 and S_2 in such a way that neither $T \cap S_1$ nor $T \cap S_2$ is empty. The distance $d = \inf_{y\in T\cap S_1,\, z\in T\cap S_2} ||y - z||$ separating $T \cap S_1$ and $T \cap S_2$ must be positive; otherwise, there would be two sequences $y_n \in T \cap S_1$ and $z_n \in T \cap S_2$ with $||y_n - z_n|| < 1/n$. Because T is compact, there is a subsequence y_{n_k} of y_n that converges to a point of T. By passing to a subsubsequence if necessary, we can assume that z_{n_k} converges as well. The limits of these two convergent subsequences coincide. The fact that the common limit belongs to the open set S_1 and the boundary of S_2 or vice versa contradicts the disjointness of S_1 and S_2.

Now consider the sequence x_n in the statement of the proposition. For large enough n, we have $||x_{n+1} - x_n|| < d/4$. As the sequence x_n bounces back and forth between limit points in S_1 and S_2, it must enter the closed set $W = \{y : \inf_{z\in T} ||y - z|| \geq d/4\}$ infinitely often. But this means that W contains a limit point of x_n. Because W is disjoint from $T \cap S_1$ and $T \cap S_2$, and these two sets are postulated to contain all of the limit points of x_n, this contradiction implies that T is connected.

Since a finite set with more than one point is necessarily disconnected, T can be a finite set only if it consists of a single point. Finally, a bounded sequence with only a single limit point has that point as its limit. □

We are now in a position to prove convergence of Jacobi's method via the strategy of Michel Crouzeix [1]. Let us tackle eigenvalues first.

Proposition 8.2.2. *Suppose that Ω is an $m \times m$ symmetric matrix. The classical Jacobi strategy generates a sequence of rotations U_n and a sequence of similar matrices Ω_n related to Ω by*

$$\Omega_n = U_n^t \cdots U_1^t \Omega U_1 \cdots U_n.$$

With the rotations U_n chosen as described above, $\lim_{n\to\infty} \Omega_n$ exists and equals a diagonal matrix D whose entries are the eigenvalues of Ω in some order.

Proof. If Jacobi's method gives a diagonal matrix in a finite number of iterations, there is nothing to prove. Otherwise, let D_n be the diagonal part of Ω_n. We have already argued that the off-diagonal part $\Omega_n - D_n$ of Ω_n tends to the zero matrix $\mathbf{0}$. Because $||D_n||_E \leq ||\Omega_n||_E = ||\Omega||_E$, the sequence D_n is bounded in R^{m^2}. Let D_{n_k} be a convergent subsequence with

limit D, not necessarily assumed to represent the eigenvalues of Ω. Owing to the similarity of the matrix Ω_{n_k} to Ω, we find

$$\begin{aligned}\det(D-\lambda I) &= \lim_{k\to\infty}\det(D_{n_k}-\lambda I)\\ &= \lim_{k\to\infty}\det(\Omega_{n_k}-\lambda I)\\ &= \det(\Omega-\lambda I).\end{aligned}$$

Thus, D possesses the same eigenvalues, counting multiplicities, as Ω. But the eigenvalues of D are just its diagonal entries.

To rule out more than one limit point D, we apply Proposition 8.2.1, noting that there are only a finite number of permutations of the eigenvalues of Ω. According to equation (3) and its immediate sequel, if U_n is a rotation through an angle θ_n in the plane of entries i and j, then the diagonal entries of D_{n+1} and D_n satisfy

$$d_{n+1,kk}-d_{nkk}=\begin{cases}0 & k\neq i,j\\ -\omega_{nij}\tan\theta_n & k=i\\ +\omega_{nij}\tan\theta_n & k=j,\end{cases}$$

where $\Omega_n=(\omega_{nkl})$. Because $|\theta_n|\leq\pi/4$ and $|\omega_{nij}|\leq||\Omega_n-D_n||_E$, it follows that $\lim_{n\to\infty}||D_{n+1}-D_n||_E=0$. □

Proposition 8.2.3. *If all of the eigenvalues λ_i of the matrix Ω are distinct, then the sequence $O_n=U_1\cdots U_n$ of orthogonal matrices constructed in Proposition 8.2.2 converges to the matrix of eigenvectors of the limiting diagonal matrix D.*

Proof. Mimicking the strategy of Proposition 8.2.2, we show that the sequence O_n is bounded, that it possesses only a finite number of limit points, and that

$$\lim_{n\to\infty}||O_{n+1}-O_n||_E=0.$$

The sequence O_n is bounded because $||O_n||_E^2=\operatorname{tr}(O_n^tO_n)=\operatorname{tr}(I)$. Suppose O_n has a convergent subsequence O_{n_k} with limit O. Then $D=O^t\Omega O$ holds because of Proposition 8.2.2. This implies that the columns of O are the orthonormal eigenvectors of Ω ordered consistently with the eigenvalues appearing in D. The eigenvectors are unique up to sign. Thus, O can be one of only 2^m possibilities.

To prove $\lim_{n\to\infty}||O_{n+1}-O_n||_E=0$, again suppose that U_n is a rotation through an angle θ_n in the plane of entries i and j. This angle is defined by $\tan(2\theta_n)=2\omega_{nij}/(\omega_{njj}-\omega_{nii})$. Because Ω_n converges to D and the entries of D are presumed unique,

$$\begin{aligned}|\omega_{njj}-\omega_{nii}| &> \frac{1}{2}\min_{k\neq l}|d_{kk}-d_{ll}|\\ &> 0\end{aligned}$$

for all sufficiently large n. In view of the fact that $\lim_{n\to\infty} \omega_{nij} = 0$, this implies that $\lim_{n\to\infty} \theta_n = 0$, which in turn yields $\lim_{n\to\infty} U_n = I$. The inequality

$$\begin{aligned} ||O_{n+1} - O_n||_E &= ||O_n(U_{n+1} - I)||_E \\ &\leq ||O_n||_E ||U_{n+1} - I||_E \end{aligned}$$

completes the proof. □

8.3 The Rayleigh Quotient

Sometimes it is helpful to characterize the eigenvalues and eigenvectors of an $m \times m$ symmetric matrix A in terms of the extrema of the Rayleigh quotient

$$R(x) = \frac{x^t A x}{x^t x}$$

defined for $x \neq \mathbf{0}$. This was the case, for example, in computing the norm $||A||_2$. To develop the properties of the Rayleigh quotient, let A have eigenvalues $\lambda_1 \leq \cdots \leq \lambda_m$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_m$. Because any x can be written as a unique linear combination $\sum_i c_i u_i$, the Rayleigh quotient can be expressed as

$$R(x) = \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2}. \tag{4}$$

This representation clearly yields the inequality $R(x) \leq \lambda_m$ and the equality $R(u_m) = \lambda_m$. Hence, $R(x)$ is maximized by $x = u_m$ and correspondingly minimized by $x = u_1$. The Courant–Fischer theorem is a notable generalization of these results.

Proposition 8.3.1 (Courant–Fischer)**.** *Let V_k be a k-dimensional subspace of R^m. Then*

$$\begin{aligned} \lambda_k &= \min_{V_k} \max_{x \in V_k,\, x \neq \mathbf{0}} R(x) \\ &= \max_{V_{m-k+1}} \min_{x \in V_{m-k+1},\, x \neq \mathbf{0}} R(x). \end{aligned}$$

The minimum in the first characterization of λ_k is attained for the subspace spanned by $u_1, \ldots, u_k$, and the maximum in the second characterization of λ_k is attained for the subspace spanned by $u_k, \ldots, u_m$.

Proof. If U_k is the subspace spanned by $u_1, \ldots, u_k$, then it is clear that

$$\lambda_k = \max_{x \in U_k,\, x \neq \mathbf{0}} R(x).$$

If V_k is an arbitrary subspace of dimension k, then there must be some nontrivial vector $x \in V_k$ orthogonal to $u_1, \ldots, u_{k-1}$. For this $x = \sum_{i=k}^m c_i u_i$,

we find

$$\begin{aligned} R(x) &= \frac{\sum_{i=k}^m \lambda_i c_i^2}{\sum_{i=k}^m c_i^2} \\ &\geq \frac{\lambda_k \sum_{i=k}^m c_i^2}{\sum_{i=k}^m c_i^2} \\ &= \lambda_k. \end{aligned}$$

This proves that $\max_{x \in V_k,\, x \neq \mathbf{0}} R(x) \geq \lambda_k$. The second characterization of λ_k follows from the first characterization applied to $-A$, whose eigenvalues are $-\lambda_m \leq \cdots \leq -\lambda_1$. □

The next proposition applies the Courant–Fischer theorem to the problem of estimating how much the eigenvalues of a symmetric matrix change under a symmetric perturbation of the matrix.

Proposition 8.3.2. *Let the $m \times m$ symmetric matrices A and $B = A + \Delta A$ have ordered eigenvalues $\lambda_1 \leq \cdots \leq \lambda_m$ and $\mu_1 \leq \cdots \leq \mu_m$, respectively. Then the inequality*

$$|\lambda_k - \mu_k| \leq ||\Delta A||_2$$

holds for all $k \in \{1, \ldots, m\}$.

Proof. Suppose that U_k is the subspace of R^m spanned by the eigenvectors $u_1, \ldots, u_k$ of A corresponding to $\lambda_1, \ldots, \lambda_k$. If V_k is an arbitrary subspace of dimension k, then the identity $R_B(x) = R_A(x) + R_{\Delta A}(x)$ and the Courant–Fischer theorem imply that

$$\begin{aligned} \mu_k &= \min_{V_k} \max_{x \in V_k,\, x \neq \mathbf{0}} R_B(x) \\ &\leq \max_{x \in U_k,\, x \neq \mathbf{0}} R_B(x) \\ &\leq \lambda_k + \max_{x \in U_k,\, x \neq \mathbf{0}} R_{\Delta A}(x) \\ &\leq \lambda_k + \max_{x \neq \mathbf{0}} R_{\Delta A}(x) \\ &\leq \lambda_k + ||\Delta A||_2. \end{aligned}$$

If we reverse the roles of A and B, the inequality $\lambda_k \leq \mu_k + ||\Delta A||_2$ follows similarly. □

Finally, it is worth generalizing the above analysis to some nonsymmetric matrices. Suppose that A and B are symmetric matrices with B positive definite. An eigenvalue λ of $B^{-1}A$ satisfies $B^{-1}Ax = \lambda x$ for some $x \neq \mathbf{0}$. This identity is equivalent to the identity $Ax = \lambda Bx$. Taking the inner product of this latter identity with x suggests examining the generalized Rayleigh quotient [4]

$$R(x) = \frac{x^t A x}{x^t B x}. \tag{5}$$

For instance, it is easy to prove that the maximum and minimum eigenvalues of $B^{-1}A$ coincide with the maximum and minimum values of (5). Furthermore, the Courant–Fischer theorem carries over.

Another useful perspective on this subject is gained by noting that B has a symmetric square root $B^{1/2}$ defined in terms of its diagonalization $B = UDU^t$ by $B^{1/2} = UD^{1/2}U^t$. The eigenvalue equation $B^{-1}Ax = \lambda x$ is then equivalent to

$$B^{-\frac{1}{2}}AB^{-\frac{1}{2}}y = \lambda y$$

for $y = B^{1/2}x$. This lands us back in the territory of symmetric matrices and the ordinary Rayleigh quotient. The extended Courant–Fischer theorem now follows directly from the standard Courant–Fischer theorem.

8.4 Problems

1. For symmetric matrices A and B, define $A \trianglerighteq 0$ to mean that A is nonnegative definite and $A \trianglerighteq B$ to mean that $A - B \trianglerighteq 0$. Show that $A \trianglerighteq B$ and $B \trianglerighteq C$ imply $A \trianglerighteq C$. Also show that $A \trianglerighteq B$ and $B \trianglerighteq A$ imply $A = B$. Thus, $\trianglerighteq$ induces a partial order on the set of symmetric matrices.

2. Find the eigenvalues and eigenvectors of the matrix

$$\Omega = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

of R.S. Wilson by Jacobi's method. You may use the appropriate subroutine in Press et al. [4].

3. Find the eigenvalues and eigenvectors of the rotation matrix (1). Note that the eigenvalues are complex conjugates.

4. Consider the reflection matrix

$$\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Show that it is orthogonal with determinant -1. Find its eigenvalues and eigenvectors.

5. Let Ω be an $m \times m$ symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_m$. If $|\lambda_m| > |\lambda_i|$ for every $i \neq m$, then the power method can be used to find λ_m and u_m. For any $x_0 \neq 0$, define a sequence of vectors x_n and a sequence of associated numbers σ_n by $x_n = \Omega x_{n-1}$ and $\sigma_n = x_n^t x_{n+1}/x_n^t x_n$. Provided $x_0^t u_m \neq 0$, demonstrate that $\lim_{n\to\infty} \sigma_n = \lambda_m$ and that either $\lim_{n\to\infty} x_n/|x_n||_2 = \pm u_m$ or $\lim_{n\to\infty}(-1)^n x_n/||x_n||_2 = \pm u_m$.

6. Suppose the $m \times m$ symmetric matrix Ω has eigenvalues

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{m-1} < \lambda_m.$$

The iterative scheme $x_{n+1} = (\Omega - \eta_n I)x_n$ can be used to approximate either λ_1 or λ_m. Consider the criterion

$$\sigma_n = \frac{x_{n+1}^t \Omega x_{n+1}}{x_{n+1}^t x_{n+1}}.$$

Choosing η_n to maximize σ_n causes $\lim_{n\to\infty} \sigma_n = \lambda_m$, while choosing η_n to minimize σ_n causes $\lim_{n\to\infty} \sigma_n = \lambda_1$. If $\tau_k = x_n^t \Omega^k x_n$, then show that the extrema of σ_n as a function of η are given by the roots of the quadratic equation

$$0 = \det \begin{pmatrix} 1 & \eta & \eta^2 \\ \tau_0 & \tau_1 & \tau_2 \\ \tau_1 & \tau_2 & \tau_3 \end{pmatrix}.$$

7. Apply the algorithm of the previous problem to find the largest and smallest eigenvalue of the matrix in Problem 2.

8. Show that the extended Rayleigh quotient (5) has gradient

$$\frac{2[A - R(x)B]x}{x^t B x}.$$

Argue that the eigenvalues and eigenvectors of $B^{-1}A$ are the stationary values and stationary points, respectively, of $R(x)$.

9. In the notation of Problem 1, show that two positive definite matrices A and B satisfy $A \rhd B$ if and only they satisfy $B^{-1} \rhd A^{-1}$. If $A \rhd B$, then prove that $\det A \geq \det B$ and $\operatorname{tr} A \geq \operatorname{tr} B$.

10. In Proposition 8.3.2 suppose the matrix ΔA is nonnegative definite. Prove that $\lambda_k \leq \mu_k$ for all k.

11. Let A and B be $m \times m$ symmetric matrices. Denote the smallest and largest eigenvalues of the convex combination $\alpha A + (1-\alpha)B$ by $\lambda_1[\alpha A + (1-\alpha)B]$ and $\lambda_m[\alpha A + (1-\alpha)B]$, respectively. For $\alpha \in [0, 1]$, demonstrate that

$$\lambda_1[\alpha A + (1-\alpha)B] \geq \alpha\lambda_1[A] + (1-\alpha)\lambda_1[B]$$
$$\lambda_m[\alpha A + (1-\alpha)B] \leq \alpha\lambda_m[A] + (1-\alpha)\lambda_m[B].$$

12. Given the assumptions of the previous problem, show that the smallest and largest eigenvalues satisfy

$$\lambda_1[A + B] \geq \lambda_1[A] + \lambda_1[B]$$
$$\lambda_m[A + B] \leq \lambda_m[A] + \lambda_m[B].$$

13. One of the simplest ways of showing that a symmetric matrix is nonnegative definite is to show that it is the covariance matrix of a random vector. Use this insight to prove that if the symmetric matrices $A = (a_{ij})$ and $B = (b_{ij})$ are nonnegative definite, then the matrix

$C = (c_{ij})$ with $c_{ij} = a_{ij}b_{ij}$ is also nonnegative definite [7]. (Hint: Take independent random vectors X and Y with covariance matrices A and B and form the random vector Z with components $Z_i = X_iY_i$.)

14. Continuing Problem 13, suppose that the $n \times n$ symmetric matrices A and B have entries $a_{ij} = i(n - j + 1)$ and $b_{ij} = \sum_{k=1}^{i} \sigma_k^2$ for $j \geq i$ and $\sigma_k^2 \geq 0$. Show that A and B are nonnegative definite [7]. (Hint: For A, consider the order statistics from a random sample of the uniform distribution on $[0, 1]$.)

References

[1] Ciarlet PG (1989) *Introduction to Numerical Linear Algebra and Optimization*. Cambridge University Press, Cambridge

[2] Golub GH, Van Loan CF (1989) *Matrix Computations*, 2nd ed. Johns Hopkins University Press, Baltimore, MD

[3] Hämmerlin G, Hoffmann K-H (1991) *Numerical Mathematics*. Springer-Verlag, New York

[4] Hestenes MR (1981) *Optimization Theory: The Finite Dimensional Case*. Robert E Krieger Publishing, Huntington, NY

[5] Isaacson E, Keller HB (1966) *Analysis of Numerical Methods*. Wiley, New York

[6] Lang S (1971) *Linear Algebra*, 2nd ed. Addison-Wesley, Reading, MA

[7] Olkin I (1985) A probabilistic proof of a theorem of Schur. *Amer Math Monthly* 92:50-51

[8] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge

[9] Rao CR (1973) *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York

[10] Thisted RA (1988) *Elements of Statistical Computing*. Chapman & Hall, New York

9
Splines

9.1 Introduction

Splines are used for interpolating functions. Before the advent of computer graphics, a draftsman would draw a smooth curve through a set of points plotted on graph paper by forcing a flexible strip to pass over the points. The strip, made of wood, metal, or plastic, typically was held in place by weights as the draftsman drew along its edge. Subject to passing through the interpolating points, the strip or spline would minimize its stress by straightening out as much as possible. Beyond the terminal points on the left and right, the spline would be straight.

Mathematical splines are idealizations of physical splines. For simplicity we will deal only with the most commonly used splines, cubic splines. These are piecewise cubic polynomials that interpolate a tabulated function $f(x)$ at certain data points $x_0 < x_1 < \cdots < x_n$ called nodes or knots. There are, of course, many ways of interpolating a function. For example, Lagrange's interpolation formula provides a polynomial $p(x)$ of degree n that agrees with $f(x)$ at the nodes. Unfortunately, interpolating polynomials can behave poorly even when fitted to slowly varying functions. (See Problem 1 for a discussion of the classical example of Runge.) Splines minimize average squared curvature and consequently perform better than interpolating polynomials.

The program of this chapter is to investigate a few basic properties of cubic splines, paying particular attention to issues of computing. We then show how splines can be employed in nonparametric regression. For a much

fuller exposition of splines, readers can consult one or more of the books [1, 2, 3, 5, 8].

9.2 Definition and Basic Properties

We start with a formal definition that uniquely determines a spline.

Definition. Let the values $f(x_i) = f_i$ of the function $f(x)$ be given at the points $x_0 < x_1 < \cdots < x_n$. A natural, cubic, interpolatory spline $s(x)$ is a function on the interval $[x_0, x_n]$ possessing the following properties:

(a) $s(x)$ is a cubic polynomial on each node-to-node interval $[x_i, x_{i+1}]$;
(b) $s(x_i) = f_i$ at each node x_i;
(c) the second derivative $s''(x)$ exists and is continuous throughout the entire interval $[x_0, x_n]$;
(d) at the terminal nodes, $s''(x_0) = s''(x_n) = 0$.

For brevity we simply call $s(x)$ a spline.

Proposition 9.2.1. *There is exactly one function $s(x)$ on $[x_0, x_n]$ satisfying the above properties.*

Proof. For notational convenience, let

$$\begin{aligned} h_i &= x_{i+1} - x_i \\ \sigma_i &= s''(x_i) \\ s_i(x) &= s(x), \qquad x \in [x_i, x_{i+1}]. \end{aligned}$$

Note that the second derivatives σ_i are as yet unknown. Because $s_i(x)$ is a cubic polynomial, $s_i''(x)$ is a linear polynomial that can be expressed as

$$s_i''(x) = \sigma_i \frac{x_{i+1} - x}{h_i} + \sigma_{i+1} \frac{x - x_i}{h_i}. \tag{1}$$

The function $s''(x)$ pieced together in this fashion is clearly continuous on $[x_0, x_n]$. Integrating equation (1) twice gives

$$\begin{aligned} s_i(x) = \frac{\sigma_i}{6h_i}(x_{i+1} - x)^3 + \frac{\sigma_{i+1}}{6h_i}(x - x_i)^3 \\ + \; c_1(x - x_i) + c_2(x_{i+1} - x). \end{aligned} \tag{2}$$

The constants of integration c_1 and c_2 can be determined from the interpolation conditions

$$\begin{aligned} f_i &= s_i(x_i) = \frac{\sigma_i}{6} h_i^2 + c_2 h_i \\ f_{i+1} &= s_i(x_{i+1}) = \frac{\sigma_{i+1}}{6} h_i^2 + c_1 h_i. \end{aligned}$$

Solving for c_1 and c_2 and substituting the results in equation (2) produce

$$s_i(x) = \frac{\sigma_i}{6h_i}(x_{i+1} - x)^3 + \frac{\sigma_{i+1}}{6h_i}(x - x_i)^3 + \Big(\frac{f_{i+1}}{h_i} - \frac{\sigma_{i+1}h_i}{6}\Big)(x - x_i) + \Big(\frac{f_i}{h_i} - \frac{\sigma_i h_i}{6}\Big)(x_{i+1} - x). \tag{3}$$

Since it satisfies the interpolation conditions, $s(x)$ as defined by (3) is continuous on $[x_0, x_n]$.

Choosing the σ_i appropriately also will guarantee that $s'(x)$ is continuous. Differentiating equation (2) yields

$$s_i'(x) = -\frac{\sigma_i}{2h_i}(x_{i+1} - x)^2 + \frac{\sigma_{i+1}}{2h_i}(x - x_i)^2 + \frac{f_{i+1} - f_i}{h_i} - \frac{h_i}{6}(\sigma_{i+1} - \sigma_i). \tag{4}$$

Continuity is achieved when $s_{i-1}'(x_i) = s_i'(x_i)$. In terms of equation (4), this is equivalent to

$$\frac{1}{6}h_{i-1}\sigma_{i-1} + \frac{1}{3}(h_{i-1} + h_i)\sigma_i + \frac{1}{6}h_i\sigma_{i+1} = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}. \tag{5}$$

This is a system of $n-1$ equations for the $n+1$ unknown $\sigma_0, \ldots, \sigma_n$. However, two of these unknowns σ_0 and σ_n are 0 by assumption. If by good fortune the $(n-1) \times (n-1)$ matrix of coefficients multiplying the remaining unknowns is invertible, then we can solve for $\sigma_1, \ldots, \sigma_{n-1}$ uniquely. Invertibility follows immediately from the strict diagonal dominance conditions

$$\begin{aligned} \frac{1}{3}(h_0 + h_1) &> \frac{1}{6}h_1 \\ \frac{1}{3}(h_{i-1} + h_i) &> \frac{1}{6}h_{i-1} + \frac{1}{6}h_i \qquad i = 2, \ldots, n-2 \\ \frac{1}{3}(h_{n-1} + h_n) &> \frac{1}{6}h_{n-1}. \end{aligned}$$

This completes the proof because, as already indicated, the coefficients $\sigma_1, \ldots, \sigma_{n-1}$ uniquely determine the spline $s(x)$. □

To solve for $\sigma_1, \ldots, \sigma_{n-1}$, one can use functional iteration as described in Chapter 6. In practice, it is better to exploit the fact that the matrix of coefficients is tridiagonal. To do so, define

$$d_i = \frac{6}{h_i}\Big(\frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}\Big)$$

and rewrite the system (5) as

$$\frac{h_{i-1}}{h_i}\sigma_{i-1} + 2(1 + \frac{h_{i-1}}{h_i})\sigma_i + \sigma_{i+1} = d_i. \tag{6}$$

Now set

$$\sigma_{i-1} = \rho_i \sigma_i + \tau_i, \tag{7}$$

where ρ_i and τ_i are constants to be determined. In view of $\sigma_0 = 0$, we take $\rho_1 = \tau_1 = 0$. In general, substitution of equation (7) in equation (6) leads to

$$\sigma_i = -\frac{\sigma_{i+1}}{\frac{h_{i-1}}{h_i}\rho_i + 2(1 + \frac{h_{i-1}}{h_i})} + \frac{d_i - \frac{h_{i-1}}{h_i}\tau_i}{\frac{h_{i-1}}{h_i}\rho_i + 2(1 + \frac{h_{i-1}}{h_i})}.$$

This has the form of equation (7) and suggests computing

$$\begin{aligned}
\rho_{i+1} &= -\frac{1}{\frac{h_{i-1}}{h_i}\rho_i + 2(1 + \frac{h_{i-1}}{h_i})} \\
\tau_{i+1} &= \frac{d_i - \frac{h_{i-1}}{h_i}\tau_i}{\frac{h_{i-1}}{h_i}\rho_i + 2(1 + \frac{h_{i-1}}{h_i})}
\end{aligned}$$

recursively beginning at $i = 0$. Once the constants ρ_i and τ_i are available, then the σ_i can be computed in order from equation (7) beginning at $i = n$.

The next proposition validates the minimum curvature property of natural cubic splines.

Proposition 9.2.2. *Let $s(x)$ be the spline interpolating the function $f(x)$ at the nodes $x_0 < x_1 < \cdots < x_n$. If $g(x)$ is any other twice continuously differentiable function interpolating $f(x)$ at these nodes, then*

$$\int_{x_0}^{x_n} g''(x)^2 dx \geq \int_{x_0}^{x_n} s''(x)^2 dx, \tag{8}$$

with equality only if $g(x) = s(x)$ throughout $[x_0, x_n]$.

Proof. If $\int_{x_0}^{x_n} g''(x)^2 dx = \infty$, then there is nothing to prove. Therefore, assume the contrary and consider the identity

$$\begin{aligned}
\int_{x_0}^{x_n} [g''(x) - s''(x)]^2 dx = &\int_{x_0}^{x_n} g''(x)^2 dx - 2\int_{x_0}^{x_n} [g''(x) - s''(x)]s''(x)dx \\
&- \int_{x_0}^{x_n} s''(x)^2 dx.
\end{aligned} \tag{9}$$

Let us prove that the second integral on the right-hand side of equation (9) vanishes. Decomposing this integral and integrating each piece by parts give

$$\begin{aligned}
&\int_{x_0}^{x_n} [g''(x) - s''(x)]s''(x)dx \\
&\quad = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} [g''(x) - s''(x)]s''(x)dx
\end{aligned}$$

$$= \sum_{i=0}^{n-1} \Big\{ [g'(x) - s'(x)]s''(x)|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} [g'(x) - s'(x)]s'''(x)dx \Big\}.$$

Since $s(x)$ is a piecewise cubic polynomial, $s'''(x)$ equals some constant α_i on $[x_i, x_{i+1}]$. Thus, we find

$$\begin{aligned} \int_{x_i}^{x_{i+1}} [g'(x) - s'(x)]s'''(x)dx\} &= \alpha_i[g(x) - s(x)]_{x_i}^{x_{i+1}} \\ &= 0 \end{aligned}$$

because $g(x)$ and $s(x)$ agree with $f(x)$ at each node. We are left with

$$\sum_{i=0}^{n-1} \Big\{ s''(x_{i+1})[g'(x_{i+1}) - s'(x_{i+1})] - s''(x_i)[g'(x_i) - s'(x_i)] \Big\},$$

which telescopes to

$$s''(x_n)[g'(x_n) - s'(x_n)] - s''(x_0)[g'(x_0) - s'(x_0)].$$

By assumption, $s''(x_0) = s''(x_n) = 0$.

This proves our contention about the vanishing of the second integral on the right-hand side of equation(9) and allows us to write

$$\int_{x_0}^{x_n} g''(x)^2 dx = \int_{x_0}^{x_n} s''(x)^2 dx + \int_{x_0}^{x_n} [g''(x) - s''(x)]^2 dx. \qquad (10)$$

Inequality (8) is now obvious. If equality obtains in inequality (8), then the continuous function $g''(x) - s''(x)$ is identically 0. This implies that $s(x) = g(x) + c_0 + c_1 x$ for certain constants c_0 and c_1. Because $s(x)$ and $g(x)$ both interpolate $f(x)$ at x_0 and x_1, it follows that $c_0 = c_1 = 0$. □

Remark. The curvature of a function $g(x)$ is technically the function $\kappa(x) = g''(x)/[1 + g'(x)^2]^{\frac{3}{2}}$. For $|g'(x)| \ll 1$, we recover $g''(x)$. Proposition 9.2.2 should be interpreted in this light.

The final proposition of this section provides bounds on the errors committed in spline approximation.

Proposition 9.2.3. *Suppose that $f(x)$ is twice continuously differentiable and $s(x)$ is the spline interpolating $f(x)$ at the nodes $x_0 < x_1 < \cdots < x_n$. If $h = \max_{0 \le i \le n-1}(x_{i+1} - x_i)$, then*

$$\begin{aligned} \max_{x_0 \le x \le x_n} |f(x) - s(x)| &\le h^{\frac{3}{2}} \Big[\int_{x_0}^{x_n} f''(y)^2 dy \Big]^{\frac{1}{2}} \\ \max_{x_0 \le x \le x_n} |f'(x) - s'(x)| &\le h^{\frac{1}{2}} \Big[\int_{x_0}^{x_n} f''(y)^2 dy \Big]^{\frac{1}{2}}. \end{aligned} \qquad (11)$$

It follows that $s(x)$ and $s'(x)$ converge uniformly to $f(x)$ and $f'(x)$ as the mesh length h goes to 0.

Proof. Any $x \in [x_0, x_n]$ lies in some interval $[x_i, x_{i+1}]$. Because $f(t)-s(t)$ vanishes at x_i and x_{i+1}, Rolle's theorem indicates that $f'(z) - s'(z) = 0$ for some $z \in [x_i, x_{i+1}]$. Hence,

$$f'(x) - s'(x) = \int_z^x [f''(y) - s''(y)]dy,$$

and consequently the Cauchy–Schwarz inequality implies

$$\begin{aligned} |f'(x) - s'(x)| &\leq \Big\{ \int_z^x [f''(y) - s''(y)]^2 dy \Big\}^{\frac{1}{2}} \Big\{ \int_z^x 1^2 dy \Big\}^{\frac{1}{2}} \\ &= \Big\{ \int_z^x [f''(y) - s''(y)]^2 dy \Big\}^{\frac{1}{2}} |x - z|^{\frac{1}{2}} \\ &\leq \Big\{ \int_{x_0}^{x_n} [f''(y) - s''(y)]^2 dy \Big\}^{\frac{1}{2}} h^{\frac{1}{2}}. \end{aligned}$$

In view of equation (10) with $g(x) = f(x)$, this gives the second inequality of (11).

To prove the first inequality, again let $x \in [x_i, x_{i+1}]$. Then

$$\begin{aligned} |f(x) - s(x)| &= \Big| \int_{x_i}^x [f'(y) - s'(y)]dy \Big| \\ &\leq \int_{x_i}^x \max_{x_0 \leq z \leq x_n} |f'(z) - s'(z)|dy \\ &\leq h \max_{x_0 \leq z \leq x_n} |f'(z) - s'(z)|. \end{aligned}$$

Substituting in this inequality the second inequality of (11) yields the first inequality of (11). □

Remark. Better error bounds are available when $f(x)$ possesses more derivatives and the nodes are uniformly spaced [3]. For instance, if the fourth derivative $f^{(4)}(x)$ exists and is continuous, and the uniform spacing is h, then

$$\max_{x_0 \leq x \leq x_n} |f(x) - s(x)| \leq \frac{h^4}{16} \max_{x_0 \leq x \leq x_n} |f^{(4)}(x)|. \tag{12}$$

9.3 Applications to Differentiation and Integration

Splines can be quite useful in numerical differentiation and integration. For example, equation (4) of Proposition 9.2.1 offers an accurate method of numerically differentiating $f(x)$ at any point $x \in [x_0, x_n]$. To integrate $f(x)$, we note that equation (3) implies

$$\int_{x_i}^{x_{i+1}} s_i(x)dx$$

$$= \frac{\sigma_i}{24}h_i^3 + \frac{\sigma_{i+1}}{24}h_i^3 + \Big(\frac{f_{i+1}}{h_i} - \frac{\sigma_{i+1}h_i}{6}\Big)\frac{h_i^2}{2} + \Big(\frac{f_i}{h_i} - \frac{\sigma_i h_i}{6}\Big)\frac{h_i^2}{2}$$
$$= \frac{f_i + f_{i+1}}{2}h_i - \frac{\sigma_i + \sigma_{i+1}}{24}h_i^3.$$

It follows that

$$\int_{x_0}^{x_n} f(x)dx \approx \int_{x_0}^{x_n} s(x)dx$$
$$= \sum_{i=0}^{n-1}\Big[\frac{f_i + f_{i+1}}{2}h_i - \frac{\sigma_i + \sigma_{i+1}}{24}h_i^3\Big].$$

According to inequality (12), the error committed in the approximation $\int_{x_0}^{x_n} f(x)dx \approx \int_{x_0}^{x_n} s(x)ds$ is bounded above by

$$\frac{h^4}{16}(x_n - x_0)\max_{x_0 \le x \le x_n}|f^{(4)}(x)|$$

for nodes with uniform spacing h.

9.4 Application to Nonparametric Regression

In parametric regression, one minimizes a weighted sum of squares

$$\sum_{i=0}^{n} w_i[y_i - g(x_i)]^2 \tag{13}$$

over a particular class of functions $g(x)$, taking the observations y_i and the weights $w_i > 0$ as given. For instance, in polynomial regression, the relevant class consists of all polynomials of a certain degree d or less. In time series analysis, the class typically involves linear combinations of a finite number of sines and cosines. However, often there is no convincing rationale for restricting attention to a narrow class of candidate regression functions. This has prompted statisticians to look at wider classes of functions.

At first glance, some restriction on the smoothness of the regression functions seems desirable. This is a valuable insight, but one needs to exercise caution because there exist many infinitely differentiable functions reducing the weighted sum of squares (13) to 0. For example, the unique polynomial of degree n interpolating the observed values y_i at the points x_i achieves precisely this. Smoothness per se is insufficient. Control of the overall size of the derivatives of the regression function is also important. One criterion incorporating these competing aims is the convex combination

$$J_\alpha(g) = \alpha\sum_{i=0}^{n} w_i[y_i - g(x_i)]^2 + (1-\alpha)\int_{x_0}^{x_n} g''(x)^2 dx \tag{14}$$

for $0 < \alpha < 1$. Minimizing $J_\alpha(g)$ reaches a compromise between minimizing the weighted sum of squares and minimizing the average squared curvature of the regression function. For α near 1, the weighted sum of squares predominates. For α near 0, the average squared curvature takes precedence. One immediate consequence of Proposition 9.2.2 is that the class of relevant functions collapses to the class of splines. For if $g(x)$ is twice continuously differentiable, then the spline $s(x)$ that interpolates $g(x)$ at the nodes x_i contributes the same weighted sum of squares and a reduced integral term; in other words, $J_\alpha(s) \leq J_\alpha(g)$.

To find the spline $s(x)$ minimizing J_α, we take the approach of de Boor [1] and extend the notation of Proposition 9.2.1. The system of $n-1$ equations displayed in (5) can be summarized by defining the vectors

$$\begin{aligned}\sigma &= (\sigma_1, \ldots, \sigma_{n-1})^t \\ f &= (f_0, \ldots, f_n)^t = [s(x_0), \ldots, s(x_n)]^t \\ y &= (y_0, \ldots, y_n)^t,\end{aligned}$$

the $(n-1) \times (n-1)$ tridiagonal matrix R with entries

$$r_{ij} = \frac{1}{6} \begin{cases} h_{i-1} & j = i-1 \\ 2(h_{i-1} + h_i) & j = i \\ h_i & j = i+1 \\ 0 & \text{otherwise,} \end{cases}$$

and the $(n-1) \times (n+1)$ tridiagonal matrix Q with entries

$$q_{ij} = \begin{cases} \frac{1}{h_{i-1}} & j = i-1 \\ -(\frac{1}{h_{i-1}} + \frac{1}{h_i}) & j = i \\ \frac{1}{h_i} & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

In this notation, the system of equations (5) is expressed as $R\sigma = Qf$. If we also let W be the diagonal matrix with ith diagonal entry w_i, then the weighted sum of squares (13) becomes $(y-f)^t W(y-f)$.

The integral contribution to $J_\alpha(s)$ can be represented in matrix notation by observing that equation (1) implies

$$\begin{aligned}&\int_{x_i}^{x_{i+1}} s''(x)^2 dx \\ &\quad = \frac{1}{h_i^2} \int_{x_i}^{x_{i+1}} [\sigma_i(x_{i+1} - x) + \sigma_{i+1}(x - x_i)]^2 dx \\ &\quad = \frac{1}{h_i^2} \frac{\sigma_i^2 h_i^3}{3} + \frac{2\sigma_i \sigma_{i+1}}{h_i^2} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) dx + \frac{1}{h_i^2} \frac{\sigma_{i+1}^2 h_i^3}{3} \\ &\quad = \frac{h_i}{3} \sigma_i^2 + 2h_i \sigma_i \sigma_{i+1} \int_0^1 (1-z) z dz + \frac{h_i}{3} \sigma_{i+1}^2 \\ &\quad = \frac{h_i}{3} (\sigma_i^2 + \sigma_i \sigma_{i+1} + \sigma_{i+1}^2).\end{aligned}$$

Taking into account $\sigma_0 = \sigma_n = 0$, we infer that

$$\begin{aligned}\int_{x_0}^{x_n} s''(x)^2 dx &= \frac{1}{3}\sum_{i=0}^{n-1} h_i(\sigma_i^2 + \sigma_i\sigma_{i+1} + \sigma_{i+1}^2) \\ &= \frac{1}{6}\sum_{i=1}^{n-1}[h_{i-1}\sigma_{i-1}\sigma_i + 2\sigma_i^2(h_{i-1}+h_i) + h_i\sigma_i\sigma_{i+1}] \\ &= \sigma^t R\sigma.\end{aligned}$$

This shows that the symmetric, invertible matrix R is positive definite. Furthermore, because $\sigma = R^{-1}Qf$, the criterion $J_\alpha(s)$ reduces to

$$\begin{aligned}J_\alpha(s) &= \alpha(y-f)^t W(y-f) + (1-\alpha)\sigma^t R\sigma \\ &= \alpha(y-f)^t W(y-f) + (1-\alpha)f^t Q^t R^{-1} Qf. \qquad (15)\end{aligned}$$

Based on the identity (15), it is possible to minimize $J_\alpha(s)$ as a function of f. At the minimum point of $J_\alpha(s)$, its gradient with respect to f satisfies

$$-2\alpha W(y-f) + 2(1-\alpha)Q^t R^{-1} Qf = \mathbf{0}. \qquad (16)$$

Solving for the optimal f yields

$$\hat{f} = [\alpha W + (1-\alpha)Q^t R^{-1} Q]^{-1}\alpha W y.$$

Alternatively, equation (16) can be rewritten as

$$-2\alpha W(y-f) + 2(1-\alpha)Q^t\sigma = \mathbf{0}. \qquad (17)$$

Thus, the optimal σ determines the optimal f through

$$y - \hat{f} = \Big(\frac{1-\alpha}{\alpha}\Big)W^{-1}Q^t\hat{\sigma}. \qquad (18)$$

Multiplying equation (17) by QW^{-1} gives

$$-2\alpha Qy + 2\alpha R\sigma + 2(1-\alpha)QW^{-1}Q^t\sigma = \mathbf{0}$$

with solution

$$\hat{\sigma} = [\alpha R + (1-\alpha)QW^{-1}Q^t]^{-1}\alpha Qy.$$

This solution for $\hat{\sigma}$ has the advantage that the positive definite matrix $\alpha R + (1-\alpha)QW^{-1}Q^t$ inherits a banding pattern from R and Q. Solving linear equations involving banded matrices is more efficiently accomplished via their Cholesky decompositions than via the sweep operator. See the problems of Chapter 7 for a brief discussion of the Cholesky decomposition of a positive definite matrix. Once $\hat{\sigma}$ is available, equation (18) not only determines $\hat{f}$ but also determines the weighted residual sum of squares

$$(y-\hat{f})^t W(y-\hat{f}) = \Big(\frac{1-\alpha}{\alpha}\Big)^2\hat{\sigma}^t QW^{-1}Q^t\hat{\sigma}.$$

9.5 Problems

1. Consider the function $f(x) = (1+25x^2)^{-1}$ on $[-1, 1]$. Runge's example [4] involves fitting an interpolating polynomial $p_n(x)$ to $f(x)$ at $n+1$ equally spaced nodes

$$x_i = -1 + ih, \qquad i = 0, 1, \ldots, n,$$
$$h = \frac{2}{n}$$

using Lagrange's formula

$$p_n(x) = \sum_{i=0}^{n} f(x_i) \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}.$$

Compare the fit of $p_n(x)$ to $f(x)$ to that of the natural, cubic, interpolating spline $s_n(x)$. In this comparison pay particular attention to the point $-1 + n^{-1}$ as n increases. Please feel free to use relevant subroutines from [6] to carry out your computations.

2. Show that Proposition 9.2.1 remains valid if the condition

(d*) $s'(x_0) = f'(x_0)$ and $s'(x_n) = f'(x_n)$

replaces condition (d) in the definition of a cubic spline. Show that Proposition 9.2.2 also carries over if $g(x)$ as well as $s(x)$ satisfies (d*).

3. For nodes $x_0 < x_1 < \cdots < x_n$ and function values $f_i = f(x_i)$, develop a quadratic interpolating spline $s(x)$ satisfying the conditions:

(a) $s(x)$ is a quadratic polynomial on each interval $[x_i, x_{i+1}]$;
(b) $s(x_i) = f_i$ at each node x_i;
(c) the first derivative $s'(x)$ exists and is continuous throughout the entire interval $[x_0, x_n]$.

To simplify your theory, write

$$s(x) = a_i + b_i(x - x_i) + c_i(x - x_i)(x - x_{i+1})$$

for $x \in [x_i, x_{i+1}]$. Derive explicit expressions for the a_i and b_i from property (b). Using property (c), prove that

$$c_i = \frac{b_i - b_{i-1}}{x_{i+1} - x_i} - c_{i-1}\frac{x_i - x_{i-1}}{x_{i+1} - x_i}$$

for $i = 1, \ldots, n-1$. What additional information do you require to completely determine the spline?

4. Given the nodes $x_0 < x_1 < \cdots < x_n$, let V be the vector space of functions that are twice continuously differentiable at each node x_i and cubic polynomials on $(-\infty, x_0)$, (x_n, ∞), and each of the intervals (x_i, x_{i+1}). Show that any function $s(x) \in V$ can be uniquely

represented as

$$s(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=0}^{n} c_i (x - x_i)_+^3,$$

where $(x - x_i)_+^3$ is 0 for $x \leq x_i$ and $(x - x_i)^3$ otherwise. Conclude that this vector space has dimension $n + 5$.

5. Continuing Problem 4, consider the vector subspace $W \subset V$ whose functions are linear on $(-\infty, x_0)$ and (x_n, ∞). Prove that W is just the subspace of natural cubic splines and has dimension $n + 1$.

6. Let $s(x)$ be the natural cubic spline interpolating the function $f(x)$ at the three equally spaced nodes $x_0 < x_1 < x_2$. Explicitly evaluate the integral $\int_{x_0}^{x_2} s(x)dx$ and the derivatives $s'(x_i)$ in terms of the spacing $h = x_2 - x_1 = x_1 - x_0$ and the function values $f_i = f(x_i)$.

7. In the spline model for nonparametric regression, show that the positive definite matrix $\alpha R + (1 - \alpha)QW^{-1}Q^t$ is banded. How many subdiagonals display nonzero entries?

8. Continuing Problems 4 and 5, let

$$f_i(x) = a_{i0} + a_{i1} x + \sum_{j=0}^{n} c_{ij} (x - x_j)_+^3, \qquad i = 0, \ldots, n$$

be a basis of the vector space W. If $s(x) = \sum_{i=0}^{n} \beta_i f_i(x)$, then show that

$$\int_{x_0}^{x_n} s''(x)t''(x)dx = 6 \sum_{i=0}^{n} \sum_{j=0}^{n} \beta_i c_{ij} t(x_j)$$

for any $t(x) \in V$. (Hints: Integrate by parts as in Proposition 9.2.2, and use the fact that $\sum_{j=0}^{n} c_{ij} = 0$ by virtue of Problem 5.)

9. Mindful of Problems 4, 5, and 8, let $s(x) = \sum_{i=0}^{n} \beta_i f_i(x) \in W$ be the spline minimizing the functional $J_\alpha(g)$ defined in equation (14). Prove that $s(x)$ satisfies

$$\begin{aligned} 0 &= -\alpha \sum_{j=0}^{n} w_j [y_j - s(x_j)] t(x_j) + (1 - \alpha) \int_{x_0}^{x_n} s''(x)t''(x)dx \\ &= -\alpha \sum_{j=0}^{n} w_j [y_j - \sum_{i=0}^{n} \beta_i f_i(x_j)] t(x_j) + 6(1 - \alpha) \sum_{i=0}^{n} \sum_{j=0}^{n} \beta_i c_{ij} t(x_j) \end{aligned}$$

for any function $t(x) \in W$. Because the constants $t(x_j)$ are arbitrary, demonstrate that this provides the system of linear equations

$$\alpha w_j y_j = \alpha w_j \sum_{i=0}^{n} \beta_i f_i(x_j) + 6(1 - \alpha) \sum_{i=0}^{n} \beta_i c_{ij}$$

determining the β_j. Summarize this system of equations as the single vector equation $\alpha W y = \alpha W F^t \beta + 6(1-\alpha) C^t \beta$ by defining appropriate

matrices. Because the symmetric matrix FC^t has entry

$$\sum_{k=0}^{n} f_i(x_k)c_{jk} = \int_{x_0}^{x_n} f_i''(x)f_j''(x)dx \tag{19}$$

in row i and column j, argue finally that the solution

$$\hat{\beta} = [\alpha FWF^t + 6(1-\alpha)FC^t]^{-1}\alpha FWy \tag{20}$$

is well defined. This approach to minimizing $J_\alpha(g)$ can exploit any of several different bases for W [2]. (Hint: Adopting the usual calculus of variations tactic, evaluate the derivative $J'_\alpha(s+\epsilon t)|_{\epsilon=0}$. Equality (19) follows from Problem 8.)

10. It is possible to give a Bayesian interpretation to the spline solution (20) of Problem 9 [7]. Suppose in the notation of Problem 9 that

$$y_j = \sum_{i=0}^{n} \beta_i f_i(x_j) + \frac{1}{\sqrt{w_j}}\epsilon_j,$$

where the errors ϵ_j are independent, univariate normals with common mean 0 and common variance σ^2. Assuming that β has a multivariate normal prior with mean $\mathbf{0}$ and covariance

$$\sigma^2\Omega = \sigma^2[6(1-\alpha)\alpha^{-1}FC^t]^{-1},$$

demonstrate that the posterior mean of β is given by (20).

References

[1] de Boor C (1978) *A Practical Guide to Splines*. Springer-Verlag, New York

[2] Eubank R (1990) *Smoothing Splines and Nonparametric Regression*. Marcel Dekker, New York

[3] Hämmerlin G, Hoffmann K-H (1991) *Numerical Mathematics*. Springer-Verlag, New York

[4] Isaacson E, Keller HB (1966) *Analysis of Numerical Methods*. Wiley, New York

[5] Powell MJD (1981) *Approximation Theory and Methods*. Cambridge University Press, Cambridge

[6] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge

[7] Silverman BW (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J Roy Stat Soc B* 47:1–52

[8] Wahba G (1990) *Spline Functions for Observational Data*. CBMS–NSF Regional Conference Series, Society for Industrial and Applied Mathematics, Philadelphia

10
The EM Algorithm

10.1 Introduction

Maximum likelihood is the dominant form of estimation in applied statistics. Because closed-form solutions to likelihood equations are the exception rather than the rule, numerical methods for finding maximum likelihood estimates are of paramount importance. In this chapter we study maximum likelihood estimation by the EM algorithm [2, 5, 6]. In specific problems, the EM algorithm is more a prescription for an algorithm than a detailed algorithm. At the heart of every EM algorithm is some notion of missing data. Data can be missing in the ordinary sense of a failure to record certain observations on certain cases. Data can also be missing in a theoretical sense. We can think of the E, or expectation, step of the algorithm as filling in the missing data. Once the missing data are reconstructed, then the parameters are estimated in the M, or maximization, step. Because the M step usually involves maximizing a much simpler function than the likelihood of the observed data, we can often solve the M step analytically. The price we pay for this simplification is that the EM algorithm is iterative. Reconstructing the missing data is bound to be slightly wrong if the parameters do not already equal their maximum likelihood estimates.

One of the advantages of the EM algorithm is its numerical stability. As we shall see, the EM algorithm leads to a steady increase in the likelihood of the observed data. Thus, the EM algorithm avoids wildly overshooting or undershooting the maximum of the likelihood along its current direc-

tion of search. Besides this desirable feature, the EM handles parameter constraints gracefully. Constraint satisfaction is by definition built into the solution of the M step. In contrast, competing methods of maximization must incorporate special techniques to cope with parameter constraints.

A negative feature of the EM algorithm is its often excruciatingly slow convergence rate in a neighborhood of the optimal point. This rate directly reflects the amount of missing data in a problem. Under fairly mild assumptions, the EM algorithm is guaranteed to converge to a stationary point of the likelihood function. In some very contrived examples, it converges to a saddle point, but this rarely happens in practice. Convergence to a local maximum is more likely to occur. The global maximum can usually be reached by starting the parameters at good but suboptimal estimates such as method-of-moments estimates or by choosing multiple random starting points. In general, almost all maximum likelihood algorithms have trouble distinguishing global from local maximum points.

10.2 General Definition of the EM Algorithm

A sharp distinction is drawn in the EM algorithm between the observed, incomplete data Y and the unobserved, complete data X of a statistical experiment [2, 5, 9]. Some function $t(X) = Y$ collapses X onto Y. For instance, if we represent X as (Y, Z), with Z as the missing data, then t is simply projection onto the Y-component of X. It should be stressed that the missing data can consist of more than just observations missing in the ordinary sense. In fact, the definition of X is left up to the intuition and cleverness of the statistician. The general idea is to choose X so that maximum likelihood becomes trivial for the complete data.

The complete data are assumed to have a probability density $f(X \mid \theta)$ that is a function of a parameter vector θ as well as of X. In the E step of the EM algorithm, we calculate the conditional expectation

$$Q(\theta \mid \theta_n) = \mathrm{E}[\ln f(X \mid \theta) \mid Y, \theta_n].$$

Here θ_n is the current estimated value of θ.In the M step, we maximize $Q(\theta \mid \theta_n)$ with respect to θ. This yields the new parameter estimate θ_{n+1}, and we repeat this two-step process until convergence occurs. Note that θ and θ_n play fundamentally different roles in $Q(\theta \mid \theta_n)$.

The essence of the EM algorithm is that maximizing $Q(\theta \mid \theta_n)$ leads to an increase in the loglikelihood $\ln g(Y \mid \theta)$ of the observed data. This assertion is proved in the following theoretical section, which can be omitted by readers interested primarily in practical applications of the EM algorithm.

10.3 Ascent Property of the EM Algorithm

The information inequality at the heart of the EM algorithm is a consequence of Jensen's inequality, which relates convex functions to expectations. Recall that a twice differentiable function $h(w)$ is convex on an interval (a, b) if and only if $h''(w) \geq 0$ for all w in (a, b). If the defining inequality is strict, then $h(w)$ is said to be strictly convex.

Proposition 10.3.1 (Jensen's Inequality**).** *Assume that the values of the random variable W are confined to the possibly infinite interval (a, b). If $h(w)$ is convex on (a, b), then $\mathrm{E}[h(W)] \geq h[\mathrm{E}(W)]$, provided both expectations exist. For a strictly convex function $h(w)$, equality holds in Jensen's inequality if and only if $W = \mathrm{E}(W)$ almost surely.*

Proof. Put $u = \mathrm{E}(W)$. For w in (a, b) we have

$$\begin{aligned} h(w) &= h(u) + h'(u)(w-u) + h''(v)\frac{(w-u)^2}{2} \\ &\geq h(u) + h'(u)(w-u) \end{aligned}$$

for some v between u and w. Note that v is in (a, b). Now substitute the random variable W for the point w and take expectations. It follows that

$$\mathrm{E}[h(W)] \geq h(u) + h'(u)[\mathrm{E}(W) - u] = h(u).$$

If $h(w)$ is strictly convex, then the neglected term $h''(v)(w-u)^2/2$ is positive whenever $w \neq u$. Figure 10.1 gives an alternative proof that does not depend on the existence of $h''(w)$. □

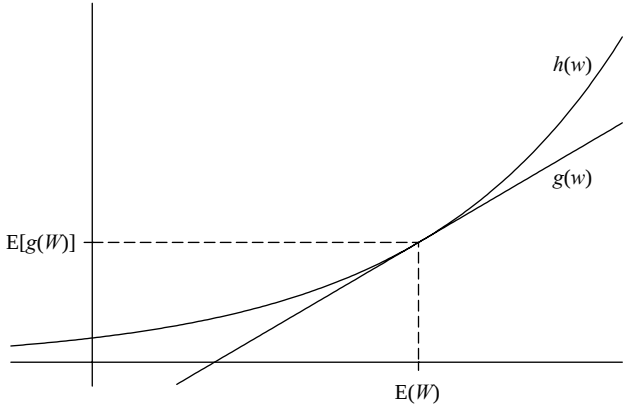

FIGURE 10.1. Geometric proof of Jensen's inequality. The line $g(w)$ is tangent to the convex function $h(w)$ at $w = \mathrm{E}(W)$. By convexity, $h(w) \geq g(w)$ for all w and thus $\mathrm{E}[h(W)] \geq \mathrm{E}[g(W)]$. However, the linearity of $g(w)$ implies that $\mathrm{E}[g(W)] = g[\mathrm{E}(W)] = h[\mathrm{E}(W)]$.

Proposition 10.3.2 (Information Inequality**).** *Let f and g be probability densities with respect to a measure μ. Suppose $f > 0$ and $g > 0$ almost everywhere relative to μ. If E_f denotes expectation with respect to the probability measure $f d\mu$, then $\mathrm{E}_f(\ln f) \geq \mathrm{E}_f(\ln g)$, with equality only if $f = g$ almost everywhere relative to μ.*

Proof. Because $-\ln(w)$ is a strictly convex function on $(0, \infty)$, Jensen's inequality applied to the random variable g/f implies

$$\begin{aligned}
\mathrm{E}_f(\ln f) - \mathrm{E}_f(\ln g) &= \mathrm{E}_f\left(-\ln \frac{g}{f}\right) \\
&\geq -\ln \mathrm{E}_f\left(\frac{g}{f}\right) \\
&= -\ln \int \frac{g}{f} f d\mu \\
&= -\ln \int g d\mu \\
&= 0.
\end{aligned}$$

Equality holds only if $g/f = \mathrm{E}_f(g/f)$ almost everywhere relative to μ. But $\mathrm{E}_f(g/f) = 1$. □

Reverting to the notation $Q(\theta \mid \theta_n) = \mathrm{E}[\ln f(X \mid \theta) \mid Y = y, \theta_n]$ of the EM algorithm, let us next prove that

$$Q(\theta_n \mid \theta_n) - \ln g(y \mid \theta_n) \geq Q(\theta \mid \theta_n) - \ln g(y \mid \theta)$$

for all θ and θ_n, where $g(y \mid \theta)$ is the probability density of the observed data $Y = y$. With this end in mind, note that both $f(x \mid \theta)/g(y \mid \theta)$ and $f(x \mid \theta_n)/g(y \mid \theta_n)$ are conditional densities of X on $\{x : t(x) = y\}$ with respect to some measure μ_y. The information inequality now indicates that

$$\begin{aligned}
Q(\theta \mid \theta_n) - \ln g(y \mid \theta) &= \mathrm{E}(\ln \left[\frac{f(X \mid \theta)}{g(Y \mid \theta)}\right] \mid Y = y, \theta_n) \\
&\leq \mathrm{E}(\ln \left[\frac{f(X \mid \theta_n)}{g(Y \mid \theta_n)}\right] \mid Y = y, \theta_n) \\
&= Q(\theta_n \mid \theta_n) - \ln g(y \mid \theta_n).
\end{aligned}$$

Thus, the difference $\ln g(y \mid \theta) - Q(\theta \mid \theta_n)$ attains its minimum when $\theta = \theta_n$. If we choose θ_{n+1} to maximize $Q(\theta \mid \theta_n)$, then it follows that

$$\begin{aligned}
\ln g(y \mid \theta_{n+1}) &= Q(\theta_{n+1} \mid \theta_n) + [\ln g(y \mid \theta_{n+1}) - Q(\theta_{n+1} \mid \theta_n)] \\
&\geq Q(\theta_n \mid \theta_n) + [\ln g(y \mid \theta_n) - Q(\theta_n \mid \theta_n)] \\
&= \ln g(y \mid \theta_n).
\end{aligned}$$

In other words, maximizing $Q(\theta \mid \theta_n)$ forces an increase in $\ln g(y \mid \theta)$. This fundamental property is summarized in the next proposition.

Proposition 10.3.3. *Suppose that $g(y \mid \theta)$ and $f(x \mid \theta)$ are the probability densities of the observed and complete data, respectively. Then the EM iterates obey*

$$\ln g(y \mid \theta_{n+1}) \geq \ln g(y \mid \theta_n),$$

with strict inequality when $f(x \mid \theta_{n+1})/g(y \mid \theta_{n+1})$ and $f(x \mid \theta_n)/g(y \mid \theta_n)$ are different conditional densities or when the surrogate function $Q(\theta \mid \theta_n)$ satisfies

$$Q(\theta_{n+1} \mid \theta_n) > Q(\theta_n \mid \theta_n).$$

10.3.1 Technical Note

The above proof is a little vague as to the meaning of the conditional density $f(x \mid \theta)/g(y \mid \theta)$ and its associated measure μ_y. Commonly the complete data decomposes as $X = (Y, Z)$, where Z is considered the missing data and $t(Y, Z) = Y$ is projection onto the observed data. Suppose (Y, Z) has joint density $f(y, z \mid \theta)$ relative to a product measure $\omega \times \mu\,(y, z)$; ω and μ are typically Lebesgue measure or counting measure. In this framework, we define $g(y \mid \theta) = \int f(y, z, \theta) d\mu(z)$ and set $\mu_y = \mu$. The function $g(y \mid \theta)$ serves as a density relative to ω. To check that these definitions make sense, it suffices to prove that $\int h(y, z) f(y, z \mid \theta)/g(y \mid \theta) d\mu(z)$ is a version of the conditional expectation $\mathrm{E}[h(Y, Z) \mid Y = y]$ for every well-behaved function $h(y, z)$. This assertion can be verified by showing

$$\mathrm{E}\{1_S(Y)\,\mathrm{E}[h(Y, Z) \mid Y]\} = \mathrm{E}[1_S(Y) h(Y, Z)]$$

for every measurable set S. With

$$\mathrm{E}[h(Y, Z) \mid Y = y] = \int h(y, z) \frac{f(y, z \mid \theta)}{g(y \mid \theta)} d\mu(z),$$

we calculate

$$\begin{aligned} \mathrm{E}\{1_S(Y)\,\mathrm{E}[h(Y, Z) \mid Y]\} &= \int_S \int h(y, z) \frac{f(y, z \mid \theta)}{g(y \mid \theta)} d\mu(z) g(y \mid \theta) d\omega(y) \\ &= \int_S \int h(y, z) f(y, z \mid \theta) d\mu(z) d\omega(y) \\ &= \mathrm{E}[1_S(Y) h(Y, Z)]. \end{aligned}$$

Thus in this situation, $f(x \mid \theta)/g(y \mid \theta)$ is indeed the conditional density of X given $Y = y$.

10.4 Allele Frequency Estimation

The ABO and Rh genetic loci are usually typed in matching blood donors to blood recipients. The ABO locus exhibits the three alleles A, B, and O

and the four observable phenotypes A, B, AB, and O. These phenotypes arise because each person inherits two alleles, one from his mother and one from his father, and the alleles A and B are genetically dominant to allele O. Dominance amounts to a masking of the O allele by the presence of an A or B allele. For instance, a person inheriting an A allele from one parent and an O allele from the other parent is said to have genotype A/O and is indistinguishable from a person inheriting an A allele from both parents. This second person is said to have genotype A/A.

The EM algorithm for estimating the population frequencies or proportions of the three alleles involves an interplay between observed phenotypes and underlying unobserved genotypes. As just noted, both genotypes A/O and A/A correspond to the same phenotype A. Likewise, phenotype B corresponds to either genotype B/O or genotype B/B. Phenotypes AB and O correspond to the single genotypes A/B and O/O, respectively.

As a concrete example, Clarke et al. [1] noted that among their population sample of $n = 521$ duodenal ulcer patients, a total of $n_A = 186$ had phenotype A, $n_B = 38$ had phenotype B, $n_{AB} = 13$ had phenotype AB, and $n_O = 284$ had phenotype O. If we want to estimate the frequencies p_A, p_B, and p_O of the three different alleles from this sample, then we can employ the EM algorithm with the four phenotype counts as the observed data Y and the underlying six genotype counts $n_{A/A}$, $n_{A/O}$, $n_{B/B}$, $n_{B/O}$, $n_{A/B} = n_{AB}$, and $n_{O/O} = n_O$ as the complete data X [8]. Note that the allele frequencies are nonnegative and satisfy the constraint $p_A + p_B + p_O = 1$. Furthermore, the classical Hardy-Weinberg law of population genetics specifies that each genotype frequency equals the product of the corresponding allele frequencies with an extra factor of 2 thrown in to account for ambiguity in parental source when the two alleles differ. For example, genotype A/A has frequency p_A^2, and genotype A/O has frequency $2p_Ap_O$.

With these preliminaries in mind, the complete data loglikelihood becomes

$$\begin{aligned}\ln f(X \mid p) &= n_{A/A} \ln p_A^2 + n_{A/O} \ln(2p_Ap_O) + n_{B/B} \ln p_B^2 \\ &\quad + n_{B/O} \ln(2p_Bp_O) + n_{AB} \ln(2p_Ap_B) + n_O \ln p_O^2 \\ &\quad + \ln \binom{n}{n_{A/A}\, n_{A/O}\, n_{B/B}\, n_{B/O}\, n_{AB}\, n_O}. \end{aligned} \tag{1}$$

In the E step of the EM algorithm we take the expectation of $\ln f(X \mid p)$ conditional on the observed counts n_A, n_B, n_{AB}, and n_O and the current parameter vector $p_m = (p_{mA}, p_{mB}, p_{mO})^t$. It is obvious that

$$\begin{aligned}\mathrm{E}(n_{AB} \mid Y, p_m) &= n_{AB} \\ \mathrm{E}(n_O \mid Y, p_m) &= n_O.\end{aligned}$$

A moment's reflection also yields

$$n_{mA/A} = \mathrm{E}(n_{A/A} \mid Y, p_m) = n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}$$
$$n_{mA/O} = \mathrm{E}(n_{A/O} \mid Y, p_m) = n_A \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}}.$$

The conditional expectations $n_{mB/B}$ and $n_{mB/O}$ are given by similar expressions.

The M step of the EM algorithm maximizes the $Q(p \mid p_m)$ function derived from (1) by replacing $n_{A/A}$ by $n_{mA/A}$, and so forth. Maximization of $Q(p \mid p_m)$ can be accomplished by introducing a Lagrange multiplier and finding a stationary point of the unconstrained function

$$H(p, \lambda) = Q(p \mid p_m) + \lambda(p_A + p_B + p_O - 1)$$

as explained in Chapter 14. Setting the partial derivatives of $H(p, \lambda)$,

$$\frac{\partial}{\partial p_A} H(p, \lambda) = \frac{2n_{mA/A}}{p_A} + \frac{n_{mA/O}}{p_A} + \frac{n_{AB}}{p_A} + \lambda$$
$$\frac{\partial}{\partial p_B} H(p, \lambda) = \frac{2n_{mB/B}}{p_B} + \frac{n_{mB/O}}{p_B} + \frac{n_{AB}}{p_B} + \lambda$$
$$\frac{\partial}{\partial p_O} H(p, \lambda) = \frac{n_{mA/O}}{p_O} + \frac{n_{mB/O}}{p_O} + \frac{2n_O}{p_O} + \lambda,$$
$$\frac{\partial}{\partial \lambda} H(p, \lambda) = p_A + p_B + p_O - 1,$$

equal to 0 provides the unique stationary point of $H(p, \lambda)$. The solution of the resulting equations is

$$p_{m+1,A} = \frac{2n_{mA/A} + n_{mA/O} + n_{AB}}{2n}$$
$$p_{m+1,B} = \frac{2n_{mB/B} + n_{mB/O} + n_{AB}}{2n}$$
$$p_{m+1,O} = \frac{n_{mA/O} + n_{mB/O} + 2n_O}{2n}.$$

In other words, the EM update is identical to a form of gene counting in which the unknown genotype counts are imputed based on the current allele frequency estimates.

Table 10.1 shows the progress of the EM iterates starting from the initial guesses $p_{0A} = 0.3$, $p_{0B} = 0.2$, and $p_{0O} = 0.5$. The EM updates are simple enough to carry out on a pocket calculator. Convergence occurs quickly in this example.

TABLE 10.1. Iterations for ABO Duodenal Ulcer Data

Iteration m	p_{mA}	p_{mB}	p_{mO}
0	0.3000	0.2000	0.5000
1	0.2321	0.0550	0.7129
2	0.2160	0.0503	0.7337
3	0.2139	0.0502	0.7359
4	0.2136	0.0501	0.7363
5	0.2136	0.0501	0.7363

10.5 Transmission Tomography

The purpose of transmission tomography is to reconstruct the local attenuation properties of the object being imaged. Attenuation is to be roughly equated with density. In medical applications, material such as bone is dense and stops or deflects X-rays better than soft tissue. With enough radiation, even small gradations in soft tissue can be detected. The traditional method of image reconstruction in transmission tomography relies on Fourier analysis and the Radon transform [3]. An alternative to this deterministic approach is to pose an explicitly stochastic model that permits parameter estimation by maximum likelihood [4]. The EM algorithm immediately suggests itself in this context.

The stochastic model depends on dividing the object of interest into small nonoverlapping regions of constant attenuation called pixels. Typically the pixels are squares. The attenuation attributed to pixel j constitutes parameter θ_j of the model. Since there may be thousands of pixels, implementation of maximum likelihood algorithms such as scoring or Newton's method is out of the question. Each observation Y_i is generated by beaming a stream of X-rays or high-energy photons from an X-ray source toward some detector on the opposite side of the object. The observation (or projection) Y_i counts the number of photons detected along the ith line of flight. Naturally, only a fraction of the photons are successfully transmitted from source to detector. If l_{ij} is the length of the segment of projection line i intersecting pixel j, then we claim that the probability of a photon escaping attenuation along projection line i is the exponentiated line integral $\exp(-\sum_j l_{ij}\theta_j)$.

This result can be demonstrated by letting $t \to (1-t)u + tv$, $t \in [0,1]$, be a parametric representation of projection line i from the source position u to the detector position v. If $\theta(t)$ denotes the attenuation value at point $(1-t)u + tv$ of this line, and if $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = 1$ is a partition of $[0,1]$ with mesh size $\delta = \max_k(t_{k+1} - t_k)$, then the probability

that a photon escapes attenuation is approximately

$$\begin{aligned}\prod_{k=0}^{m-1}[1-\theta(t_k)(t_{k+1}-t_k)] &= e^{\sum_{k=0}^{m-1}\ln[1-\theta(t_k)(t_{k+1}-t_k)]} \\ &= e^{-\sum_{k=0}^{m-1}\theta(t_k)(t_{k+1}-t_k)[1+O(\delta)]}.\end{aligned}$$

In the limit as $\delta \to 0$, we recover the exponentiated line integral formula.

In the absence of the intervening object, the number of photons generated and ultimately detected follows a Poisson distribution. Let the mean of this distribution be d_i for projection line i. Because random thinning of a Poisson random variable gives a Poisson random variable, the number Y_i is Poisson distributed with mean $d_i \exp(-\sum_j l_{ij}\theta_j)$. Owing to the Poisson nature of X-ray generation, the different projections will be independent even if collected simultaneously. This fact enables us to write the loglikelihood of the observed data $Y_i = y_i$ as the finite sum

$$\sum_i \left[-d_i e^{-\sum_j l_{ij}\theta_j} - y_i \sum_j l_{ij}\theta_j + y_i \ln d_i - \ln y_i! \right]. \tag{2}$$

The missing data in this model correspond to the number of photons X_{ij} entering each pixel j along each projection line i. These random variables supplemented by the observations Y_i constitute the complete data. If projection line i does not intersect pixel j, then $X_{ij} = 0$. Although X_{ij} and $X_{ij'}$ are not independent, the collection $\{X_{ij}\}_j$ indexed by projection i is independent of the collection $\{X_{i'j}\}_j$ indexed by another projection i'. This allows us to work projection by projection in writing the complete data likelihood. We will therefore temporarily drop the projection subscript i and relabel pixels, starting with pixel 1 adjacent to the source and ending with pixel $m-1$ adjacent to the detector. In this notation X_1 is the number of photons leaving the source, X_j is the number of photons entering pixel j, and $X_m = Y$ is the number of photons detected.

By assumption X_1 follows a Poisson distribution with mean d. Conditional on $X_1, \ldots, X_j$, the random variable X_{j+1} is binomially distributed with X_j trials and success probability $e^{-l_j\theta_j}$. In other words, each of the X_j photons entering pixel j behaves independently and has a chance $e^{-l_j\theta_j}$ of avoiding attenuation in pixel j. It follows that the complete data loglikelihood for the current projection is

$$\begin{aligned}&-d + X_1 \ln d - \ln X_1! \\ &\quad + \sum_{j=1}^{m-1}\left[\ln\binom{X_j}{X_{j+1}} + X_{j+1}\ln e^{-l_j\theta_j} + (X_j - X_{j+1})\ln(1-e^{-l_j\theta_j})\right].\end{aligned} \tag{3}$$

To perform the E step of the EM algorithm, we need only compute the conditional expectations $\mathrm{E}(X_j \mid X_m = y, \theta)$, $j = 1, \ldots, m$. The conditional expectations of other terms such as $\ln\binom{X_j}{X_{j+1}}$ appearing in (3) are irrelevant in the subsequent M step.

Reasoning as above, we infer that the unconditional mean of X_j is

$$\begin{aligned}\mu_j &= \mathrm{E}(X_j) \\ &= de^{-\sum_{k=1}^{j-1} l_k\theta_k}\end{aligned}$$

and that the distribution of X_m conditional on X_j is binomial with X_j trials and success probability

$$\frac{\mu_m}{\mu_j} = e^{-\sum_{k=j}^{m-1} l_k\theta_k}.$$

Hence, the joint probability of X_j and X_m reduces to

$$\Pr(X_j = x_j, X_m = x_m) = e^{-\mu_j}\frac{\mu_j^{x_j}}{x_j!}\binom{x_j}{x_m}\Big(\frac{\mu_m}{\mu_j}\Big)^{x_m}\Big(1-\frac{\mu_m}{\mu_j}\Big)^{x_j-x_m},$$

and the conditional probability of X_j given X_m becomes

$$\begin{aligned}\Pr(X_j = x_j \mid X_m = x_m) &= \frac{e^{-\mu_j}\frac{\mu_j^{x_j}}{x_j!}\binom{x_j}{x_m}(\frac{\mu_m}{\mu_j})^{x_m}(1-\frac{\mu_m}{\mu_j})^{x_j-x_m}}{e^{-\mu_m}\frac{\mu_m^{x_m}}{x_m!}} \\ &= e^{-(\mu_j-\mu_m)}\frac{(\mu_j-\mu_m)^{x_j-x_m}}{(x_j-x_m)!}.\end{aligned}$$

In other words, conditional on X_m, the difference $X_j - X_m$ follows a Poisson distribution with mean $\mu_j - \mu_m$. This implies in particular that

$$\begin{aligned}\mathrm{E}(X_j \mid X_m) &= \mathrm{E}(X_j - X_m \mid X_m) + X_m \\ &= \mu_j - \mu_m + X_m.\end{aligned}$$

Reverting to our previous notation, it is now possible to assemble the function $Q(\theta \mid \theta_n)$ of the E step. Define

$$\begin{aligned}M_{ij} &= d_i(e^{-\sum_{k\in S_{ij}} l_{ik}\theta_{nk}} - e^{-\sum_k l_{ik}\theta_{nk}}) + y_i \\ N_{ij} &= d_i(e^{-\sum_{k\in S_{ij}\cup\{j\}} l_{ik}\theta_{nk}} - e^{-\sum_k l_{ik}\theta_{nk}}) + y_i,\end{aligned}$$

where S_{ij} is the set of pixels between the source and pixel j along projection i. If j' is the next pixel after pixel j along projection i, then

$$\begin{aligned}M_{ij} &= \mathrm{E}(X_{ij} \mid Y_i = y_i, \theta_n) \\ N_{ij} &= \mathrm{E}(X_{ij'} \mid Y_i = y_i, \theta_n).\end{aligned}$$

In view of expression (3), we find

$$Q(\theta \mid \theta_n) = \sum_i \sum_j \Big[-N_{ij}l_{ij}\theta_j + (M_{ij} - N_{ij})\ln(1 - e^{-l_{ij}\theta_j})\Big]$$

up to an irrelevant constant.

If we try to maximize $Q(\theta \mid \theta_n)$ by setting its partial derivatives equal to 0, we get for pixel j the equation

$$-\sum_i N_{ij} l_{ij} + \sum_i \frac{(M_{ij} - N_{ij}) l_{ij}}{e^{l_{ij}\theta_j} - 1} = 0. \tag{4}$$

This is an intractable transcendental equation in the single variable θ_j, and the M step must be solved numerically, say by Newton's method. It is straightforward to check that the left-hand side of equation (4) is strictly decreasing in θ_j and has exactly one positive solution. Thus, the EM approach in this problem has the advantages of decoupling the parameters in the likelihood equations and of satisfying the natural boundary constraints $\theta_j \geq 0$. We will return to this model in Chapter 12.

10.6 Problems

1. Let $f(x)$ be a real-valued function whose Hessian matrix $(\frac{\partial^2}{\partial x_i \partial x_j} f)$ is positive definite throughout some convex open set U of R^m. For $u \neq \mathbf{0}$ and $x \in U$, show that the function $t \to f(x + tu)$ of the real variable t is strictly convex on $\{t : x + tu \in U\}$. Use this fact to demonstrate that $f(x)$ can have at most one local minimum point on any convex subset of U.

2. Apply the result of the last problem to show that the loglikelihood of the observed data in the ABO example is strictly concave and therefore possesses a single global maximum. Why does the maximum occur on the interior of the feasible region?

3. The entropy of a probability density $p(x)$ on R^m is defined by

$$-\int p(x) \ln p(x) dx. \tag{5}$$

Among all densities with a fixed mean $\mu = \int x p(x) dx$ and covariance $\Omega = \int (x - \mu)(x - \mu)^t p(x) dx$, prove that the multivariate normal has maximum entropy. (Hint: Apply Proposition 10.3.2.)

4. In statistical mechanics, entropy is employed to characterize the equilibrium distribution of many independently behaving particles. Let $p(x)$ be the probability density that a particle is found at position x in phase space R^m, and suppose that each position x is assigned an energy $u(x)$. If the average energy $U = \int u(x) p(x) dx$ per particle is fixed, then Nature chooses $p(x)$ to maximize entropy as defined in (5). Show that if constants α and β exist satisfying

$$\int \alpha e^{\beta u(x)} dx = 1$$

$$\int u(x) \alpha e^{\beta u(x)} dx = U,$$

then $p(x) = \alpha e^{\beta u(x)}$ does indeed maximize entropy subject to the average energy constraint. The density $p(x)$ is the celebrated Maxwell–Boltzmann density.

5. In the EM algorithm [2], suppose that the complete data X possesses a regular exponential density

$$f(x \mid \theta) = a(x)e^{b(\theta)+s(x)^t\theta}$$

relative to some measure ν. Prove that the unconditional mean of the sufficient statistic $s(X)$ is given by the negative gradient $-db(\theta)^t$ and that the EM update is characterized by the condition

$$\mathrm{E}[s(X) \mid Y, \theta_n] = -db(\theta_{n+1})^t.$$

6. Consider an i.i.d. sample drawn from a bivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2)^t$ and covariance matrix

$$\Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Suppose through some random accident that the first p observations are missing their first component, the next q observations are missing their second component, and the last r observations are complete. Design an EM algorithm for estimating the five mean and variance parameters, taking as complete data the original data before the accidental loss.

7. In a genetic linkage experiment, 197 animals are randomly assigned to four categories according to the multinomial distribution with cell probabilities $\pi_1 = \frac{1}{2} + \frac{\theta}{4}$, $\pi_2 = \frac{1-\theta}{4}$, $\pi_3 = \frac{1-\theta}{4}$ and $\pi_4 = \frac{\theta}{4}$. If the corresponding observations are

$$\begin{aligned} y &= (y_1, y_2, y_3, y_4)^t \\ &= (125, 18, 20, 34)^t, \end{aligned}$$

then devise an EM algorithm and use it to estimate $\hat{\theta} = .6268$ [7]. (Hint: Split the first category into two so that there are five categories for the complete data.)

8. The standard linear regression model can be written in matrix notation as $X = A\beta + U$. Here X is the $r \times 1$ vector of dependent variables, A is the $r \times s$ design matrix, β is the $s \times 1$ vector of regression coefficients, and U is the $r \times 1$ normally distributed error vector with mean $\mathbf{0}$ and covariance $\sigma^2 I$. The dependent variables are right censored if for each i there is a constant c_i such that only $Y_i = \min\{c_i, X_i\}$ is observed. The EM algorithm offers a vehicle for estimating the parameter vector $\theta = (\beta^t, \sigma^2)^t$ in the presence of censoring [2, 9]. Show that

$$\begin{aligned} \beta_{n+1} &= (A^t A)^{-1} A^t \,\mathrm{E}(X \mid Y, \theta_n) \\ \sigma_{n+1}^2 &= \frac{1}{r}\,\mathrm{E}[(X - A\beta_{n+1})^t (X - A\beta_{n+1}) \mid Y, \theta_n]. \end{aligned}$$

To compute the conditional expectations appearing in these formulas, let a_i be the ith row of A and define

$$H(v) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}}}{\frac{1}{\sqrt{2\pi}} \int_v^\infty e^{-\frac{w^2}{2}} dw}.$$

For a censored observation $y_i = c_i < \infty$, prove that

$$\mathrm{E}(X_i \mid Y_i = c_i, \theta_n) = a_i\beta_n + \sigma_n H\Big(\frac{c_i - a_i\beta_n}{\sigma_n}\Big)$$

$$\mathrm{E}(X_i^2 \mid Y_i = c_i, \theta_n) = (a_i\beta_n)^2 + \sigma_n^2 + \sigma_n(c_i + a_i\beta_n) H\Big(\frac{c_i - a_i\beta_n}{\sigma_n}\Big).$$

Use these formulas to complete the specification of the EM algorithm.

9. Let $x_1, \ldots, x_m$ be i.i.d. observations drawn from a mixture of two normal densities with means μ_1 and μ_2 and common variance σ^2. These three parameters together with the proportion α of observations taken from population 1 can be estimated by an EM algorithm. If

$$p_{ni} = \frac{\frac{\alpha_n}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(x_i - \mu_{n1})^2}{2\sigma_n^2}}}{\frac{\alpha_n}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(x_i - \mu_{n1})^2}{2\sigma_n^2}} + \frac{(1-\alpha_n)}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(x_i - \mu_{n2})^2}{2\sigma_n^2}}}$$

is the current posterior probability that observation i is taken from population 1, then derive the EM algorithm whose updates are

$$\begin{aligned}
\alpha_{n+1} &= \frac{1}{m} \sum_{i=1}^m p_{ni} \\
\mu_{n+1,1} &= \frac{\sum_{i=1}^m p_{ni} x_i}{\sum_{i=1}^m p_{ni}} \\
\mu_{n+1,2} &= \frac{\sum_{i=1}^m q_{ni} x_i}{\sum_{i=1}^m q_{ni}} \\
\sigma_{n+1}^2 &= \frac{1}{m} \sum_{i=1}^m \Big[p_{ni}(x_i - \mu_{n+1,1})^2 + q_{ni}(x_i - \mu_{n+1,2})^2 \Big],
\end{aligned}$$

where $q_{ni} = 1 - p_{ni}$.

10. Consider the data from *The London Times* [10] during the years 1910–1912 given in Table 10.2. The two columns labeled "Deaths i" refer to the number of deaths to women 80 years and older reported by day. The columns labeled "Frequency n_i" refer to the number of days with i deaths. A Poisson distribution gives a poor fit to these data, possibly because of different patterns of deaths in winter and summer. A mixture of two Poissons provides a much better fit. Under the Poisson

TABLE 10.2. Death Notices from *The London Times*

Deaths i	**Frequency** n_i	**Deaths** i	**Frequency** n_i
0	162	5	61
1	267	6	27
2	271	7	8
3	185	8	3
4	111	9	1

admixture model, the likelihood of the observed data is

$$\prod_{i=0}^{9}\left[\alpha e^{-\mu_1}\frac{\mu_1^i}{i!}+(1-\alpha)e^{-\mu_2}\frac{\mu_2^i}{i!}\right]^{n_i},$$

where α is the admixture parameter and μ_1 and μ_2 are the means of the two Poisson distributions.

Formulate an EM algorithm for this model. Let $\theta=(\alpha,\mu_1,\mu_2)^t$ and

$$z_i(\theta)=\frac{\alpha e^{-\mu_1}\mu_1^i}{\alpha e^{-\mu_1}\mu_1^i+(1-\alpha)e^{-\mu_2}\mu_2^i}$$

be the posterior probability that a day with i deaths belongs to Poisson population 1. Show that the EM algorithm is given by

$$\begin{aligned}\alpha_{m+1}&=\frac{\sum_i n_i z_i(\theta_m)}{\sum_i n_i}\\ \mu_{m+1,1}&=\frac{\sum_i n_i i z_i(\theta_m)}{\sum_i n_i z_i(\theta_m)}\\ \mu_{m+1,2}&=\frac{\sum_i n_i i[1-z_i(\theta_m)]}{\sum_i n_i[1-z_i(\theta_m)]}.\end{aligned}$$

From the initial estimates $\alpha_0=0.3$, $\mu_{0,1}=1$ and $\mu_{0,2}=2.5$, compute via the EM algorithm the maximum likelihood estimates $\hat{\alpha}=0.3599$, $\hat{\mu}_1=1.2561$, and $\hat{\mu}_2=2.6634$. Note how slowly the EM algorithm converges in this example.

11. In the transmission tomography model it is possible to approximate the solution of equation (4) to good accuracy in certain situations. Verify the expansion

$$\frac{1}{e^s-1}=\frac{1}{s}-\frac{1}{2}+\frac{s}{12}+O(s^2).$$

Using the approximation $1/(e^s-1)\approx 1/s-1/2$ for $s=l_{ij}\theta_j$, show that

$$\theta_{n+1,j}=\frac{\sum_i(M_{ij}-N_{ij})}{\frac{1}{2}\sum_i(M_{ij}+N_{ij})l_{ij}}$$

results. Can you motivate this result heuristically?

12. Show that the observed data loglikelihood (2) for the transmission tomography model is concave. State a necessary condition for strict concavity in terms of the number of pixels and the number of projections.

References

[1] Clarke CA, Price Evans DA, McConnell RB, Sheppard PM (1959) Secretion of blood group antigens and peptic ulcers. *Brit Med J* 1:603–607

[2] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Stat Soc B* 39:1–38

[3] Herman GT (1980) *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. Springer-Verlag, New York

[4] Lange K, Carson R (1984) EM reconstruction algorithms for emission and transmission tomography. *J Computer Assist Tomography* 8:306–316

[5] Little RJA, Rubin DB (1987) *Statistical Analysis with Missing Data*. Wiley, New York

[6] McLachlan GJ, Krishnan T (1997) *The EM Algorithm and Extensions*. Wiley, New York

[7] Rao CR (1975) *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York

[8] Smith CAB (1957) Counting methods in genetical statistics. *Ann Hum Genet* 21:254–276

[9] Tanner MA (1993) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed. Springer-Verlag, New York

[10] Titterington DM, Smith AFM, Makov UE (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York

11
Newton's Method and Scoring

11.1 Introduction

The EM algorithm is not the only method of finding maximum likelihood estimates. Many of the alternative algorithms such as Newton's method and scoring share a similar structure superficially quite distinct from the EM algorithm. In iterating toward the maximum point $\hat{\theta}$, these algorithms rely on quadratic approximations to the loglikelihood $L(\theta)$. Newton's method is the gold standard for speed of convergence in a neighborhood of $\hat{\theta}$. To motivate Newton's method, we define $dL(\theta)$ to be the row vector of first partial derivatives of $L(\theta)$ and $d^2L(\theta)$ to be the Hessian matrix of second partial derivatives of $L(\theta)$. In mathematics, these functions are termed the first and second differentials of $L(\theta)$. In statistics, the gradient $\nabla L(\theta) = dL(\theta)^t$ of $L(\theta)$ is traditionally called the score, and $-d^2L(\theta)$ is called the observed information. One major advantage of maximizing the loglikelihood rather than the likelihood is that loglikelihood, score, and observed information are all additive functions of independent observations.

11.2 Newton's Method

A second-order Taylor expansion around the current point θ_n gives

$$\begin{aligned} L(\theta) &\approx L(\theta_n) + dL(\theta_n)(\theta - \theta_n) \\ &\quad + \frac{1}{2}(\theta - \theta_n)^t d^2L(\theta_n)(\theta - \theta_n). \end{aligned} \tag{1}$$

In Newton's method one maximizes the right-hand side of (1) by equating its gradient $dL(\theta_n)^t + d^2L(\theta_n)(\theta - \theta_n)$ to $\mathbf{0}$ and solving for the next iterate

$$\theta_{n+1} = \theta_n - d^2L(\theta_n)^{-1}dL(\theta_n)^t.$$

Obviously, any stationary point of $L(\theta)$ is a fixed point of Newton's algorithm.

There are two potential problems with Newton's method. First, it can be expensive computationally to evaluate the observed information. Second, far from $\hat{\theta}$, Newton's method is equally happy to head uphill or down. In other words, Newton's method is not an ascent algorithm in the sense that $L(\theta_{n+1}) > L(\theta_n)$. To generate an ascent algorithm, we can replace the observed information $-d^2L(\theta_n)$ by a positive definite approximating matrix A_n. With this substitution, the proposed increment $\Delta\theta_n = A_n^{-1}dL(\theta_n)^t$, if sufficiently contracted, forces an increase in $L(\theta)$. For a nonstationary point, this assertion follows from the first-order Taylor expansion

$$\begin{aligned} L(\theta_n + \alpha\Delta\theta_n) - L(\theta_n) &= dL(\theta_n)\alpha\Delta\theta_n + o(\alpha) \\ &= \alpha dL(\theta_n)A_n^{-1}dL(\theta_n)^t + o(\alpha), \end{aligned}$$

where the error ratio $o(\alpha)/\alpha$ tends to 0 as the positive contraction constant α tends to 0. Thus, a positive definite modification of the observed information combined with some form of backtracking leads to an ascent algorithm. The simplest form of backtracking is step-halving. If the initial increment $\Delta\theta_n$ does not produce an increase in $L(\theta)$, then try $\Delta\theta_n/2$. If $\Delta\theta_n/2$ fails, then try $\Delta\theta_n/4$, and so forth.

11.3 Scoring

One can approximate the observed information a variety of ways. The method of steepest ascent replaces the observed information by the identity matrix I. The usually more efficient scoring algorithm replaces the observed information by the expected information $J(\theta) = \mathrm{E}[-d^2L(\theta)]$. The alternative representation $J(\theta) = \mathrm{Var}[dL(\theta)^t]$ of $J(\theta)$ as a covariance matrix shows that it is nonnegative definite [14]. An extra dividend of scoring is that the inverse matrix $J(\hat{\theta})^{-1}$ immediately supplies the asymptotic variances and covariances of the maximum likelihood estimate $\hat{\theta}$ [14]. Scoring shares this benefit with Newton's method since the observed information is under natural assumptions asymptotically equivalent to the expected information.

It is possible to compute $J(\theta)$ explicitly for exponential families of densities following the approach of Jennrich and Moore [10]. (See also [1, 3, 9, 13], where the connections between scoring and iteratively reweighted least squares are emphasized.) Such densities take the form

$$f(x \mid \theta) = g(x)e^{\beta(\theta) + h(x)^t\gamma(\theta)} \tag{2}$$

relative to some measure ν. Most of the distributional families commonly encountered in statistics are exponential families. The score and expected information can be expressed in terms of the mean vector $\mu(\theta) = \mathrm{E}[h(X)]$ and covariance matrix $\Sigma(\theta) = \mathrm{Var}[h(X)]$ of the sufficient statistic $h(X)$. If $d\gamma(\theta)$ is the matrix of partial derivatives of the column vector $\gamma(\theta)$, then the first differential amounts to

$$dL(\theta) = d\beta(\theta) + h(x)^t d\gamma(\theta). \tag{3}$$

If $\gamma(\theta)$ is linear in θ, then $J(\theta) = -d^2 L(\theta) = -d^2\beta(\theta)$, and scoring coincides with Newton's method. If, in addition, $J(\theta)$ is positive definite, then $L(\theta)$ is strictly concave and possesses at most one local maximum.

The score conveniently has vanishing expectation because

$$\begin{aligned}\mathrm{E}[dL(\theta)] &= \int \frac{df(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta) d\nu(x) \\ &= d \int f(x \mid \theta) d\nu(x)\end{aligned}$$

and $\int f(x,\theta) d\nu(x) = 1$. (Differentiation under the expectation sign is incidentally permitted for exponential families [12].) For an exponential family, this fact can be restated as

$$d\beta(\theta) + \mu(\theta)^t d\gamma(\theta) = 0. \tag{4}$$

Subtracting equation (4) from equation (3) yields the alternative representation

$$dL(\theta) = [h(x) - \mu(\theta)]^t d\gamma(\theta) \tag{5}$$

of the first differential. From this it follows directly that the expected information is given by

$$\begin{aligned}J(\theta) &= \mathrm{Var}[dL(\theta)^t] \\ &= d\gamma(\theta)^t \Sigma(\theta) d\gamma(\theta).\end{aligned} \tag{6}$$

To eliminate $d\gamma(\theta)$ in equations (5) and (6), note that

$$\begin{aligned}d\mu(\theta) &= \int h(x) df(x \mid \theta) d\nu(x) \\ &= \int h(x) dL(\theta) f(x \mid \theta) d\nu(x) \\ &= \int h(x)[h(x) - \mu(\theta)]^t d\gamma(\theta) f(x \mid \theta) d\nu(x) \\ &= \Sigma(\theta) d\gamma(\theta).\end{aligned}$$

When $\Sigma(\theta)$ is invertible, this calculation implies $d\gamma(\theta) = \Sigma(\theta)^{-1} d\mu(\theta)$, which in view of (5) and (6) yields

$$\begin{aligned}dL(\theta) &= [h(x) - \mu(\theta)]^t \Sigma(\theta)^{-1} d\mu(\theta) \\ J(\theta) &= d\mu(\theta)^t \Sigma(\theta)^{-1} d\mu(\theta).\end{aligned} \tag{7}$$

TABLE 11.1. Score and Information for Some Exponential Families

Distribution	$L(\theta)$	$dL(\theta)$	$J(\theta)$
Binomial	$x \ln \frac{p}{1-p} + m \ln(1-p)$	$\frac{x-np}{p(1-p)} dp$	$\frac{n}{p(1-p)} dp^t dp$
Multinomial	$\sum_i x_i \ln p_i$	$\sum_i \frac{x_i}{p_i} dp_i$	$\sum_i \frac{n}{p_i} dp_i^t dp_i$
Poisson	$-\mu + x \ln \mu$	$-d\mu + \frac{x}{\mu} d\mu$	$\frac{1}{\mu} d\mu^t d\mu$
Exponential	$-\ln \mu - \frac{x}{\mu}$	$-\frac{1}{\mu} d\mu + \frac{x}{\mu^2} d\mu$	$\frac{1}{\mu^2} d\mu^t d\mu$

In fact, the representations in (7) hold even when $\Sigma(\theta)$ is not invertible provided a generalized inverse $\Sigma(\theta)^-$ is substituted for $\Sigma(\theta)^{-1}$ [10]. By definition, a generalized inverse satisfies $\Sigma(\theta)\Sigma(\theta)^-\Sigma(\theta) = \Sigma(\theta)$.

Table 11.1 displays the loglikelihood, score vector, and expected information matrix for some commonly applied exponential families. In this table, x represents a single observation from the binomial, Poisson, and exponential families. For the multinomial family with m categories, $x = (x_1, \ldots, x_m)$ gives the category-by-category counts. The quantity μ denotes the mean of x in each case. For the binomial family, we express the mean $\mu = np$ in terms of the number of trials n and the success probability p per trial. Similar conventions hold for the multinomial family.

The multinomial distribution deserves special comment. An easy calculation shows that the covariance matrix $\Sigma(\theta)$ has entries

$$n[1_{\{i=j\}} p_i(\theta) - p_i(\theta)p_j(\theta)].$$

In this case the matrix $\Sigma(\theta)$ is not invertible, but the diagonal matrix with ith diagonal entry $1/[np_i(\theta)]$ does provide a generalized inverse. Taking this fact into account, straightforward calculations validate the multinomial family entries of Table 11.1.

In the ABO allele frequency estimation problem studied in Chapter 10, scoring can be implemented by taking as basic parameters p_A and p_B and expressing $p_O = 1 - p_A - p_B$. Scoring then leads to the same maximum likelihood point $(\hat{p}_A, \hat{p}_B, \hat{p}_O) = (.2136, .0501, .7363)$ as the EM algorithm. The quicker convergence of scoring here—four iterations as opposed to five starting from $(.3, .2, .5)$—is often more dramatic in other problems. Scoring also has the advantage over EM of immediately providing asymptotic standard deviations of the parameter estimates. These are $(.0135, .0068, .0145)$ for the estimates $(\hat{p}_A, \hat{p}_B, \hat{p}_O)$.

TABLE 11.2. AIDS Data from Australia during 1983–1986

Quarter	Deaths	Quarter	Deaths	Quarter	Deaths
1	0	6	4	11	20
2	1	7	9	12	25
3	2	8	18	13	37
4	3	9	23	14	45
5	1	10	31	—	—

11.4 Generalized Linear Models

The generalized linear model [13] deals with exponential families (2) in which the sufficient statistic $h(X)$ is X and the mean μ of X completely determines the distribution of X. In many applications it is natural to postulate that $\mu(\theta) = q(z^t\theta)$ is a monotone function q of some linear combination of known covariates z. The inverse of q is called the link function. In this setting, $d\mu(\theta) = q'(z^t\theta)z^t$. It follows from equation (7) that if $x_1, \ldots, x_m$ are independent observations with corresponding covariates $z_1, \ldots, z_m$, then the score and expected information can be written as

$$dL(\theta)^t = \sum_{i=1}^{m} \frac{x_i - \mu_i(\theta)}{\sigma_i^2} q'(z_i^t\theta) z_i$$

$$J(\theta) = \sum_{i=1}^{m} \frac{1}{\sigma_i^2} q'(z_i^t\theta)^2 z_i z_i^t,$$

where $\sigma_i^2 = \mathrm{Var}(X_i)$.

Table 11.2 contains quarterly data on AIDS deaths in Australia that illustrate the application of a generalized linear model [7, 15]. A simple plot of the data suggests exponential growth. A plausible model therefore involves Poisson distributed observations x_i with means $\mu_i(\theta) = e^{\theta_1 + i\theta_2}$. Because this parameterization renders scoring equivalent to Newton's method, scoring gives the quick convergence noted in Table 11.3.

TABLE 11.3. Scoring Iterates for the AIDS Model

Iteration	Step-Halves	θ_1	θ_2
1	0	0.0000	0.0000
2	3	−1.3077	0.4184
3	0	0.6456	0.2401
4	0	0.3744	0.2542
5	0	0.3400	0.2565
6	0	0.3396	0.2565

11.5 The Gauss–Newton Algorithm

In nonlinear regression with normally distributed errors, the scoring algorithm metamorphoses into the Gauss–Newton algorithm. Suppose that the m independent observations $x_1, \ldots, x_m$ are normally distributed with means $\mu_i(\phi)$ and variances σ^2/w_i, where the w_i are known constants. To estimate the mean parameter vector ϕ and the variance parameter σ^2 by scoring, we first write the loglikelihood up to a constant as the function

$$L(\theta) = -\frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi)]^2$$

of the parameters $\theta = (\phi^t, \sigma^2)^t$. Straightforward differentiations and integrations yield the score

$$dL(\theta)^t = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi)]d\mu_i(\phi)^t \\ -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi)]^2 \end{pmatrix}$$

and the expected information

$$J(\theta) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{m} w_i d\mu_i(\phi)^t d\mu_i(\phi) & 0 \\ 0 & \frac{m}{2\sigma^4} \end{pmatrix}.$$

Scoring updates ϕ by

$$\phi_{n+1} = \phi_n + \Big[\sum_{i=1}^{m} w_i d\mu_i(\phi_n)^t d\mu(\phi_n)\Big]^{-1}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi_n)]d\mu_i(\phi_n)^t \quad (8)$$

and σ^2 by

$$\sigma^2_{n+1} = \frac{1}{m}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi_n)]^2.$$

The iterations (8) on ϕ can be carried out blithely neglecting those on σ^2.

Another way of deriving the Gauss–Newton update (8) is to minimize the weighted sum of squares $S(\phi) = \frac{1}{2}\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi)]^2$ by Newton's method, omitting the contribution

$$-\sum_{i=1}^{m} w_i[x_i - \mu_i(\phi)]d^2\mu_i(\phi)$$

to the Hessian $d^2S(\phi)$. With this omission, the approximate Hessian

$$d^2S(\phi) \approx \sum_{i=1}^{m} w_i d\mu_i(\phi)^t d\mu(\phi)$$

is nonnegative definite. If the weighted residuals $w_i[x_i - \mu_i(\phi)]$ are small or the regression functions $\mu_i(\theta)$ are nearly linear, then the Gauss–Newton algorithm shares the fast convergence of Newton's method.

11.6 Quasi-Newton Methods

Quasi-Newton methods of maximum likelihood update the current approximation A_n to the observed information $-d^2L(\theta_n)$ by a low-rank perturbation satisfying a secant condition. The secant condition originates from the first-order Taylor approximation

$$dL(\theta_n)^t - dL(\theta_{n+1})^t \approx d^2L(\theta_{n+1})(\theta_n - \theta_{n+1}).$$

If we set

$$\begin{aligned} g_n &= dL(\theta_n)^t - dL(\theta_{n+1})^t \\ s_n &= \theta_n - \theta_{n+1}, \end{aligned}$$

then the secant condition is $-A_{n+1}s_n = g_n$. The unique, symmetric, rank-one update to A_n satisfying the secant condition is furnished by Davidon's formula [6]

$$A_{n+1} = A_n - c_n v_n v_n^t \tag{9}$$

with constant c_n and vector v_n specified by

$$\begin{aligned} c_n &= \frac{1}{(g_n + A_n s_n)^t s_n} \\ v_n &= g_n + A_n s_n. \end{aligned} \tag{10}$$

Until recently, symmetric rank-two updates such as those associated with Davidon, Fletcher, and Powell (DFP) or with Broyden, Fletcher, Goldfarb, and Shanno (BFGS) were considered superior to the more parsimonious update (9). However, numerical analysts [4, 11] are now beginning to appreciate the virtues of Davidon's formula. To put it into successful practice, one must usually monitor A_n for positive definiteness. An immediate concern is that the constant c_n is undefined when the inner product $(g_n + A_n s_n)^t s_n = 0$. In such situations or when $(g_n + A_n s_n)^t s_n$ is small compared to $\|g_n + A_n s_n\|_2 \|s_n\|_2$, one can ignore the secant requirement and simply take $A_{n+1} = A_n$.

If A_n is positive definite and $c_n \leq 0$, then A_{n+1} is certainly positive definite. If $c_n > 0$, then it may be necessary to shrink c_n to maintain positive definiteness. In order for A_{n+1} to be positive definite, it is necessary that $\det A_{n+1} > 0$. In view of formula (9) of Chapter 7,

$$\det A_{n+1} = (1 - c_n v_n^t A_n^{-1} v_n) \det A_n,$$

and $\det A_{n+1} > 0$ requires $1 - c_n v_n^t A_n^{-1} v_n > 0$. Conversely, the condition

$$1 - c_n v_n^t A_n^{-1} v_n > 0 \tag{11}$$

is also sufficient to ensure positive definiteness of A_{n+1}. This fact can be most easily demonstrated by invoking the Sherman–Morrison formula

$$[A_n - c_n v_n v_n^t]^{-1} = A_n^{-1} + \frac{c_n}{1 - c_n v_n^t A_n^{-1} v_n} A_n^{-1} v_n [A_n^{-1} v_n]^t. \tag{12}$$

Formula (12) shows that $[A_n - c_n v_n v_n^t]^{-1}$ exists and is positive definite under condition (11). Since the inverse of a positive definite matrix is positive definite, it follows that $A_n - c_n v_n v_n^t$ is positive definite as well.

The above analysis suggests the possibility of choosing c_n so that not only does A_{n+1} remain positive definite, but $\det A_{n+1}$ always exceeds a small constant $\epsilon > 0$. This strategy can be realized by replacing c_n by

$$\min \left\{ c_n, \left(1 - \frac{\epsilon}{\det A_n}\right) \frac{1}{v_n^t A_n^{-1} v_n} \right\}$$

in updating A_n. An even better strategy that monitors the condition number of A_n is sketched in Problem 12.

In successful applications of quasi-Newton methods, choice of the initial matrix A_1 is critical. Setting $A_1 = I$ is convenient, but often poorly scaled for a particular problem. A better choice is $A_1 = J(\theta_1)$ when the expected information matrix $J(\theta_1)$ is available. In some problems, $J(\theta)$ is expensive to compute and manipulate for general θ but cheap to compute and manipulate for special θ. The special θ can furnish good starting points for a quasi-Newton search. For instance, $J(\theta)$ can be diagonal in certain circumstances. We will return to the delicate issue of selecting an initial approximate Hessian when discussing extensions of the EM algorithm in Chapter 12.

It is noteworthy that the strategy of updating the approximation A_n to $-d^2 L(\theta_n)$ can be reformulated to update the approximation $H_n = A_n^{-1}$ to $-d^2 L(\theta_n)^{-1}$ instead. Restating the secant condition $-A_{n+1} s_n = g_n$ as the inverse secant condition $-H_{n+1} g_n = s_n$ leads to the symmetric rank-one update

$$H_{n+1} = H_n - b_n w_n w_n^t, \tag{13}$$

where $b_n = [(s_n + H_n g_n)^t g_n]^{-1}$ and $w_n = s_n + H_n g_n$. This strategy has the advantage of avoiding explicit inversion of A_n in calculating the quasi-Newton direction $\Delta\theta_n = H_n dL(\theta_n)^t$. However, monitoring positive definiteness of H_n forces us to invert it. Whichever strategy one adopts, monitoring positive definiteness is most readily accomplished by carrying forward simultaneously A_n and $H_n = A_n^{-1}$ and applying the Sherman–Morrison formula to update each.

The final result of this chapter increases our confidence in the symmetric rank-one update [8].

Proposition 11.6.1. *Let* $L(\theta) = d + e^t\theta + \frac{1}{2}\theta^t F\theta$ *be a strictly concave quadratic function on* R^m. *In the quasi-Newton scheme*

$$\theta_n = \theta_{n-1} + H_{n-1}dL(\theta_{n-1})^t \tag{14}$$

with H_n *defined by equation (13), suppose that the constants* $b_1, \ldots, b_m$ *are well defined and that the vectors* $s_1, \ldots, s_m$ *are linearly independent. Then* $H_{m+1} = -F^{-1}$, *and* θ_{m+1} *is the maximum of* $L(\theta)$.

Proof. First we observe that the exact Taylor expansion $g_i = Fs_i$ holds for all i. Given this fact, let us prove inductively that $-H_i g_j = s_j$ for $1 \leq j \leq i-1$. Our claim is true for $i = 2$ by design, so suppose it is true for an arbitrary $i \geq 2$. Again, $-H_{i+1}g_i = s_i$ holds by design. If $1 \leq j < i$, then equation (13), the identity $g_i = Fs_i$, and the induction hypothesis together imply that

$$\begin{aligned} -H_{i+1}g_j &= -H_i g_j + b_i w_i w_i^t g_j \\ &= s_j + b_i w_i (s_i^t + g_i^t H_i) g_j \\ &= s_j + b_i w_i (s_i^t g_j - g_i^t s_j) \\ &= s_j + b_i w_i (s_i^t F s_j - s_i^t F s_j) \\ &= s_j. \end{aligned}$$

If we let $S = (s_1, \ldots, s_m)$ and $G = (g_1, \ldots, g_m)$, then H_{m+1} satisfies the identity $-H_{m+1}G = S$. But F satisfies $G = FS$. It follows that G is invertible and $-H_{m+1} = SG^{-1} = F^{-1}$. The last assertion of the proposition is left to Problem 2. □

11.7 Problems

1. Verify the score and information entries in Table 11.1.
2. Show that Newton's method converges in one iteration to the maximum of

$$L(\theta) = d + e^t\theta + \frac{1}{2}\theta^t F\theta$$

if the symmetric matrix F is negative definite. Note the relevance of this result to Proposition 11.6.1 and to the one-step convergence of the Gauss–Newton algorithm (8) when the regression functions $\mu_i(\phi)$ are linear.

3. Prove that the inverse matrix $\Sigma(\theta)^{-1}$ appearing in (7) can be replaced by a generalized inverse $\Sigma(\theta)^-$ [10]. (Hint: Show that the difference $h(X) - \mu(\theta)$ is almost surely in the range of $\Sigma(\theta)$ and hence that

$$\Sigma(\theta)\Sigma(\theta)^-[h(X) - \mu(\theta)] = h(X) - \mu(\theta)$$

TABLE 11.4. Ingot Data for a Quantal Response Model

Trials n_i	**Observation** x_i	**Covariate** z_{i1}	**Covariate** z_{i2}
55	0	1	7
157	2	1	14
159	7	1	27
16	3	1	57

almost surely. To validate the claim about the range of $h(X) - \mu(\theta)$, let P denote perpendicular projection onto the range of $\Sigma(\theta)$. Then show that $\mathrm{E}(\|(I-P)[h(X)-\mu(\theta)]\|_2^2) = 0$.)

4. A quantal response model involves independent binomial observations $x_1, \ldots, x_m$ with n_i trials and success probability $\pi_i(\theta)$ per trial for the ith observation. If z_i is a covariate vector and θ a parameter vector, then the specification

$$\pi_i(\theta) = \frac{e^{z_i^t\theta}}{1 + e^{z_i^t\theta}}$$

gives a generalized linear model. Estimate $\hat{\theta} = (-5.132, 0.0677)^t$ for the ingot data of Cox [5] displayed in Table 11.4.

5. In robust regression it is useful to consider location-scale families with densities of the form

$$\frac{c}{\sigma} e^{-\rho(\frac{x-\mu}{\sigma})}, \qquad x \in (-\infty, \infty). \tag{15}$$

Here $\rho(r)$ is a strictly convex, even function, decreasing to the left of 0 and symmetrically increasing to the right of 0. Without loss of generality, one can take $\rho(0) = 0$. The normalizing constant c is determined by $c\int_{-\infty}^{\infty} e^{-\rho(r)}dr = 1$. Show that a random variable X with density (15) has mean μ and variance

$$\mathrm{Var}(X) = c\sigma^2 \int_{-\infty}^{\infty} r^2 e^{-\rho(r)} dr.$$

If μ depends on a parameter vector ϕ, demonstrate that the score corresponding to a single observation $X = x$ amounts to

$$dL(\theta)^t = \begin{pmatrix} \frac{1}{\sigma}\rho'(\frac{x-\mu}{\sigma}) d\mu(\phi) \\ -\frac{1}{\sigma} + \rho'(\frac{x-\mu}{\sigma})\frac{x-\mu}{\sigma^2} \end{pmatrix}$$

for $\theta = (\phi^t, \sigma)^t$. Finally, prove that the expected information $J(\theta)$ is block diagonal with upper left block

$$\frac{c}{\sigma^2} \int_{-\infty}^{\infty} \rho''(r) e^{-\rho(r)} dr\, d\mu(\phi)^t d\mu(\phi)$$

and lower right block

$$\frac{c}{\sigma^2}\int_{-\infty}^{\infty}\rho''(r)r^2e^{-\rho(r)}dr+\frac{1}{\sigma^2}.$$

6. In the context of Problem 5, take $\rho(r)=\ln\cosh^2(\frac{r}{2})$. Show that this corresponds to the logistic distribution with density

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

Compute the integrals

$$\begin{aligned}\frac{\pi^2}{3}&=c\int_{-\infty}^{\infty}r^2e^{-\rho(r)}dr\\ \frac{1}{3}&=c\int_{-\infty}^{\infty}\rho''(r)e^{-\rho(r)}dr\\ \frac{1}{3}+\frac{\pi^2}{9}&=c\int_{-\infty}^{\infty}\rho''(r)r^2e^{-\rho(r)}dr+1\end{aligned}$$

determining the variance and expected information of the density (15) for this choice of $\rho(r)$.

7. Continuing Problems 5 and 6, compute the normalizing constant c and the three integrals determining the variance and expected information for Huber's function

$$\rho(r)=\begin{cases}\frac{r^2}{2} & |r|\le k\\ k|r|-\frac{k^2}{2} & |r|>k\end{cases}.$$

8. A family of discrete density functions $p_n(\theta)$ defined on $\{0,1,\ldots\}$ and indexed by a parameter $\theta>0$ is said to be a power series family if for all n

$$p_n(\theta)=\frac{c_n\theta^n}{g(\theta)}, \tag{16}$$

where $c_n\ge 0$, and where $g(\theta)=\sum_{k=0}^{\infty}c_k\theta^k$ is the appropriate normalizing constant. If $x_1,\ldots,x_m$ are independent observations from the discrete density (16), then show that the maximum likelihood estimate of θ is a root of the equation

$$\frac{1}{m}\sum_{i=1}^{m}x_i=\frac{\theta g'(\theta)}{g(\theta)}.$$

Prove that the expected information in a single observation is

$$J(\theta)=\frac{\sigma^2(\theta)}{\theta^2},$$

where $\sigma^2(\theta)$ is the variance of the density (16).

9. In the Gauss–Newton algorithm (8), the matrix

$$\sum_{i=1}^{m} w_i d\mu_i(\phi_n)^t d\mu(\phi_n)$$

can be singular or nearly so. To cure this ill, Marquardt suggested substituting

$$A_n = \sum_{i=1}^{m} w_i d\mu_i(\phi_n)^t d\mu(\phi_n) + \lambda I$$

for it and iterating according to

$$\phi_{n+1} = \phi_n + A_n^{-1} \sum_{i=1}^{m} w_i [x_i - \mu_i(\phi_n)] d\mu_i(\phi_n)^t. \tag{17}$$

Prove that the increment $\Delta\phi_n = \phi_{n+1} - \phi_n$ proposed in equation (17) minimizes the criterion

$$\frac{1}{2} \sum_{i=1}^{m} w_i [x_i - \mu_i(\phi_n) - d\mu(\phi_n)\Delta\phi_n]^2 + \frac{\lambda}{2} \|\Delta\phi_n\|_2^2.$$

10. Consider the quadratic function

$$L(\theta) = -(1,1)\theta - \frac{1}{2}\theta^t \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \theta$$

defined on R^2. Compute the iterates of the quasi-Newton scheme (14) starting from $\theta_1 = (0,0)^t$ and $H_1 = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

11. Let A be a positive definite matrix. Prove [2] that

$$\operatorname{tr}(A) - \ln\det(A) \geq \ln[\operatorname{cond}_2(A)]. \tag{18}$$

(Hint: Express $\operatorname{tr}(A) - \ln\det(A)$ in terms of the eigenvalues of A. Then use the inequalities $\lambda - \ln\lambda \geq 1$ and $\lambda > 2\ln\lambda$ for all $\lambda > 0$.)

12. In Davidon's symmetric rank-one update (9), it is possible to control the condition number of A_{n+1} by shrinking the constant c_n. Suppose a moderately sized number d is chosen. Due to inequality (18), one can avoid ill-conditioning in the matrices A_n by imposing the constraint $\operatorname{tr}(A_n) - \ln\det(A_n) \leq d$. To see how this fits into the updating scheme (9), verify that

$$\begin{aligned} \ln\det(A_{n+1}) &= \ln\det(A_n) + \ln(1 - c_n v_n^t A_n^{-1} v_n) \\ \operatorname{tr}(A_{n+1}) &= \operatorname{tr}(A_n) - c_n \|v_n\|_2^2. \end{aligned}$$

Employing these results, deduce that $\operatorname{tr}(A_{n+1}) - \ln\det(A_{n+1}) \leq d$ provided c_n satisfies

$$-c_n\|v_n\|_2^2 - \ln(1 - c_n v_n^t A_n^{-1} v_n) \leq d - \operatorname{tr}(A_n) + \ln\det(A_n).$$

References

[1] Bradley EL (1973) The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J Amer Stat Assoc* 68: 199–200

[2] Byrd RH, Nocedal J (1989) A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM J Numer Anal* 26:727–739

[3] Charnes A, Frome EL, Yu PL (1976) The equivalence of generalized least squares and maximum likelihood in the exponential family. *J Amer Stat Assoc* 71:169–171

[4] Conn AR, Gould NIM, Toint PL (1991) Convergence of quasi-Newton matrices generated by the symmetric rank one update. *Math Prog* 50:177–195

[5] Cox DR (1970) *Analysis of Binary Data*. Methuen, London

[6] Davidon WC (1959) Variable metric methods for minimization. *AEC Research and Development Report ANL–5990*, Argonne National Laboratory, USA

[7] Dobson AJ (1990) *An Introduction to Generalized Linear Models*. Chapman & Hall, London

[8] Fiacco AV, McCormick GP (1968) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley, New York

[9] Green PJ (1984) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J Roy Stat Soc B* 46:149–192

[10] Jennrich RI, Moore RH (1975) Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Statistical Computing Section: Amer Stat Assoc* 57–65

[11] Khalfan HF, Byrd RH, Schnabel RB (1993) A theoretical and experimental study of the symmetric rank-one update. *SIAM J Optimization* 3:1–24

[12] Lehmann EL (1986) *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York

[13] Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J Roy Stat Soc A* 135:370–384

[14] Rao CR (1973) *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York

[15] Whyte BM, Gold J, Dobson AJ, Cooper DA (1987) Epidemiology of acquired immunodeficiency syndrome in Australia. *Med J Aust* 147:65–69

12
Variations on the EM Theme

12.1 Introduction

The triumvirate of the EM algorithm, Newton's method, and scoring never totally dominates the maximum likelihood world. There are many departures from and variations on these general themes. Among the alternatives are cyclic coordinate ascent, the EM gradient algorithm, accelerated EM algorithms, and majorization methods. Our brief introduction to these alternatives emphasizes by way of example their complex web of connections and their fascinating relations to the EM algorithm. The next chapter explores local and global convergence patterns for the standard optimization methods.

12.2 Iterative Proportional Fitting

In some problems it pays to update only a subset of the parameters at a time. This point is best illustrated by a specific example taken from the contingency table literature [2, 11]. Consider a three-way contingency table with first-order interactions. If the three factors are indexed by i, j, and k and have r, s, and t levels, respectively, then a loglinear model for the observed data y_{ijk} consists in defining an exponentially parameterized mean

$$\mu_{ijk} = e^{\lambda+\lambda_i^1+\lambda_j^2+\lambda_k^3+\lambda_{ij}^{12}+\lambda_{ik}^{13}+\lambda_{jk}^{23}}$$

for each cell ijk. To ensure that all parameters are identifiable, we make the usual assumption that a parameter set summed over one of its indices yields 0. For instance, $\lambda^1_\cdot = \sum_i \lambda^1_i = 0$ and $\lambda^{12}_{i\cdot} = \sum_j \lambda^{12}_{ij} = 0$. The overall effect λ is permitted to be nonzero.

If we postulate independent Poisson distributions for the random variables Y_{ijk} underlying the observed values y_{ijk}, then the loglikelihood is

$$L = \sum_i \sum_j \sum_k (y_{ijk} \ln \mu_{ijk} - \mu_{ijk}). \tag{1}$$

Maximizing L with respect to λ can be accomplished by setting

$$\frac{\partial}{\partial \lambda} L = \sum_i \sum_j \sum_k (y_{ijk} - \mu_{ijk}) = 0.$$

This tells us that whatever the other parameters are, λ should be adjusted so that $\mu_{\cdots} = y_{\cdots} = m$ is the total sample size. In other words, if $\mu_{ijk} = e^\lambda \omega_{ijk}$, then λ is chosen so that $e^\lambda = m/\omega_{\cdots}$. With this proviso, the loglikelihood becomes up to an irrelevant constant

$$L = \sum_i \sum_j \sum_k y_{ijk} \ln \frac{\omega_{ijk}}{\omega_{\cdots}},$$

which is just the loglikelihood of a multinomial distribution with probability $\omega_{ijk}/\omega_{\cdots}$ attached to cell ijk. Thus, for purposes of maximum likelihood estimation, we might as well stick with the Poisson sampling model.

Unfortunately, no closed-form solution to the Poisson likelihood equations exists satisfying the complicated linear constraints. The resolution of this dilemma lies in refusing to update all of the parameters simultaneously. Suppose that we consider only the parameters λ, λ^1_i, λ^2_j, and λ^{12}_{ij} pertinent to the first two factors. If in equation (1) we let

$$\begin{aligned} \mu_{ij} &= e^{\lambda + \lambda^1_i + \lambda^2_j + \lambda^{12}_{ij}} \\ \alpha_{ijk} &= e^{\lambda^3_k + \lambda^{13}_{ik} + \lambda^{23}_{jk}}, \end{aligned}$$

then setting

$$\begin{aligned} \frac{\partial}{\partial \lambda^{12}_{ij}} L &= \sum_k (y_{ijk} - \mu_{ijk}) \\ &= y_{ij\cdot} - \mu_{ij\cdot} \\ &= y_{ij\cdot} - \mu_{ij} \alpha_{ij\cdot} \\ &= 0 \end{aligned}$$

leads to $\mu_{ij} = y_{ij\cdot}/\alpha_{ij\cdot}$. The constraint $\sum_k (y_{ijk} - \mu_{ijk}) = 0$ implies that the other partial derivatives

$$\frac{\partial}{\partial \lambda} L = y_{\cdots} - \mu_{\cdots}$$

$$\frac{\partial}{\partial \lambda_i^1} L = y_{i\cdot\cdot} - \mu_{i\cdot\cdot}$$
$$\frac{\partial}{\partial \lambda_j^2} L = y_{\cdot j\cdot} - \mu_{\cdot j\cdot}$$

vanish as well, and we have found a stationary point of the loglikelihood provided appropriate parameters λ, λ_i^1, λ_j^2, and λ_{ij}^{12} exist consistent with the choices $\mu_{ij} = y_{ij\cdot}/\alpha_{ij\cdot}$.

Given the means μ_{ij}, the model parameters are determined from the logarithms $\ln \mu_{ij} = \lambda + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12}$ via

$$\lambda = \frac{1}{rs} \sum_i \sum_j \ln \mu_{ij}$$
$$\lambda_i^1 = \frac{1}{s} \sum_j \ln \mu_{ij} - \lambda$$
$$\lambda_j^2 = \frac{1}{r} \sum_i \ln \mu_{ij} - \lambda$$
$$\lambda_{ij}^{12} = \ln \mu_{ij} - \lambda - \lambda_i^1 - \lambda_j^2.$$

It is easy to check that these updated parameters satisfy the relevant constraints.

At the second stage, the parameter set λ, λ_i^1, λ_k^3, and λ_{ik}^{13} is updated, holding the remaining parameters fixed. At the third stage, the parameter set λ, λ_j^2, λ_k^3, and λ_{jk}^{23} is updated, holding the remaining parameters fixed. These three successive stages constitute one iteration of the iterative proportional fitting algorithm. Each stage either leaves all parameters unchanged or increases the loglikelihood.

If some observations y_{ijk} are missing in this model, then one can formulate an EM algorithm that fills in the missing data by replacing each missing observation by its current mean μ_{ijk}. This simple imputation procedure is a direct consequence of the assumed independence of the Y_{ijk}. Once the missing data are filled in, iterative proportional fitting can commence as just described. Meng and Rubin dub this the ECM algorithm [20]. In fact, there are two versions of the ECM algorithm. One version redoes the E step after every stage, and the other redoes the E step only after all stages within an iteration are finished.

12.3 EM Gradient Algorithm

As noted in our earlier transmission tomography example, the M step of the EM algorithm cannot always be solved exactly. In such cases one can approximately maximize the E-step function $Q(\theta \mid \theta_n)$ by one step of

Newton's method. The EM gradient algorithm [15] iterates according to

$$\begin{aligned}\theta_{n+1} &= \theta_n - d^{20}Q(\theta_n \mid \theta_n)^{-1}d^{10}Q(\theta_n \mid \theta_n)^t \\ &= \theta_n - d^{20}Q(\theta_n \mid \theta_n)^{-1}dL(\theta_n)^t, \end{aligned} \tag{2}$$

where $d^{10}Q(\theta \mid \theta_n)$ and $d^{20}Q(\theta \mid \theta_n)$ indicate the first and second differentials of $Q(\theta \mid \theta_n)$ with respect to its left variable θ. The substitution of the score $dL(\theta_n)^t$ for $d^{10}Q(\theta_n \mid \theta_n)^t$ in (2) is valid because $L(\theta) - Q(\theta \mid \theta_n)$ attains its minimum at $\theta = \theta_n$. The EM gradient algorithm and the EM algorithm enjoy the same rate of convergence approaching the maximum likelihood point $\hat{\theta}$. Furthermore, in the vicinity of $\hat{\theta}$, the EM gradient algorithm also satisfies the ascent condition $L(\theta_{n+1}) > L(\theta_n)$ [15].

12.3.1 Application to the Dirichlet Distribution

As an example, consider parameter estimation for the Dirichlet distribution [13]. This distribution is derived by taking independent gamma random variables $X_1, \ldots, X_k$ and forming the vector of proportions

$$Y = (Y_1, \ldots, Y_k)^t$$

defined componentwise by $Y_i = X_i/(\sum_{j=1}^k X_j)$. If the random variable X_i has density $x_i^{\theta_i - 1}e^{-x_i}\Gamma(\theta_i)^{-1}$ on $(0, \infty)$, then standard arguments show that Y has the Dirichlet density

$$\frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k y_i^{\theta_i - 1} \tag{3}$$

on the simplex $\{y = (y_1, \ldots, y_k)^t : y_1 > 0, \ldots, y_k > 0, \sum_{i=1}^k y_i = 1\}$ endowed with the uniform measure. In the context of the EM algorithm, the random vector Y constitutes the observed data, and the underlying random vector $X = (X_1, \ldots, X_k)^t$ constitutes the complete data.

Suppose now $Y_1 = y_1, \ldots, Y_m = y_m$ are randomly sampled vectors from the Dirichlet distribution. To estimate the underlying parameter vector $\theta = (\theta_1, \ldots, \theta_k)^t$ by the EM algorithm, let $X_1, \ldots, X_m$ be the corresponding complete data random vectors. It is immediately evident that

$$\begin{aligned} Q(\phi \mid \theta_n) = &-m\sum_{j=1}^k \ln \Gamma(\phi_j) + \sum_{j=1}^k (\phi_j - 1)\sum_{i=1}^m \mathrm{E}(\ln X_{ij} \mid Y_i = y_i, \theta_n) \\ &- \sum_{i=1}^m \sum_{j=1}^k \mathrm{E}(X_{ij} \mid Y_i = y_i, \theta_n), \end{aligned} \tag{4}$$

where X_{ij} is the jth component of the vector X_i. Owing to the presence of the terms $\ln \Gamma(\phi_j)$ in (4), the M step appears intractable. However, the EM gradient algorithm can be readily implemented since the score vector

has entries

$$\frac{\partial}{\partial \phi_j} L(\phi) = m\psi\Big(\sum_{i=1}^{k} \phi_i\Big) - m\psi(\phi_j) + \sum_{i=1}^{m} \ln y_{ij},$$

and the Hessian matrix $d^{20}Q(\phi \mid \phi)$ is diagonal with jth diagonal entry $-m\psi'(\phi_j)$, where $\psi(t) = \frac{d}{dt} \ln \Gamma(t)$ and $\psi'(t) = \frac{d^2}{dt^2} \ln \Gamma(t)$ are the digamma and trigamma functions, respectively. Because the Dirichlet distribution belongs to a regular exponential family, it is unnecessary in this example to evaluate the conditional expectations of the E step. In fact, the negative definite Hessian $-d^{20}Q(\theta_n \mid \theta_n)$ collapses to the expected information of the complete data, and the EM gradient algorithm coincides with an earlier gradient algorithm proposed by Titterington [24, 25]. Lange [15] compares the EM gradient algorithm to Newton's method [22] for a specific example of Dirichlet modeled data.

12.4 Bayesian EM

If a prior $\pi(\theta)$ is imposed on the parameter vector θ, then $L(\theta) + \ln \pi(\theta)$ is the logposterior function. Its maximum occurs at the posterior mode. The posterior mode can be found by defining the surrogate function

$$\begin{aligned} Q(\theta \mid \theta_n) &= \mathrm{E}[\ln f(X \mid \theta) + \ln \pi(\theta) \mid Y, \theta_n] \\ &= \mathrm{E}[\ln f(X \mid \theta) \mid Y, \theta_n] + \ln \pi(\theta). \end{aligned}$$

Thus, in the E step of the Bayesian algorithm, one simply adds the logprior to the usual surrogate function. The difference

$$L(\theta) + \ln \pi(\theta) - Q(\theta \mid \theta_n) = L(\theta) - \mathrm{E}[\ln f(X \mid \theta) \mid Y, \theta_n]$$

again attains its minimum at $\theta = \theta_n$. Thus, the M-step strategy of maximizing $Q(\theta \mid \theta_n)$ forces an increase in the logposterior function. Because the logprior often complicates the M step, the EM gradient algorithm tends to be even more valuable in computing posterior modes than in computing maximum likelihood estimates. We will explore the Bayesian version of the EM algorithm when we revisit transmission tomography later in this chapter.

12.5 Accelerated EM

We now consider the question of how to accelerate the often excruciatingly slow convergence of the EM algorithm. In contrast, Newton's method enjoys exceptionally quick convergence in a neighborhood of the maximum point. This suggests amending the EM algorithm so that it resembles Newton's method. Because the EM algorithm typically performs well far from

the maximum likelihood point, hybrid algorithms that begin as pure EM and gradually make the transition to Newton's method are apt to perform best. We now describe one such algorithm based on quasi-Newton approximations [12, 16].

If we let H_n approximate $-d^2L(\theta_n)^{-1}$, the inverse of the observed information matrix, then a quasi-Newton scheme employing H_n iterates according to $\theta_{n+1} = \theta_n + H_n dL(\theta_n)^t$. Updating H_n can be based on the inverse secant condition $-H_{n+1}g_n = s_n$, where $g_n = dL(\theta_n) - dL(\theta_{n+1})$ and $s_n = \theta_n - \theta_{n+1}$. This is just the usual quasi-Newton procedure described earlier. However, one can do better. First of all, the EM algorithm already provides the natural approximation $-d^{20}Q(\theta_n \mid \theta_n)^{-1}$ to $-d^2L(\theta_n)^{-1}$. To improve on this baseline approximation, one can add to it an estimate of the difference

$$d^{20}Q(\theta_n \mid \theta_n)^{-1} - d^2L(\theta_n)^{-1}.$$

If B_n accurately approximates $d^{20}Q(\theta_n \mid \theta_n)^{-1} - d^2L(\theta_n)^{-1}$, then

$$H_n = B_n - d^{20}Q(\theta_n \mid \theta_n)^{-1}$$

should furnish a good approximation to $-d^2L(\theta_n)^{-1}$. The inverse secant condition for B_{n+1} is

$$-B_{n+1}g_n = s_n - d^{20}Q(\theta_{n+1} \mid \theta_{n+1})^{-1}g_n. \tag{5}$$

Davidon's symmetric rank-one update can be used to construct B_{n+1} from B_n.

Given the availability of B_n, the next iterate in the quasi-Newton search can be expressed as

$$\theta_{n+1} = \theta_n + B_n dL(\theta_n)^t - d^{20}Q(\theta_n \mid \theta_n)^{-1}dL(\theta_n)^t. \tag{6}$$

Equation (6) can be simplified by noting that the term

$$-d^{20}Q(\theta_n \mid \theta_n)^{-1}dL(\theta_n)^t \tag{7}$$

is the EM gradient increment and as such closely approximates the ordinary EM increment $\Delta_{EM}\theta_n$. Thus, the algorithm

$$\theta_{n+1} = \theta_n + B_n dL(\theta_n)^t + \Delta_{EM}\theta_n. \tag{8}$$

should be as effective as algorithm (6). Replacing the EM gradient increment (7) by $\Delta_{EM}\theta_n$ also simplifies the inverse secant condition (5). With the understanding that $d^{20}Q(\theta_n \mid \theta_n)^{-1} \approx d^{20}Q(\theta_{n+1} \mid \theta_{n+1})^{-1}$, the inverse secant condition becomes

$$-B_{n+1}g_n = s_n + \Delta_{EM}\theta_n - \Delta_{EM}\theta_{n+1}.$$

Thus, quasi-Newton acceleration can be phrased entirely in terms of the score $dL(\theta)^t$ and ordinary EM increments when the M step of the EM algorithm is solvable [12].

In practice, we need some initial approximation B_1. The choice $B_1 = \mathbf{0}$ works well because it guarantees that the first iterate of the accelerated algorithm is either the EM or EM gradient iterate. There is also the issue of whether θ_{n+1} actually increases the loglikelihood. If it does not, one can reduce the contribution of $B_n dL(\theta_n)^t$ by a step-halving tactic until

$$\theta_{n+1} = \theta_n + \frac{1}{2^k} B_n dL(\theta_n)^t + \Delta_{EM}\theta_n$$

does lead to an increase in the loglikelihood [16]. Alternatively, Jamshidian and Jennrich [12] recommend conducting a limited line search along the direction implied by the update (8). If this search is unsuccessful, then they suggest resetting $B_n = \mathbf{0}$ and beginning the approximation process anew. Regardless of these details, formulating the acceleration scheme in terms of approximating $-d^2L(\theta_n)^{-1}$ avoids time-consuming matrix inversions. This will be a boon in high-dimensional applications such as medical imaging.

12.6 EM Algorithms Without Missing Data

The EM algorithm transfers maximization from the loglikelihood $L(\theta)$ to the surrogate function $Q(\theta \mid \theta_n)$. The key ingredient in making this transfer successful is the fact that $L(\theta) - Q(\theta \mid \theta_n)$ attains its minimum at $\theta = \theta_n$. Thus, determining the next iterate θ_{n+1} to maximize $Q(\theta \mid \theta_n)$ forces an increase in $L(\theta)$. The EM derives its numerical stability from this ascent property. Optimization transfer also tends to substitute simple optimization problems for difficult optimization problems. Simplification usually relies on one or more of the following devices: (a) avoidance of large matrix inversions, (b) linearization, (c) separation of parameters, and (d) graceful handling of equality and inequality constraints.

In the remainder of this chapter, we will explore some examples of how to construct surrogate (or majorization) functions $Q(\theta \mid \theta_n)$ without explicitly invoking notions of missing data [1, 4]. These constructions depend on inequalities derived from convexity. As demonstrated in the next chapter, a unified convergence theory can be erected to cover the classical EM algorithm and its generalizations based on surrogate optimization functions.

12.6.1 Quadratic Lower Bound Principle

Böhning and Lindsay [3] introduced a lower bound algorithm under the assumption that a negative definite matrix B can be found such that the matrix difference $d^2L(\theta) - B$ is nonnegative definite for all θ. In this situation, the quadratic function

$$Q(\theta \mid \theta_n) = L(\theta_n) + dL(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^t B(\theta - \theta_n)$$

serves as a surrogate for $L(\theta)$. Because

$$L(\theta) = L(\theta_n) + dL(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^t d^2 L(\overline{\theta})(\theta - \theta_n)$$

for some intermediate point $\overline{\theta}$ between θ and θ_n, we find that

$$\begin{aligned} L(\theta) - Q(\theta \mid \theta_n) &= \frac{1}{2}(\theta - \theta_n)^t [d^2 L(\overline{\theta}) - B](\theta - \theta_n) \\ &\geq 0 \end{aligned}$$

and that $L(\theta) - Q(\theta \mid \theta_n)$ has its minimum at $\theta = \theta_n$. It follows just as in the case of the EM algorithm that if θ_{n+1} maximizes $Q(\theta \mid \theta_n)$, then $L(\theta_{n+1}) \geq L(\theta_n)$. Setting $d^{10}Q(\theta \mid \theta_n) = \mathbf{0}$ produces the maximum point

$$\theta_{n+1} = \theta_n - B^{-1} dL(\theta_n)^t.$$

Böhning and Lindsay [3] consider the problem of logistic regression with a large vector z_i of predictors for each observation y_i, $i = 1, \ldots, m$. If the y_i are realizations of independent Bernoulli trials with success probabilities

$$p_i = \frac{e^{z_i^t \theta}}{1 + e^{z_i^t \theta}},$$

then the loglikelihood and the observed information are

$$\begin{aligned} L(\theta) &= \sum_{i=1}^m [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] \\ -d^2 L(\theta) &= \sum_{i=1}^m p_i(1 - p_i) z_i z_i^t. \end{aligned}$$

Because $p_i(1 - p_i) \leq 1/4$, the negative definite matrix $B = -\frac{1}{4} \sum_{i=1}^m z_i z_i^t$ yields a difference $d^2 L(\theta) - B$ that is positive definite. In this example, the EM gradient algorithm has the advantage of requiring a single matrix inversion instead of repeated matrix inversions. It can be faster than Newton's method despite taking more iterations.

12.6.2 Elliptically Symmetric Densities and L_p Regression

Dutter and Huber [10] introduced an optimization transfer principle for elliptically symmetric densities

$$\frac{e^{-\frac{1}{2}\kappa(\delta^2)}}{(2\pi)^{\frac{k}{2}} \det(\Omega)^{\frac{1}{2}}} \tag{9}$$

defined for $y \in R^k$, where $\delta^2 = (y - \mu)^t \Omega^{-1} (y - \mu)$ denotes the Mahalanobis distance between y and μ and $\kappa(s)$ is some strictly increasing, strictly concave function. The multivariate t provides a typical example of an elliptically symmetric distribution that can profitably be substituted for the multivariate normal distribution in robust estimation [18, 19].

For a sequence $y_1, \ldots, y_m$ of independent observations from the elliptically symmetric density (9) with covariance matrices $\Omega_1, \ldots, \Omega_m$ and means $\mu_1, \ldots, \mu_m$, the surrogate function is the multivariate normal loglikelihood

$$Q(\theta \mid \theta_n) = -\frac{1}{2} \sum_{i=1}^{m} [w_i \delta_i^2(\theta) + \ln \det \Omega_i(\theta)],$$

with weights $w_i = \kappa'[\delta_i^2(\theta_n)]$. The fact that $L(\theta) - Q(\theta \mid \theta_n)$ has its minimum at θ_n follows from the fact that $\kappa'(s_n)s - \kappa(s)$ has its minimum at $s = s_n$. The array of techniques from linear algebra for maximizing the normal distribution can be brought to bear on maximizing $Q(\theta \mid \theta_n)$. It is a nontrivial fact that the Dutter–Huber algorithm usually coincides with the classical EM algorithm for normal/independent distributions [5, 19].

For $0 < p \leq 2$ and independent univariate observations $y_1, \ldots, y_m$ with unit variances, the choice $\kappa(s) = s^{p/2}$ leads to least L_p regression. The Dutter–Huber procedure minimizes $\sum_{i=1}^{m} w_i(\theta_n)[y_i - \mu_i(\theta)]^2$ at each iteration with weights $w_i(\theta_n) = |y_i - \mu_i(\theta_n)|^{p-2}$. In other words, least L_p regression can be done by iteratively reweighted least squares. This algorithm, originally proposed by Schlossmacher [23] and Merle and Spath [21], is unfortunately plagued by infinite weights for those observations with zero residuals. To avoid this difficulty, we can redefine the weights to be

$$w_i(\theta_n) = \frac{1}{\epsilon + |y_i - \mu_i(\theta_n)|^{2-p}}$$

for a small $\epsilon > 0$ [19]. This corresponds to the choice

$$\kappa'(s) = \frac{p}{2(\epsilon + s^{1-p/2})}$$

and also leads to a maximum likelihood algorithm that increases the loglikelihood at each iteration. The revised algorithm for $p = 1$ minimizes the criterion

$$\sum_{i=1}^{m} \{|y_i - \mu_i(\theta)| - \epsilon \ln[\epsilon + |y_i - \mu_i(\theta)|]\}, \tag{10}$$

which obviously tends to $\sum_{i=1}^{m} |y_i - \mu_i(\theta)|$ as $\epsilon \to 0$.

12.6.3 Transmission Tomography Revisited

As noted in Chapter 10, the loglikelihood in transmission tomography can be written succinctly as

$$L(\theta) = -\sum_i f_i(l_i^t \theta),$$

where $f_i(s) = d_i e^{-s} + y_i s$ and $l_i^t \theta = \sum_j l_{ij} \theta_j$ is the inner product of the attenuation parameter vector θ and the vector of intersection lengths l_i for projection i. Following the lead of De Pierro [6] in emission tomography,

one can devise an alternative EM algorithm based on a convexity argument [17]. First define admixture constants

$$\alpha_{ij} = \frac{l_{ij}\theta_{nj}}{l_i^t\theta_n}. \tag{11}$$

Since $\sum_j \alpha_{ij} = 1$ and each $f_i(s)$ is strictly convex, the right-hand side of the inequality

$$\begin{aligned} L(\theta) &= -\sum_i f_i\Big(\sum_j \alpha_{ij}\frac{\theta_j}{\theta_{nj}} l_i^t\theta_n\Big) \\ &\geq -\sum_i \sum_j \alpha_{ij} f_i\Big(\frac{\theta_j}{\theta_{nj}} l_i^t\theta_n\Big) \\ &= Q(\theta \mid \theta_n) \end{aligned} \tag{12}$$

defines the surrogate function $Q(\theta \mid \theta_n)$. Equality occurs in inequality (12) when $\theta_j = \theta_{nj}$ for all j. Thus, the function $L(\theta) - Q(\theta \mid \theta_n)$ attains its minimum of 0 when $\theta = \theta_n$. By construction, maximization of $Q(\theta \mid \theta_n)$ separates into a sequence of one-dimensional problems, each of which can be solved approximately by one step of Newton's method.

This algorithm is easier to orchestrate than the ordinary EM algorithm because it uses only full line integrals and avoids partial line integrals. It also requires far fewer exponentiations than the ordinary EM algorithm. In practice, these advantages make it decisively faster [17].

The images produced by maximum likelihood estimation in transmission tomography look grainy. Geman and McClure [8] recommend incorporating a Gibbs prior that enforces image smoothness. A Gibbs prior $\pi(\theta)$ can be written as

$$\ln \pi(\theta) = -\gamma \sum_{\{j,k\}\epsilon N} w_{jk}\psi(\theta_j - \theta_k),$$

where γ and the weights w_{jk} are positive constants, N is a set of unordered pairs $\{j,k\}$ defining a neighborhood system, and $\psi(r)$ is called a potential function. For instance, if the pixels are squares, we might define the weights by $w_{jk} = 1$ for orthogonal nearest neighbors and $w_{jk} = 1/\sqrt{2}$ for diagonal nearest neighbors. The constant γ scales the overall strength assigned to the prior.

Choice of the potential function $\psi(r)$ is the most crucial feature of the Gibbs prior. It is convenient to assume that $\psi(r)$ is even and strictly convex. Strict convexity leads to strict concavity of the log posterior $L(\theta) + \ln \pi(\theta)$ and permits simple modification of the alternative EM algorithm based on the $Q(\theta \mid \theta_n)$ function defined by inequality (12). Many potential functions exist satisfying these conditions. One simple example is $\psi(r) = r^2$. Because this choice tends to deter the formation of boundaries, Green [9] has suggested the gentler alternative $\psi(r) = \ln[\cosh(r)]$, which grows for large $|r|$ linearly rather than quadratically.

One adverse consequence of introducing a prior is that it couples the parameters in the M step of the EM algorithm for finding the posterior mode. One can decouple the parameters by exploiting the convexity and evenness of the potential function $\psi(r)$ through the inequality

$$\begin{aligned}\psi(\theta_j - \theta_k) &= \psi\Big(\frac{1}{2}\Big[2\theta_j - \theta_{nj} - \theta_{nk}\Big] + \frac{1}{2}\Big[-2\theta_k + \theta_{nj} + \theta_{nk}\Big]\Big) \\ &\leq \frac{1}{2}\psi(2\theta_j - \theta_{nj} - \theta_{nk}) + \frac{1}{2}\psi(2\theta_k - \theta_{nj} - \theta_{nk}),\end{aligned}$$

which is strict unless $\theta_j + \theta_k = \theta_{nj} + \theta_{nk}$ [6]. This inequality allows us to redefine the surrogate function as

$$\begin{aligned}Q(\theta \mid \theta_n) = &-\sum_i \sum_j \alpha_{ij} f_i\Big(\frac{\theta_j}{\theta_{nj}} l_i^t \theta_n\Big) \\ &- \frac{\gamma}{2} \sum_{\{j,k\}\epsilon N} w_{jk}[\psi(2\theta_j - \theta_{nj} - \theta_{nk}) + \psi(2\theta_k - \theta_{nj} - \theta_{nk})],\end{aligned}$$

where $f_i(s) = d_i e^{-s} + y_i s$ and the admixture constants a_{ij} are given by (11). The parameters are once again separated in the M step, and the difference $L(\theta) + \ln \pi(\theta) - Q(\theta \mid \theta_n)$ assumes its minimum at $\theta = \theta_n$. Maximizing $Q(\theta \mid \theta_n)$ therefore drives the logposterior uphill and eventually leads to the posterior mode.

12.7 Problems

1. Consider the coronary disease data [11, 14] displayed in the three-way contingency Table 12.1. Using iterative proportional fitting, find the maximum likelihood estimates for the loglinear model with first-order interactions. Perform a chi-square test to decide whether this model fits the data better than the model postulating independence of the three factors.
2. As noted in the text, the loglinear model for categorical data can be interpreted as assuming independent Poisson distributions for the various categories with category i having mean $\mu_i(\theta) = e^{l_i^t \theta}$, where l_i is a vector whose entries are 0's or 1's. Calculate the observed information $-d^2 L(\theta) = \sum_i e^{l_i^t \theta} l_i l_i^t$ in this circumstance, and deduce that it is nonnegative definite. In the presence of linear constraints on θ, show that any maximum likelihood estimate of θ is necessarily unique provided the vector subspace of possible θ is included in the linear span of the l_i. (Hint: Expand the loglikelihood $L(\theta)$ to second order around the maximum likelihood point.)
3. In Example 12.3.1, digamma and trigamma functions must be evaluated. Show that these functions satisfy the recurrence relations

$$\psi(t) = -t^{-1} + \psi(t+1)$$

TABLE 12.1. Coronary Disease Data

Disease Status	Cholesterol Level	Blood Pressure				
		1	2	3	4	Total
Coronary	1	2	3	3	4	12
	2	3	2	1	3	9
	3	8	11	6	6	31
	4	7	12	11	11	41
Total		20	28	21	24	93
No Coronary	1	117	121	47	22	307
	2	85	98	43	20	246
	3	119	209	68	43	439
	4	67	99	46	33	245
Total		388	527	204	118	1237

$$\psi'(t) = t^{-2} + \psi'(t+1).$$

Thus, if $\psi(t)$ and $\psi'(t)$ can be accurately evaluated via asymptotic expansions for large t, then they can be accurately evaluated for small t. For example, it is known that $\psi(t) = \ln t - (2t)^{-1} + O(t^{-2})$ and $\psi'(t) = t^{-1} + (\sqrt{2}t)^{-2} + O(t^{-3})$ as $t \to \infty$.

4. Compute the score vector and the observed and expected information matrices for the Dirichlet distribution (3). Explicitly invert the expected information using the Sherman–Morrison formula.

5. Suppose that $\kappa'(s) = \frac{1}{2(\epsilon+\sqrt{s})}$. Show that $\kappa(s) = \sqrt{s} - \epsilon \ln(\epsilon + \sqrt{s}) + c$, where c is a constant, and that $\kappa(s)$ is strictly increasing and strictly concave on $[0, \infty)$.

6. Suppose that the complete data in the EM algorithm involve N binomial trials with success probability θ per trial. Here N can be random or fixed. If M trials result in success, then the complete data likelihood can be written as $\theta^M(1-\theta)^{N-M}c$, where c is an irrelevant constant. The E step of the EM algorithm amounts to forming

$$Q(\theta \mid \theta_n) = \mathrm{E}(M \mid Y, \theta_n) \ln \theta + \mathrm{E}(N - M \mid Y, \theta_n) \ln(1-\theta) + \ln c.$$

The binomial trials are hidden because only a function Y of them is directly observed. Show in this setting that the EM update is given by either of the two equivalent expressions

$$\begin{aligned} \theta_{n+1} &= \frac{\mathrm{E}(M \mid Y, \theta_n)}{\mathrm{E}(N \mid Y, \theta_n)} \\ &= \theta_n + \frac{\theta_n(1-\theta_n)}{\mathrm{E}(N \mid Y, \theta_n)} \frac{\partial}{\partial \theta} L(\theta_n), \end{aligned}$$

where $L(\theta)$ is the loglikelihood of the observed data Y [26]. This is a gradient form of the EM algorithm, but is it the EM gradient algorithm?

7. As an example of the hidden binomial trials theory sketched in Problem 6, consider a random sample of twin pairs. Let u of these pairs consists of male pairs, v consist of female pairs, and w consist of opposite-sex pairs. A simple model to explain these data involves a random Bernoulli choice for each pair dictating whether it consists of identical or nonidentical twins. Suppose that identical twins occur with probability p and nonidentical twins with probability $1-p$. Once the decision is made as to whether or not the twins are identical, sexes are assigned to the twins. If the twins are identical, one assignment of sex is made. If the twins are nonidentical, two independent assignments of sex are made. Suppose boys are chosen with probability q and girls with probability $1-q$. Model these data as hidden binomial trials. Using the result of the previous problem, give the EM algorithm for estimating p and q.

8. In the spirit of Problem 6, formulate a model for hidden Poisson or exponential trials [26]. If the number of trials is N and the mean per trial is θ, then show that the EM update in the Poisson case is

$$\theta_{n+1} = \theta_n + \frac{\theta_n}{\mathrm{E}(N \mid Y, \theta_n)} \frac{\partial}{\partial \theta} L(\theta_n)$$

and in the exponential case is

$$\theta_{n+1} = \theta_n + \frac{\theta_n^2}{\mathrm{E}(N \mid Y, \theta_n)} \frac{\partial}{\partial \theta} L(\theta_n),$$

where $L(\theta)$ is the loglikelihood of the observed data Y.

9. In least L_1 regression, show that the maximum likelihood estimate satisfies the equality

$$\sum_{i=1}^{m} \mathrm{sgn}[y_i - \mu_i(\theta)] d\mu_i(\theta)^t = \mathbf{0},$$

provided no residual $y_i - \mu_i(\theta) = 0$ and the regression functions $\mu_i(\theta)$ are differentiable. What is the corresponding equality for the modified criterion (10)?

10. Show that the EM gradient version of the alternative EM algorithm for transmission tomography iterates according to

$$\theta_{n+1,j} = \theta_{nj} \frac{\sum_i l_{ij}[d_i e^{-l_i^t \theta_n}(1 + l_i^t \theta_n) - y_i]}{\sum_i l_{ij} l_i^t \theta_n d_i e^{-l_i^t \theta_n}}$$

in the absence of a smoothing prior.

11. Continuing Problem 10, demonstrate that the exact solution of the one-dimensional likelihood equation

$$\frac{\partial}{\partial \theta_j} Q(\theta \mid \theta_n) = 0$$

exists and is positive when $\sum_i l_{ij} d_i > \sum_i l_{ij} y_i$. Why does this condition usually hold?

12. Prove that the function $\psi(r) = \ln[\cosh(r)]$ is even, strictly convex, infinitely differentiable, and asymptotic to $|r|$ as $|r| \to \infty$.

13. De Pierro [7] has suggested an approach to maximizing functions of the form $L(\theta) = -\sum_{i=1}^p f_i(c_i^t\theta)$, where $c_i^t\theta = \sum_{j=1}^q c_{ij}\theta_j$ denotes an inner product and each function $f_i(r)$ is strictly convex. In this setting we impose no nonnegativity conditions on either the constants c_{ij} or the parameter components θ_j. If each $f_i(r)$ is twice continuously differentiable, then show that $L(\theta)$ has observed information matrix $-d^2L(\theta) = \sum_{i=1}^p f_i''(c_i^t\theta)c_i c_i^t$, and consequently $L(\theta)$ is strictly concave provided each $f_i''(r)$ is strictly positive and the c_i span the space inhabited by θ.

Now assume nonnegative constants λ_{ij} are given with $\lambda_{ij} > 0$ when $c_{ij} \neq 0$ and with $\sum_{j=1}^q \lambda_{ij} = 1$. If $S_i = \{j : \lambda_{ij} > 0\}$, then demonstrate the inequality

$$\begin{aligned} -\sum_{i=1}^p f_i(c_i^t\theta) &\geq -\sum_{i=1}^p \sum_{j\in S_i} \lambda_{ij} f_i\Big[\frac{c_{ij}}{\lambda_{ij}}(\theta_j - \theta_{nj}) + c_i^t\theta_n\Big] \\ &= Q(\theta \mid \theta_n), \end{aligned} \tag{13}$$

with equality when $\theta = \theta_n$. Prove that $Q(\theta \mid \theta_n)$ attains its maximum when

$$\sum_{i\in T_j} f_i'\Big[\frac{c_{ij}}{\lambda_{ij}}(\theta_j - \theta_{nj}) + c_i^t\theta_n\Big]c_{ij} = 0 \tag{14}$$

holds for all j, where $T_j = \{i : \lambda_{ij} > 0\}$. Thus, $Q(\theta \mid \theta_n)$ serves as a surrogate optimization function in which the parameters are separated. Check that one step of Newton's method provides the approximate solution

$$\theta_{n+1,j} = \theta_{nj} - \Big[\sum_{i\in T_j} f_i''(c_i^t\theta_n)\frac{c_{ij}^2}{\lambda_{ij}}\Big]^{-1} \sum_{i\in T_j} f_i'(c_i^t\theta_n)c_{ij}. \tag{15}$$

Three reasonable choices for the constants λ_{ij} are $\lambda_{ij} = |c_{ij}|^2/||c_i||_2^2$, $\lambda_{ij} = |c_{ij}|/||c_i||_1$, and $\lambda_{ij} = 1_{\{c_{ij}\neq 0\}}/(\sum_k 1_{\{c_{ik}\neq 0\}})$.

14. Continuing Problem 13, suppose $f_i(r) = (y_i - r)^2$. Prove that maximizing the function $Q(\theta \mid \theta_n)$ leads to

$$\theta_{n+1,j} = \theta_{nj} + \frac{\sum_{i\in T_j}(y_i - c_i^t\theta_n)c_{ij}}{\sum_{i\in T_j}\frac{c_{ij}^2}{\lambda_{ij}}},$$

which is a special case of the update (15). Thus, we have found a novel iterative method of calculating linear regression coefficients requiring no matrix inversion.

15. Continuing Problem 13, show that the functions

$$\begin{aligned} f_i(r) &= -y_i r + n_i \ln(1 + e^r) \\ f_i(r) &= -n_i r + y_i \ln(1 + e^r) \\ f_i(r) &= -y_i r + e^r \\ f_i(r) &= \nu_i r + y_i e^{-r} \end{aligned}$$

are strictly convex and provide up to sign loglikelihoods for the binomial, negative binomial, Poisson, and gamma densities, respectively. For the binomial, let y_i be the number of successes in n_i trials. For the negative binomial, let y_i be the number of trials until n_i successes. In both cases, $p = e^r/(e^r + 1)$ is the success probability. For the Poisson, let e^r be the mean. Finally, for the gamma, let e^r be the scale parameter, assuming the shape parameter ν_i is fixed. For each density, equation (14) determining $\theta_{n+1,j}$ appears analytically intractable, but presumably the update (15) is viable. This problem has obvious implications for logistic and Poisson regression.

16. Continuing Problem 13, suppose $f_i(r) = |y_i - r|$. These nondifferentiable functions correspond to least L_1 regression. Show that the maximum of the surrogate function $Q(\theta \mid \theta_n)$ defined in (13) has jth component $\theta_{n+1,j}$ minimizing

$$\begin{aligned} s(\theta_j) &= \sum_{i \in T_j} w_i |d_i - \theta_j| \\ w_i &= |c_{ij}| \\ d_i &= \theta_{nj} + (y_i - c_i^t \theta_n) \frac{\lambda_{ij}}{c_{ij}}, \end{aligned}$$

where $T_j = \{i : \lambda_{ij} > 0\}$. If we assume that $T_j = \{1, \ldots, p\}$ and that $d_1 < d_2 < \ldots < d_p$ for the sake of simplicity, then show that $s(\theta_j)$ is minimized by choosing $\theta_{n+1,j} = d_i$, where i is the largest integer in $\{1, \ldots, p\}$ such that $\sum_{k<i} w_k - \sum_{k \geq i} w_k < 0$. This, of course, is the median of the discrete random variable taking the value d_i with probability proportional to w_i. (Hint: Examine $s'(\theta_j)$ on each of the intervals (d_{i-1}, d_i), or invoke the usual characterization of a median as solving a least L_1 problem.)

17. In the spirit of Problem 13, show that the function

$$Q(\theta \mid \theta_n) = \sum_{i=1}^{p} y_i c_i^t \theta - n \ln \Big[\sum_{j=1}^{q} g_j(\theta_j) \Big],$$

where

$$g_j(\theta_j) = \sum_{i \in T_j} \lambda_{ij} e^{\frac{c_{ij}}{\lambda_{ij}}(\theta_j - \theta_{nj}) + c_i^t \theta_n},$$

serves as a surrogate maximization function for a loglinear model with cell i having count y_i and probability proportional to $e^{c_i^t\theta}$. Although this choice does not separate parameters, it yields a simple, one-step Newton update. Prove that $-d^{20}Q(\theta \mid \theta_n)$ is a nonnegative definite matrix that can be expressed as a rank-one perturbation of a diagonal matrix. Thus, $-d^{20}Q(\theta \mid \theta_n)$ can be straightforwardly inverted by the Sherman–Morrison formula. (Hint: For nonnegative definiteness, prove that $-\Delta\theta^t d^{20}Q(\theta \mid \theta_n)\Delta\theta \geq 0$ follows from the Cauchy-Schwarz inequality.)

References

[1] Becker MP, Yang I, Lange K (1997) EM algorithms without missing data. *Stat Methods Med Res* 6:37–53

[2] Bishop YMM, Feinberg SE, Holland PW (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA

[3] Böhning D, Lindsay BG (1988) Monotonicity of quadratic approximation algorithms. *Ann Math Stat* 40:641–663

[4] de Leeuw J (1994) Block relaxation algorithms in statistics. *Information Systems and Data Analysis,* edited by Bock HH, Lenski W, Richter MM, Springer-Verlag, Berlin, pp 308–325

[5] Dempster AP, Laird NM, Rubin DB (1980) Iteratively reweighted least squares for linear regression when the errors are normal/independent distributed. in *Multivariate Analysis-V*, Krishnaiah PR, editor, North Holland, Amsterdam, pp 35–57

[6] De Pierro AR (1993) On the relation between the ISRA and EM algorithm for positron emission tomography. *IEEE Trans Med Imaging* 12:328–333

[7] De Pierro AR (1995) A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans Med Imaging* 14:132–137

[8] Geman S, McClure D (1985) Bayesian image analysis: An application to single photon emission tomography. *Proc Stat Comput Sec*, Amer Stat Assoc, Washington, DC, pp 12–18

[9] Green P (1990) Bayesian reconstruction for emission tomography data using a modified EM algorithm. *IEEE Trans Med Imaging* 9:84–94

[10] Huber PJ (1981) *Robust Statistics*, Wiley, New York

[11] Everitt BS (1977) *The Analysis of Contingency Tables*. Chapman & Hall, London

[12] Jamshidian M, Jennrich RI (1995) Acceleration of the EM algorithm by using quasi-Newton methods. *J Roy Stat Soc B* 59:569–587

[13] Kingman JFC (1993) *Poisson Processes*. Oxford University Press, Oxford

[14] Ku HH, Kullback S (1974) Log-linear models in contingency table analysis. *Biometrics* 10:452–458

[15] Lange K (1995) A gradient algorithm locally equivalent to the EM algorithm. *J Roy Stat Soc B* 57:425–437

[16] Lange K (1995) A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* 5:1–18
[17] Lange K, Fessler JA (1995) Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Trans Image Processing* 4:1430–1438
[18] Lange K, Little RJA, Taylor JMG (1989) Robust statistical modeling using the t distribution. *J Amer Stat Assoc* 84:881–896
[19] Lange K, Sinsheimer JS (1993) Normal/independent distributions and their applications in robust regression. *J Comp Graph Stat* 2:175–198
[20] Meng X-L, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80:267–278
[21] Merle G, Spath H (1974) Computational experiences with discrete L_p approximation. *Computing* 12:315–321 (1974)
[22] Narayanan A (1991) Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution. *Appl Stat* 40:365–374
[23] Schlossmacher EJ (1973) An iterative technique for absolute deviations curve fitting. *J Amer Stat Assoc* 68:857–859
[24] Titterington DM (1984) Recursive parameter estimation using incomplete data. *J Roy Stat Soc B* 46:257–267
[25] Titterington DM, Smith AFM, Makov UE (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York
[26] Weeks DE, Lange K (1989) Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis. *IMA J Math Appl Med Biol* 6:209–232

13

Convergence of Optimization Algorithms

13.1 Introduction

Proving convergence of the various maximization algorithms is a delicate exercise. In general, it is helpful to consider local and global convergence patterns separately. The local convergence rate of an algorithm provides a useful benchmark for comparing it to other algorithms. On this basis, Newton's method wins hands down. However, the tradeoffs are subtle. Besides the sheer number of iterations until convergence, the computational complexity and numerical stability of an algorithm are critically important. The EM algorithm is often the epitome of numerical stability and computational simplicity. Scoring lies somewhere between Newton's method and the EM algorithm. It tends to converge more quickly than the EM algorithm and to behave more stably than Newton's method. Quasi-Newton methods also occupy this intermediate zone. Because the issues are complex, all of these algorithms survive and prosper in certain computational niches.

The following short overview of convergence manages to cover only some highlights. Quasi-Newton methods are given especially short shrift. The efforts of a generation of numerical analysts in understanding quasi-Newton methods defy easy summary or digestion. Interested readers can consult one of the helpful references [2, 4, 9, 11]. We emphasize EM and related algorithms, partially because a fairly coherent theory for them can be reviewed in a few pages.

13.2 Calculus Preliminaries

As a prelude to our study of convergence, let us review some ideas from advanced calculus. A function $f : R^m \to R^m$ is differentiable at a point $\theta \in R^m$ if and only if an $m \times m$ matrix A exists such that

$$||f(\theta + \Delta\theta) - f(\theta) - A\Delta\theta|| = o(||\Delta\theta||)$$

as $||\Delta\theta|| \to 0$. Because of the equivalence of vector norms, any norm $||\cdot||$ will do in this definition. The matrix A is typically written $df(\theta)$. Its ith row consists of the partial derivatives of the ith component $f_i(\theta)$ of $f(\theta)$. To avoid certain pathologies, we usually make the simplifying assumption that all first partial derivatives of $f(\theta)$ exist and are continuous. This continuity assumption guarantees that the differential of $f(\theta)$ exists for all θ and can be identified with the Jacobian matrix $df(\theta)$. Differentiability of $f(\theta)$ obviously entails continuity of $f(\theta)$.

We will need an analog of the mean value theorem. This analog can be best developed by introducing the vector-valued integral $\int_a^b g(t)dt$ of a continuous vector-valued function $g : [a, b] \to R^m$. The components of $\int_a^b g(t)dt$ are just the integrals $\int_a^b g_i(t)dt$ of the components $g_i(t)$ of $g(t)$. If $a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$ is a partition of $[a, b]$, the Riemann sum $\sum_{i=1}^n g(t_i)(t_i - t_{i-1})$ approximates the integral $\int_a^b g(t)dt$ and satisfies the norm inequality

$$\Big|\Big| \sum_{i=1}^n g(t_i)(t_i - t_{i-1}) \Big|\Big| \leq \sum_{i=1}^n ||g(t_i)||(t_i - t_{i-1}).$$

Passing to the limit as the mesh size $\max_i(t_i - t_{i-1}) \to 0$, one can readily verify that $||\int_a^b g(t)dt|| \leq \int_a^b ||g(t)||dt$. Applying this inequality, the fundamental theorem of calculus, and the chain rule leads to the bound

$$\begin{aligned} ||f(\phi) - f(\theta)|| &= \Big|\Big| \int_0^1 df[\theta + t(\phi - \theta)](\phi - \theta)dt \Big|\Big| \\ &\leq \int_0^1 ||df[\theta + t(\phi - \theta)](\phi - \theta)||dt \\ &\leq \int_0^1 ||df[\theta + t(\phi - \theta)]|| \cdot ||\phi - \theta||dt \\ &\leq \sup_{t \in [0,1]} ||df[\theta + t(\phi - \theta)]|| \cdot ||\phi - \theta||. \end{aligned} \tag{1}$$

The mean value inequality (1) can be improved. Suppose that along the line $\{\theta + t(\phi - \theta) : t \in [0, 1]\}$ the differential df satisfies the inequality

$$||df(\beta) - df(\gamma)|| \leq \lambda||\beta - \gamma|| \tag{2}$$

for some constant $\lambda > 0$. This is the case if the second differential $d^2 f$ exists and is continuous, for then inequality (2) follows from an analog of

inequality (1). Assuming the truth of (2), we find that

$$\begin{aligned}
&||f(\phi) - f(\theta) - df(\theta)(\phi - \theta)|| \\
&\quad = \Big|\Big| \int_0^1 \{df[\theta + t(\phi - \theta)] - df(\theta)\}(\phi - \theta)dt \Big|\Big| \\
&\quad \leq \int_0^1 ||df[\theta + t(\phi - \theta)] - df(\theta)|| \cdot ||\phi - \theta|| dt \\
&\quad \leq \lambda ||\phi - \theta||^2 \int_0^1 t dt \\
&\quad = \frac{\lambda}{2} ||\phi - \theta||^2. \qquad (3)
\end{aligned}$$

13.3 Local Convergence

Local convergence of many optimization algorithms hinges on the following result of Ostrowski [12]:

Proposition 13.3.1. *Let the map $f : R^m \to R^m$ have fixed point θ_∞. If $f(\theta)$ is differentiable at θ_∞, and the spectral radius $\rho[df(\theta_\infty)]$ of its differential satisfies $\rho[df(\theta_\infty)] < 1$, then the iteration scheme $\theta_{n+1} = f(\theta_n)$ is locally attracted to θ_∞.*

Proof. As mentioned in our discussion of matrix norms, there exists a vector norm $||\theta||$ such that the induced matrix norm $||df(\theta_\infty)||$ comes arbitrarily close to $\rho[df(\theta_\infty)]$. Accordingly, choose an appropriate norm with $||df(\theta_\infty)|| = \sigma < 1$ and then a constant $\epsilon > 0$ with $\epsilon + \sigma < 1$. Because $f(\theta)$ is differentiable, there is a sphere $S = \{\theta : ||\theta - \theta_\infty|| < \delta\}$ such that

$$||f(\theta) - f(\theta_\infty) - df(\theta_\infty)(\theta - \theta_\infty)|| \leq \epsilon ||\theta - \theta_\infty||$$

for $\theta \in S$. It follows that $\theta \in S$ implies

$$\begin{aligned}
||f(\theta) - \theta_\infty|| &= ||f(\theta) - f(\theta_\infty)|| \\
&\leq ||f(\theta) - f(\theta_\infty) - df(\theta_\infty)(\theta - \theta_\infty)|| + ||df(\theta_\infty)(\theta - \theta_\infty)|| \\
&\leq (\epsilon + \sigma)||\theta - \theta_\infty||.
\end{aligned}$$

One can now argue inductively that if $\theta_1 \in S$, then all iterates $\theta_{n+1} = f(\theta_n)$ belong to S, and that

$$||\theta_{n+1} - \theta_\infty|| \leq (\sigma + \epsilon)^n ||\theta_1 - \theta_\infty||.$$

In other words, θ_n converges to θ_∞ at least as fast as $(\sigma + \epsilon)^n \to 0$. □

Our intention is to apply Ostrowski's result to iteration maps of the type

$$f(\theta) = \theta + A(\theta)^{-1} dL(\theta)^t, \qquad (4)$$

where $L(\theta)$ is a loglikelihood and $A(\theta)$ is the corresponding observed information, expected information, or $-d^{20}Q(\theta \mid \theta)$ matrix of the EM gradient

algorithm. Obviously, our first order of business is to compute the differential $df(\theta_\infty)$ at a local maximum θ_∞ of $L(\theta)$. If θ_∞ is a strict local maximum, then ordinarily $d^2L(\theta_\infty)$ is negative definite. We make this assumption as well as the assumption that $d^2L(\theta)$ is continuous in a neighborhood of θ_∞. Thus, the iteration map is certainly well defined in some neighborhood of θ_∞ for Newton's method. For scoring and the EM algorithm, we likewise assume that $A(\theta)$ is continuously invertible in a neighborhood of θ_∞.

Because $dL(\theta_\infty)^t = \mathbf{0}$, the differential $d[A(\theta_\infty)^{-1}]$ is irrelevant to determining $df(\theta_\infty)$. Thus, it is plausible to conjecture that

$$df(\theta_\infty) = I + A(\theta_\infty)^{-1}d^2L(\theta_\infty).$$

This claim can be verified by noting that the algorithm definition (4) and the facts $f(\theta_\infty) = \theta_\infty$ and $dL(\theta_\infty)^t = \mathbf{0}$ together imply the inequality

$$\begin{aligned}
&||f(\theta) - f(\theta_\infty) - [I + A(\theta_\infty)^{-1}d^2L(\theta_\infty)](\theta - \theta_\infty)|| \\
&\quad = ||A(\theta)^{-1}dL(\theta)^t - A(\theta_\infty)^{-1}d^2L(\theta_\infty)(\theta - \theta_\infty)|| \\
&\quad \le ||A(\theta)^{-1}[dL(\theta)^t - dL(\theta_\infty)^t - d^2L(\theta_\infty)(\theta - \theta_\infty)]|| \\
&\qquad + ||A(\theta)^{-1}[A(\theta_\infty) - A(\theta)]A(\theta_\infty)^{-1}d^2L(\theta_\infty)(\theta - \theta_\infty)|| \\
&\quad \le ||A(\theta)^{-1}||o(||\theta - \theta_\infty||) \\
&\qquad + ||A(\theta)^{-1}|| \cdot ||A(\theta_\infty) - A(\theta)|| \cdot ||A(\theta_\infty)^{-1}|| \cdot ||d^2L(\theta_\infty)|| \cdot ||\theta - \theta_\infty||.
\end{aligned}$$

Because $||A(\theta_\infty) - A(\theta)|| \to 0$ as $||\theta - \theta_\infty|| \to 0$ and $||A(\theta)^{-1}||$ is bounded in a neighborhood of θ_∞, the overall error in the linear approximation of $f(\theta)$ around θ is consequently $o(||\theta - \theta_\infty||)$.

Calculation of the differential $df(\theta_\infty)$ of an EM map at a local maximum θ_∞ is equally interesting. Recall that $f(\theta)$ provides the unique ϕ maximizing $Q(\phi \mid \theta)$. At the stationary point $f(\theta)$, we find

$$d^{10}Q[f(\theta) \mid \theta] = \mathbf{0}. \tag{5}$$

If $d^{20}Q(\phi \mid \theta)$ is invertible around $f(\theta)$, then the implicit function theorem implies that $f(\theta)$ is differentiable. Implicit differentiation of (5) produces

$$d^{20}Q[f(\theta) \mid \theta]df(\theta) + d^{11}Q[f(\theta) \mid \theta] = \mathbf{0}.$$

Solving for $df(\theta)$ at the local maximum $\theta_\infty = f(\theta_\infty)$ therefore yields

$$df(\theta_\infty) = -d^{20}Q(\theta_\infty \mid \theta_\infty)^{-1}d^{11}Q(\theta_\infty \mid \theta_\infty). \tag{6}$$

To simplify $d^{11}Q(\theta_\infty \mid \theta_\infty)$, observe that

$$dL(\theta) - d^{10}Q(\theta \mid \theta) = \mathbf{0} \tag{7}$$

holds for all θ since $L(\phi) - Q(\phi \mid \theta)$ attains its minimum at $\phi = \theta$. Taking differentials in (7) gives

$$d^2L(\theta) - d^{20}Q(\theta \mid \theta) - d^{11}Q(\theta \mid \theta) = \mathbf{0}. \tag{8}$$

Equation (8) can be solved for $d^{11}Q(\theta \mid \theta)$ at θ_∞ and the result substituted in (6). This yields

$$\begin{aligned} df(\theta_\infty) &= -d^{20}Q(\theta_\infty \mid \theta_\infty)^{-1}[d^2L(\theta_\infty) - d^{20}Q(\theta_\infty \mid \theta_\infty)] \\ &= I - d^{20}Q(\theta_\infty \mid \theta_\infty)^{-1}d^2L(\theta_\infty), \end{aligned} \tag{9}$$

which is precisely the differential computed for the EM gradient algorithm. It is noteworthy that this calculation totally ignores the missing data context of the ordinary EM algorithm.

We are now in a position to prove local convergence of the EM and EM gradient algorithms.

Proposition 13.3.2. *Both the EM algorithm and the EM gradient algorithm are locally attracted to a local maximum θ_∞ at a linear rate determined by the spectral radius of $I - d^{20}Q(\theta_\infty \mid \theta_\infty)^{-1}d^2L(\theta_\infty)$.*

Proof. Let $f(\theta)$ be the iteration map. According to Proposition 13.3.1, it suffices to show that all eigenvalues of the differential $df(\theta_\infty)$ lie on the half-open interval $[0, 1)$. But $df(\theta_\infty)$ has eigenvalues determined by the stationary values of the Rayleigh quotient

$$\begin{aligned} R(v) &= \frac{v^t[d^2L(\theta_\infty) - d^{20}Q(\theta_\infty \mid \theta_\infty)]v}{-v^t d^{20}Q(\theta_\infty \mid \theta_\infty)v} \\ &= 1 - \frac{v^t d^2L(\theta_\infty)v}{v^t d^{20}Q(\theta_\infty \mid \theta_\infty)v}. \end{aligned} \tag{10}$$

Because both $d^2L(\theta_\infty)$ and $d^{20}Q(\theta_\infty \mid \theta_\infty)$ are negative definite, $R(v) < 1$ for all vectors $v \neq \mathbf{0}$. On the other hand, $R(v) \geq 0$ since the difference $d^2L(\theta_\infty) - d^{20}Q(\theta_\infty \mid \theta_\infty)$ is nonnegative definite. □

The next proposition validates local convergence of Newton's method.

Proposition 13.3.3. *Newton's method is locally attracted to a local maximum θ_∞ at a rate faster than linear. If the observed information $-d^2L(\theta)$ satisfies*

$$||d^2L(\phi) - d^2L(\theta)|| \leq \lambda||\phi - \theta|| \tag{11}$$

in some neighborhood of θ_∞, then the Newton iterates θ_n satisfy

$$||\theta_{n+1} - \theta_\infty|| \leq 2\lambda||d^2L(\theta_\infty)^{-1}|| \cdot ||\theta_n - \theta_\infty||^2 \tag{12}$$

close to θ_∞.

Proof. If $f(\theta)$ represents the Newton iteration map, then

$$\begin{aligned} df(\theta_\infty) &= I - d^2L(\theta_\infty)^{-1}d^2L(\theta_\infty) \\ &= \mathbf{0}. \end{aligned}$$

Hence, Proposition 13.3.1 implies local attraction to θ_∞ at a rate faster than linear. If, in addition, inequality (11) holds, then inequality (3) is

true; inequalities (3) and (11) together imply

$$\begin{aligned}
&||\theta_{n+1} - \theta_\infty|| \\
&\quad = ||\theta_n - d^2L(\theta_n)^{-1}dL(\theta_n)^t - \theta_\infty|| \\
&\quad \le || - d^2L(\theta_n)^{-1}[dL(\theta_n)^t - dL(\theta_\infty)^t - d^2L(\theta_\infty)(\theta_n - \theta_\infty)]|| \\
&\qquad + ||d^2L(\theta_n)^{-1}[d^2L(\theta_n) - d^2L(\theta_\infty)](\theta_n - \theta_\infty)|| \\
&\quad \le \left(\frac{\lambda}{2} + \lambda\right)||d^2L(\theta_n)^{-1}|| \cdot ||\theta_n - \theta_\infty||^2,
\end{aligned}$$

which ultimately implies inequality (12) by virtue of the assumed continuity and invertibility of $d^2L(\theta)$ near θ_∞. □

Local convergence of the scoring algorithm is not guaranteed by Proposition 13.3.1 because nothing prevents an eigenvalue of

$$df(\theta_\infty) = I + J(\theta_\infty)^{-1}d^2L(\theta_\infty)$$

from falling below -1. Scoring with a fixed partial step,

$$\theta_{n+1} = \theta_n + \alpha J(\theta_n)^{-1}dL(\theta_n)^t,$$

will converge locally for $\alpha > 0$ sufficiently small. In practice, no adjustment is usually necessary. For reasonably large sample sizes, the expected information matrix $J(\theta_\infty)$ approximates the observed information matrix $-d^2L(\theta_\infty)$ well, and the spectral radius of $df(\theta_\infty)$ is nearly 0.

Algorithms such as iterative proportional fitting and the ECM algorithm update different subsets of the parameters in sequence. As a simple prototype of such algorithms, we now briefly examine the local convergence properties of cyclic coordinate ascent. Here $L(\theta)$ is maximized along each coordinate direction in turn. Let e_i be the vector whose ith coordinate equals 1 and whose other coordinates equal 0. The first step of cyclic coordinate ascent chooses the scalar t_1 to maximize $t \to L(\theta + te_1)$. The ith step inductively chooses t_i to maximize $t \to L(\theta + \sum_{j=1}^{i-1} t_j e_j + te_i)$. When all coordinates have been updated, one iteration is complete.

Suppose we let $f_i(\theta) = \theta_i + t_i$ denote the ith component of the algorithm map $f(\theta)$ for cyclic coordinate ascent. To compute the differential of $f(\theta)$ at a local maximum θ_∞, note that

$$0 = \frac{\partial}{\partial\theta_i}L([f_1(\theta), \ldots, f_i(\theta), \theta_{i+1}, \ldots, \theta_m]^t) \tag{13}$$

holds for each i. Taking the differential of equation (13) and invoking the identity $f(\theta_\infty) = \theta_\infty$, we infer that

$$\begin{aligned}
0 &= \sum_{k=1}^{i} \frac{\partial^2}{\partial\theta_k\partial\theta_i}L(\theta_\infty)\frac{\partial}{\partial\theta_j}f_k(\theta_\infty), \quad j \le i \\
0 &= \sum_{k=1}^{i} \frac{\partial^2}{\partial\theta_k\partial\theta_i}L(\theta_\infty)\frac{\partial}{\partial\theta_j}f_k(\theta_\infty) + \frac{\partial^2}{\partial\theta_j\partial\theta_i}L(\theta_\infty), \quad j > i. \qquad (14)
\end{aligned}$$

This linear system of equations for $df(\theta_\infty)$ can by solved by introducing the lower triangular part T of the symmetric matrix $d^2L(\theta_\infty)$. By definition T includes the diagonal D of $d^2L(\theta_\infty)$. With this notation, the system (14) can be restated as $Tdf(\theta_\infty) = D - T^t$ and explicitly solved in the form $df(\theta_\infty) = T^{-1}(D - T^t)$.

Local convergence of cyclic coordinate ascent hinges on whether the spectral radius ρ of the matrix $T^{-1}(T^t - D)$ satisfies $\rho < 1$. Suppose that λ is an eigenvalue of $T^{-1}(D - T^t)$ with eigenvector u. These can be complex. The equality $T^{-1}(D - T^t)u = \lambda u$ implies $(1 - \lambda)Tu = (T + T^t - D)u$. Premultiplying this by the conjugate transpose u^* gives

$$\frac{1}{1-\lambda} = \frac{u^*Tu}{u^*(T + T^t - D)u}.$$

Hence, the real part of $1/(1 - \lambda)$ satisfies

$$\begin{aligned} Re\Big(\frac{1}{1-\lambda}\Big) &= \frac{u^*(T + T^t)u}{2u^*(T + T^t - D)u} \\ &= \frac{1}{2}\left[1 + \frac{u^*Du}{u^*d^2L(\theta_\infty)u}\right] \\ &> \frac{1}{2} \end{aligned}$$

for $d^2L(\theta_\infty)$ negative definite. If $\lambda = \alpha + \beta i$, then the last inequality entails

$$\frac{1-\alpha}{(1-\alpha)^2 + \beta^2} > \frac{1}{2},$$

which is equivalent to $|\lambda|^2 = \alpha^2 + \beta^2 < 1$. Hence, the spectral radius $\rho < 1$.

13.4 Global Convergence

In studying global convergence, we must carefully specify the parameter domain U. Because local convergence analysis depends only on the properties of the loglikelihood $L(\theta)$ in a neighborhood of a local maximum θ_∞, we made the harmless assumption that $U = R^m$. Now we take U to be any open, convex subset of R^m. To avoid colliding with the boundary of U, we assume that $L(\theta)$ is coercive in that sense that the set $\{\theta \in U : L(\theta) \geq c\}$ is compact for every constant c. Coerciveness implies that $L(\theta)$ attains its maximum somewhere in U and that $L(\theta)$ tends to $-\infty$ as θ approaches the boundary of U or $||\theta||$ approaches ∞. It is also convenient to continue assuming when necessary that $L(\theta)$ and $Q(\theta \mid \theta_n)$ and their various first and second differentials are jointly continuous in θ and θ_n and that the expected information $J(\theta)$ is continuous in θ. Finally, we demand that the Hessian matrix $d^{20}Q(\theta \mid \theta_n)$ be negative definite. Among other things, this implies that the EM gradient algorithm iterates are well defined, except possibly for the question of whether they fall in U.

A major impediment to establishing global convergence of the various maximum likelihood algorithms is the possible failure of the ascent property $L(\theta_{n+1}) \geq L(\theta_n)$ enjoyed by the EM algorithm. Provided the matrix $A(\theta)$ is positive definite, the iteration map (4) is guaranteed to point locally uphill. Hence, if we elect the natural strategy of instituting a line search along the direction $A(\theta_n)^{-1}dL(\theta_n)^t$ emanating from θ_n, we can certainly find a θ_{n+1} that increases $L(\theta)$. It is tempting to choose θ_{n+1} to be the maximum point of $L(\theta)$ on this line. Without further restrictions on $L(\theta)$, this choice is ambiguous. For one thing, the maximum point may not be unique. Even if it is, it may be hard to distinguish it from other local maxima on the line. We can finesse this dilemma by simply assuming that the observed information $-d^2L(\theta)$ is positive definite. This implies that the loglikelihood $L(\theta)$ is strictly concave.

The strong assumption of strict concavity can be circumvented in the EM gradient algorithm by taking θ_{n+1} to the maximum of $Q(\theta \mid \theta_n)$ rather than of $L(\theta)$ along the search line. Recall that maximizing $Q(\theta \mid \theta_n)$ forces an increase in $L(\theta)$. Under the weaker assumption that $d^{20}Q(\theta \mid \theta_n)$ is negative definite, θ_{n+1} again exists and is unique. In cyclic coordinate ascent, $L(\theta)$ is maximized along each coordinate direction in turn. In this situation, we require that $-\frac{\partial^2}{\partial\theta_i^2}L(\theta) > 0$ hold for all i and θ.

Besides avoiding ambiguity in the choice of the next iterate θ_{n+1}, full or partial strict concavity assumptions on $L(\theta)$ or $Q(\theta \mid \theta_n)$ ensure that the iteration map $M(\theta_n) = \theta_{n+1}$ is continuous. While not absolutely necessary for establishing global convergence, continuity of $M(\theta)$ simplifies verification of convergence. Weaker assumptions than continuity are explored in [9].

Proposition 13.4.1. *Assume that the differentiability, coerciveness, and strict concavity assumptions posited above are in force. Combining a line search with Newton's method, scoring, or the EM gradient algorithm leads to a continuous iteration map $M(\theta)$ with the ascent property $L[M(\theta)] \geq L(\theta)$. Continuity and the ascent condition also hold for the iteration maps of cyclic coordinate ascent and the EM algorithm.*

Proof. By construction, all of the previously mentioned algorithms are ascent algorithms. To dispose of continuity, consider first the iteration map $M(\theta)$ of the EM algorithm. As noted earlier, the implicit function theorem implies that $M(\theta)$ is continuously differentiable when $d^{20}Q[M(\theta) \mid \theta]$ is invertible. Invertibility of this matrix follows immediately from its assumed negative definiteness.

Now consider the iteration map resulting from a line search combined with algorithm (4). The matrix $A(\theta)$ is continuous for both Newton's method and scoring by assumption. Because the observed information is positive definite by assumption and the expected information is the covariance of the score, $A(\theta)$ is also positive definite for both algorithms. Thus,

the search direction $d(\theta) = A(\theta)^{-1}dL(\theta)^t$ is well defined and continuous. The revised iteration map $M(\theta) = \theta + t(\theta)d(\theta)$ is determined by the scalar $t(\theta) \geq 0$ maximizing L along the line $t \to \theta + td(\theta)$.

To prove continuity of the iteration map $M(\theta)$ for Newton's method or scoring, one must show that $\lim_{n\to\infty} M(\phi_n) = M(\phi)$ for any sequence ϕ_n possessing limit ϕ. (The sequence ϕ_n is not generated by $\phi_{n+1} = M(\phi_n)$ in this context.) The coerciveness property of $L(\theta)$ and the ascent condition imply that all $M(\phi_n)$ belong to the same compact set. Let ψ be the limit of a subsequence $M(\phi_{n_k})$. If $d(\phi) = \mathbf{0}$, then ϕ is a stationary point of $L(\theta)$. Due to the strict concavity assumption, $L(\theta)$ has only one stationary point, and this point coincides with its maximum point. Hence,

$$\begin{aligned} L(\psi) &= \lim_{k\to\infty} L[M(\phi_{n_k})] \\ &\geq \lim_{k\to\infty} L(\phi_{n_k}) \\ &= L(\phi) \end{aligned}$$

implies $\psi = \phi = M(\phi)$ when $d(\phi) = \mathbf{0}$.

If $d(\phi) \neq \mathbf{0}$ and $\psi \neq \phi$, then

$$\begin{aligned} \frac{\psi - \phi}{||\psi - \phi||} &= \lim_{k\to\infty} \frac{M(\phi_{n_k}) - \phi_{n_k}}{||M(\phi_{n_k}) - \phi_{n_k}||} \\ &= \lim_{k\to\infty} \frac{d(\phi_{n_k})}{||d(\phi_{n_k})||} \\ &= \frac{d(\phi)}{||d(\phi)||}. \end{aligned}$$

Thus, $\psi = \phi + \frac{||\psi-\phi||}{||d(\phi)||}d(\phi)$, and taking limits in

$$\begin{aligned} L[M(\phi_{n_k})] &= L[\phi_{n_k} + t(\phi_{n_k})d(\phi_{n_k})] \\ &\geq L[\phi_{n_k} + sd(\phi_{n_k})] \end{aligned}$$

implies $L(\psi) \geq L[\phi + sd(\phi)]$ for all feasible $s \geq 0$. This again proves that $\psi = M(\phi)$.

The proof of continuity for the EM gradient algorithm follows in exactly the same manner with $Q[\theta + sd(\theta) \mid \theta]$ substituted for $L[\theta + sd(\theta)]$ throughout. For cyclic coordinate ascent, the above proof demonstrates continuity for the map $M_i(\theta)$ changing the ith coordinate. Because

$$M(\theta) = M_m \circ \cdots \circ M_1(\theta)$$

is a composition of continuous maps, it is continuous as well. □

The preceding proof uses the fact that the maximum point of $L(\theta)$ is a fixed point of $M(\theta)$. In general, any stationary point θ_∞ of $L(\theta)$ is a fixed point of $M(\theta)$. For the line search algorithms, this fact follows because the search direction $d(\theta_\infty) = \mathbf{0}$. For the EM algorithm, it follows from equality (7). Conversely, the strict concavity assumptions imply that any

fixed point of $M(\theta)$ is a stationary point of $L(\theta)$. Furthermore, we can make the stronger assertion that stationary points and only stationary points give equality in the ascent inequality $L(M[\theta]) \geq L(\theta)$.

With these facts in mind, we now state and prove a version of Liapunov's theorem for discrete dynamical systems [9].

Proposition 13.4.2 (Liapunov's Theorem**).** *Let Γ be the set of limit points generated by the sequence $\theta_{n+1} = M(\theta_n)$ starting from some initial θ_1. Then Γ is contained in the set S of stationary points of $L(\theta)$.*

Proof. The sequence θ_n occurs in the compact set

$$\{\theta \in U : L(\theta) \geq L(\theta_1)\}.$$

Consider a typical limit point $\phi = \lim_{k\to\infty} \theta_{n_k}$. Since the sequence $L(\theta_n)$ is monotone increasing and bounded above, $\lim_{n\to\infty} L(\theta_n)$ exists. Hence, taking limits in the inequality $L[M(\theta_{n_k})] \geq L(\theta_{n_k})$ and using the continuity of $M(\theta)$ and $L(\theta)$, we infer that $L[M(\phi)] = L(\phi)$. Thus, ϕ is a fixed point of $M(\theta)$ and consequently also a stationary point of $L(\theta)$. □

The next two propositions are adapted from [10].

Proposition 13.4.3. *The set of limit points Γ of $\theta_{n+1} = M(\theta_n)$ is compact and connected.*

Proof. Γ is a closed subset of the compact set $\{\theta \in U : L(\theta) \geq L(\theta_1)\}$ and is therefore itself compact. According to Proposition 8.2.1, Γ is connected provided $\lim_{n\to\infty} \|\theta_{n+1} - \theta_n\| = 0$. If this sufficient condition fails, then the compactness of $\{\theta \in U : L(\theta) \geq L(\theta_1)\}$ makes it possible to extract a subsequence θ_{n_k} such that $\lim_{k\to\infty} \theta_{n_k} = \phi$ and $\lim_{k\to\infty} \theta_{n_k+1} = \psi$ both exist, but $\psi \neq \phi$. However, the continuity of $M(\theta)$ entails $\psi = M(\phi)$ while the ascent condition implies $L(\psi) = L(\phi) = \lim_{n\to\infty} L(\theta_n)$. The equality $L(\psi) = L(\phi)$ forces the contradictory conclusion that ϕ is a fixed point of $M(\theta)$. Hence, the sufficient condition $\lim_{n\to\infty} \|\theta_{n+1} - \theta_n\| = 0$ for connectivity holds. □

Proposition 13.4.4. *Assume that the differentiability, coerciveness, and strict concavity assumptions are true and that all stationary points of $L(\theta)$ are isolated. Then any sequence of iterates $\theta_{n+1} = M(\theta_n)$ generated by the iteration map $M(\theta)$ of one of the algorithms discussed possesses a limit, and that limit is a stationary point of $L(\theta)$. If $L(\theta)$ is strictly concave, then $\lim_{n\to\infty} \theta_n$ is the maximum likelihood point.*

Proof. In the compact set $\{\theta \in U : L(\theta) \geq L(\theta_1)\}$ there can only be a finite number of stationary points. Since the set of limit points Γ is a connected subset of this finite set of stationary points, Γ reduces to a single point. □

Two comments on Proposition 13.4.4 are in order. First, except when full strict concavity prevails, the proposition offers no guarantee that the

limit θ_∞ of the sequence θ_n furnishes a global maximum. Problems 4 and 7 contain counterexamples of de Leeuw [1] and Wu [13] exhibiting convergence to a saddle point in cyclic coordinate ascent and the EM algorithm. Fortunately, in practice, ascent algorithms almost always converge to at least a local maximum of the loglikelihood. Second, if the set S of stationary points is not discrete, then there exists a sequence $\phi_n \in S$ converging to $\phi_\infty \in S$. Because the surface of the unit sphere in R^m is compact, we can extract a subsequence such that

$$\lim_{k\to\infty} \frac{\phi_{n_k} - \phi_\infty}{||\phi_{n_k} - \phi_\infty||} = \omega$$

exists and is nontrivial. Taking limits in

$$\begin{aligned}\mathbf{0} &= \frac{1}{||\phi_{n_k} - \phi_\infty||}[dL(\phi_{n_k}) - dL(\phi_\infty)] \\ &= \int_0^1 d^2L[\phi_\infty + t(\phi_{n_k} - \phi_\infty)]\frac{\phi_{n_k} - \phi_\infty}{||\phi_{n_k} - \phi_\infty||}dt\end{aligned}$$

then produces $\mathbf{0} = d^2L(\phi_\infty)\omega$. In other words, the observed information at ϕ_∞ is singular. If one can rule out such degeneracies, then all stationary points are isolated. Interested readers can consult the literature on Morse functions for further commentary on this subject [5].

13.5 Problems

1. Define $f : R^2 \to R$ by $f(\mathbf{0}) = 0$ and $f(\theta) = \theta_1^3/(\theta_1^2 + \theta_2^2)$ for $\theta \neq \mathbf{0}$. Show that $f(\theta)$ is differentiable along every straight line in R^2 but lacks a differential at $\mathbf{0}$.
2. For $f : R^2 \to R^2$ given by $f_1(\theta) = \theta_1^3$ and $f_2(\theta) = \theta_1^2$, show that no $\bar{\theta}$ exists on the line segment from $\mathbf{0} = (0,0)^t$ to $\mathbf{1} = (1,1)^t$ such that
$$f(\mathbf{1}) - f(\mathbf{0}) = df(\bar{\theta})(\mathbf{1} - \mathbf{0}).$$
3. Let $f : U \to R$ be a continuously differentiable function defined on an open, connected set $U \subset R^m$. Suppose that $df(\theta) = \mathbf{0}$ for all $\theta \in U$. Show that $f(\theta)$ is constant on U. (Hint: There is a polygonal path between any two points along which one can integrate $df(\theta)$.)
4. Demonstrate that cyclic coordinate ascent either diverges or converges to a saddle point of the function $f : R^2 \to R$ defined by
$$f(\theta, \phi) = 2\theta\phi - (\theta - \phi)^2.$$
This function of de Leeuw [1] has no maximum.
5. Consider a Poisson distributed random variable Y with mean $a\theta + b$, where a and b are known positive constants and $\theta \geq 0$ is a parameter to be estimated. An EM algorithm for estimating θ can be concocted

that takes as complete data independent Poisson random variables U and V with means $a\theta$ and b and sum $U+V=Y$. If $Y=y$ is observed, then show that the EM iterates are defined by

$$\theta_{n+1}=\frac{y\theta_n}{a\theta_n+b}.$$

Show that these iterates converge monotonely to the maximum likelihood estimate $\max\{0,(y-b)/a\}$. When $y=b$, verify that convergence to the boundary value 0 occurs at a rate slower than linear [3]. (Hint: When $y=b$ check that $\theta_{n+1}=b\theta_1/(na\theta_1+b)$.)

6. The sublinear convergence of the EM algorithm exhibited in the previous problem occurs in other problems. Here is a conceptually harder example by Robert Jennrich. Suppose that $W_1,\ldots,W_m$, and B are independent, normally distributed random variables with 0 means. Let σ_w^2 be the common variance of the W's and σ_b^2 be the variance of B. If the values y_i of the linear combinations $Y_i=B+W_i$ are observed, then show that the EM algorithm amounts to

$$\sigma_{n+1,b}^2=\left(\frac{m\sigma_{nb}^2\bar{y}}{m\sigma_{nb}^2+\sigma_{nw}^2}\right)^2+\frac{\sigma_{nb}^2\sigma_{nw}^2}{m\sigma_{nb}^2+\sigma_{nw}^2}$$
$$\sigma_{n+1,w}^2=\frac{(m-1)}{m}s_y^2+\left(\frac{\sigma_{nw}^2\bar{y}}{m\sigma_{nb}^2+\sigma_{nw}^2}\right)^2+\frac{\sigma_{nb}^2\sigma_{nw}^2}{m\sigma_{nb}^2+\sigma_{nw}^2},$$

where $\bar{y}=\frac{1}{m}\sum_{i=1}^m y_i$ and $s_y^2=\frac{1}{m-1}\sum_{i=1}^m(y_i-\bar{y})^2$ are the sample mean and variance. Although one can formally calculate the maximum likelihood estimates $\hat{\sigma}_w^2=s_y^2$ and $\hat{\sigma}_b^2=\bar{y}^2-s_y^2/m$, these are only valid provided $\hat{\sigma}_b^2\geq 0$. If for instance $\bar{y}=0$, then the EM iterates will converge to $\sigma_w^2=(m-1)s_y^2/m$ and $\sigma_b^2=0$. Show that convergence is sublinear when $\bar{y}=0$.

7. Suppose the data displayed in Table 13.1 constitute a random sample from a bivariate normal distribution with both means 0, variances σ_1^2 and σ_2^2, and correlation coefficient ρ. The asterisks indicate missing values. Specify the EM algorithm for estimating σ_1^2, σ_2^2, and ρ. Show that the observed loglikelihood has global maxima at $\rho=\pm\frac{1}{2}$ and $\sigma_1^2=\sigma_2^2=\frac{8}{3}$ and a saddle point at $\rho=0$ and $\sigma_1^2=\sigma_2^2=\frac{5}{2}$. If the EM algorithm starts with $\rho=0$, prove that convergence to the saddle point occurs [13].

TABLE 13.1. Bivariate Normal Data for the EM Algorithm

Obs	Obs	Obs	Obs	Obs	Obs
(1,1)	(1,-1)	(-1,1)	(-1,-1)	(2,∗)	(2,∗)
(-2,∗)	(-2,∗)	(∗,2)	(∗,2)	(∗,-2)	(∗,-2)

8. Suppose the EM gradient iterates θ_n converge to a local maximum θ_∞ of the loglikelihood $L(\theta)$. Under the hypotheses of the text, prove that for all sufficiently large n, either $\theta_n = \theta_\infty$ or $L(\theta_{n+1}) > L(\theta_n)$ [7]. (Hints: Show that

$$\begin{aligned} &L(\theta_{n+1}) - L(\theta_n) \\ &\quad = \frac{1}{2}(\theta_{n+1} - \theta_n)^t [d^2 L(\phi_n) - 2d^{20} Q(\theta_n \mid \theta_n)](\theta_{n+1} - \theta_n), \end{aligned}$$

where ϕ_n lies on the line segment between θ_n and θ_{n+1}. Then use a continuity argument, noting that $d^2 L(\theta_\infty) - d^{20} Q(\theta_\infty \mid \theta_\infty)$ is nonnegative definite and $d^{20} Q(\theta_\infty \mid \theta_\infty)$ is negative definite.)

9. Let $M(\theta)$ be the EM algorithm or EM gradient algorithm map. Consider the modified algorithm $M_t(\theta) = \theta + t[M(\theta) - \theta]$ for $t > 0$. At a local maximum θ_∞, show that the spectral radius ρ_t of the differential $dM_t(\theta_\infty) = (1-t)I + tdM(\theta_\infty)$ satisfies $\rho_t < 1$ when $0 < t < 2$. Hence, Ostrowski's theorem implies local attraction of $M_t(\theta)$ to θ_∞. If the largest and smallest eigenvalues of $dM(\theta_\infty)$ are $\omega_{\max}$ and $\omega_{\min}$, then prove that ρ_t is minimized by taking $t = [1 - (\omega_{\min} + \omega_{\max})/2]^{-1}$. In practice, the eigenvalues of $dM(\theta_\infty)$ are impossible to predict in advance of knowing θ_∞, but for problems with a high proportion of missing data, the value $t = 2$ often works well [7]. (Hint: To every eigenvalue ω of $dM(\theta_\infty)$, there corresponds an eigenvalue $\omega_t = 1 - t + t\omega$ of $dM_t(\theta_\infty)$ and vice versa.)

10. In the notation of Chapters 10 and 12, prove the EM algorithm formula

$$d^2 L(\theta) = d^{20} Q(\theta \mid \theta) + \operatorname{Var}[d \ln f(X \mid \theta) \mid Y, \theta]$$

of Louis [8].

11. In our exposition of least L_1 regression in Chapter 12, we considered a modified iteration scheme that minimizes the criterion

$$\sum_{i=1}^m \{|y_i - \mu_i(\theta)| - \epsilon \ln[\epsilon + |y_i - \mu(\theta)|]\}. \tag{15}$$

For a sequence of constants ϵ_n tending to 0, let θ_n be a corresponding sequence minimizing (15). If θ_∞ is a limit point of this sequence and the regression functions $\mu_i(\theta)$ are continuous, then show that θ_∞ minimizes $\sum_{i=1}^m |y_i - \mu_i(\theta)|$. If, in addition, the minimum point θ_∞ of $\sum_{i=1}^m |y_i - \mu_i(\theta)|$ is unique and $\lim_{||\theta|| \to \infty} \sum_{i=1}^m |\mu_i(\theta)| = \infty$, then prove that $\lim_{n \to \infty} \theta_n = \theta_\infty$. (Hints: For the first assertion, take limits in

$$\sum_{i=1}^m h_\epsilon[s_i(\theta_n)] \le \sum_{i=1}^m h_\epsilon[s_i(\theta)],$$

where $s_i(\theta) = y_i - \mu_i(\theta)$ and $h_\epsilon(s) = |s| - \epsilon \ln(\epsilon + |s|)$. Note that $h_\epsilon(s)$ is jointly continuous in ϵ and s. For the second assertion, it suffices

that the sequence θ_n be confined to a bounded set. To prove this fact, demonstrate and use the inequalities

$$\begin{aligned}\sum_{i=1}^m h_\epsilon(s_i) &\geq \frac{1}{2}\sum_{i=1}^m 1_{\{|s_i|\geq 1\}}|s_i| \\ &\geq \frac{1}{2}\sum_{i=1}^m |\mu_i(\theta)| - \frac{1}{2}\sum_{i=1}^m |y_i| - \frac{m}{2} \\ \sum_{i=1}^m h_\epsilon(s_i) &\leq \sum_{i=1}^m [|s_i| - \epsilon \ln \epsilon] \\ &\leq \sum_{i=1}^m |s_i| + \frac{m}{e}\end{aligned}$$

for $0 \leq \epsilon < \frac{1}{2}$.)

12. Newton's method for finding the reciprocal of a number can be generalized to find the inverse of a matrix [6]. Consider the iteration scheme

$$B_{n+1} = 2B_n - B_n A B_n$$

for some fixed $m \times m$ matrix A. Prove that

$$A^{-1} - B_{n+1} = (A^{-1} - B_n)A(A^{-1} - B_n)$$

and therefore that

$$||A^{-1} - B_{n+1}|| \leq ||A|| \cdot ||A^{-1} - B_n||^2$$

for every matrix norm. Conclude that the sequence B_n converges at a quadratic rate to A^{-1} if B_1 is sufficiently close to A^{-1}.

13. Continuing Problem 12, consider iterating according to

$$B_{n+1} = B_n \sum_{i=0}^j (I - AB_n)^i \tag{16}$$

to find A^{-1} [6]. Problem 12 is the special case $j = 1$. Verify the alternative representation

$$B_{n+1} = \sum_{i=0}^j (I - B_n A)^i B_n,$$

and use it to prove that B_{n+1} is symmetric whenever A and B_n are. Also show by induction that

$$A^{-1} - B_{n+1} = (A^{-1} - B_n)[A(A^{-1} - B_n)]^j.$$

From this last identity deduce the norm inequality

$$\|A^{-1} - B_{n+1}\| \leq \|A\|^j \|A^{-1} - B_n\|^{j+1}.$$

Thus, the algorithm converges at a cubic rate when $j = 2$, at a quartic rate when $j = 3$, and so forth.

14. Problems 12 and 13 provide another method of accelerating the EM gradient algorithm. Denote the loglikelihood of the observed data by $L(\theta)$ and the result of the E step by $Q(\theta \mid \theta_n)$. To accelerate the EM gradient algorithm, we can replace the positive definite matrix $B(\theta)^{-1} = -d^{20}Q(\theta \mid \theta)$ by a matrix that better approximates the observed information $A(\theta) = -d^2L(\theta)$. Note that often $d^{20}Q(\theta \mid \theta)$ is diagonal and therefore trivial to invert. Now consider the formal expansion

$$\begin{aligned} A^{-1} &= (B^{-1} + A - B^{-1})^{-1} \\ &= \{B^{-\frac{1}{2}}[I - B^{\frac{1}{2}}(B^{-1} - A)B^{\frac{1}{2}}]B^{-\frac{1}{2}}\}^{-1} \\ &= B^{\frac{1}{2}} \sum_{i=0}^{\infty} [B^{\frac{1}{2}}(B^{-1} - A)B^{\frac{1}{2}}]^i B^{\frac{1}{2}}. \end{aligned}$$

If we truncate this series after a finite number of terms, then we recover the first iterate of (16) in the disguised form

$$S_j = B^{\frac{1}{2}} \sum_{i=0}^{j} [B^{\frac{1}{2}}(B^{-1} - A)B^{\frac{1}{2}}]^i B^{\frac{1}{2}}.$$

The accelerated algorithm

$$\theta_{n+1} = \theta_n + S_j(\theta_n) dL(\theta_n)^t \tag{17}$$

has several desirable properties.

(a) Show that S_j is positive definite and hence that (17) is an ascent algorithm. (Hint: Use the fact that $B^{-1} - A$ is nonnegative definite.)

(b) Algorithm (17) has differential

$$I + S_j(\theta_\infty) d^2L(\theta_\infty) = I - S_j(\theta_\infty) A(\theta_\infty)$$

at a local maximum θ_∞. If $d^2L(\theta_\infty)$ is negative definite, then prove that all eigenvalues of this differential lie on $[0, 1)$. (Hint: The eigenvalues are determined by the stationary points of the Rayleigh quotient $v^t[A^{-1}(\theta_\infty) - S_j(\theta_\infty)]v / v^t A^{-1}(\theta_\infty)v$.)

(c) If ρ_j is the spectral radius of the differential, then demonstrate that $\rho_j \leq \rho_{j-1}$, with strict inequality when $B^{-1}(\theta_\infty) - A(\theta_\infty)$ is positive definite.

In other words, the accelerated algorithm (17) is guaranteed to converge faster than the EM gradient algorithm. It will be particularly useful for maximum likelihood problems with many parameters because it entails no matrix inversion or multiplication, just matrix times

vector multiplication. When $j = 1$, it takes the simple form

$$\theta_{n+1} = \theta_n + [2B(\theta_n) - B(\theta_n)A(\theta_n)B(\theta_n)]dL(\theta_n)^t.$$

15. In Problem 13 of Chapter 12, we considered maximizing functions of the form $L(\theta) = -\sum_{i=1}^p f_i(c_i^t\theta)$ with each f_i strictly convex. If we choose nonnegative constants λ_{ij} such that $\sum_j \lambda_{ij} = 1$ and $\lambda_{ij} > 0$ when $c_{ij} \neq 0$, then the function

$$Q(\theta \mid \theta_n) = -\sum_{i=1}^p \sum_{j \in S_i} \lambda_{ij} f_i\Big[\frac{c_{ij}}{\lambda_{ij}}(\theta_j - \theta_{nj}) + c_i^t\theta_n\Big]$$

serves as a surrogate function for an EM algorithm without missing data. Here the set $S_i = \{j : \lambda_{ij} > 0\}$. We suggested that a reasonable choice for λ_{ij} might be $\lambda_{ij} = |c_{ij}|^\alpha/||c_i||_\alpha^\alpha$, where $||c_i||_\alpha^\alpha = \sum_j |c_{ij}|^\alpha$ and $\alpha > 0$. It would be helpful to determine the α yielding the fastest rate of convergence. As pointed out in Proposition 13.3.2, the rate of convergence is given by the maximum of the Rayleigh quotient (10). This fact suggests that we should choose α to minimize $-v^t d^{20}Q(\theta \mid \theta)v$ over all unit vectors v. This appears to be a difficult problem. A simpler problem is to minimize $\mathrm{tr}[-d^{20}Q(\theta \mid \theta)]$. Show that this substitute problem has solution $\alpha = 1$ regardless of the point θ selected. (Hint: Multiply the inequality

$$\Big(\sum_{j \in S_i} |c_{ij}|\Big)^2 \leq \sum_{j \in S_i} \frac{c_{ij}^2}{\lambda_{ij}}$$

by $f_i''(c_i^t\theta)$ and sum on i.)

References

[1] de Leeuw J (1994) Block relaxation algorithms in statistics. *Information Systems and Data Analysis,* edited by Bock HH, Lenski W, Richter MM, Springer-Verlag, Berlin, pp 308–325

[2] Dennis JE Jr, Schnabel RB (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs, NJ

[3] Fessler JA, Clinthorne NH, Rogers WL (1993) On complete-data spaces for PET reconstruction algorithms. *IEEE Trans Nuclear Sci* 40:1055–1061

[4] Gill PE, Murray W, Wright MH (1981) *Practical Optimization.* Academic Press, New York

[5] Guillemin V, Pollack A (1974) *Differential Topology.* Prentice-Hall, Englewood Cliffs, NJ

[6] Householder AS (1975) *The Theory of Matrices in Numerical Analysis.* Dover, New York

[7] Lange K (1995) A gradient algorithm locally equivalent to the EM algorithm. *J Roy Stat Soc B* 57:425–437

[8] Louis TA (1982) Finding the observed information matrix when using the EM algorithm. *J Roy Stat Soc B* 44:226–233
[9] Luenberger DG (1984) *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, MA
[10] Meyer RR (1976) Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J Computer System Sci* 12:108–121
[11] Nocedal J (1991) Theory of algorithms for unconstrained optimization. *Acta Numerica 1991*: 199–242
[12] Ortega JM (1990) *Numerical Analysis: A Second Course*. Society for Industrial and Applied Mathematics, Philadelphia
[13] Wu CF (1983) On the convergence properties of the EM algorithm. *Ann Stat* 11:95–103

14
Constrained Optimization

14.1 Introduction

Many optimization problems in statistics involve constraints. For instance, in testing nested hypotheses by the likelihood ratio method, maximum likelihood estimation must be carried out with equality constraints. Inequality constraints also commonly occur. Variance components and Poisson intensities are nonnegative; probabilities must be confined to the unit interval. In designing algorithms for constrained optimization, we first need to sharpen our theoretical understanding. The natural place to start (and for us to end) is with linear equality and inequality constraints. Fortunately, many nonlinear constraints can be transformed to linear constraints by an appropriate change of parameters.

In this chapter, we bow to convention and discuss minimization rather than maximization. This choice should cause readers little trouble; if we want to maximize $f(\theta)$, then we minimize $-f(\theta)$. Most nonlinear programming algorithms revolve around the minimization of positive definite quadratic functions subject to linear equality constraints. As we shall see, this class of problems can be solved explicitly by the methods of linear algebra. Once we pass to nonquadratic functions or to linear inequality constraints, things become more complicated.

At first glance, dealing with nonquadratic functions may appear to be more of a barrier to algorithm design than dealing with linear inequality constraints. Painful experience has taught numerical analysts otherwise. Suppose that we want to minimize the nonquadratic objective function

$f(\theta)$ subject to the linear equality constraints $v_i^t\theta = d_i$ for some finite set of column vectors v_i. We proceed just as in unconstrained optimization and expand $f(\theta)$ in a second-order Taylor series

$$f(\theta) \approx f(\theta_k) + df(\theta_k)(\theta - \theta_k) + \frac{1}{2}(\theta - \theta_k)^t d^2 f(\theta_k)(\theta - \theta_k)$$

about the current approximation θ_k to the constrained minimum. We then minimize the quadratic approximation to $f(\theta)$ subject to the linear equality constraints to get the next iterate θ_{k+1}.

Quite apart from the addition of linear inequality constraints, several complications can occur. First, in most problems it is expensive to compute the Hessian $d^2 f(\theta_k)$. The remedy is to substitute a positive definite approximation A_k to $d^2 f(\theta_k)$. For instance, in some statistical applications, the expected information is straightforward to calculate and can be substituted for the observed information. Alternatively, we can start with some convenient value of A_0, say the identity matrix, and update A_k by the quasi-Newton methods described in Chapter 11. With quasi-Newton updates, the quadratic approximation to $f(\theta)$ should improve dramatically as θ_k approaches the minimum point θ_∞. If we are fortunate, the θ_k will then exhibit the fast convergence of Newton's method. Another complication is that θ_{k+1} may actually increase $f(\theta)$. In this circumstance, we view $\delta_k = \theta_{k+1} - \theta_k$ as a direction along which to search $f(\theta)$. By forcing A_k to be positive definite, the function $s \mapsto f(\theta_k + s\delta_k)$ for $s \in [0, 1]$ will decrease in a neighborhood of $s = 0$. Some variant of step halving is usually adequate to ensure a decrease in $f(\theta)$.

To handle linear inequality constraints, one can replace $f(\theta)$ by a surrogate function $R(\theta \mid \theta_k)$ at each iteration. This seemingly odd proposal combines features of the EM algorithm and the logarithmic barrier method of optimization [6, 7]. If we choose $R(\theta \mid \theta_k)$ just right, then the minimum θ_{k+1} of $R(\theta \mid \theta_k)$ will simultaneously satisfy the descent property $f(\theta_{k+1}) < f(\theta_k)$ and the linear inequality constraints. The resulting algorithm is by no stretch of the imagination the fastest available, but it is among the easiest to program and understand. It also ties in well with our brief exposition of Lagrange multipliers. Readers interested in pursuing other methods of optimization and tactics for handling nonlinear constraints are urged to consult the references [1, 2, 3, 4, 5, 8, 9]. The chapter ends with a brief explanation of how standard errors are computed in the presence of linear equality constraints.

14.2 Necessary and Sufficient Conditions for a Minimum

Before we derive necessary and sufficient conditions for a point to be the global minimum of an objective function $f(\theta)$, we need to attend to certain

details about convex sets. In the current context, it is easy to spot several convex sets. For instance, the feasible region

$$\{\theta : u_i^t\theta \geq c_i,\ i \in S_{ineq},\ v_j^t\theta = d_j,\ j \in S_{eq}\}$$

is convex. If the function $f(\theta)$ is convex, then the set $\{\theta : f(\theta) \leq c\}$ is also convex for any constant c.

Proposition 14.2.1 (Projection Theorem)**.** *Suppose C is a nonempty, closed, convex subset of R^n. Given a point $\theta \in R^n$, there is a unique, closest point $P(\theta) \in C$ to θ. Furthermore, we have $[P(\theta) - \theta]^t[\phi - P(\theta)] \geq 0$ for every $\phi \in C$, and $\|P(\theta) - P(\omega)\|_2 \leq \|\theta - \omega\|_2$ for every θ and ω.*

Proof. Obviously, $P(\theta) = \theta$ for $\theta \in C$. For $\theta \notin C$, we will show that $P(\theta)$ exists and is unique by employing the easily verified parallelogram law

$$\|\alpha + \beta\|_2^2 + \|\alpha - \beta\|_2^2 = 2\Big(\|\alpha\|_2^2 + \|\beta\|_2^2\Big).$$

If $d = \inf\{\|\theta - \phi\|_2 : \phi \in C\}$, then there exists a sequence $\phi_k \in C$ such that $\lim_{k\to\infty} \|\theta - \phi_k\|_2 = d$. By virtue of the parallelogram law,

$$\|\phi_j - \phi_k\|_2^2 + \|\phi_j + \phi_k - 2\theta\|_2^2 = 2\Big(\|\phi_j - \theta\|_2^2 + \|\theta - \phi_k\|_2^2\Big),$$

or

$$\frac{1}{4}\|\phi_j - \phi_k\|_2^2 = \frac{1}{2}\Big(\|\theta - \phi_j\|_2^2 + \|\theta - \phi_k\|_2^2\Big) - \|\theta - \frac{1}{2}\Big(\phi_j + \phi_k\Big)\|_2^2.$$

Because C is convex, $\frac{1}{2}(\phi_j + \phi_k) \in C$, and it follows from the definition of d that

$$\frac{1}{4}\|\phi_j - \phi_k\|_2^2 \leq \frac{1}{2}\Big(\|\theta - \phi_j\|_2^2 + \|\theta - \phi_k\|_2^2\Big) - d^2. \tag{1}$$

Letting j and k tend to ∞ confirms that ϕ_m is a Cauchy sequence whose limit ϕ_∞ must lie in the closed set C. In view of the continuity of the norm, we have $\|\theta - \phi_\infty\|_2 = d$. To prove uniqueness, suppose there is a second point $\omega_\infty \in C$ satisfying $\|\theta - \omega_\infty\|_2 = d$. Then substituting ϕ_∞ for ϕ_j and ω_∞ for ϕ_k in inequality (1) shows that $\omega_\infty = \phi_\infty$.

To prove the inequality $[P(\theta) - \theta]^t[\phi - P(\theta)] \geq 0$ for $\phi \in C$, note that the definition of $P(\theta)$ and the convexity of C imply

$$\begin{aligned}\|P(\theta) - \theta\|_2^2 &\leq \|(1-s)P(\theta) + s\phi - \theta\|_2^2 \\ &= \|P(\theta) - \theta\|_2^2 + 2s[P(\theta) - \theta]^t[\phi - P(\theta)] + s^2\|\phi - P(\theta)\|_2^2\end{aligned}$$

for any $s \in (0, 1]$. Canceling $\|P(\theta) - \theta\|_2^2$ from both sides and dividing by s yields

$$0 \leq 2[P(\theta) - \theta]^t[\phi - P(\theta)] + s\|\phi - P(\theta)\|_2^2.$$

Sending s to 0 now gives the result.

Finally, to prove the inequality $\|P(\theta) - P(\omega)\|_2 \leq \|\theta - \omega\|_2$, we add the inequalities

$$[P(\theta) - \theta]^t[P(\omega) - P(\theta)] \geq 0$$
$$[P(\omega) - \omega]^t[P(\theta) - P(\omega)] \geq 0$$

and rearrange. This gives

$$\begin{aligned}[P(\theta) - P(\omega)]^t[P(\theta) - P(\omega)] &\leq [\theta - \omega]^t[P(\theta) - P(\omega)] \\ &\leq \|\theta - \omega\|_2\|P(\theta) - P(\omega)\|_2\end{aligned}$$

by virtue of the Cauchy–Schwarz inequality. Dividing by $\|P(\theta) - P(\omega)\|_2$ now completes the proof. □

For most closed, convex sets C, it is impossible to give an explicit formula for the projection operator P. A notable exception is projection onto the range of a matrix A of full column rank. This puts us back in the familiar terrain of least squares estimation, where $P(\theta) = A(A^tA)^{-1}A^t\theta$. One can easily check that the projection matrix $P_A = A(A^tA)^{-1}A^t$ satisfies $P_A^t = P_A$ and $P_A^2 = P_A$. The idempotent property $P(P(\theta)) = P(\theta)$ obviously extends to projection onto any closed, convex set. Projection onto a Cartesian product $\prod_{i=1}^n [a_i, b_i]$ yields the simple formula

$$P(\theta)_i = \begin{cases} a_i & \theta_i < a_i \\ \theta_i & \theta_i \in [a_i, b_i] \\ b_i & \theta_i > b_i \end{cases}$$

for the ith coordinate of $P(\theta)$. This formula holds even when $a_i = -\infty$, $a_i = b_i$, or $b_i = \infty$. Problem 1 gives two other examples where $P(\theta)$ can be calculated explicitly.

Our next proposition is technical preparation for Proposition 14.2.3, the main theoretical result of this section.

Proposition 14.2.2 (Farkas)**.** *Let $v_1, \ldots, v_m$ be nontrivial column vectors in R^n. Then a necessary and sufficient condition that $b^t\theta \geq 0$ for all θ in the set $S = \{\theta \in R^n : v_i^t\theta \geq 0,\ i = 1, \ldots, m\}$ is that $b = \sum_{i=1}^m c_i v_i$ for nonnegative scalars $c_1, \ldots, c_m$.*

Proof. The condition is clearly sufficient. To demonstrate its necessity, consider the set $C = \{\sum_{i=1}^m c_i v_i : c_i \geq 0,\ i = 1, \ldots, m\}$. Because it contains the origin, C is nonempty. We assert that C is a closed, convex cone. Now C is a convex cone if and only if $au + bv \in C$ whenever $a, b \geq 0$ and $u, w \in C$. This property obviously follows from the definition of C. Proving that C is closed is more subtle. If the vectors $v_1, \ldots, v_m$ are independent, then we reason as follows: Suppose the sequence $u_k = \sum_{i=1}^m c_{ki} v_i \in C$ converges to u_∞. The coefficients c_{ki} are the unique coordinates of u_k in the finite-dimensional subspace spanned by the v_i. This subspace is closed, and if a sequence of points converges in it, then the coordinates of the points must

converge as well. Because $\lim_{k\to\infty} c_{ki} = c_{\infty i}$ exists and is nonnegative for each i, the point $u_\infty = \sum_{i=1}^m c_{\infty i} v_i \in C$.

We now prove that C is closed by induction on m. The case $m = 1$ is true because a single vector v_1 is linearly independent. Assume that the claim holds for $m-1$ vectors. If the vectors $v_1, \ldots, v_m$ are linearly independent, then we are done. If the vectors $v_1, \ldots, v_m$ are linearly dependent, then there exist scalars $a_1, \ldots, a_m$, not all 0, such that $\sum_{i=1}^m a_i v_i = \mathbf{0}$. Without loss of generality, we can assume that $a_i < 0$ for at least one index i. Now express $u \in C$ as

$$u = \sum_{i=1}^m c_i v_i = \sum_{i=1}^m (c_i + r a_i) v_i$$

for an arbitrary scalar r. If we increase r gradually from 0, then there is a first value at which $c_j + r a_j = 0$ for some index j. This shows that C can be decomposed as the union

$$C = \bigcup_{j=1}^m \Big\{ \sum_{i \neq j} d_i v_i : d_i \geq 0,\ i \neq j \Big\}.$$

Each of the sets $\{\sum_{i\neq j} d_i v_i : d_i \geq 0,\ i \neq j\}$ is closed by the induction hypothesis. Since a finite union of closed sets is closed, C itself is closed.

Next suppose $b^t\theta \geq 0$ for all $\theta \in S$, but $b \notin C$. By Proposition 14.2.1, there is a unique closest point $P(b) \in C$ satisfying $[P(b) - b]^t[u - P(b)] \geq 0$ for every $u \in C$. Hence,

$$\begin{aligned} u^t[P(b) - b] &\geq P(b)^t[P(b) - b] \\ &> b^t[P(b) - b]. \end{aligned}$$

It is now evident that if we put $w = P(b) - b$ and choose a scalar s such that

$$P(b)^t w > s > b^t w,$$

then $u^t w > s > b^t w$ holds for all $u \in C$. In particular, $\mathbf{0} \in C$ gives $0 = \mathbf{0}^t w > s$. Furthermore, because $ru \in C$ whenever $u \in C$ and $r > 0$, the inequality $u^t w > s/r$ implies $u^t w \geq 0$ for all $u \in C$. In view of the fact that every $v_i \in C$, we finally conclude that $w \in S$. This conclusion forces a contradiction between the hypothesis $b^t w \geq 0$ and the inequality $0 > s > b^t w$. Thus, our assumption that $b \notin C$ must be false. □

Proposition 14.2.3 (Karush–Kuhn–Tucker)**.** *Suppose $f(\theta)$ is a continuously differentiable function on the set*

$$S = \{\theta \in R^n : g_i(\theta) \geq 0,\ i = 1, \ldots, m\},$$

where $g_i(\theta) = u_i^t\theta - c_i$. *If* $f(\theta)$ *has a local minimum at* $\phi \in S$, *then the differential* $df(\phi)$ *satisfies the Lagrange multiplier condition*

$$df(\phi) = \sum_{i=1}^{m} \lambda_i dg_i(\phi) = \sum_{i=1}^{m} \lambda_i u_i^t,$$

where each $\lambda_i \geq 0$ *and* $\sum_{i=1}^m \lambda_i g_i(\phi) = 0$. *Conversely, if* $f(\theta)$ *is twice continuously differentiable, convex, and satisfies a Lagrange multiplier condition at* $\phi \in S$, *then* ϕ *is a global minimum of* $f(\theta)$. *If* $f(\theta)$ *is strictly convex, then* ϕ *is the unique minimum of* $f(\theta)$.

Proof. Let ϕ and θ be two points of S. It clear geometrically that the direction $\delta = \theta - \phi$ leads into S from ϕ. Such a direction is said to be tangent to S at ϕ. Alternatively, we can characterize the tangent directions in terms of the active constraints $A(\phi) = \{i : g_i(\phi) = 0\}$ at ϕ. The expansion

$$g_i(\phi + r\delta) - g_i(\phi) = ru_i^t\delta$$

for r positive and $i \in A(\phi)$ makes it clear that δ is a tangent direction if and only if $u_i^t\delta \geq 0$ for all $i \in A(\phi)$. Note that the inactive constraints play no role in determining tangent directions because they do not obstruct local movement away from ϕ.

Now suppose ϕ is a local minimum, and consider the linear approximation

$$f(\phi + r\delta) - f(\phi) = rdf(\phi)\delta + o(r)$$

along a tangent direction δ. If $df(\phi)\delta < 0$, then all sufficiently small positive r entail $f(\phi + r\delta) < f(\phi)$, contradicting the fact that ϕ is a local minimum. Hence, $df(\phi)\delta \geq 0$ whenever δ is a tangent direction. The necessity of the Lagrange multiplier rule follows from Proposition 14.2.2 with $df(\phi)^t$ playing the role of the vector b.

To prove the sufficiency of the Lagrange multiplier rule for a convex function $f(\theta)$, we execute a second-order Taylor expansion

$$f(\phi + r\delta) - f(\phi) = rdf(\phi)\delta + \frac{r^2}{2}\delta^t d^2 f(\omega)\delta$$

around ϕ. Here ω lies on the line segment extending from ϕ to $\phi + r\delta$. In view of the convexity of $f(\theta)$, the quadratic form $\delta^t d^2 f(\omega)\delta \geq 0$. The Lagrange multiplier condition ensures that $df(\phi)\delta \geq 0$. Hence, $f(\phi + r\delta) \geq f(\phi)$. If $f(\theta)$ is strictly convex, then the quadratic form $\delta^t d^2 f(\omega)\delta > 0$ when $\delta \neq \mathbf{0}$. □

The Lagrange multiplier rule is often stated for equality constraints rather than inequality constraints. Since the equality constraint $g_i(\theta) = 0$ can be viewed as the pair of inequality constraints $g_i(\theta) \geq 0$ and $-g_i(\theta) \geq 0$, the Lagrange multiplier representation of $df(\phi)$ will involve separate contributions $\lambda_i^+ dg_i(\phi)$ and $-\lambda_i^- dg_i(\phi)$ for nonnegative multipliers λ_i^+ and λ_i^-. These contributions can be combined as $\lambda_i dg_i(\phi)$, with $\lambda_i = \lambda_i^+ - \lambda_i^-$ of indeterminate sign.

Example 14.2.1 (*Estimation of Multinomial Probabilities*). Consider a multinomial experiment with n trials and observed outcomes $n_1, \ldots, n_m$ over m categories. The maximum likelihood estimate of the probability p_i of category i is $\hat{p}_i = n_i/n$. To demonstrate this fact, let

$$L(p) = \binom{n}{n_1 \cdots n_m} \prod_{i=1}^{m} p_i^{n_i}$$

denote the likelihood. If $n_i = 0$ for some i, then we interpret $p_i^{n_i}$ as 1 even when $p_i = 0$. This convention makes it clear that we can increase $L(p)$ by replacing p_i by 0 and p_j by $p_j/(1-p_i)$ for $j \neq i$. Thus, for purposes of maximum likelihood estimation, we can assume that all $n_i > 0$. Given this assumption, $L(p)$ tends to 0 when any p_i tends to 0. It follows that we can further restrict our attention to the interior region where all $p_i > 0$ and maximize the loglikelihood $\ln L(p)$ subject to the equality constraint $\sum_{i=1}^{m} p_i = 1$. To find the minimum of $-\ln L(p)$, we look for a stationary point (not a minimum) of the Lagrangian

$$\mathcal{L}(p, \lambda) = -\ln \binom{n}{n_1 \cdots n_m} - \sum_{i=1}^{m} n_i \ln p_i - \lambda\Big(\sum_{i=1}^{m} p_i - 1\Big).$$

Setting the partial derivative of $\mathcal{L}(p, \lambda)$ with respect to p_i equal to 0 gives the equation

$$-\frac{n_i}{p_i} = \lambda.$$

These m equations are satisfied subject to the constraint by taking $\lambda = -n$ and $p_i = \hat{p}_i = n_i/n$. Thus, the necessary condition of Proposition 14.2.3 holds at $\hat{p}$. The sufficient condition for a minimum also holds because the entries

$$-\frac{\partial^2}{\partial p_i \partial p_j} \ln L(p) = \begin{cases} \frac{n_i}{p_i^2} & i = j \\ 0 & i \neq j \end{cases}.$$

of the observed information $-d^2 \ln L(p)$ show that $-\ln L(p)$ is strictly convex. ■

Example 14.2.2 (*A Counterexample to Sufficiency*). Convexity is crucial in demonstrating that the Lagrange multiplier condition is sufficient for a point to furnish a minimum. For example, consider the function $f(\theta) = \theta_1^3 - \theta_2$ subject to the constraint $g(\theta) = -\theta_2 \geq 0$. The Lagrange multiplier condition

$$df(\mathbf{0}) = (0, -1) = dg(\mathbf{0})$$

holds, but the origin $\mathbf{0}$ fails to minimize $f(\theta)$. Indeed, the one-dimensional slice $\theta_1 \mapsto f(\theta_1, 0)$ has a saddle point at $\theta_1 = 0$. ■

14.3 Quadratic Programming with Equality Constraints

Minimizing a quadratic function

$$q(\theta) = b^t\theta + \frac{1}{2}\theta^t A\theta$$

on R^n subject to the m linear equality constraints

$$v_i^t\theta = d_i, \quad i \in S_{eq}$$

is the computational engine of nonlinear programming. Here the symmetric matrix A is assumed positive definite. The constraints can be reexpressed as $V\theta = d$ by defining V to be the $m \times n$ matrix with ith row v_i^t and d to be the column vector with ith entry d_i.

To minimize $q(\theta)$ subject to the constraints, we introduce the Lagrangian

$$\begin{aligned}\mathcal{L}(\theta, \lambda) &= b^t\theta + \frac{1}{2}\theta^t A\theta + \sum_{i=1}^m \lambda_i[v_i^t\theta - d_i] \\ &= b^t\theta + \frac{1}{2}\theta^t A\theta + \lambda^t(V\theta - d).\end{aligned}$$

A stationary point of $\mathcal{L}(\theta, \lambda)$ is determined by the equations

$$\begin{aligned}b + A\theta + V^t\lambda &= \mathbf{0} \\ V\theta &= d,\end{aligned}$$

whose formal solution amounts to

$$\begin{pmatrix}\theta \\ \lambda\end{pmatrix} = \begin{pmatrix}A & V^t \\ V & \mathbf{0}\end{pmatrix}^{-1}\begin{pmatrix}-b \\ d\end{pmatrix}. \tag{2}$$

The next proposition shows that the indicated matrix inverse exists. Because the quadratic function $q(\theta)$ is strictly convex, the Lagrange multiplier conditions of Proposition 14.2.3 are sufficient to ensure that the calculated point provides the unique minimum.

Proposition 14.3.1. *Let A be an $n \times n$ positive definite matrix and V be an $m \times n$ matrix. Then the matrix*

$$M = \begin{pmatrix}A & V^t \\ V & \mathbf{0}\end{pmatrix}$$

has inverse

$$M^{-1} = \begin{pmatrix}A^{-1} - A^{-1}V^t(VA^{-1}V^t)^{-1}VA^{-1} & A^{-1}V^t(VA^{-1}V^t)^{-1} \\ (VA^{-1}V^t)^{-1}VA^{-1} & -(VA^{-1}V^t)^{-1}\end{pmatrix}$$

if and only if V has linearly independent rows $v_1^t, \ldots, v_m^t$.

Proof. We first show that M is invertible with the specified inverse if and only if $(VA^{-1}V^t)^{-1}$ exists. Suppose M^{-1} is given by $\begin{pmatrix} B & C^t \\ C & D \end{pmatrix}$. Then the identity

$$\begin{pmatrix} A & V^t \\ V & \mathbf{0} \end{pmatrix} \begin{pmatrix} B & C^t \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & \mathbf{0} \\ \mathbf{0} & I_m \end{pmatrix}$$

implies that $VC^t = I_m$ and $AC^t + V^tD = \mathbf{0}$. Multiplying the last equality by VA^{-1} gives $I_m = -VA^{-1}V^tD$. Thus, $(VA^{-1}V^t)^{-1}$ exists. Conversely, if $(VA^{-1}V^t)^{-1}$ exists, then one can check by direct multiplication that M has the claimed inverse.

If $(VA^{-1}V^t)^{-1}$ exists, then V must have full row rank m. Conversely, if V has full row rank m, take any nontrivial $u \in R^m$. Then the fact

$$u^tV = u_1v_1^t + \cdots u_mv_m^t \neq \mathbf{0}$$

and the positive definiteness of A imply $u^tVA^{-1}V^tu > 0$. Thus, $VA^{-1}V^t$ is also positive definite and invertible. □

It is noteworthy that the matrix M of Proposition 14.3.1 can be inverted by sweeping on its diagonal entries. Indeed, sweeping on the diagonal entries of A takes

$$\begin{pmatrix} A & V^t \\ V & \mathbf{0} \end{pmatrix} \longrightarrow \begin{pmatrix} -A^{-1} & A^{-1}V^t \\ VA^{-1} & -VA^{-1}V^t \end{pmatrix}.$$

Sweeping is now possible for the remaining diagonal entries of M since $VA^{-1}V^t$ is positive definite.

14.4 An Adaptive Barrier Method

We now consider the general problem of minimizing the twice continuously differentiable function $f(\theta)$ subject to the linear inequality constraints $g_i(\theta) = u_i^t\theta - c_i \geq 0$ for $1 \leq i \leq l$ and the linear equality constraints $v_j^t\theta = d_j$ for $1 \leq j \leq m$. Interior-point methods seek the minimum of $f(\theta)$ while remaining on the interior $U = \{\theta : g_i(\theta) > 0 \text{ for all } i,\ v_j^t\theta = d_j \text{ for all } j\}$ of the feasible region. Note that U is an open set of $\cap_j\{\theta : v_j^t\theta = d_j\}$ but not of the underlying space R^n as a whole unless $m = 0$. We assume that U is nonempty.

If θ_k is an interior point, then minimization of $f(\theta)$ can be transferred to the surrogate function

$$R(\theta \mid \theta_k) = f(\theta) - \mu \sum_{i=1}^{l} [g_i(\theta_k) \ln g_i(\theta) - u_i^t\theta] \tag{3}$$

for some constant $\mu > 0$. If we minimize $R(\theta \mid \theta_k)$ subject to the linear equality constraints, then the logarithms in the barrier function

$$f(\theta) - R(\theta \mid \theta_k) = \mu \sum_{i=1}^{l} [g_i(\theta_k) \ln g_i(\theta) - u_i^t \theta]$$

force the solution θ_{k+1} to remain within the interior of the feasible region.

Straightforward differentiation shows that

$$df(\theta) - d^{10}R(\theta \mid \theta_k) = \mu \sum_{i=1}^{l} \Big[\frac{g_i(\theta_k)}{g_i(\theta)} - 1\Big] u_i^t$$

$$d^2 f(\theta) - d^{20}R(\theta \mid \theta_k) = -\mu \sum_{i=1}^{l} \frac{g_i(\theta_k)}{g_i(\theta)^2} u_i u_i^t.$$

Thus, the barrier function $f(\theta) - R(\theta \mid \theta_k)$ is concave and attains it maximum at $\theta = \theta_k$. The now familiar EM reasoning

$$\begin{aligned} f(\theta_{k+1}) &= R(\theta_{k+1} \mid \theta_k) + f(\theta_{k+1}) - R(\theta_{k+1} \mid \theta_k) \\ &\le R(\theta_k \mid \theta_k) + f(\theta_k) - R(\theta_k \mid \theta_k) \\ &= f(\theta_k) \end{aligned}$$

shows that minimizing $R(\theta \mid \theta_k)$ drives $f(\theta)$ downhill. In other words, we have again concocted a kind of EM algorithm without missing data.

One worry is that the logarithmic barriers will prevent the θ_k from approaching a boundary when the minimum point θ_∞ lies on the boundary. However, the barrier function is adaptive in the sense that the coefficients $g_i(\theta_k)$ of the barrier terms $\ln g_i(\theta)$ change from one iteration to the next. Thus, when $g_i(\theta_\infty) = 0$, the coefficient $g_i(\theta_k)$ can tend to 0 and permit θ_k to tend to θ_∞. In fact, one can rigorously prove that $\lim_{k\to\infty} \theta_k = \theta_\infty$ whenever $f(\theta)$ is convex and possesses a unique minimum. Reference [7] contains a proof for convex programming problems in what is called standard form.

Example 14.4.1 (*Application to Multinomial Probabilities*)**.** To apply the adaptive barrier method to the multinomial example of Example 14.2.1, consider the Lagrangian

$$\mathcal{L}(p, \lambda) = -\sum_{i=1}^{m} n_i \ln p_i - \mu \sum_{i=1}^{m} p_{ki} \ln p_i - \lambda\Big(\sum_{i=1}^{m} p_i - 1\Big)$$

based on $R(p \mid p_k)$. Setting the ith partial derivative of $\mathcal{L}(p, \lambda)$ equal to 0 and multiplying the result by p_i give the equation

$$-n_i - \mu p_{ki} - \lambda p_i = 0, \tag{4}$$

which in turn can be summed on i and solved for λ. Substituting the result $\lambda = -(n + \mu)$ in equation (4) yields

$$p_{k+1,i} = \frac{n_i + \mu p_{ki}}{n + \mu}.$$

The algebraic reduction

$$\begin{aligned} p_{k+1,i} - \frac{n_i}{n} &= \frac{n_i + \mu p_{ki}}{n+\mu} - \frac{n_i}{n} \\ &= \frac{\mu}{n+\mu}\Big(p_{ki} - \frac{n_i}{n}\Big) \end{aligned}$$

shows that p_{ki} approaches n_i/n at the linear rate $\mu/(n+\mu)$. ■

In most problems, it is impossible to minimize $R(\theta \mid \theta_k)$ explicitly. If this is the case, and if θ_k is an interior point of the feasible region, then we minimize the quadratic approximation

$$R(\theta_k + \delta_k \mid \theta_k) \approx R(\theta_k \mid \theta_k) + d^{10}R(\theta_k \mid \theta_k)\delta_k + \frac{1}{2}\delta_k^t d^{20}R(\theta_k \mid \theta_k)\delta_k$$

with respect to the increment δ_k and subject to the modified linear equality constraints $v_j^t\delta_k = 0$. When only an approximation A_k to $d^2f(\theta_k)$ is available, we also substitute A_k for $d^2f(\theta_k)$ in the Hessian $d^{20}R(\theta_k \mid \theta_k)$. Because $d^{10}R(\theta_k \mid \theta_k) = df(\theta_k)$ and δ_k is a descent direction for the quadratic approximation, δ_k is a descent direction for $f(\theta)$ as well. Problem 8 sketches a proof of this fact.

One side effect of only approximating the minimum of $R(\theta \mid \theta_k)$ is the unfortunate possibility of violating one or more of the linear inequality constraints. We can rectify this defect by choosing a small positive constant ϵ and defining $\theta_{k+1} = \theta_k + s\delta_k$, where s is the largest number in $[0,1]$ consistent with $g_i(\theta_{k+1}) \geq \epsilon g_i(\theta_k)$ for all i. In situations where $g_i(\theta_k)$ is slightly positive and $g_i(\theta_k + \delta_k)$ is slightly negative, this tactic can slow convergence. It is better to redefine δ_k by subtracting off its projection onto u_i. The slightly modified direction

$$\tilde{\delta}_k = \delta_k - \frac{u_i^t\delta_k}{u_i^t u_i}u_i$$

is still apt to be a descent direction. However, it now parallels rather than crosses the boundary $g_i(\theta) = 0$.

14.5 Standard Errors

To calculate the asymptotic covariance matrix of an estimated parameter vector $\hat{\theta}$ subject to linear equality constraints $V\theta = d$, we reparameterize. Suppose that the $m \times n$ matrix V has full row rank $m < n$ and that α is a particular solution of $V\theta = d$. The Gram–Schmidt process allows us to construct an $n \times (n-m)$ matrix W with $n-m$ linearly independent columns $w_1, \ldots, w_{n-m}$ orthogonal to the rows $v_1^t, \ldots, v_m^t$ of V. Now consider the reparameterization $\theta = \alpha + W\beta$. By virtue of our choice of α and the identity $VW = \mathbf{0}$, it is clear that $V(\alpha + W\beta) = d$. Since the range of W

and the null space of V both have dimension $n-m$, it is also clear that all solutions of $V\theta = d$ are generated as image vectors $\alpha + W\beta$.

Under the preceding reparameterization, the loglikelihood $L(\alpha+W\beta)$ has score $dL(\alpha+W\beta)W$, observed information $-W^t d^2L(\alpha+W\beta)W$, and expected information $-W^t\,\mathrm{E}[d^2L(\alpha+W\beta)]W$. If we let $\mathcal{I}(\theta)$ represent either the observed information $-d^2L(\theta)$ or the expected information $\mathrm{E}[-d^2L(\theta)]$ of the original parameters, then an asymptotic covariance matrix of the estimated parameter vector $\hat{\beta}$ is $[W^t\mathcal{I}(\alpha+W\hat{\beta})W]^{-1}$. The brief calculation

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(\alpha + W\hat{\beta}) \\ &= \mathrm{Var}(W\hat{\beta}) \\ &= W\,\mathrm{Var}(\hat{\beta})W^t\end{aligned}$$

shows that $\hat{\theta}$ has asymptotic covariance matrix $W[W^t\mathcal{I}(\hat{\theta})W]^{-1}W^t$, which unfortunately appears to depend on the particular reparameterization chosen. This is an illusion, because if we replace W by WT, where T is any invertible matrix, then

$$\begin{aligned}WT[T^tW^t\mathcal{I}(\hat{\theta})WT]^{-1}T^tW^t &= WTT^{-1}[W^t\mathcal{I}(\hat{\theta})W]^{-1}(T^t)^{-1}T^tW^t \\ &= W[W^t\mathcal{I}(\hat{\theta})W]^{-1}W^t.\end{aligned}$$

Problems 10 and 11 sketch Silvey [10] and Jennrich's method of computing $W[W^t\mathcal{I}(\hat{\theta})W]^{-1}W^t$. This method relies on the sweep operator rather than Gram–Schmidt orthogonalization.

In calculating the asymptotic covariance matrix of $\hat{\theta}$ in the presence of linear inequality constraints, the traditional procedure is to ignore the inactive constraints and to append the active constraints to the existing linear equality constraints. This creates a larger constraint matrix V and a corresponding smaller matrix W orthogonal to V.

14.6 Problems

1. In the context of Proposition 14.2.1, find the projection operator $P(\theta)$ onto the closed, convex set $C \subset R^n$ for a half-space $C = \{\phi : v^t\phi \geq d\}$ and a sphere $C = \{\phi : \|\phi - \omega\|_2 \leq r\}$.

2. In Proposition 14.2.1, show that validity of the inequality

$$[P(\theta) - \theta]^t[\phi - P(\theta)] \geq 0$$

for every $\phi \in C$ is a sufficient as well as necessary condition for $P(\theta)$ to be the closest point to θ.

3. Establish the inequality

$$\Big(\prod_{i=1}^n a_i\Big)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{i=1}^n a_i$$

between the geometric and arithmetic mean of n positive numbers. Verify that equality holds if and only if all $a_i = a$. (Hints: It suffices to minimize $\frac{1}{n}\sum_{i=1}^n a_i$ subject to the constraint $\prod_{i=1}^n a_i = 1$. Replace a_i by e^{θ_i} to linearize the constraint.)

4. For $p > 1$, show that the function $f(\theta) = \sum_{i=1}^n \theta_i^p$ has its minimum at $\theta = (n^{-1}, \ldots, n^{-1})^t$ subject to the constraints $\sum_{i=1}^n \theta_i = 1$ and $\theta_i \geq 0$ for all i.

5. Prove that the minimum of the function $f(\theta) = \theta_2 + (\theta_1 + 1)^3/3$ subject to the constraints $\theta_1 \geq 1$ and $\theta_2 \geq 0$ occurs at $\theta = (1, 0)^t$. What is the Lagrange multiplier condition?

6. Find the triangle of greatest area with fixed perimeter p. (Hint: Recall that a triangle with sides a, b, and c has area $\sqrt{s(s-a)(s-b)(s-c)}$, where $s = (a+b+c)/2 = p/2$.)

7. Use the techniques of this chapter to show that the minimum of the quadratic form $\theta^t A\theta$ subject to $\|\theta\|_2^2 = 1$ coincides with the smallest eigenvalue of the symmetric matrix A. The minimum point furnishes the corresponding eigenvector. Note that you will have to use the Lagrange multiplier rule for nonlinear constraints.

8. Section 14.4 asserts that the increment δ_k giving the minimum of the quadratic approximation to the surrogate function $R(\theta \mid \theta_k)$ is a descent direction for the objective function $f(\theta)$. Prove this fact. (Hint: Show that the upper-left block of the matrix M^{-1} in Proposition 14.3.1 is nonnegative definite by applying the Cauchy–Schwarz inequality.)

9. Based on the theory of Section 14.5, show that the asymptotic covariance matrix of the maximum likelihood estimates in Example 14.2.1 has entries

$$\widehat{\mathrm{Cov}}(\hat{p}_i, \hat{p}_j) = \begin{cases} \frac{1}{n}\hat{p}_i(1-\hat{p}_i) & i = j \\ -\frac{1}{n}\hat{p}_i\hat{p}_j & i \neq j \end{cases}.$$

For n large, these are close to the true values

$$\mathrm{Cov}(\hat{p}_i, \hat{p}_j) = \begin{cases} \frac{1}{n}p_i(1-p_i) & i = j \\ -\frac{1}{n}p_ip_j & i \neq j \end{cases}.$$

(Hints: You will need to use the Sherman–Morrison formula. For the sake of simplicity, assume that all $n_i > 0$.)

10. In the notation of Section 14.5, show that the asymptotic covariance matrix $W[W^t\mathcal{I}(\hat{\theta})W]^{-1}W^t$ appears as the upper left block of the matrix inverse

$$\begin{pmatrix} \mathcal{I} & V^t \\ V & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} W[W^t\mathcal{I}W]^{-1}W^t & \mathcal{I}^{-1}V^t[V\mathcal{I}^{-1}V^t]^{-1} \\ [V\mathcal{I}^{-1}V^t]^{-1}V\mathcal{I}^{-1} & -[V\mathcal{I}^{-1}V^t]^{-1} \end{pmatrix}.$$

(Hint: Show that the matrices

$$P_1 = \mathcal{I}W[W^t\mathcal{I}W]^{-1}W^t$$
$$P_2 = V^t[V\mathcal{I}^{-1}V^t]^{-1}V\mathcal{I}^{-1}$$

satisfy $P_1^2 = P_1$, $P_2^2 = P_2$, $P_1P_2 = P_2P_1 = \mathbf{0}$, and $P_1 + P_2 = I_n$.)

11. Continuing Problem 10, it may be impossible to invert the matrix

$$M = \begin{pmatrix} \mathcal{I} & V^t \\ V & \mathbf{0} \end{pmatrix}$$

by sweeping on its diagonal entries. However, suppose that $\theta^t \mathcal{I} \theta > 0$ whenever $\theta \neq \mathbf{0}$ and $V\theta = \mathbf{0}$. Then the matrix

$$M(s) = \begin{pmatrix} \mathcal{I} + sV^tV & V^t \\ V & \mathbf{0} \end{pmatrix}$$

for sufficiently large $s > 0$ serves as a substitute for M. Show that (a) if $M(s)$ is invertible for one s, then it is invertible for all s, (b) the upper left block of $M(s)^{-1}$ is $W[W^t\mathcal{I}W]^{-1}W^t$, and (c) $M(s)$ can be inverted by sweeping on its diagonal entries if s is sufficiently large. (Hint: Write

$$M(s) = \begin{pmatrix} I_n & sV^t \\ \mathbf{0} & I_m \end{pmatrix} \begin{pmatrix} \mathcal{I} & V^t \\ V & \mathbf{0} \end{pmatrix}$$

and invert. Part (c) is a direct consequence of Theorem 6.1 of [5].)

References

[1] Beltrami EJ (1970) *An Algorithmic Approach to Nonlinear Analysis and Optimization.* Academic Press, New York

[2] Ciarlet PG (1989) *Introduction to Numerical Linear Algebra and Optimization.* Cambridge University Press, Cambridge

[3] Fang S-C, Puthenpura S (1993) *Linear Optimization and Extensions: Theory and Algorithms.* Prentice-Hall, Englewood Cliffs, NJ

[4] Gill PE, Murray W, Wright MH (1991) *Numerical Linear Algebra and Optimization, Vol 1.* Addison-Wesley, Reading, MA

[5] Hestenes MR (1981) *Optimization Theory: The Finite Dimensional Case.* Robert E Krieger Publishing, Huntington, NY

[6] Iusem AN, Teboulle M (1995) Convergence rate analysis of nonquadratic proximal methods for convex and linear programming. *Math Operations Res* 20:657–677

[7] Lange K (1994) An adaptive barrier method for convex programming. *Methods Applications Analysis* 1:392–402

[8] Luenberger DG (1984) *Linear and Nonlinear Programming,* 2nd ed. Addison-Wesley, Reading, MA

[9] Peressini AL, Sullivan FE, Uhl JJ Jr (1988) *The Mathematics of Nonlinear Programming.* Springer-Verlag, New York

[10] Silvey SD (1975) *Statistical Inference.* Chapman & Hall, London

15
Concrete Hilbert Spaces

15.1 Introduction

In this chapter we consider an infinite-dimensional generalization of Euclidean space introduced by the mathematician Hilbert. This generalization preserves two fundamental geometric notions of Euclidean space—namely, distance and perpendicularity. Both of these geometric properties depend on the existence of an inner product. In the infinite-dimensional case, however, we take the inner product of functions rather than of vectors. Our emphasis here will be on concrete examples of Hilbert spaces relevant to statistics. To keep our discussion within bounds, the principal theoretical facts are stated without proof. Relevant proofs can be found in almost any book on real or functional analysis [2, 5]. Applications of our examples to numerical integration, wavelets, and other topics appear in later chapters.

15.2 Definitions and Basic Properties

An inner product space H is a vector space over the real or complex numbers equipped with an inner product $\langle f, g \rangle$ on pairs of vectors f and g from H. If the underlying field is the real numbers, then $\langle f, g \rangle$ is always real. If the field is the complex numbers, then, in general, $\langle f, g \rangle$ is complex. An inner product satisfies the following postulates:

(a) $\langle f, g \rangle$ is linear in f for g fixed,
(b) $\langle f, g \rangle = \langle g, f \rangle^*$, where * denotes complex conjugate,

(c) $\langle f, f\rangle \geq 0$, with equality if and only $f = \mathbf{0}$.

The inner product allows one to define a vector norm $\|f\| = \langle f, f\rangle^{1/2}$ on H, just as in linear algebra. Furthermore, the Cauchy–Schwarz inequality immediately generalizes. This says that any two vectors f and g in H satisfy

$$|\langle f, g\rangle| \leq \|f\| \cdot \|g\|,$$

with equality only when f and g are linearly dependent. An inner product space is said to be complete if every Cauchy sequence converges. In other words, if for some sequence $\{f_n\}_{n=1}^{\infty}$ the norm $\|f_m - f_n\|$ can be made arbitrarily small by taking m and n large enough, then the limit of f_n exists as $n \to \infty$. A complete inner product space is called a Hilbert space.

Example 15.2.1 (*Finite-Dimensional Vector Spaces*). The Euclidean space R^m of m-dimensional vectors with real components is a Hilbert space over the real numbers with the usual inner product $\langle f, g\rangle = \sum_{i=1}^{m} f_i g_i$. If we consider the space $\mathcal{C}^m$ of m-dimensional vectors with complex components, then we get a Hilbert space over the complex numbers with inner product $\langle f, g\rangle = \sum_{i=1}^{m} f_i g_i^*$. Note that the first of these spaces is embedded in the second space in a way that preserves inner products and distances. Since the other Hilbert spaces met in this chapter also exist in compatible real and complex versions, we ordinarily omit specifying the number field. ■

Example 15.2.2 (*Space of Square-Integrable Functions*). This is the canonical example of a Hilbert space. Let μ be a measure on some Euclidean space R^m. The vector space $\mathrm{L}^2(\mu)$ of real (or complex) square-integrable functions with respect to μ is a Hilbert space over the real (or complex) numbers with inner product

$$\langle f, g\rangle = \int f(x) g(x)^* d\mu(x).$$

If μ is the uniform measure on an interval (a, b), then we denote the corresponding space of square-integrable functions by $\mathrm{L}^2(a, b)$.

It is a fairly deep fact that the set of continuous functions with compact support is dense in $\mathrm{L}^2(\mu)$ [2]. This means that every square-integrable function f can be approximated to within an arbitrarily small $\epsilon > 0$ by a continuous function g vanishing outside some bounded interval; in symbols, $\|f - g\| < \epsilon$. On the real line the step functions with compact support also constitute a dense set of $\mathrm{L}^2(\mu)$. Both of these dense sets contain countable, dense subsets. In general, a Hilbert space H with a countable, dense set is called separable. Most concrete Hilbert spaces possess this property, so we append it as an additional postulate. ■

A finite or infinite sequence $\{\psi_n\}_{n\geq 1}$ of nonzero vectors in a Hilbert space H is said to be orthogonal if $\langle \psi_m, \psi_n\rangle = 0$ for $m \neq n$. An orthogonal sequence $\{\psi_n\}_{n\geq 1}$ is orthonormal if $\|\psi_n\| = 1$ for every n. Given a

function $f \in H$, one can compute its Fourier coefficients $\langle f, \psi_n \rangle$ relative to an orthonormal sequence $\{\psi_n\}_{n \geq 1}$. The finite expansion $\sum_{n=1}^{m} \langle f, \psi_n \rangle \psi_n$ provides the best approximation to f in the sense that

$$\begin{aligned} \|f - \sum_{n=1}^{m} \langle f, \psi_n \rangle \psi_n\|^2 &= \|f\|^2 - \sum_{n=1}^{m} |\langle f, \psi_n \rangle|^2 \\ &\leq \|f - \sum_{n=1}^{m} c_n \psi_n\|^2 \end{aligned} \tag{1}$$

for any other finite sequence of coefficients $\{c_n\}_{n=1}^{m}$. Inequality (1) incidentally entails Bessel's inequality

$$\sum_{n=1}^{m} |\langle f, \psi_n \rangle|^2 \leq \|f\|^2. \tag{2}$$

An orthonormal sequence $\{\psi_n\}_{n \geq 1}$ is said to be complete (or constitute a basis for H) if

$$f = \sum_{n \geq 1} \langle f, \psi_n \rangle \psi_n$$

for every $f \in H$. (This usage of the word "complete" conflicts with the topological notion of completeness involving Cauchy sequences.) The next proposition summarizes and extends our discussion thus far.

Proposition 15.2.1. *The following statements about an orthonormal sequence $\{\psi_n\}_{n \geq 1}$ are equivalent:*

(a) *The sequence is a basis for H.*
(b) *For each $f \in H$ and $\epsilon > 0$, there is a corresponding $m(f, \epsilon)$ such that*

$$\|f - \sum_{n=1}^{m} \langle f, \psi_n \rangle \psi_n\| \leq \epsilon$$

for all $m \geq m(f, \epsilon)$.
(c) *If a vector $f \in H$ satisfies $\langle f, \psi_n \rangle = 0$ for every n, then $f = \mathbf{0}$.*
(d) *For every $f \in H$,*

$$f = \sum_{n \geq 1} \langle f, \psi_n \rangle \psi_n.$$

(e) *For every $f \in H$,*

$$\|f\|^2 = \sum_{n \geq 1} |\langle f, \psi_n \rangle|^2. \tag{3}$$

Proof. This basic characterization is proved in standard mathematical texts such as [2, 5]. □

15.3 Fourier Series

The complex exponentials $\{e^{2\pi inx}\}_{n=-\infty}^{\infty}$ provide an orthonormal basis for the space of square-integrable functions with respect to the uniform distribution on [0,1]. Indeed, the calculation

$$\begin{aligned}\int_0^1 e^{2\pi imx}e^{-2\pi inx}dx &= \begin{cases} 1 & m=n \\ \frac{e^{2\pi i(m-n)x}}{2\pi i(m-n)}\Big|_0^1 & m\neq n \end{cases} \\ &= \begin{cases} 1 & m=n \\ 0 & m\neq n \end{cases}\end{aligned}$$

shows that the sequence is orthonormal. Completeness is essentially a consequence of Fejér's theorem [1], which says that any periodic, continuous function can be uniformly approximated by a linear combination of sines and cosines. (Fejér's theorem is a special case of the more general Stone-Weierstrass theorem [2].) In dealing with a square-integrable function $f(x)$ on [0,1], it is convenient to extend it periodically to the whole real line via the equation $f(x+1)=f(x)$. (We consider only functions with period 1 in this chapter.) The Fourier coefficients of $f(x)$ are computed according to the standard recipe

$$c_n = \int_0^1 f(x)e^{-2\pi inx}dx.$$

The Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{2\pi inx}$ is guaranteed to converge to $f(x)$ in mean square. The more delicate issue of pointwise convergence is partially covered by the next proposition.

Proposition 15.3.1. *Assume that the square-integrable function $f(x)$ on [0,1] is continuous at x_0 and possesses both one-sided derivatives there. Then*

$$\lim_{m\to\infty}\sum_{n=-m}^{m} c_n e^{2\pi inx_0} = f(x_0).$$

Proof. Extend $f(x)$ to be periodic, and consider the associated periodic function

$$g(x) = \frac{f(x+x_0)-f(x_0)}{e^{-2\pi ix}-1}.$$

Applying l'Hôpital's rule yields

$$\lim_{x\to 0^+} g(x) = \frac{\frac{d}{dx}f(x_0^+)}{-2\pi i},$$

where $\frac{d}{dx}f(x_0^+)$ denotes the one-sided derivative from the right. A similar expression holds for the limit from the left. Since these two limits are finite and $\int_0^1 |f(x)|^2dx<\infty$, we have $\int_0^1 |g(x)|^2dx<\infty$ as well.

Now let d_n be the nth Fourier coefficient of $g(x)$. Because

$$f(x+x_0) = f(x_0) + (e^{-2\pi ix} - 1)g(x),$$

it follows that

$$\begin{aligned} c_n e^{2\pi i n x_0} &= \int_0^1 f(x) e^{-2\pi i n(x-x_0)} dx \\ &= \int_0^1 f(x+x_0) e^{-2\pi i n x} dx \\ &= f(x_0) 1_{\{n=0\}} + d_{n+1} - d_n. \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{n=-m}^{m} c_n e^{2\pi i n x_0} &= f(x_0) + \sum_{n=-m}^{m} (d_{n+1} - d_n) \\ &= f(x_0) + d_{m+1} - d_{-m}. \end{aligned}$$

To complete the proof, observe that

$$\lim_{|m|\to\infty} d_m = \lim_{|m|\to\infty} \int_0^1 g(x) e^{-2\pi i m x} dx = 0$$

by the Riemann–Lebesgue lemma to be proved in Proposition 17.4.1 of Chapter 17. □

Example 15.3.1 (*Bernoulli Functions*)**.** There are Bernoulli polynomials $B_n(x)$ and periodic Bernoulli functions $b_n(x)$. Let us start with the Bernoulli polynomials. These are defined by the three conditions

$$\begin{aligned} B_0(x) &= 1 \\ \frac{d}{dx} B_n(x) &= n B_{n-1}(x), \qquad n > 0 \\ \int_0^1 B_n(x) dx &= 0, \qquad n > 0. \end{aligned} \tag{4}$$

For example, we calculate recursively

$$\begin{aligned} B_1(x) &= x - \frac{1}{2} \\ B_2(x) &= 2\Big(\frac{x^2}{2} - \frac{x}{2} + \frac{1}{12}\Big). \end{aligned}$$

The Bernoulli function $b_n(x)$ coincides with $B_n(x)$ on $[0,1)$. Outside $[0,1)$, $b_n(x)$ is extended periodically. In particular, $b_0(x) = B_0(x) = 1$ for all x. Note that $b_1(x)$ is discontinuous at $x = 1$ while $b_2(x)$ is continuous. All subsequent $b_n(x)$ are continuous at $x = 1$ because

$$B_n(1) - B_n(0) = \int_0^1 \frac{d}{dx} B_n(x) dx$$

$$\begin{aligned} &= n \int_0^1 B_{n-1}(x)dx \\ &= 0 \end{aligned}$$

by assumption.

To compute the Fourier series expansion $\sum_k c_{nk} e^{2\pi ikx}$ of $b_n(x)$ for $n > 0$, note that $c_{n0} = \int_0^1 B_n(x)dx = 0$. For $k \neq 0$, we have

$$\begin{aligned} c_{nk} &= \int_0^1 b_n(x) e^{-2\pi ikx} dx \\ &= b_n(x) \frac{e^{-2\pi ikx}}{-2\pi ik}\Big|_0^1 + \frac{1}{2\pi ik} \int_0^1 \frac{d}{dx} b_n(x) e^{-2\pi ikx} dx \qquad (5) \\ &= b_n(x) \frac{e^{-2\pi ikx}}{-2\pi ik}\Big|_0^1 + \frac{n}{2\pi ik} \int_0^1 b_{n-1}(x) e^{-2\pi ikx} dx. \end{aligned}$$

From the integration-by-parts formula (5), we deduce that $b_1(x)$ has Fourier series expansion

$$-\frac{1}{2\pi i} \sum_{k \neq 0} \frac{e^{2\pi ikx}}{k}.$$

This series converges pointwise to $b_1(x)$ except at $x = 0$ and $x = 1$. For $n > 1$, the boundary terms in (5) vanish, and

$$c_{nk} = \frac{n c_{n-1,k}}{2\pi ik}. \qquad (6)$$

Formula (6) and Proposition 15.3.1 together imply that

$$b_n(x) = -\frac{n!}{(2\pi i)^n} \sum_{k \neq 0} \frac{e^{2\pi ikx}}{k^n} \qquad (7)$$

for all $n > 1$ and all x.

The constant term $B_n = B_n(0)$ is known as a Bernoulli number. One can compute B_{n-1} recursively by expanding $B_n(x)$ in a Taylor series around $x = 0$. In view of the defining properties (4),

$$\begin{aligned} B_n(x) &= \sum_{k=0}^n \frac{1}{k!} \frac{d^k}{dx^k} B_n(0) x^k \\ &= \sum_{k=0}^n \frac{1}{k!} n^{\underline{k}} B_{n-k} x^k, \end{aligned}$$

where

$$n^{\underline{k}} = n(n-1) \cdots (n-k+1)$$

denotes a falling power. The continuity and periodicity of $b_n(x)$ for $n \geq 2$ therefore imply that

$$B_n = B_n(1) = \sum_{k=0}^{n} \binom{n}{k} B_{n-k}.$$

Subtracting B_n from both sides of this equality gives the recurrence relation

$$0 = \sum_{k=1}^{n} \binom{n}{k} B_{n-k}$$

for computing B_{n-1} from $B_0, \ldots, B_{n-2}$. For instance, starting from $B_0 = 1$, we calculate $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$, and $B_4 = -1/30$. From the expansion (7), evidently $B_n = 0$ for all odd integers $n > 1$. ■

15.4 Orthogonal Polynomials

The subject of orthogonal polynomials has a distinguished history and many applications in physics and engineering [1, 3]. Although it is a little under-appreciated in statistics, subsequent chapters will illustrate that it is well worth learning. Our goal here is simply to provide some concrete examples of orthogonal polynomials. The next proposition is useful in checking that an orthonormal sequence of polynomials is complete. In applying it, note that condition (8) holds whenever the probability measure μ possesses a moment generating function. In particular, if μ is concentrated on a finite interval, then its moment generating function exists.

Proposition 15.4.1. *Let μ be a probability measure on the line R such that for some $\alpha > 0$*

$$\int e^{\alpha|x|} d\mu(x) < \infty. \tag{8}$$

Then the polynomials $1, x, x^2, \ldots$ generate an orthonormal sequence of polynomials $\{p_n(x)\}_{n \geq 0}$ that is complete in the Hilbert space $L^2(\mu)$ of square-integrable functions.

Proof. This is proved as Proposition 43.1 of [4]. □

We now discuss some concrete examples of orthogonal polynomial sequences. Because no universally accepted conventions exist for most of the classical sequences, we adopt conventions that appear best suited to the purposes of probability and statistics.

Example 15.4.1 (*Poisson–Charlier Polynomials*). Let μ be the Poisson probability measure with mean λ. This probability measure attributes mass $\mu(\{x\}) = e^{-\lambda}\lambda^x/x!$ to the nonnegative integer x. Consider the exponential

generating function

$$\begin{aligned} p(x,t) &= e^{-t}\left(1+\frac{t}{\lambda}\right)^x \\ &= \sum_{n=0}^{\infty} \frac{p_n^{(\lambda)}(x)}{n!} t^n, \end{aligned}$$

where t is a real parameter. Expanding e^{-t} and $(1+t/\lambda)^x$ in power series and equating coefficients of t^n give

$$\begin{aligned} p_n^{(\lambda)}(x) &= n! \sum_{k=0}^{n} \binom{x}{k} \lambda^{-k} \frac{(-1)^{n-k}}{(n-k)!} \\ &= \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \lambda^{-k} x^{\underline{k}}. \end{aligned}$$

The polynomial $p_n^{(\lambda)}(x)$ is the nth-degree Poisson–Charlier polynomial.

These polynomials form an orthonormal sequence if properly normalized. Indeed, on the one hand,

$$\begin{aligned} \int p(x,s)p(x,t)d\mu(x) &= \sum_{x=0}^{\infty} e^{-s}\Big(1+\frac{s}{\lambda}\Big)^x e^{-t}\Big(1+\frac{t}{\lambda}\Big)^x e^{-\lambda}\frac{\lambda^x}{x!} \\ &= e^{-(s+t+\lambda)} \sum_{x=0}^{\infty} \frac{1}{x!}\Big[\Big(1+\frac{s}{\lambda}\Big)\Big(1+\frac{t}{\lambda}\Big)\lambda\Big]^x \\ &= e^{-(s+t+\lambda)} e^{(1+\frac{s}{\lambda})(1+\frac{t}{\lambda})\lambda} \\ &= e^{\frac{st}{\lambda}} \\ &= \sum_{n=0}^{\infty} \frac{1}{\lambda^n n!} s^n t^n. \end{aligned}$$

On the other hand,

$$\int p(x,s)p(x,t)d\mu(x) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{s^m t^n}{m!n!} \int p_m^{(\lambda)}(x) p_n^{(\lambda)}(x) d\mu(x).$$

Equating coefficients of $s^m t^n$ in these two expressions shows that

$$\int p_m^{(\lambda)}(x) p_n^{(\lambda)}(x) d\mu(x) = \begin{cases} 0 & m \neq n \\ \frac{n!}{\lambda^n} & m = n. \end{cases}$$

Proposition 15.4.1 implies that the sequence $\{p_n^{(\lambda)}(x)\sqrt{\lambda^n/n!}\}_{n\geq 0}$ is a complete orthonormal sequence for the Poisson distribution. ■

Example 15.4.2 (*Hermite Polynomials*). If μ is the probability measure associated with the standard normal distribution, then

$$\mu(S) = \frac{1}{\sqrt{2\pi}} \int_S e^{-\frac{1}{2}x^2} dx$$

for any measurable set S. The Hermite polynomials have exponential generating function

$$\begin{aligned} p(x,t) &= e^{xt-\frac{1}{2}t^2} \\ &= \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n \end{aligned}$$

for t real. The fact that $H_n(x)$ is a polynomial of degree n follows from evaluating the nth partial derivative of $e^{xt-\frac{1}{2}t^2}$ with respect to t at $t = 0$. To prove that the Hermite polynomials yield an orthogonal sequence, equate coefficients of $s^m t^n$ in the formal expansion

$$\begin{aligned} \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} s^m t^n \int \frac{H_m(x)}{m!}\frac{H_n(x)}{n!} d\mu(x) &= \int p(x,s)p(x,t)d\mu(x) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{st} e^{-\frac{1}{2}(x-s-t)^2} dx \\ &= e^{st}. \end{aligned}$$

This gives

$$\int \frac{H_m(x)}{m!}\frac{H_n(x)}{n!} d\mu(x) = \begin{cases} 0 & m \neq n \\ \frac{1}{n!} & m = n, \end{cases}$$

and Proposition 15.4.1 implies that $\{H_n(x)/\sqrt{n!}\}_{n\geq 0}$ is a complete orthonormal sequence for the standard normal distribution.

An explicit expression for $H_n(x)$ can be derived by writing

$$\begin{aligned} e^{xt-\frac{1}{2}t^2} &= e^{xt} e^{-\frac{1}{2}t^2} \\ &= \Big(\sum_{i=0}^{\infty} \frac{x^i t^i}{i!}\Big)\Big(\sum_{j=0}^{\infty} \frac{(-1)^j t^{2j}}{2^j j!}\Big). \end{aligned}$$

This shows that

$$H_n(x) = \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \frac{n!(-1)^j x^{n-2j}}{2^j j!(n-2j)!}.$$

In practice, the recurrence relation for $H_n(x)$ given in Section 2.3.3 of Chapter 2 and repeated in equation (11) ahead is more useful. ■

Example 15.4.3 (*Laguerre Polynomials*). If we let μ be the probability measure associated with the gamma distribution with scale parameter 1 and shape parameter α, then

$$\mu(S) = \frac{1}{\Gamma(\alpha)} \int_S x^{\alpha-1} e^{-x} dx$$

for any measurable set $S \subset (0,\infty)$. The sequence of Laguerre polynomials $\{L_n^{(\alpha)}(x)\}_{n=0}^{\infty}$ has exponential generating function

$$\begin{aligned} p(x,t) &= \frac{1}{(1-t)^{\alpha}} e^{-\frac{tx}{1-t}} \\ &= \sum_{n=0}^{\infty} \frac{L_n^{(\alpha)}(x)}{n!} t^n \end{aligned}$$

for $t \in (-1,1)$. If we let $\alpha^{\overline{n}} = \alpha(\alpha+1)\cdots(\alpha+n-1)$ denote a rising power, then equating coefficients of $s^m t^n$ in

$$\begin{aligned} &\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} s^m t^n \int \frac{L_m^{(\alpha)}(x)}{m!}\frac{L_n^{(\alpha)}(x)}{n!} d\mu(x) \\ &\quad = \int p(x,s)p(x,t)d\mu(x) \\ &\quad = \frac{1}{(1-s)^{\alpha}(1-t)^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-\frac{sx}{1-s}-\frac{tx}{1-t}-x} dx \\ &\quad = \frac{1}{(1-s)^{\alpha}(1-t)^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-\frac{x(1-st)}{(1-s)(1-t)}} dx \\ &\quad = \frac{1}{(1-st)^{\alpha}} \\ &\quad = \sum_{n=0}^{\infty} \frac{\alpha^{\overline{n}}}{n!} s^n t^n \end{aligned}$$

shows that the sequence of polynomials

$$\left\{ \left[\frac{1}{n!\alpha^{\overline{n}}} \right]^{\frac{1}{2}} L_n^{(\alpha)}(x) \right\}_{n=0}^{\infty}$$

is orthonormal for the gamma distribution. In view of Proposition 15.4.1, it is also complete.

It is possible to find explicit expressions for the Laguerre polynomials by expanding

$$\begin{aligned} \frac{1}{(1-t)^{\alpha}} e^{-\frac{tx}{1-t}} &= \sum_{m=0}^{\infty} \frac{(-1)^m x^m}{m!} \frac{t^m}{(1-t)^{m+\alpha}} \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m x^m}{m!} \sum_{k=0}^{\infty} \binom{-m-\alpha}{k} (-1)^k t^{k+m} \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m x^m}{m!} \sum_{k=0}^{\infty} \frac{\Gamma(k+m+\alpha)}{\Gamma(m+\alpha)} \frac{t^{m+k}}{k!} \\ &= \sum_{m=0}^{\infty} (-1)^m x^m \sum_{n=m}^{\infty} \frac{\Gamma(n+\alpha)}{\Gamma(m+\alpha)} \binom{n}{m} \frac{t^n}{n!} \end{aligned}$$

$$= \sum_{n=0}^{\infty} \frac{t^n}{n!} \Gamma(n+\alpha) \sum_{m=0}^{n} \binom{n}{m} \frac{(-1)^m x^m}{\Gamma(m+\alpha)}.$$

Thus,

$$L_n^{(\alpha)}(x) = \Gamma(n+\alpha) \sum_{m=0}^{n} \binom{n}{m} \frac{(-1)^m x^m}{\Gamma(m+\alpha)}.$$

Again, this is not the most convenient form for computing. ■

Example 15.4.4 (*Beta Distribution Polynomials*). The measure μ associated with the beta distribution assigns probability

$$\mu(S) = \frac{1}{B(\alpha,\beta)} \int_S x^{\alpha-1}(1-x)^{\beta-1} dx$$

to any measurable set $S \subset (0,1)$, where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the usual normalizing constant. No miraculous generating function exists in this case, but if we abbreviate the beta density by

$$w(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

then the functions

$$\begin{aligned}\psi_n^{(\alpha,\beta)}(x) &= \frac{1}{w(x)} \frac{d^n}{dx^n} \Big[w(x) x^n (1-x)^n \Big] \\ &= x^{-\alpha+1}(1-x)^{-\beta+1} \frac{d^n}{dx^n} \Big[x^{n+\alpha-1}(1-x)^{n+\beta-1} \Big]\end{aligned}$$

will prove to be orthogonal polynomials.

We can demonstrate inductively that $\psi_n^{(\alpha,\beta)}(x)$ is a polynomial of degree n by noting first that $\psi_0^{(\alpha,\beta)}(x) = 1$ and second that

$$\psi_n^{(\alpha,\beta)}(x) = (n+\alpha-1)(1-x)\psi_{n-1}^{(\alpha,\beta+1)}(x) - (n+\beta-1)x\psi_{n-1}^{(\alpha+1,\beta)}(x). \quad (9)$$

Equality (9) follows from the definition of $\psi_n^{(\alpha,\beta)}(x)$ and the identity

$$\begin{aligned}\frac{d^n}{dx^n} \Big[x^{n+\alpha-1}(1-x)^{n+\beta-1} \Big] &= (n+\alpha-1) \frac{d^{n-1}}{dx^{n-1}} \Big[x^{n+\alpha-2}(1-x)^{n+\beta-1} \Big] \\ &\quad - (n+\beta-1) \frac{d^{n-1}}{dx^{n-1}} \Big[x^{n+\alpha-1}(1-x)^{n+\beta-2} \Big].\end{aligned}$$

To show that the polynomials $\psi_n^{(\alpha,\beta)}(x)$ are orthogonal, note that for $m \le n$ repeated integration by parts leads to

$$\begin{aligned}&\int_0^1 \psi_m^{(\alpha,\beta)}(x) \psi_n^{(\alpha,\beta)}(x) w(x) dx \\ &\quad = \int_0^1 \psi_m^{(\alpha,\beta)}(x) \frac{d^n}{dx^n} \Big[w(x) x^n (1-x)^n \Big] dx \\ &\quad = (-1)^n \int_0^1 \frac{d^n}{dx^n} \Big[\psi_m^{(\alpha,\beta)}(x) \Big] w(x) x^n (1-x)^n dx\end{aligned}$$

since all boundary contributions vanish. If $m < n$, then

$$\frac{d^n}{dx^n}\psi_m^{(\alpha,\beta)}(x) = 0,$$

and this proves orthogonality. When $m = n$,

$$\frac{d^n}{dx^n}\left[\psi_n^{(\alpha,\beta)}(x)\right] = n!c_n^{(\alpha,\beta)},$$

where $c_n^{(\alpha,\beta)}$ is the coefficient of x^n in $\psi_n^{(\alpha,\beta)}(x)$. It follows that

$$\begin{aligned}\int_0^1 \psi_n^{(\alpha,\beta)}(x)\psi_n^{(\alpha,\beta)}(x)w(x)dx &= (-1)^n n!c_n^{(\alpha,\beta)}\int_0^1 w(x)x^n(1-x)^n dx \\ &= (-1)^n n!c_n^{(\alpha,\beta)}\frac{B(\alpha+n,\beta+n)}{B(\alpha,\beta)}.\end{aligned}$$

Because the beta distribution is concentrated on a finite interval, the polynomial sequence

$$\left\{\sqrt{\frac{B(\alpha,\beta)}{(-1)^n n!c_n^{(\alpha,\beta)}B(\alpha+n,\beta+n)}}\psi_n^{(\alpha,\beta)}(x)\right\}_{n=0}^{\infty}$$

provides an orthonormal basis.

Finally, we claim that $c_n^{(\alpha,\beta)} = (-1)^n(2n+\alpha+\beta-2)^{\underline{n}}$. This assertion is certainly true when $n = 0$ because $c_0^{(\alpha,\beta)} = 1$. In general, the recurrence relation (9) and induction imply

$$\begin{aligned}c_n^{(\alpha,\beta)} &= -(n+\alpha-1)c_{n-1}^{(\alpha,\beta+1)} - (n+\beta-1)c_{n-1}^{(\alpha+1,\beta)} \\ &= -(n+\alpha-1)(-1)^{n-1}(2[n-1]+\alpha+\beta+1-2)^{\underline{n-1}} \\ &\quad -(n+\beta-1)(-1)^{n-1}(2[n-1]+\alpha+1+\beta-2)^{\underline{n-1}} \\ &= (-1)^n(2n+\alpha+\beta-2)(2n+\alpha+\beta-3)^{\underline{n-1}} \\ &= (-1)^n(2n+\alpha+\beta-2)^{\underline{n}}.\end{aligned}$$

This proves the asserted formula. ∎

The next proposition permits straightforward recursive computation of orthonormal polynomials. Always use high-precision arithmetic when applying the proposition for a particular value of x.

Proposition 15.4.2. *Let a_n and b_n be the coefficients of x^n and x^{n-1} in the nth term $p_n(x)$ of an orthonormal polynomial sequence with respect to a probability measure μ. Then*

$$p_{n+1}(x) = (A_n x + B_n)p_n(x) - C_n p_{n-1}(x), \tag{10}$$

where

$$A_n = \frac{a_{n+1}}{a_n}, \quad B_n = \frac{a_{n+1}}{a_n}\left(\frac{b_{n+1}}{a_{n+1}} - \frac{b_n}{a_n}\right), \quad C_n = \frac{a_{n+1}a_{n-1}}{a_n^2},$$

and $a_{-1} = 0$.

Proof. We will repeatedly use the fact that $\int p_n(x)q(x)d\mu(x) = 0$ for any polynomial $q(x)$ of degree $n-1$ or lower. This follows because $q(x)$ must be a linear combination of $p_0(x), \ldots, p_{n-1}(x)$. Now given the definition of A_n, it is clear that

$$p_{n+1}(x) - A_n x p_n(x) = B_n p_n(x) - C_n p_{n-1}(x) + r_{n-2}(x)$$

for as yet undetermined constants B_n and C_n and a polynomial $r_{n-2}(x)$ of degree $n-2$. If $0 \le k \le n-2$, then

$$\begin{aligned} 0 &= \int p_{n+1}(x)p_k(x)d\mu(x) \\ &= \int p_n(x)[A_n x + B_n]p_k(x)d\mu(x) - C_n \int p_{n-1}(x)p_k(x)d\mu(x) \\ &\quad + \int r_{n-2}(x)p_k(x)d\mu(x), \end{aligned}$$

and consequently $\int r_{n-2}(x)p_k(x)d\mu(x) = 0$. This forces $r_{n-2}(x) = 0$ because $r_{n-2}(x)$ is a linear combination of $p_0(x), \ldots, p_{n-2}(x)$.

If we write

$$x p_{n-1}(x) = \frac{a_{n-1}}{a_n} p_n(x) + q_{n-1}(x),$$

where $q_{n-1}(x)$ is a polynomial of degree $n-1$, then

$$\begin{aligned} 0 &= \int p_{n+1}(x)p_{n-1}(x)d\mu(x) \\ &= A_n \int p_n(x)x p_{n-1}(x)d\mu(x) + B_n \int p_n(x)p_{n-1}(x)d\mu(x) \\ &\quad - C_n \int p_{n-1}^2(x)d\mu(x) \\ &= A_n \frac{a_{n-1}}{a_n} \int p_n^2(x)d\mu(x) + A_n \int p_n(x)q_{n-1}(x)d\mu(x) - C_n \\ &= A_n \frac{a_{n-1}}{a_n} - C_n. \end{aligned}$$

This gives C_n. Finally, equating coefficients of x^n in equation (10) yields $b_{n+1} = A_n b_n + B_n a_n$, and this determines B_n. □

After tedious calculations, Proposition 15.4.2 translates into the following recurrence relations for the orthogonal polynomials considered in Examples 15.4.1 through 15.4.4:

$$\begin{aligned} p_{n+1}^{(\lambda)}(x) &= \Big(\frac{x}{\lambda} - \frac{n}{\lambda} - 1\Big)p_n^{(\lambda)}(x) - \frac{n}{\lambda} p_{n-1}^{(\lambda)}(x) \\ H_{n+1}(x) &= xH_n(x) - nH_{n-1}(x) \\ L_{n+1}^{(\alpha)}(x) &= (2n + \alpha - x)L_n^{(\alpha)}(x) - n(n+\alpha-1)L_{n-1}^{(\alpha)}(x) \\ \psi_{n+1}^{(\alpha,\beta)}(x) &= \frac{(2n+\alpha+\beta)(2n+\alpha+\beta-1)}{n+\alpha+\beta-1} \end{aligned} \tag{11}$$

$$\times \Big[\frac{(n+1)(n+\alpha)}{2n+\alpha+\beta} - \frac{n(n+\alpha-1)}{2n+\alpha+\beta-2} - x\Big]\psi_n^{(\alpha,\beta)}(x)$$
$$- \frac{n(2n+\alpha+\beta)(n+\alpha-1)(n+\beta-1)}{(n+\alpha+\beta-1)(2n+\alpha+\beta-2)}\psi_{n-1}^{(\alpha,\beta)}(x).$$

15.5 Problems

1. Find the Fourier series of the function $|x|$ defined on $[-1/2, 1/2]$ and extended periodically to the whole real line. At what points of $[-1/2, 1/2]$ does the Fourier series converge pointwise to $|x|$?

2. Let $f(x)$ be a periodic function on the real line whose kth derivative is piecewise continuous for some positive integer k. Show that the Fourier coefficients c_n of $f(x)$ satisfy

$$|c_n| \leq \frac{\int_0^1 |f^{(k)}(x)|dx}{|2\pi n|^k}$$

for $n \neq 0$.

3. Suppose that the periodic function $f(x)$ is square-integrable on $[0, 1]$. Prove the assertions: (a) $f(x)$ is an even (respectively odd) function if and only if its Fourier coefficients c_n are even (respectively odd) functions of n, (b) $f(x)$ is real and even if and only if the c_n are real and even, and (c) $f(x)$ is even (odd) if and only if it is even (odd) around $1/2$. By even around $1/2$ we mean $f(1/2 + x) = f(1/2 - x)$.

4. Demonstrate that

$$\frac{\pi^2}{12} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^2}$$
$$\frac{\pi^4}{90} = \sum_{k=1}^{\infty} \frac{1}{k^4}.$$

5. Show that the Bernoulli polynomials satisfy the identity

$$B_n(x) = (-1)^n B_n(1-x)$$

for all n and $x \in [0, 1]$. Conclude from this identity that $B_n(1/2) = 0$ for n odd.

6. Continuing Problem 5, show inductively for $n \geq 1$ that $B_{2n}(x)$ has exactly one simple zero in $(0, 1/2)$ and one in $(1/2, 1)$, while $B_{2n+1}(x)$ has precisely the simple zeros 0, $1/2$, and 1.

7. Demonstrate that the Bernoulli polynomials satisfy the identity

$$B_n(x+1) - B_n(x) = nx^{n-1}.$$

Use this result to verify that the sum of the nth powers of the first m integers can be expressed as

$$\sum_{k=1}^{m} k^n = \frac{1}{n+1}\Big[B_{n+1}(m+1) - B_{n+1}(1)\Big].$$

(Hint: Prove the first assertion by induction or by expanding $B_n(x)$ in a Taylor series around the point 1.)

8. If μ is the probability measure associated with the binomial distribution with m trials and success probability p, then

$$\mu(\{x\}) = \binom{m}{x} p^x q^{m-x},$$

where $0 \le x \le m$ and $q = 1 - p$. The exponential generating function

$$(1+tq)^x (1-tp)^{m-x} = \sum_{n=0}^{\infty} \frac{K_n^{(m,p)}(x)}{n!} t^n$$

defines the Krawtchouk polynomials. Show that

$$K_n^{(m,p)}(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} p^k q^{n-k} \underline{x^{n-k}} (m-x)^{\underline{k}}$$

and that the normalized polynomials

$$p_n(x) = \frac{K_n^{(m,p)}(x)}{(pq)^{\frac{n}{2}} \binom{m}{n}^{\frac{1}{2}} n!}$$

constitute an orthonormal basis for the binomial distribution.

9. Show that

$$H_n(x) = (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n} e^{-\frac{1}{2}x^2}.$$

(Hint: Expand the left-hand side of the identity

$$e^{-\frac{1}{2}(x-t)^2} = \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n e^{-\frac{1}{2}x^2}$$

in a Taylor series and equate coefficients of t^n.)

10. Verify that

$$L_n^{(\alpha)}(x) = e^x x^{-\alpha+1} \frac{d^n}{dx^n} \Big(e^{-x} x^{n+\alpha-1}\Big).$$

11. Validate the recurrences listed in (11). (Hint: Note that the coefficient of x^{n-1} in $\psi_n^{(\alpha,\beta)}(x)$ is $(-1)^{n-1} n(n+\alpha-1)(2n+\alpha+\beta-3)^{\underline{n-1}}$.)

References

[1] Dym H, McKean HP (1972) *Fourier Series and Integrals.* Academic Press, New York
[2] Hewitt E, Stromberg K (1965) *Real and Abstract Analysis.* Springer-Verlag, New York
[3] Hochstadt H (1986) *The Functions of Mathematical Physics.* Dover, New York
[4] Parthasarathy KR (1977) *Introduction to Probability and Measure.* Springer-Verlag, New York
[5] Rudin W (1973) *Functional Analysis.* McGraw-Hill, New York

16
Quadrature Methods

16.1 Introduction

The numerical calculation of one-dimensional integrals, or quadrature, is one of the oldest branches of numerical analysis. Long before calculus was invented, Archimedes found accurate approximations to π by inscribed and circumscribed polygons to a circle of unit radius. In modern applied mathematics and statistics, quadrature is so pervasive that even hand-held calculators are routinely programmed to perform it. Nonetheless, gaining a theoretical understanding of quadrature is worth the effort. In many scientific problems, large numbers of quadratures must be carried out quickly and accurately.

This chapter focuses on the two dominant methods of modern quadrature, Romberg's algorithm [6, 7] and Gaussian quadrature [6, 11, 12, 13]. To paraphrase Henrici [7], Romberg's algorithm uses an approximate knowledge of the error in integration to approximately eliminate that error. Gaussian quadrature is ideal for integration against standard probability densities such as the normal or gamma. Both methods work extremely well for good integrands such as low-degree polynomials. In fact, Gaussian quadrature is designed to give exact answers in precisely this situation. Gaussian quadrature also has the virtue of handling infinite domains of integration gracefully. In spite of these advantages, Romberg's algorithm is usually the preferred method of quadrature for a wide variety of problems. It is robust and simple to code. At its heart is the trapezoidal rule, which can be adapted to employ relatively few function evaluations for smooth

regions of an integrand and many evaluations for rough regions. Whatever the method of quadrature chosen, numerical integration is an art. We briefly describe some tactics for taming bad integrands.

16.2 Euler–Maclaurin Sum Formula

As a prelude to Romberg's algorithm, we discuss the summation formula of Euler and Maclaurin [3, 5, 10, 15]. Besides providing insight into the error of the trapezoidal method of quadrature, this formula is an invaluable tool in asymptotic analysis. Our applications to harmonic series and Stirling's formula illustrate this fact.

Proposition 16.2.1. *Suppose $f(x)$ has $2m$ continuous derivatives on the interval $[1, n]$ for some positive integer n. Then*

$$\sum_{k=1}^{n} f(k) = \int_1^n f(x)dx + \frac{1}{2}\Big[f(n) + f(1)\Big] + \sum_{j=1}^{m} \frac{B_{2j}}{(2j)!} f^{(2j-1)}(x)|_1^n - \frac{1}{(2m)!}\int_1^n b_{2m}(x) f^{(2m)}(x)dx, \tag{1}$$

where B_k is a Bernoulli number and $b_k(x)$ is a Bernoulli function. The remainder in this expansion is bounded by

$$\Big|\frac{1}{(2m)!}\int_1^n b_{2m}(x) f^{(2m)}(x)dx\Big| \le C_{2m}\int_1^n |f^{(2m)}(x)|dx, \tag{2}$$

where

$$C_{2m} = \frac{2}{(2\pi)^{2m}}\sum_{k=1}^{\infty}\frac{1}{k^{2m}}.$$

Proof. Consider an arbitrary function $g(x)$ defined on $[0, 1]$ with $2m$ continuous derivatives. In view of the definition of the Bernoulli polynomials in Chapter 15, repeated integration by parts gives

$$\begin{aligned}\int_0^1 g(x)dx &= \int_0^1 B_0(x)g(x)dx \\ &= B_1(x)g(x)|_0^1 - \int_0^1 B_1(x)g'(x)dx \\ &= \sum_{i=1}^{2m}\frac{(-1)^{i-1}B_i(x)}{i!}g^{(i-1)}(x)|_0^1 \\ &\quad + \frac{(-1)^{2m}}{(2m)!}\int_0^1 B_{2m}(x)g^{(2m)}(x)dx.\end{aligned}$$

This formula can be simplified by noting that (a) $B_{2m}(x) = b_{2m}(x)$ on $[0, 1]$, (b) $B_1(x) = x - 1/2$, (c) $B_i(0) = B_i(1) = B_i$ when $i > 1$, and

(d) $B_i = 0$ when $i > 1$ and i is odd. Hence,

$$\int_0^1 g(x)dx = \frac{1}{2}\Big[g(1) + g(0)\Big] - \sum_{j=1}^{m} \frac{B_{2j}}{(2j)!} g^{(2j-1)}(x)\Big|_0^1 + \frac{1}{(2m)!}\int_0^1 b_{2m}(x) g^{(2m)}(x)dx.$$

If we apply this result successively to $g(x) = f(x+k)$ for $k = 1, \ldots, n-1$ and add the results, then cancellation of successive terms produces formula (1). The bound (2) follows immediately from the Fourier series representation of $b_{2m}(x)$ noted in Chapter 15. □

Example 16.2.1 (*Harmonic Series*). The harmonic series $\sum_{k=1}^{n} k^{-1}$ can be approximated by taking $f(x)$ to be x^{-1} in Proposition 16.2.1. For example with $m = 2$, we find that

$$\begin{aligned}\sum_{k=1}^{n} \frac{1}{k} &= \int_1^n \frac{1}{x}dx + \frac{1}{2}\Big[\frac{1}{n} + 1\Big] + \frac{B_2}{2}\Big[1 - \frac{1}{n^2}\Big] \\ &\quad + \frac{B_4}{4!}\Big[3! - \frac{3!}{n^4}\Big] - \frac{1}{4!}\int_1^n b_4(x)\frac{4!}{x^5}dx \\ &= \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + \int_n^\infty b_4(x)\frac{1}{x^5}dx \\ &= \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + O\Big(\frac{1}{n^4}\Big),\end{aligned}$$

where

$$\begin{aligned}\gamma &= \frac{1}{2} + \frac{1}{12} - \frac{1}{120} - \int_1^\infty b_4(x)\frac{1}{x^5}dx \\ &\approx 0.5772\end{aligned} \tag{3}$$

is Euler's constant. ■

Example 16.2.2 (*Stirling's Formula*). If we let $f(x)$ be the function $\ln x = \frac{d}{dx}[x\ln x - x]$ and $m = 2$ in Proposition 16.2.1, then we recover Stirling's formula

$$\begin{aligned}\ln n! &= \sum_{k=1}^{n} \ln k \\ &= \int_1^n \ln x\, dx + \frac{1}{2}\ln n + \frac{B_2}{2}\Big[\frac{1}{n} - 1\Big] \\ &\quad + \frac{B_4}{4!}\Big[\frac{2!}{n^3} - 2!\Big] + \frac{1}{4!}\int_1^n b_4(x)\frac{3!}{x^4}dx \\ &= n\ln n - n + \frac{1}{2}\ln n + s + \frac{1}{12n} - \frac{1}{360n^3} - \frac{1}{4}\int_n^\infty b_4(x)\frac{1}{x^4}dx \\ &= (n + \frac{1}{2})\ln n - n + s + \frac{1}{12n} + O\Big(\frac{1}{n^3}\Big),\end{aligned}$$

where

$$s = 1 - \frac{1}{12} + \frac{1}{360} + \frac{1}{4}\int_1^\infty b_4(x)\frac{1}{x^4}dx = \ln\sqrt{2\pi} \tag{4}$$

was determined in Chapter 4. ■

Example 16.2.3 (*Error Bound for the Trapezoidal Rule*)**.** The trapezoidal rule is one simple mechanism for integrating a function $f(x)$ on a finite interval $[a, b]$. If we divide the interval into n equal subintervals of length $h = (b-a)/n$, then the value of the integral of $f(x)$ between $a+kh$ and $a+(k+1)h$ is approximately $\frac{h}{2}\{f(a+kh)+f(a+[k+1]h)\}$. Summing these approximate values over all subintervals therefore gives

$$\int_a^b f(x)dx \approx h\Big[\frac{1}{2}g(0) + g(1) + \cdots + g(n-1) + \frac{1}{2}g(n)\Big] \tag{5}$$

for $g(t) = f(a+th)$. If we abbreviate the trapezoidal approximation on the right of (5) by $T(h)$, then Proposition 16.2.1 implies that

$$\begin{aligned} T(h) &= h\int_0^n g(t)dt + \frac{hB_2}{2}g'(t)|_0^n + \frac{hB_4}{4!}g^{(3)}(t)|_0^n - \frac{h}{4!}\int_0^n b_4(t)g^{(4)}(t)dt \\ &= \int_a^b f(x)dx + \frac{h^2}{12}\Big[f'(b) - f'(a)\Big] - \frac{h^4}{720}\Big[f^{(3)}(b) - f^{(3)}(a)\Big] \\ &\quad - \frac{h^4}{4!}\int_a^b b_4\Big(\frac{x-a}{h}\Big)f^{(4)}(x)dx \\ &= \int_a^b f(x)dx + \frac{h^2}{12}\Big[f'(b) - f'(a)\Big] + O(h^4). \end{aligned}$$

In practice, it is inconvenient to suppose that $f'(x)$ is known, so the error committed in using the trapezoidal rule is $O(h^2)$. If $f(x)$ possesses $2k$ continuous derivatives, then a slight extension of the above argument indicates that the trapezoidal approximation satisfies

$$T(h) \;=\; \int_a^b f(x)dx + c_1h^2 + c_2h^4 + \cdots + c_{k-1}h^{2(k-1)} + O(h^{2k}) \tag{6}$$

for constants $c_1, \ldots, c_{k-1}$ that depend on $f(x)$, a, and b but not on h. ■

16.3 Romberg's Algorithm

Suppose in the trapezoidal rule, we halve the integration step h. Then the error estimate (6) becomes

$$T\Big(\frac{1}{2}h\Big) = \int_a^b f(x)dx + \frac{c_1}{4}h^2 + \frac{c_2}{4^2}h^4 + \cdots + \frac{c_{k-1}}{4^{k-1}}h^{2(k-1)} + O(h^{2k}).$$

Romberg recognized that forming the linear combination

$$\begin{aligned}\frac{4T(\frac{1}{2}h)-T(h)}{3} &= T\Big(\frac{1}{2}h\Big)-\frac{1}{3}\Big[T(h)-T\Big(\frac{1}{2}h\Big)\Big] \\ &= \int_a^b f(x)dx + d_2h^4+\cdots+d_{k-1}h^{2(k-1)}+O(h^{2k})\end{aligned} \tag{7}$$

eliminates the h^2 error term, where $d_2,\ldots,d_{k-1}$ are new constants that can be easily calculated. For h small, decreasing the error in estimating $\int_a^b f(x)dx$ from $O(h^2)$ to $O(h^4)$ is a striking improvement in accuracy.

Even more interesting is the fact that this tactic can be iterated. Suppose we compute the trapezoidal approximations $T_{m0}=T[2^{-m}(b-a)]$ for several consecutive integers m beginning with $m=0$. In essence, we double the number of quadrature points and halve h at each stage. When $f(x)$ has $2k$ continuous derivatives, the natural inductive generalization of the refinement (7) is provided by the sequence of refinements

$$\begin{aligned}T_{mn} &= \frac{4^nT_{m,n-1}-T_{m-1,n-1}}{4^n-1} \\ &= T_{m,n-1}-\frac{1}{4^n-1}[T_{m-1,n-1}-T_{m,n-1}]\end{aligned} \tag{8}$$

for $n\le\min\{m,k-1\}$. From this recursive definition, it follows that $\lim_{m\to\infty}T_{m0}=\int_a^b f(x)dx$ implies $\lim_{m\to\infty}T_{mn}=\int_a^b f(x)dx$ for every n. Furthermore, if

$$T_{m,n-1}=\int_a^b f(x)dx+\gamma_n4^{-mn}+\cdots+\gamma_{k-1}4^{-m(k-1)}+O(4^{-mk})$$

for appropriate constants $\gamma_n,\ldots,\gamma_{k-1}$, then

$$T_{mn}=\int_a^b f(x)dx+\delta_{n+1}4^{-m(n+1)}+\cdots+\delta_{k-1}4^{-m(k-1)}+O(4^{-mk})$$

for appropriate new constants $\delta_{n+1},\ldots,\delta_{k-1}$. In other words, provided the condition $n+1\le k$ holds, the error drops from $O(4^{-mn})$ to $O(4^{-m(n+1)})$ in going from $T_{m,n-1}$ to T_{mn}.

It is convenient to display the trapezoidal approximations to a definite integral $\int_a^b f(x)dx$ as the first column of the Romberg array

$$\begin{pmatrix} T_{00} & & & & & \\ T_{10} & T_{11} & & & & \\ T_{20} & T_{21} & T_{22} & & & \\ T_{30} & T_{31} & T_{32} & T_{33} & & \\ T_{40} & T_{41} & T_{42} & T_{43} & T_{44} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

Romberg's algorithm fills in an entry of this array based on the two entries immediately to the left and diagonally above and to the left of the given

TABLE 16.1. Romberg Table for E(Y)

m	T_{m0}	T_{m1}	T_{m2}	T_{m3}
0	0.4207355	—	—	—
1	0.4500805	0.4598622	—	—
2	0.4573009	0.4597077	0.4596974	—
3	0.4590990	0.4596983	0.4596977	0.4596977
∞	0.4596977	0.4596977	0.4596977	0.4596977

entry. Depending on the smoothness of the integrand, the columns to the right of the first column converge more and more rapidly to $\int_a^b f(x)dx$. In practice, convergence can occur so quickly that computing entries T_{mn} with n beyond 2 or 3 is wasted effort.

Example 16.3.1 (*Numerical Examples of Romberg's Algorithm***).** For two simple examples, let X be uniformly distributed on $[0, 1]$, and define the random variables $Y = \sin X$ and $Z = \sqrt{1 - X^2}$. The explicit values

$$\mathrm{E}(Y) = \int_0^1 \sin x \, dx = 1 - \cos(1)$$

$$\mathrm{E}(Z) = \int_0^1 \sqrt{1 - y^2} \, dy = \frac{1}{2} \arcsin(1)$$

available for the means of Y and Z offer an opportunity to assess the performance of Romberg's algorithm. Table 16.1 clearly demonstrates the accelerated convergence possible for a smooth integrand. Because the derivative of $\sqrt{1 - y^2}$ is singular at $y = 1$, the slower convergence seen in Table 16.2 is to be expected. ■

TABLE 16.2. Romberg Table for E(Z)

m	T_{m0}	T_{m1}	T_{m2}	T_{m3}
0	0.50000	—	—	—
1	0.68301	0.74402	—	—
2	0.74893	0.77090	0.77269	—
3	0.77245	0.78030	0.78092	0.78105
4	0.78081	0.78360	0.78382	0.78387
5	0.78378	0.78476	0.78484	0.78486
6	0.78482	0.78517	0.78520	0.78521
7	0.78520	0.78532	0.78533	0.78533
8	0.78533	0.78537	0.78537	0.78537
9	0.78537	0.78539	0.78539	0.78539
10	0.78539	0.78539	0.78540	0.78540
∞	0.78540	0.78540	0.78540	0.78540

16.4 Adaptive Quadrature

A crude, adaptive version of the trapezoidal rule can be easily constructed [4]. In the first stage of the adaptive algorithm, the interval of integration is split at its midpoint $c = (a + b)/2$, and the two approximations

$$S_0 = \frac{(b-a)}{2}\Big[f(a) + f(b)\Big]$$
$$S_1 = \frac{(b-a)}{4}\Big[f(a) + 2f(c) + f(b)\Big]$$

are compared. If $|S_0 - S_1| < \epsilon(b-a)$ for $\epsilon > 0$ small, then $\int_a^b f(x)dx$ is set equal to S_1. If the test $|S_0 - S_1| < \epsilon(b-a)$ fails, then the integrals $\int_a^c f(x)dx$ and $\int_c^b f(x)dx$ are separately computed and the results added. This procedure is made recursive by computing an integral via the trapezoidal rule at each stage or splitting the integral for further processing. Obviously, a danger in the adaptive algorithm is that the two initial approximations S_0 and S_1 (or similar approximations at some subsequent early stage) agree by chance.

16.5 Taming Bad Integrands

We illustrate by way of example some of the usual tactics for improving integrands.

Example 16.5.1 (*Subtracting off a Singularity*). Sometimes one can subtract off the singular part of an integrand and integrate that part analytically. The example

$$\begin{aligned}\int_0^1 \frac{e^x}{\sqrt{x}}dx &= \int_0^1 \frac{e^x - 1}{\sqrt{x}}dx + \int_0^1 \frac{1}{\sqrt{x}}dx \\ &= \int_0^1 \frac{e^x - 1}{\sqrt{x}}dx + 2\end{aligned}$$

is fairly typical. The remaining integrand

$$\frac{e^x - 1}{\sqrt{x}} = \sqrt{x}\sum_{k=1}^{\infty}\frac{x^{k-1}}{k!}$$

is well behaved at $x = 0$. In the vicinity of this point, it is wise to evaluate the integrand by a few terms of its series expansion to avoid roundoff errors in the subtraction $e^x - 1$. ■

Example 16.5.2 (*Splitting an Interval and Changing Variables*). Consider the problem of computing the expectation of $\ln X$ for a beta

distributed random variable X. The necessary integral

$$\mathrm{E}(\ln X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \ln x \, x^{\alpha-1}(1-x)^{\beta-1} dx$$

has singularities at $x = 0$ and $x = 1$. We attack this integral by splitting it into parts over $[0, 1/2]$ and $[1/2, 1]$. The integral over $[1/2, 1]$ can be tamed by making the change of variables $y^n = 1 - x$. This gives

$$\int_{\frac{1}{2}}^1 \ln x \, x^{\alpha-1}(1-x)^{\beta-1} dx = n \int_0^{\frac{1}{\sqrt[n]{2}}} \ln(1-y^n)\,(1-y^n)^{\alpha-1} y^{n\beta-1} dy.$$

Provided n is chosen so large that $n\beta - 1 \geq 0$, the transformed integral is well behaved. The integral over $[0, 1/2]$ is handled similarly by making the change of variables $y^n = x$. In this case

$$\int_0^{\frac{1}{2}} \ln x \, x^{\alpha-1}(1-x)^{\beta-1} dx = n^2 \int_0^{\frac{1}{\sqrt[n]{2}}} \ln y \, y^{n\alpha-1}(1-y^n)^{\beta-1} dy.$$

Because of the presence of the singularity in $\ln y$ at $y = 0$, it is desirable that $n\alpha - 1 > 0$. Indeed, then the factor $y^{n\alpha-1}$ yields $\lim_{y\to 0} y^{n\alpha-1} \ln y = 0$. ■

Example 16.5.3 (*Infinite Integration Limit*)**.** The integral (3) appearing in the definition of Euler's constant occurs over the infinite interval $[1, \infty)$. If we make the change of variable $y^{-1} = x$, then

$$\int_1^\infty b_4(x) \frac{1}{x^5} dx = \int_0^1 b_4(y^{-1}) y^3 dy.$$

The transformed integral is still challenging to evaluate because the integrand $b_4(y^{-1})y^3$ has limited smoothness and rapidly oscillates in the vicinity of $y = 0$. Fortunately, some of the sting of rapid oscillation is removed by the damping factor y^3. Problem 4 asks readers to evaluate Euler's constant by quadrature. ■

16.6 Gaussian Quadrature

Gaussian quadrature is ideal for evaluating integrals against certain probability measures μ. If $f(x)$ is a smooth function, then it is natural to consider approximations of the sort

$$\int_{-\infty}^\infty f(x) d\mu(x) \approx \sum_{i=0}^k w_i f(x_i), \tag{9}$$

where the finite sum ranges over fixed points x_i with attached positive weights w_i. In the trapezoidal rule, μ is a uniform measure, and the points are uniformly spaced. In Gaussian quadrature, μ is typically nonuniform,

and the points cluster in regions of high probability. Since polynomials are quintessentially smooth, Gaussian quadrature requires that the approximation (9) be exact for all polynomials of sufficiently low degree.

If the probability measure μ possesses an orthonormal polynomial sequence $\{\psi_n(x)\}_{n=0}^{\infty}$, then the $k+1$ points $x_0, \ldots, x_k$ of formula (9) are taken to be the roots of the polynomial $\psi_{k+1}(x)$. Assuming for the moment that these roots are distinct and real, we have the following remarkable result.

Proposition 16.6.1. *If μ is not concentrated at a finite number of points, then there exist positive weights w_i such that the quadrature formula (9) is exact whenever $f(x)$ is any polynomial of degree $2k+1$ or lower.*

Proof. Let us first prove the result for a polynomial $f(x)$ of degree k or lower. If $l_i(x)$ denotes the polynomial

$$l_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$$

of degree k, then

$$f(x) = \sum_{i=0}^{k} l_i(x) f(x_i)$$

is the interpolating-polynomial representation of $f(x)$. The condition

$$\int f(x) d\mu(x) = \sum_{i=0}^{k} \int l_i(x) d\mu(x) f(x_i)$$

now determines the weights $w_i = \int l_i(x) d\mu(x)$, which obviously do not depend on $f(x)$.

Now let $f(x)$ be any polynomial of degree $2k+1$ or lower. The division algorithm for polynomials [2] implies that

$$f(x) = p(x)\psi_{k+1}(x) + q(x)$$

for polynomials $p(x)$ and $q(x)$ of degree k or lower. On the one hand, because $p(x)$ is orthogonal to $\psi_{k+1}(x)$,

$$\int f(x) d\mu(x) = \int q(x) d\mu(x). \tag{10}$$

On the other hand, because the x_i are roots of $\psi_{k+1}(x)$,

$$\sum_{i=0}^{k} w_i f(x_i) = \sum_{i=0}^{k} w_i q(x_i). \tag{11}$$

In view of the first part of the proof and the fact that $q(x)$ is a polynomial of degree k or lower, we also have

$$\int q(x) d\mu(x) = \sum_{i=0}^{k} w_i q(x_i). \tag{12}$$

Equations (10), (11), and (12) taken together imply that formula (9) is exact.

Finally, if $f(x) = l_i(x)^2$, then the calculation

$$\begin{aligned} w_i &= \sum_{j=0}^{k} w_j l_i(x_j)^2 \\ &= \int l_i(x)^2 d\mu(x) \end{aligned}$$

shows that $w_i > 0$, provided that μ is not concentrated at a finite number of points. □

We now make good on our implied promise concerning the roots of the polynomial $\psi_{k+1}(x)$.

Proposition 16.6.2. *Under the premises of Proposition 16.6.1, the roots of each polynomial $\psi_{k+1}(x)$ are real and distinct.*

Proof. If the contrary is true, then $\psi_{k+1}(x)$ changes sign fewer than $k+1$ times. Let the positions of the sign changes occur at the distinct roots $r_1 < \cdots < r_m$ of $\psi_{k+1}(x)$. Since the polynomial $\psi_{k+1}(x)\prod_{i=1}^{m}(x - r_i)$ is strictly negative or strictly positive except at the roots of $\psi_{k+1}(x)$, we infer that

$$\left| \int \psi_{k+1}(x) \prod_{i=1}^{m} (x - r_i) d\mu(x) \right| > 0.$$

However, $\prod_{i=1}^{m}(x - r_i)$ is a polynomial of lower degree than $\psi_{k+1}(x)$ and consequently must be orthogonal to $\psi_{k+1}(x)$. This contradiction shows that $\psi_{k+1}(x)$ must have at least $k + 1$ distinct changes of sign. □

Good software is available for computing the roots and weights of most classical orthogonal polynomials [13]. Newton's method permits rapid computation of the roots if initial values are chosen to take advantage of the interlacing of roots from successive polynomials. The interlacing property, which we will not prove, can be stated as follows [8]: If $y_1 < \cdots < y_k$ denote the roots of the orthogonal polynomial $\psi_k(x)$, then the next orthogonal polynomial $\psi_{k+1}(x)$ has exactly one root on each of the $k + 1$ intervals $(-\infty, y_1), (y_1, y_2), \ldots, (y_k, \infty)$. The weights associated with the $k + 1$ roots $x_0, \ldots, x_k$ of $\psi_{k+1}(x)$ can be computed in a variety of ways. For instance, in view of the identity $\int \psi_i(x) d\mu(x) = 0$ for $i > 0$, the set of linear equations

$$\begin{pmatrix} \psi_0(x_0) & \cdots & \psi_0(x_k) \\ \psi_1(x_0) & \cdots & \psi_1(x_k) \\ \vdots & & \vdots \\ \psi_k(x_0) & \cdots & \psi_k(x_k) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{pmatrix} = \begin{pmatrix} \int \psi_0(x) d\mu(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

uniquely determines the weights.

TABLE 16.3. Quadrature Approximations to $\mathrm{Var}(X_{(m+1)})$

n	$Q_{n,2}$	$Q_{n,4}$	$Q_{n,8}$	$Q_{n,16}$	$Q_{n,32}$	$Q_{n,64}$
11	0.1175	0.2380	0.2341	0.1612	0.1379	0.1372
21	0.0070	0.0572	0.1271	0.1230	0.0840	0.0735

For historical reasons, the Jacobi polynomials are employed in Gaussian quadrature rather than the beta distribution polynomials. The Jacobi polynomials are orthogonal with respect to the density $(1-x)^{\alpha-1}(1+x)^{\beta-1}$ defined on the interval $(-1,1)$. Integration against a beta distribution can be reduced to integration against a Jacobi density via the change of variables $y = (1-x)/2$. Indeed,

$$\begin{aligned}&\frac{1}{B(\alpha,\beta)}\int_0^1 f(y)y^{\alpha-1}(1-y)^{\beta-1}dy\\ &\quad=\frac{1}{2^{\alpha+\beta-1}B(\alpha,\beta)}\int_{-1}^1 f\Big(\frac{1-x}{2}\Big)(1-x)^{\alpha-1}(1+x)^{\beta-1}dx.\end{aligned}$$

Example 16.6.1 (*Variance of a Sample Median*). Consider an i.i.d. sample $X_1,\ldots,X_n$ from the standard normal distribution. Assuming n is odd and $m = \lfloor n/2 \rfloor$, the sample median $X_{(m+1)}$ has density

$$n\binom{n-1}{m}\Phi(x)^m[1-\Phi(x)]^m\phi(x),$$

where $\phi(x)$ is the standard normal density and $\Phi(x)$ is the standard normal distribution function. Since the mean of $X_{(m+1)}$ is 0, the variance of $X_{(m+1)}$ is

$$\mathrm{Var}(X_{(m+1)}) = n\binom{n-1}{m}\int_{-\infty}^{\infty} x^2\Phi(x)^m[1-\Phi(x)]^m\phi(x)dx.$$

Table 16.3 displays the Gauss–Hermite quadrature approximations Q_{nk} to $\mathrm{Var}(X_{(m+1)})$ for samples sizes $n = 11$ and $n = 21$ and varying numbers of quadrature points k. These results should be compared to the values 0.1428 for $n = 11$ and 0.0748 for $n = 21$ predicted by the asymptotic variance $1/[4n\phi(0)^2] = \pi/(2n)$ of the normal limit law for a sample median [14]. ■

16.7 Problems

1. Use the Euler-Maclaurin summation formula to establish the equality

$$\frac{1}{e-1} = \sum_{n=0}^{\infty}\frac{B_n}{n!}.$$

2. Verify the asymptotic expansion

$$\sum_{k=1}^{n} k^{\alpha} = C_{\alpha} + \frac{n^{\alpha+1}}{\alpha+1} + \frac{n^{\alpha}}{2} + \sum_{j=1}^{m} \frac{B_{2j}}{2j} \binom{\alpha}{2j-1} n^{\alpha-2j+1} + O(n^{\alpha-2m-1})$$

for a real number $\alpha \neq -1$ and some constant C_{α}, which you need not determine.

3. Find asymptotic expansions for the two sums $\sum_{k=1}^{n}(n^2+k^2)^{-1}$ and $\sum_{k=1}^{n}(-1)^k/k$ valid to $O(n^{-3})$.

4. Check by quadrature that the asserted values in expressions (3) and (4) are accurate.

5. After the first application of Romberg's acceleration algorithm in equation (7), show that the integral $\int_a^b f(x)dx$ is approximated by

$$\frac{h}{6}\Big(f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \cdots + 2f_{2n-2} + 4f_{2n-1} + f_{2n}\Big),$$

where $f_k = f[a + k(b-a)/(2n)]$. This is Simpson's rule.

6. Show that the refinement T_{mn} in Romberg's algorithm exactly equals $\int_a^b f(x)dx$ when $f(x)$ is a polynomial of degree $2n+1$ or lower.

7. For an integrand $f(x)$ with four continuous derivatives, let $Q(h)$ be the trapezoidal approximation to the integral $\int_a^b f(x)dx$ with integration step $h = (b-a)/(6n)$. Based on $Q(2h)$ and $Q(3h)$, construct a quadrature formula $R(h)$ such that

$$\int_a^b f(x)dx - R(h) = O(h^4).$$

8. Numerical differentiation can be improved by the acceleration technique employed in Romberg's algorithm. If $f(x)$ has $2k$ continuous derivatives, then the central difference formula

$$\begin{aligned} D_0(h) &= \frac{f(x+h) - f(x-h)}{2h} \\ &= \sum_{j=0}^{k-1} f^{(2j+1)}(x) \frac{h^{2j}}{(2j+1)!} + O(h^{2k-1}) \end{aligned}$$

follows from an application of Taylor's theorem. Show that the inductively defined quantities

$$\begin{aligned} D_j(h) &= \frac{4^j D_{j-1}(\frac{1}{2}h) - D_{j-1}(h)}{4^j - 1} \\ &= D_{j-1}\Big(\frac{1}{2}h\Big) - \frac{1}{4^j - 1}\Big[D_{j-1}(h) - D_{j-1}\Big(\frac{1}{2}h\Big)\Big] \end{aligned}$$

satisfy

$$f'(x) = D_j(h) + O(h^{2j+2})$$

for $j = 0, \ldots, k-1$. Verify that

$$D_1(h) = \frac{1}{6}\Big[8f\Big(x + \frac{1}{2}h\Big) - 8f\Big(x - \frac{1}{2}h\Big) - f(x+h) + f(x-h)\Big].$$

Finally, try this improvement of central differencing on a few representative functions such as $\sin(x)$, e^x, and $\ln \Gamma(x)$.

9. Discuss what steps you would take to compute the following integrals accurately [1]:

$$\int_1^\infty \frac{\ln x}{x\sqrt{1+x}}dx, \quad \int_\epsilon^{\frac{\pi}{2}} \frac{\sin x}{x^2}dx, \quad \int_0^{\frac{\pi}{2}} \frac{1-\cos x}{x^2\sqrt{x}}dx.$$

In the middle integral, $\epsilon > 0$ is small.

10. For a probability measure μ concentrated on a finite interval $[a, b]$, let

$$Q_k(f) = \sum_{i=0}^{k} w_i^{(k)} f(x_i^{(k)})$$

be the sequence of Gaussian quadrature operators that integrate polynomials of degree $2k+1$ or lower exactly based on the roots of the orthonormal polynomial $\psi_{k+1}(x)$. Prove that

$$\lim_{k\to\infty} Q_k(f) = \int_a^b f(x)d\mu(x)$$

for any continuous function $f(x)$. (Hint: Apply the Weierstrass approximation theorem [9].)

11. Describe and implement Newton's method for computing the roots of the Hermite polynomials. How can you use the interlacing property of the roots and the symmetry properties of the Hermite polynomials to reduce computation time? How is Proposition 15.4.2 relevant to the computation of the derivatives required by Newton's method? To avoid overflows you should use the orthonormal version of the polynomials.

12. Continuing Problem 11, describe and implement a program for the computation of the weights in Gauss–Hermite quadrature.

13. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a gamma density with scale parameter 1 and shape parameter α. Describe and implement a numerical scheme for computing the expected values of the order statistics $X_{(1)}$ and $X_{(n)}$.

14. In light of Problem 13 of Chapter 2, describe and implement a numerical scheme for computing the bivariate normal distribution function.

References

[1] Acton FS (1996) *Real Computing Made Real: Preventing Errors in Scientific and Engineering Calculations.* Princeton University Press, Princeton, NJ

[2] Birkhoff G, MacLane S (1965) *A Survey of Modern Algebra*, 3rd ed. Macmillan, New York
[3] Boas RP Jr (1977) Partial sums of infinite series, and how they grow. *Amer Math Monthly* 84:237–258
[4] Ellis TMR, Philips IR, Lahey TM (1994) *Fortran 90 Programming.* Addison-Wesley, Wokingham, England
[5] Graham RL, Knuth DE, Patashnik O (1988) *Concrete Mathematics: A Foundation for Computer Science.* Addison-Wesley, Reading, MA
[6] Hämmerlin G, Hoffmann K-H (1991) *Numerical Mathematics.* Springer-Verlag, New York
[7] Henrici (1982) *Essentials of Numerical Analysis with Pocket Calculator Demonstrations.* Wiley, New York
[8] Hochstadt H (1986) *The Functions of Mathematical Physics.* Dover, New York
[9] Hoffman K (1975) *Analysis in Euclidean Space.* Prentice-Hall, Englewood Cliffs, NJ
[10] Isaacson E, Keller HB (1966) *Analysis of Numerical Methods.* Wiley, New York
[11] Körner TW (1988) *Fourier Analysis.* Cambridge University Press, Cambridge
[12] Powell MJD (1981) *Approximation Theory and Methods.* Cambridge University Press, Cambridge
[13] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[14] Sen PK, Singer JM (1993) *Large Sample Methods in Statistics: An Introduction with Applications.* Chapman and Hall, London
[15] Wilf HS (1978) *Mathematics for the Physical Sciences.* Dover, New York

17
The Fourier Transform

17.1 Introduction

The Fourier transform is one of the most productive tools of the mathematical sciences. It crops up again and again in unexpected applications to fields as diverse as differential equations, numerical analysis, probability theory, number theory, quantum mechanics, optics, medical imaging, and signal processing [3, 5, 7, 8, 9]. One explanation for its wide utility is that it turns complex mathematical operations like differentiation and convolution into simple operations like multiplication. Readers most likely are familiar with the paradigm of transforming a mathematical equation, solving it in transform space, and then inverting the solution. Besides its operational advantages, the Fourier transform often has the illuminating physical interpretation of decomposing a temporal process into component processes with different frequencies.

In this chapter, we review the basic properties of the Fourier transform and touch lightly on its applications to Edgeworth expansions. Because of space limitations, our theoretical treatment of Fourier analysis is necessarily superficial. At this level it is difficult to be entirely rigorous without invoking some key facts from real analysis. Readers unfamiliar with the facts cited will have to take them on faith or turn to one of the many available texts on real analysis. In mitigation of these theoretical excursions, some topics from elementary probability are repeated for the sake of completeness.

TABLE 17.1. Fourier Transform Pairs

Function	Transform	Function	Transform
(a) $af(x) + bg(x)$	$a\hat{f}(y) + b\hat{g}(y)$	(e) $f(x)^*$	$\hat{f}(-y)^*$
(b) $f(x - x_0)$	$e^{iyx_0}\hat{f}(y)$	(f) $ixf(x)$	$\frac{d}{dy}\hat{f}(y)$
(c) $e^{iy_0x}f(x)$	$\hat{f}(y + y_0)$	(g) $\frac{d}{dx}f(x)$	$-iy\hat{f}(y)$
(d) $f(\frac{x}{a})$	$\lvert a\rvert\hat{f}(ay)$	(h) $f * g(x)$	$\hat{f}(y)\hat{g}(y)$

17.2 Basic Properties

The Fourier transform can be defined on a variety of function spaces. For our purposes, it suffices to consider complex-valued, integrable functions whose domain is the real line. The Fourier transform of such a function $f(x)$ is defined according to the recipe

$$\hat{f}(y) = \int_{-\infty}^{\infty} e^{iyx} f(x)dx$$

for all real numbers y. Note that by the adjective "integrable" we mean $\int_{-\infty}^{\infty} |f(x)|dx < \infty$. In the sequel we usually omit the limits of integration. If $f(x)$ is a probability density, then the Fourier transform $\hat{f}(y)$ coincides with the characteristic function of $f(x)$.

Proposition 17.2.1. *Table 17.1 summarizes the operational properties of the Fourier transform. In the table, a, b, x_0, and y_0 are constants, and the functions $f(x)$, $xf(x)$, $\frac{d}{dx}f(x)$, and $g(x)$ are assumed integrable as needed. In entry (g), $f(x)$ is taken to be absolutely continuous.*

Proof. All entries in the table except (f) through (h) are straightforward to verify. To prove (f), note that $\frac{d}{dy}\hat{f}(y)$ is the limit of the difference quotients

$$\frac{\hat{f}(y+u) - \hat{f}(y)}{u} = \int \frac{e^{iux} - 1}{u} e^{iyx} f(x)dx.$$

The integrand on the right is bounded above in absolute value by

$$\begin{aligned}\left|\frac{e^{iux} - 1}{u} e^{iyx} f(x)\right| &= \left|\int_0^x e^{iuz}dz\right| |f(x)| \\ &\leq |xf(x)|.\end{aligned}$$

Hence, the dominated convergence theorem permits one to interchange limit and integral signs as u tends to 0.

To verify property (g), we first observe that the absolute continuity of $f(x)$ is a technical condition permitting integration by parts and application of the fundamental theorem of calculus. Now the integration-by-parts formula

$$\int_c^d e^{iyx} \frac{d}{dx} f(x) dx = e^{iyx} f(x)|_c^d - iy \int_c^d e^{iyx} f(x)\, dx$$

proves property (g) provided we can demonstrate that

$$\lim_{c \to -\infty} f(c) = \lim_{d \to \infty} f(d) = 0.$$

Since $\frac{d}{dx} f(x)$ is assumed integrable, the reconstruction

$$f(d) - f(0) = \int_0^d \frac{d}{dx} f(x)\, dx$$

implies that $\lim_{d \to \infty} f(d)$ exists. This right limit is necessarily 0 because $f(x)$ is integrable. The left limit $\lim_{c \to -\infty} f(c) = 0$ follows by the same argument.

The proof of property (h) we defer until we define convolution. □

17.3 Examples

Before developing the theory of the Fourier transform further, it is useful to pause and calculate some specific transforms.

Example 17.3.1 (*Uniform Distribution*). If $f(x)$ is the uniform density on the interval $[a, b]$, then

$$\begin{aligned}
\hat{f}(y) &= \frac{1}{b-a} \int_a^b e^{iyx} dx \\
&= \frac{e^{iyx}}{(b-a)iy} \Big|_a^b \\
&= \frac{e^{i\frac{1}{2}(a+b)y} \left[e^{\frac{1}{2}i(b-a)y} - e^{-\frac{1}{2}i(b-a)y} \right]}{\frac{1}{2}(b-a)y2i} \\
&= e^{i\frac{1}{2}(a+b)y} \frac{\sin[\frac{1}{2}(b-a)y]}{\frac{1}{2}(b-a)y}.
\end{aligned}$$

When $a = -b$, this reduces to $\hat{f}(y) = \sin(by)/(by)$. ■

Example 17.3.2 (*Gaussian Distribution*). To find the Fourier transform of the Gaussian density $f(x)$ with mean 0 and variance σ^2, we derive and solve a differential equation. Indeed, integration by parts and property

(f) of Table 17.1 imply that

$$\begin{aligned}\frac{d}{dy}\hat{f}(y) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int e^{iyx}ixe^{-\frac{x^2}{2\sigma^2}}dx\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int e^{iyx}(-i\sigma^2)\frac{d}{dx}e^{-\frac{x^2}{2\sigma^2}}dx\\ &= \frac{-i\sigma^2}{\sqrt{2\pi\sigma^2}}e^{iyx}e^{-\frac{x^2}{2\sigma^2}}\Big|_{-\infty}^{\infty} - \frac{\sigma^2 y}{\sqrt{2\pi\sigma^2}}\int e^{iyx}e^{-\frac{x^2}{2\sigma^2}}dx\\ &= -\sigma^2 y\hat{f}(y).\end{aligned}$$

The unique solution to this differential equation with initial value $\hat{f}(0) = 1$ is $\hat{f}(y) = e^{-\sigma^2 y^2/2}$. ■

Example 17.3.3 (*Gamma Distribution*). We can also derive the Fourier transform of the gamma density $f(x)$ with shape parameter α and scale parameter β by solving a differential equation. Now integration by parts and property (f) yield

$$\begin{aligned}\frac{d}{dy}\hat{f}(y) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{(iy-\beta)x}ixx^{\alpha-1}dx\\ &= \frac{i\beta^\alpha}{(iy-\beta)\Gamma(\alpha)}\int_0^\infty x^\alpha\frac{d}{dx}e^{(iy-\beta)x}dx\\ &= \frac{i\beta^\alpha}{(iy-\beta)\Gamma(\alpha)}x^\alpha e^{(iy-\beta)x}\Big|_0^\infty - \frac{i\alpha\beta^\alpha}{(iy-\beta)\Gamma(\alpha)}\int_0^\infty e^{(iy-\beta)x}x^{\alpha-1}dx\\ &= -\frac{i\alpha}{(iy-\beta)}\hat{f}(y).\end{aligned}$$

The solution to this differential equation with initial condition $\hat{f}(0) = 1$ is clearly $\hat{f}(y) = [\beta/(\beta - iy)]^\alpha$. ■

Example 17.3.4 (*Bilateral Exponential*). The exponential density $f(x) = e^{-x}1_{[0,\infty)}(x)$ reduces to the special case $\alpha = \beta = 1$ of the last example. Since the bilateral exponential density $e^{-|x|}/2$ can be expressed as $[f(x) + f(-x)]/2$, property (d) of Table 17.1 shows that it has Fourier transform

$$\begin{aligned}\frac{1}{2}\left[\hat{f}(y) + \hat{f}(-y)\right] &= \frac{1}{2(1-iy)} + \frac{1}{2(1+iy)}\\ &= \frac{1}{1+y^2}.\end{aligned}$$

Up to a normalizing constant, this is the standard Cauchy density. ■

Example 17.3.5 (*Hermite Polynomials*). The Hermite polynomial $H_n(x)$ can be expressed as

$$H_n(x) = (-1)^n e^{\frac{1}{2}x^2}\frac{d^n}{dx^n}e^{-\frac{1}{2}x^2}. \tag{1}$$

Indeed, if we expand the left-hand side of the identity

$$e^{-\frac{1}{2}(x-t)^2} = e^{-\frac{1}{2}x^2}e^{xt-\frac{1}{2}t^2} = \sum_{n=0}^{\infty} \frac{H_n(x)e^{-\frac{1}{2}x^2}}{n!}t^n$$

in a Taylor series about $t = 0$, then the coefficient of $t^n/n!$ is

$$\begin{aligned} H_n(x)e^{-\frac{1}{2}x^2} &= \frac{d^n}{dt^n}e^{-\frac{1}{2}(x-t)^2}\Big|_{t=0} \\ &= (-1)^n \frac{d^n}{dx^n}e^{-\frac{1}{2}(x-t)^2}\Big|_{t=0} \\ &= (-1)^n \frac{d^n}{dx^n}e^{-\frac{1}{2}x^2}. \end{aligned}$$

Example 17.3.2 and repeated application of property (g) of Table 17.1 therefore yield

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int e^{iyx} H_n(x)e^{-\frac{1}{2}x^2}dx &= \frac{1}{\sqrt{2\pi}} \int e^{iyx}(-1)^n \frac{d^n}{dx^n}e^{-\frac{1}{2}x^2}dx \qquad (2) \\ &= (iy)^n e^{-\frac{1}{2}y^2}. \end{aligned}$$

This Fourier transform will appear in our subsequent discussion of Edgeworth expansions. ■

17.4 Further Theory

We now delve more deeply into the theory of the Fourier transform.

Proposition 17.4.1 (Riemann–Lebesgue**).** *If the function $f(x)$ is integrable, then its Fourier transform $\hat{f}(y)$ is bounded, continuous, and tends to 0 as $|y|$ tends to ∞.*

Proof. The transform $\hat{f}(y)$ is bounded because

$$\begin{aligned} |\hat{f}(y)| &= \Big| \int e^{iyx} f(x)dx \Big| \\ &\leq \int |e^{iyx}||f(x)|dx \qquad (3) \\ &= \int |f(x)|dx. \end{aligned}$$

To prove continuity, let $\lim_{n\to\infty} y_n = y$. Then the sequence of functions $g_n(x) = e^{iy_n x} f(x)$ is bounded in absolute value by $|f(x)|$ and satisfies

$$\lim_{n\to\infty} g_n(x) = e^{iyx} f(x).$$

Hence, the dominated convergence theorem implies that

$$\lim_{n\to\infty} \int g_n(x)dx = \int e^{iyx} f(x)dx.$$

To prove the last assertion, we use the fact that the space of step functions with bounded support is dense in the space of integrable functions. Thus, given any $\epsilon > 0$, there exists a step function

$$g(x) = \sum_{j=1}^{m} c_j 1_{[x_{j-1}, x_j)}(x)$$

vanishing off some finite interval and satisfying $\int |f(x) - g(x)|dx < \epsilon$. The Fourier transform $\hat{g}(y)$ has the requisite behavior at ∞ because Example 17.3.1 allows us to calculate

$$\hat{g}(y) = \sum_{j=1}^{m} c_j e^{i\frac{1}{2}(x_{j-1}+x_j)y} \frac{\sin[\frac{1}{2}(x_j - x_{j-1})y]}{\frac{1}{2}y},$$

and this finite sum clearly tends to 0 as $|y|$ tends to ∞. The original transform $\hat{f}(y)$ exhibits the same behavior because the bound (3) entails the inequality

$$\begin{aligned} |\hat{f}(y)| &\leq |\hat{f}(y) - \hat{g}(y)| + |\hat{g}(y)| \\ &\leq \epsilon + |\hat{g}(y)|. \end{aligned}$$

This completes the proof. □

The Fourier transform can be inverted. If $g(x)$ is integrable, then

$$\check{g}(y) = \frac{1}{2\pi} \int e^{-iyx} g(x)dx$$

supplies the inverse Fourier transform of $g(x)$. This terminology is justified by the next proposition.

Proposition 17.4.2. *Let $f(x)$ be a bounded, continuous function. If $f(x)$ and $\hat{f}(y)$ are both integrable, then*

$$f(x) = \frac{1}{2\pi} \int e^{-iyx} \hat{f}(y)dy. \tag{4}$$

Proof. Consider the identities

$$\begin{aligned} \frac{1}{2\pi} \int e^{-iyx} e^{-\frac{y^2}{2\sigma^2}} \hat{f}(y)dy &= \frac{1}{2\pi} \int e^{-iyx} e^{-\frac{y^2}{2\sigma^2}} \int e^{iyu} f(u)du\, dy \\ &= \int f(u) \frac{1}{2\pi} \int e^{iy(u-x)} e^{-\frac{y^2}{2\sigma^2}} dy\, du \\ &= \int f(u) \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\sigma^2(u-x)^2}{2}} du \\ &= \frac{1}{\sqrt{2\pi}} \int f(x + \frac{v}{\sigma}) e^{-\frac{v^2}{2}} dv, \end{aligned}$$

which involve Example 17.3.2 and the change of variables $u = x + v/\sigma$. As σ tends to ∞, the last integral tends to

$$\frac{1}{\sqrt{2\pi}} \int f(x) e^{-\frac{v^2}{2}} dv = f(x),$$

while the original integral tends to

$$\frac{1}{2\pi} \int e^{-iyx} \lim_{\sigma \to \infty} e^{-\frac{y^2}{2\sigma^2}} \hat{f}(y) dy = \frac{1}{2\pi} \int e^{-iyx} \hat{f}(y) dy.$$

Equating these two limits yields the inversion formula (4). □

Example 17.4.1 (*Cauchy Distribution*). If $f(x)$ is the bilateral exponential density, then Proposition 17.4.2 and Example 17.3.4 show that the Cauchy density $1/[\pi(1+x^2)]$ has Fourier transform $e^{-|y|}$. ■

Proposition 17.4.3 (Parseval–Plancherel). *If either of the integrable functions $f(x)$ or $g(x)$ obeys the further assumptions of Proposition 17.4.2, then*

$$\int f(x) g(x)^* dx = \frac{1}{2\pi} \int \hat{f}(y) \hat{g}(y)^* dy. \tag{5}$$

In particular, when $f(x) = g(x)$,

$$\int |f(x)|^2 dx = \frac{1}{2\pi} \int |\hat{f}(y)|^2 dy. \tag{6}$$

Proof. If $\hat{f}(y)$ satisfies the assumptions of Proposition 17.4.2, then

$$\begin{aligned} \int f(x) g(x)^* dx &= \int \frac{1}{2\pi} \int e^{-iyx} \hat{f}(y) dy \, g(x)^* dx \\ &= \int \hat{f}(y) \frac{1}{2\pi} \int e^{-iyx} g(x)^* dx \, dy \\ &= \int \hat{f}(y) \frac{1}{2\pi} \Big[\int e^{iyx} g(x) dx \Big]^* dy \\ &= \frac{1}{2\pi} \int \hat{f}(y) \hat{g}(y)^* dy. \end{aligned}$$

With obvious modifications the same proof works if $\hat{g}(y)$ satisfies the assumptions of Proposition 17.4.2. □

Example 17.4.2 (*Computation of an Integral*). Let $f(x) = 1_{[-b,b]}(x)/(2b)$ be the uniform density. Then the Parseval–Plancherel relation (6) implies

$$\begin{aligned} \frac{1}{2b} &= \frac{1}{(2b)^2} \int_{-b}^{b} 1 \, dx \\ &= \frac{1}{2\pi} \int \Big| \frac{\sin(by)}{by} \Big|^2 dy. \end{aligned}$$

■

There are several definitions of the Fourier transform that differ only in how the factor of 2π and the sign of the argument of the complex exponential are assigned. We have chosen the definition that coincides with the characteristic function of a probability density. For some purposes the alternative definitions

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi}} \int e^{iyx} f(x)dx$$
$$\check{g}(x) = \frac{1}{\sqrt{2\pi}} \int e^{-iyx} g(y)dy$$

of the Fourier transform and its inverse are better. Obviously, the transform and its inverse are now more symmetrical. Also, the Parseval–Plancherel relation (5) simplifies to

$$\int f(x)g(x)^* dx = \int \hat{f}(y)\hat{g}(y)^* dy.$$

In other words, the Fourier transform now preserves inner products and norms on a subspace of the Hilbert space $\mathrm{L}^2(-\infty, \infty)$ of square-integrable functions. Such a transformation is said to be unitary. One can show that this subspace is dense in $\mathrm{L}^2(-\infty, \infty)$ and therefore that the Fourier transform extends uniquely to a unitary transformation from $\mathrm{L}^2(-\infty, \infty)$ onto itself [3, 4, 9]. Proof of these theoretical niceties would take us too far afield. Let us just add that norm preservation forces the Fourier transform of a function to be unique.

Our final theoretical topic is convolution. If $f(x)$ and $g(x)$ are integrable functions, then their convolution $f * g(x)$ is defined by

$$f * g(x) = \int f(x-u)g(u)du.$$

Doubtless, readers will recall that if $f(x)$ and $g(x)$ are the densities of independent random variables U and V, then $f * g(x)$ is the density of the sum $U + V$. The fundamental properties of convolution valid in the context of density functions carry over to the more general setting of integrable functions.

Proposition 17.4.4. *The convolution of two integrable functions $f(x)$ and $g(x)$ is integrable with Fourier transform $\hat{f}(y)\hat{g}(y)$. Furthermore, convolution is a commutative, associative, and linear operation.*

Proof. Integrability of $f * g(x)$ is a consequence of the calculation

$$\begin{aligned}\int |f * g(x)|dx &= \int \Big| \int f(x-u)g(u)du \Big| dx \\ &\leq \int \int |f(x-u)||g(u)|du\, dx \\ &= \int \int |f(x-u)|dx|g(u)|du\end{aligned}$$

$$\begin{aligned} &= \int \int |f(x)|dx|g(u)|du \\ &= \int |f(x)|dx \int |g(u)|du. \end{aligned}$$

The product form of the Fourier transform follows from

$$\begin{aligned} \int e^{iyx} f * g(x)dx &= \int e^{iyx} \int f(x-u)g(u)du\,dx \\ &= \int \int e^{iy(x-u)} f(x-u)dx e^{iyu} g(u)du \\ &= \int \int e^{iyx} f(x)dx e^{iyu} g(u)du \\ &= \hat{f}(y)\hat{g}(y). \end{aligned}$$

The remaining assertions are easy consequences of the result just established and the uniqueness of the Fourier transform. □

Example 17.4.3 (*Convolution of Cauchy Densities*)**.** Let $c_1, \ldots, c_n$ be positive constants and $X_1, \ldots, X_n$ an i.i.d. sequence of random variables with common Cauchy density $1/[\pi(1+x^2)]$ and characteristic function $e^{-|y|}$. Then the sum $c_1X_1 + \cdots + c_nX_n$ has characteristic function $e^{-c|y|}$ and Cauchy density $c/[\pi(c^2+x^2)]$, where $c = c_1 + \cdots + c_n$. ■

17.5 Edgeworth Expansions

An Edgeworth expansion is an asymptotic approximation to a density or distribution function [1, 4, 6]. The main ideas can be best illustrated by considering the proof of the central limit theorem for i.i.d. random variables $X_1, X_2, \ldots$ with common mean μ, variance σ^2, and density $f(x)$. Assuming that $f(x)$ possesses a moment generating function, we can write its characteristic function as

$$\hat{f}(y) = \sum_{j=0}^{\infty} \frac{\mu_j}{j!}(iy)^j = \exp\left[\sum_{j=1}^{\infty} \frac{\kappa_j}{j!}(iy)^j\right],$$

where μ_j and κ_j are the jth moment and jth cumulant of $f(x)$, respectively. The moment series for $\hat{f}(y)$ follows from repeated application of property (f) of Table 17.1 with y set equal to 0. Cumulants are particularly handy in this context because Proposition 17.4.4 implies that the jth cumulant of the sum $S_n = \sum_{i=1}^n X_i$ is just $n\kappa_j$. The identities $\kappa_1 = \mu_1$ and $\kappa_2 = \sigma^2$ hold in general. For notational convenience, we let $\rho_j = \kappa_j/\sigma^j$.

Owing to properties (b), (d), and (h) of Table 17.1, the characteristic function of the standardized sum $T_n = (S_n - n\mu)/(\sigma\sqrt{n})$ reduces to

$$\begin{aligned} e^{-\frac{i\sqrt{n}\mu y}{\sigma}} \hat{f}\Big(\frac{y}{\sigma\sqrt{n}}\Big)^n &= \exp\Big[-\frac{y^2}{2} + n\sum_{j=3}^{\infty} \frac{\rho_j}{n^{\frac{j}{2}} j!}(iy)^j\Big] \\ &= e^{-\frac{y^2}{2}} e^{\frac{\rho_3(iy)^3}{6\sqrt{n}} + \frac{\rho_4(iy)^4}{24n} + O(n^{-3/2})} \\ &= e^{-\frac{y^2}{2}}\Big[1 + \frac{\rho_3(iy)^3}{6\sqrt{n}} + \frac{\rho_4(iy)^4}{24n} + \frac{\rho_3^2(iy)^6}{72n} + O(n^{-\frac{3}{2}})\Big]. \end{aligned}$$

Formal inversion of this Fourier transform taking into account equation (2) yields the asymptotic expansion

$$\phi(x)\Big[1 + \frac{\rho_3 H_3(x)}{6\sqrt{n}} + \frac{\rho_4 H_4(x)}{24n} + \frac{\rho_3^2 H_6(x)}{72n} + O(n^{-\frac{3}{2}})\Big] \tag{7}$$

for the density of T_n. Here $\phi(x)$ denotes the standard normal density. To approximate the distribution function of T_n in terms of the standard normal distribution function $\Phi(x)$, note that integration of the Hermite polynomial identity (1) gives

$$\int_{-\infty}^{x} \phi(u) H_n(u)\, du = -\phi(x) H_{n-1}(x).$$

Applying this fact to the integration of expression (7) yields the asymptotic expansion

$$\Phi(x) - \phi(x)\Big[\frac{\rho_3 H_2(x)}{6\sqrt{n}} + \frac{\rho_4 H_3(x)}{24n} + \frac{\rho_3^2 H_5(x)}{72n} + O(n^{-\frac{3}{2}})\Big] \tag{8}$$

for the distribution function of T_n. It is worth stressing that the formal manipulations leading to the expansions (7) and (8) can be made rigorous [4, 6].

Both of the expansions (7) and (8) suggest that the rate of convergence of T_n to normality is governed by the $O(n^{-1/2})$ correction term. However, this pessimistic impression is misleading at $x = 0$ because $H_3(0) = 0$. (A quick glance at the recurrence relation (11) of Chapter 15 confirms that $H_n(x)$ is even for n even and odd for n odd.) The device known as exponential tilting exploits this peculiarity [1].

If we let $K(t) = \sum_{j=1}^{\infty} \kappa_j t^j/j!$ be the cumulant generating function of $f(x)$ and $g(x)$ be the density of S_n, then we tilt $g(x)$ to the density $e^{tx - nK(t)} g(x)$. Because $e^{nK(t)}$ is the moment generating function of S_n, we find that

$$\int e^{sx} e^{tx - nK(t)} g(x)\, dx = e^{nK(s+t) - nK(t)}.$$

This calculation confirms that $e^{tx - nK(t)} g(x)$ is a probability density with moment generating function $e^{nK(s+t) - nK(t)}$. The tilted density has mean $nK'(t)$ and variance $nK''(t)$. We can achieve an arbitrary mean x_0 by

choosing t_0 to be the solution of the equation $nK'(t_0) = x_0$. In general, this equation must be solved numerically. Once t_0 is chosen, we can approximate the standardized tilted density

$$\sqrt{nK''(t_0)}e^{t_0[\sqrt{nK''(t_0)}x+x_0]-nK(t_0)}g\Big(\sqrt{nK''(t_0)}x + x_0\Big)$$

at $x = 0$ by the asymptotic expansion (7). To order $O(n^{-1})$ this gives

$$\sqrt{nK''(t_0)}e^{t_0x_0-nK(t_0)}g(x_0) = \phi(0)\Big[1 + O(n^{-1})\Big]$$

or

$$g(x_0) = \frac{e^{-t_0x_0+nK(t_0)}}{\sqrt{2\pi nK''(t_0)}}\Big[1 + O(n^{-1})\Big]. \tag{9}$$

This result is also called a saddlepoint approximation.

Further terms can be included in the saddlepoint approximation if we substitute the appropriate normalized cumulants

$$\rho_j(t_0) = \frac{nK^{(j)}(t_0)}{[nK''(t_0)]^{\frac{j}{2}}}$$

of the tilted density in the Edgeworth expansion (7). Once we determine the required coefficients $H_3(0) = 0$, $H_4(0) = 3$, and $H_6(0) = -15$ from recurrence relation (11) of Chapter 15, it is obvious that

$$g(x_0) = \frac{e^{-t_0x_0+nK(t_0)}}{\sqrt{2\pi nK''(t_0)}}\Big[1 + \frac{3\rho_4(t_0) - 5\rho_3^2(t_0)}{24n} + O(n^{-\frac{3}{2}})\Big]. \tag{10}$$

Example 17.5.1 (*Spread of a Random Sample from the Exponential*). If $X_1, \ldots, X_{n+1}$ are independent, exponentially distributed random variables with common mean 1, then the spread $X_{(n+1)} - X_{(1)}$ has density

$$n\sum_{k=1}^{n}(-1)^{k-1}\binom{n-1}{k-1}e^{-kx}. \tag{11}$$

This is also the density of the sum $Y_1+\cdots+Y_n$ of independent, exponentially distributed random variables Y_j with respective means $1, 1/2, \ldots, 1/n$. (A proof of these obscure facts is sketched in Problem I.13.13 of [4]). For the sake of comparison, we compute the Edgeworth approximation (7) and the two saddlepoint approximations (9) and (10) to the exact density (11). Because the Y_j have widely different variances, a naive normal approximation based on the central limit theorem is apt to be poor.

A brief calculation shows that the cumulant generating function of the sum $Y_1+\cdots+Y_n$ is $nK(t) = -\sum_{j=1}^{n}\ln(1-t/j)$. The equation $nK'(t_0) = x_0$ becomes $\sum_{j=1}^{n} 1/(j-t_0) = x_0$, which obviously must be solved numerically.

TABLE 17.2. Saddlepoint Approximations to the Spread Density

x_0	**Exact** $g(x_0)$	**Error** (7)	**Error** (9)	**Error** (10)
.50000	.00137	−.04295	−.00001	−.00001
1.00000	.05928	−.04010	−.00027	−.00024
1.50000	.22998	.04979	.00008	.00004
2.00000	.36563	.09938	.00244	.00211
2.50000	.37974	.05913	.00496	.00439
3.00000	.31442	−.00245	.00550	.00499
3.50000	.22915	−.03429	.00423	.00394
4.00000	.15508	−.03588	.00242	.00235
4.50000	.10046	−.02356	.00098	.00103
5.00000	.06340	−.00992	.00012	.00022
5.50000	.03939	−.00136	−.00026	−.00017
6.00000	.02424	.00167	−.00037	−.00029
6.50000	.01483	.00209	−.00035	−.00029
7.00000	.00904	.00213	−.00028	−.00024
7.50000	.00550	.00220	−.00021	−.00018
8.00000	.00334	.00203	−.00014	−.00013
8.50000	.00203	.00160	−.00010	−.00009
9.00000	.00123	.00112	−.00006	−.00006

The kth cumulant of the tilted density is

$$nK^{(k)}(t_0) = (k-1)! \sum_{j=1}^{n} \frac{1}{(j-t_0)^k}.$$

Table 17.2 displays the exact density (11) and the errors (exact values minus approximate values) committed in using the Edgeworth expansion (7) and the two saddlepoint expansions (9) and (10) when $n = 10$. Note that we apply equation (7) at the standardized point

$$x = [x_0 - nK'(0)]/\sqrt{nK''(0)}$$

and divide the result of the approximation to the standardized density by $\sqrt{nK''(0)}$.

Both saddlepoint expansions clearly outperform the Edgeworth expansion except very close to the mean $\sum_{j=1}^{10} 1/j \approx 2.93$. Indeed, it is remarkable how well the saddlepoint expansions do considering how far this example is from the ideal of a sum of i.i.d. random variables. The refined saddlepoint expansion (10) is an improvement over the ordinary saddlepoint expansion (9) in the tails of the density but not necessarily in the center. Daniels [2] considers variations on this problem involving pure birth processes. ■

17.6 Problems

1. Verify the first five entries in Table 17.1.

2. For an even integrable function $f(x)$, show that

$$\hat{f}(y) = 2\int_0^\infty \cos(yx)f(x)dx,$$

and for an odd integrable function $g(x)$, show that

$$\hat{g}(y) = 2i\int_0^\infty \sin(yx)g(x)dx.$$

Conclude that (a) $\hat{f}(y)$ is even, (b) $\hat{f}(y)$ is real if $f(x)$ is real, and (c) $\hat{g}(y)$ is odd.

3. If $f(x)$ is integrable, then define

$$\mathrm{S}f(x) = f(\ln x)$$

for $x > 0$. Show that S is a linear mapping satisfying

(a) $\int_0^\infty |\mathrm{S}f(x)|x^{-1}dx < \infty$,
(b) $\mathrm{S}(f * g)(x) = \int_0^\infty \mathrm{S}f(xz^{-1})\mathrm{S}g(z)z^{-1}dz$,
(c) $\hat{f}(y) = \int_0^\infty \mathrm{S}f(x)x^{iy}x^{-1}dx$.

The function $\int_0^\infty h(x)x^{iy}x^{-1}dx$ defines the Mellin transform of $h(x)$.

4. Suppose $f(x)$ is integrable and $\hat{f}(y) = c\sqrt{2\pi}f(y)$ for some constant c. Prove that either $f(x) = 0$ for all x or c is drawn from the set $\{1, i, -1, -i\}$ of fourth roots of unity. (Hint: Take the Fourier transform of $f(x)$.)

5. Compute

$$\lim_{\alpha \to 0+} \int e^{-\alpha x^2} \frac{\sin(\lambda x)}{x} dx$$

for λ real. (Hint: Use the Parseval–Plancherel relation.)

6. Find a random variable X symmetrically distributed around 0 such that X cannot be represented as $X = Y - Z$ for i.i.d. random variables Y and Z. (Hint: Assuming Y and Z possess a density function, demonstrate that the Fourier transform of the density of X must be nonnegative.)

7. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables that has common density $f(x)$ and is independent of the integer-valued random variable $N \geq 0$. If N has generating function

$$G(s) = \sum_{n=0}^{\infty} \Pr(N = n)s^n,$$

then show that the density of the random sum $\sum_{i=1}^{N} X_i$ has Fourier transform $G[\hat{f}(y)]$.

8. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Show that the saddlepoint approximation (9) to the density of $S_n = \sum_{j=1}^n X_j$ is exact.

9. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with mean 1. Show that the saddlepoint approximation (9) to the density of $S_n = \sum_{j=1}^n X_j$ is exact up to Stirling's approximation.

10. Let $X_1, \ldots, X_n$ be a random sample from a member of an exponential family of densities $f(x|\theta) = h(x)e^{\theta u(x) - \gamma(\theta)}$. Show that the saddlepoint approximation (9) to the density of $S_n = \sum_{j=1}^n u(X_j)$ at x_0 reduces to

$$\frac{e^{n[\gamma(\theta_0+\theta)-\gamma(\theta)]-\theta_0 x_0}}{\sqrt{2\pi n\gamma''(\theta_0+\theta)}}\Big[1 + O(n^{-1})\Big],$$

where θ_0 satisfies the equation $n\gamma'(\theta_0 + \theta) = x_0$.

11. Compute the Edgeworth approximation (8) to the distribution function of a sum of i.i.d. Poisson random variables with unit means. Compare your results for sample sizes $n = 4$ and $n = 8$ to the exact distribution function and, if available, to the values in Table 4.2 of [1]. Note that in computing $\Pr(S_n \le z)$ in this discrete case it is wise to incorporate a continuity correction by applying the Edgeworth approximation at the point $x = (z - n\mu + 1/2)/(\sigma\sqrt{n})$.

References

[1] Barndorff-Nielsen OE, Cox DR (1989) *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London

[2] Daniels HE (1982) The saddlepoint approximation for a general birth process. *J Appl Prob* 19:20–28

[3] Dym H, McKean HP (1972) *Fourier Series and Integrals.* Academic Press, New York

[4] Feller W (1971) *An Introduction to Probability Theory and its Applications, Vol 2*, 2nd ed. Wiley, New York

[5] Folland GB (1992) *Fourier Analysis and its Applications.* Wadsworth and Brooks/Cole, Pacific Grove, CA

[6] Hall P (1992) *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York

[7] Körner TW (1988) *Fourier Analysis.* Cambridge University Press, Cambridge

[8] Lighthill MJ (1958) *An Introduction to Fourier Analysis and Generalized Functions.* Cambridge University Press, Cambridge

[9] Rudin W (1973) *Functional Analysis.* McGraw-Hill, New York

18
The Finite Fourier Transform

18.1 Introduction

In previous chapters we have met Fourier series and the Fourier transform. These are both incarnations of Fourier analysis on a commutative group, namely the unit circle and the real line under addition. In this chapter we study Fourier analysis in the even simpler setting of the additive group of integers modulo a fixed positive integer n [6, 12]. Here, for obvious reasons, the Fourier transform is called the finite Fourier transform. Although the finite Fourier transform has many interesting applications in abstract algebra, combinatorics, number theory, and complex variables [8], we view it mainly as a tool for approximating Fourier series. Computation of finite Fourier transforms is done efficiently by an algorithm known as the fast Fourier transform [1, 3, 5, 9, 13, 15]. Although it was discovered by Gauss, the fast Fourier transform has come into prominence only with the advent of modern computing. As an indication of its critical role in many scientific and engineering applications, it is often implemented in hardware rather than software.

In this chapter we first study the operational properties of the finite Fourier transform. With minor differences these parallel the properties of Fourier series and the ordinary Fourier transform. We then derive the fast Fourier transform for any highly composite number n. In many applications n is a power of 2, but this choice is hardly necessary. Once we have developed the fast Fourier transform, we discuss applications to time series [1, 2, 4, 9, 11] and other areas of statistics.

18.2 Basic Properties

Periodic sequences $\{c_j\}_{j=-\infty}^{\infty}$ of period n constitute the natural domain of the finite Fourier transform. The transform of such a sequence is defined by

$$\hat{c}_k = \frac{1}{n}\sum_{j=0}^{n-1} c_j e^{-2\pi i \frac{jk}{n}}. \tag{1}$$

From this definition it follows immediately that the finite Fourier transform is linear and maps periodic sequences into periodic sequences with the same period. The inverse transform turns out to be

$$\check{d}_j = \sum_{k=0}^{n-1} d_k e^{2\pi i \frac{jk}{n}}. \tag{2}$$

It is fruitful to view each of these operations as a matrix times vector multiplication. Thus, if we let $u_n = e^{2\pi i/n}$ denote the principal nth root of unity, then the finite Fourier transform represents multiplication by the matrix (u_n^{-kj}/n) and the inverse transform multiplication by the matrix (u_n^{jk}). To warrant the name "inverse transform," the second matrix should be the inverse of the first. Indeed, we have

$$\begin{aligned}\sum_{l=0}^{n-1} u_n^{jl}\frac{1}{n}u_n^{-kl} &= \frac{1}{n}\sum_{l=0}^{n-1} u_n^{(j-k)l}\\ &= \begin{cases}\frac{1}{n}\frac{1-u_n^{(j-k)n}}{1-u_n^{j-k}} & j \neq k \bmod n\\ \frac{1}{n}n & j = k \bmod n\end{cases}\\ &= \begin{cases}0 & j \neq k \bmod n\\ 1 & j = k \bmod n.\end{cases}\end{aligned}$$

More symmetry in the finite Fourier transform (1) and its inverse (2) can be achieved by replacing the factor $1/n$ in the transform by the factor $1/\sqrt{n}$. The inverse transform then includes the $1/\sqrt{n}$ factor as well, and the matrix $(u_n^{-kj}/\sqrt{n})$ is unitary.

We modify periodic sequences of period n by convolution, translation, reversion, and stretching. The convolution of two periodic sequences c_j and d_j is the sequence

$$c * d_k = \sum_{j=0}^{n-1} c_{k-j}d_j = \sum_{j=0}^{n-1} c_j d_{k-j}$$

with the same period. The translate of the periodic sequence c_j by index r is the periodic sequence $T_r c_j$ defined by $T_r c_j = c_{j-r}$. Thus, the operator T_r translates a sequence r places to the right. The reversion operator R takes a sequence c_j into $Rc_j = c_{-j}$. Finally, the stretch operator S_r interpolates

$r-1$ zeros between every pair of adjacent entries of a sequence c_j. In symbols,

$$S_r c_j = \begin{cases} c_{\frac{j}{r}} & r \mid j \\ 0 & r \not| j, \end{cases}$$

where $r \mid j$ indicates r divides j without remainder. The sequence $S_r c_j$ has period rn, not n. For instance, if $n = 2$ and $r = 2$, the periodic sequence $\ldots, 1, 2, 1, 2 \ldots$ becomes $\ldots, 1, 0, 2, 0, 1, 0, 2, 0, \ldots$.

Proposition 18.2.1. *The finite Fourier transform satisfies the rules:*

(a) $\widehat{c * d}_k = n\hat{c}_k\hat{d}_k$,
(b) $\widehat{T_r c}_k = u_n^{-rk}\hat{c}_k$,
(c) $\widehat{Rc}_k = R\hat{c}_k = \widehat{c^*}_k^*$,
(d) $\widehat{S_r c}_k = \frac{\hat{c}_k}{r}$.

In (d) the transform on the left has period rn.

Proof. To prove rule (d), note that

$$\begin{aligned} \widehat{S_r c}_k &= \frac{1}{rn}\sum_{j=0}^{rn-1} S_r c_j u_{rn}^{-jk} \\ &= \frac{1}{rn}\sum_{l=0}^{n-1} c_{\frac{lr}{r}} u_{rn}^{-lrk} \\ &= \frac{1}{rn}\sum_{l=0}^{n-1} c_l u_n^{-lk} \\ &= \frac{\hat{c}_k}{r}. \end{aligned}$$

Verification of rules (a) through (c) is left to the reader. □

18.3 Derivation of the Fast Fourier Transform

The naive approach to computing the finite Fourier transform (1) takes $3n^2$ arithmetic operations (additions, multiplications, and complex exponentiations). The fast Fourier transform accomplishes the same task in $O(n \log n)$ operations when n is a power of 2. Proposition 18.2.1 lays the foundation for deriving this useful and clever result.

Consider a sequence c_j of period n, and suppose n factors as $n = rs$. For $k = 0, 1, \ldots, r-1$, define related sequences $c_j^{(k)}$ according to the recipe $c_j^{(k)} = c_{jr+k}$. Each of these secondary sequences has period s. We now argue

that we can recover the primary sequence through

$$c_j = \sum_{k=0}^{r-1} T_k S_r c_j^{(k)}. \tag{3}$$

In fact, $T_k S_r c_j^{(k)} = 0$ unless $r \mid j-k$. The condition $r \mid j-k$ occurs for exactly one value of k between 0 and $r-1$. For the chosen k,

$$\begin{aligned} T_k S_r c_j^{(k)} &= c_{\frac{j-k}{r}}^{(k)} \\ &= c_{\frac{j-k}{r} r+k} \\ &= c_j. \end{aligned}$$

In view of properties (b) and (d) of Proposition 18.2.1, taking the finite Fourier transform of equation (3) gives

$$\begin{aligned} \hat{c}_j &= \sum_{k=0}^{r-1} u_n^{-kj} \widehat{S_r c^{(k)}}_j \\ &= \sum_{k=0}^{r-1} u_n^{-kj} \frac{1}{r} \widehat{c^{(k)}}_j. \end{aligned} \tag{4}$$

Now let $Op(n)$ denote the number of operations necessary to compute a finite Fourier transform of period n. From equation (4) it evidently takes $3r$ operations to compute each $\hat{c}_j$ once the $\widehat{c^{(k)}}_j$ are computed. Since there are n numbers $\hat{c}_j$ to compute and r sequences $c_j^{(k)}$, it follows that

$$Op(n) = 3nr + rOp(s). \tag{5}$$

If r is prime but s is not, then the same procedure can be repeated on each $c_j^{(k)}$ to further reduce the amount of arithmetic. A simple inductive argument based on (5) that splits off one prime factor at a time yields

$$Op(n) = 3n(p_1 + \cdots + p_d),$$

where $n = p_1 \cdots p_d$ is the prime factorization of n. In particular, if $n = 2^d$, then $Op(n) = 6n \log_2 n$. In this case, it is noteworthy that all computations can be done in place without requiring computer storage beyond that allotted to the original vector $(c_0, \ldots, c_{n-1})^t$ [3, 9, 13, 15].

18.4 Approximation of Fourier Series Coefficients

The finite Fourier transform can furnish approximations to the Fourier coefficients of a periodic function $f(x)$. If $f(x)$ has period 1, then its kth Fourier coefficient c_k can be approximated by

$$c_k = \int_0^1 f(x) e^{-2\pi i k x} dx$$

$$\begin{aligned}
&\approx \frac{1}{n}\sum_{j=0}^{n-1} f\Big(\frac{j}{n}\Big)e^{-2\pi i\frac{jk}{n}} \\
&= \hat{b}_k,
\end{aligned}$$

where $b_j = f(j/n)$ and n is some large, positive integer. Because the transformed values $\hat{b}_k$ are periodic, only n of them are distinct, say $\hat{b}_{-n/2}$ through $\hat{b}_{n/2-1}$ for n even.

An important question is how well $\hat{b}_k$ approximates c_k. To assess the error, suppose that $\sum_k |c_k| < \infty$ and that the Fourier series of $f(x)$ converges to $f(x)$ at the points j/n for $j = 0, \ldots, n-1$. The calculation

$$\begin{aligned}
\hat{b}_k &= \frac{1}{n}\sum_{j=0}^{n-1} u_n^{-jk} f\Big(\frac{j}{n}\Big) \\
&= \frac{1}{n}\sum_{j=0}^{n-1} u_n^{-jk} \sum_m c_m u_n^{jm} \\
&= \sum_m c_m \frac{1}{n}\sum_{j=0}^{n-1} u_n^{j(m-k)} \\
&= \sum_m c_m \begin{cases} 1 & m = k \bmod n \\ 0 & m \neq k \bmod n \end{cases}
\end{aligned}$$

implies that

$$\begin{aligned}
\hat{b}_k - c_k &= \sum_{l\neq 0} c_{ln+k} \qquad (6)\\
&= \cdots + c_{-2n+k} + c_{-n+k} + c_{n+k} + c_{2n+k} + \cdots.
\end{aligned}$$

If the Fourier coefficients c_j decline sufficiently rapidly to 0 as $|j|$ tends to ∞, then the error $\hat{b}_k - c_k$ will be small for $-n/2 \leq k \leq n/2 - 1$. Problems (6), (7), and (8) explore this question in more depth.

Example 18.4.1 (*Number of Particles in a Branching Process*)**.** In a branching process the probability that there are k particles at generation j is given by the coefficient p_{jk} of s^k in the probability generating function $P_j(s) = \sum_{k=0}^{\infty} p_{jk}s^k$ [10].The generating function $P_j(s)$ is calculated from an initial progeny generating function $P_1(s) = P(s) = \sum_{k=0}^{\infty} p_k s^k$ by taking its j-fold functional composition

$$P_j(s) = \overbrace{P(P(\cdots(P(}^{j \text{ times}} s))\cdots)). \qquad (7)$$

The progeny generating function $P(s)$ summarizes the distribution of the number of progeny left at generation 1 by the single ancestral particle at generation 0. In general, it is impossible to give explicit expressions for

the p_{jk}. However, these can be easily computed numerically by the finite Fourier transform. If we extend $P_j(s)$ to the unit circle by the formula

$$P_j(e^{2\pi it}) = \sum_{k=0}^{\infty} p_{jk} e^{2\pi ikt},$$

then we can view the p_{jk} as Fourier series coefficients and recover them as discussed above. Fortunately, evaluation of $P_j(e^{2\pi it})$ at the points $t = m/n$, $m = 0, \ldots, n-1$, is straightforward under the functional composition rule (7). As a special case, consider $P(s) = \frac{1}{2} + \frac{s^2}{2}$. Then the algebraically formidable

$$\begin{aligned} P_4(s) = &\frac{24,305}{32,768} + \frac{445}{4,096}s^2 + \frac{723}{8,192}s^4 + \frac{159}{4,096}s^6 + \frac{267}{16,384}s^8 \\ &+ \frac{19}{4,096}s^{10} + \frac{11}{8,192}s^{12} + \frac{1}{4,096}s^{14} + \frac{1}{32,768}s^{16} \end{aligned}$$

can be derived by symbolic algebra programs such as Maple. Alternatively, the finite Fourier transform approximation

$$\begin{aligned} P_4(s) = &\, 0.74172974 + 0.10864258s^2 + 0.08825684s^4 + 0.03881836s^6 \\ &+ \, 0.01629639s^8 + 0.00463867s^{10} + 0.00134277s^{12} \\ &+ \, 0.00024414s^{14} + 0.00003052s^{16}[6bp] \end{aligned}$$

is trivial to compute and is exact up to machine precision if we take the period $n > 16$. ∎

Example 18.4.2 (*Differentiation of an Analytic Function*)**.** If the function $f(x)$ has a power series expansion $\sum_{j=0}^{\infty} a_j x^j$ converging in a disc $\{x : |x| < r\}$ centered at 0 in the complex plane, then we can approximate the derivatives $f^{(j)}(0) = j!a_j$ by evaluating $f(x)$ on the boundary of a small circle of radius $h < r$. This is accomplished by noting that the periodic function $t \to f(he^{2\pi it})$ has Fourier series expansion

$$f(he^{2\pi it}) = \sum_{j=0}^{\infty} a_j h^j e^{2\pi ijt}.$$

Thus, if we take the finite Fourier transform $\hat{b}_k$ of the sequence $b_j = f(hu_n^j)$, equation (6) mutates into

$$\hat{b}_k - a_k h^k = \sum_{l=1}^{\infty} a_{ln+k} h^{ln+k} = O\Big(h^{n+k}\Big)$$

for $0 \leq k \leq n-1$ under fairly mild conditions on the coefficients a_j. Rearranging this equation gives the derivative approximation

$$f^{(k)}(0) = \frac{k!\hat{b}_k}{h^k} + O\Big(h^n\Big) \tag{8}$$

highlighted in [8].

The two special cases

$$f'(0) = \frac{1}{2h}\Big[f(h) - f(-h)\Big] + O\Big(h^2\Big)$$
$$f''(0) = \frac{1}{2h^2}\Big[f(h) - f(ih) + f(-h) - f(-ih)\Big] + O\Big(h^4\Big)$$

of equation (8) when $n = 2$ and $n = 4$, respectively, deserve special mention. If h is too small, subtractive cancellation causes roundoff error in both of these equations. For a real analytic function $f(x)$, there is an elegant variation of the central difference approximation $f'(0) \approx \frac{1}{2h}\Big[f(h) - f(-h)\Big]$ that eliminates roundoff error. To derive this improved approximation, we define $g(x) = f(ix)$ and exploit the fact that $g'(0) = if'(0)$. Because the coefficients a_j are real, the identity $f(-ih) = f(ih)^*$ holds and allows us to deduce that

$$\begin{aligned} f'(0) &= \frac{1}{i}g'(0) \\ &= \frac{1}{2ih}\Big[g(h) - g(-h)\Big] + O\Big(h^2\Big) \\ &= \frac{1}{2ih}\Big[f(ih) - f(-ih)\Big] + O\Big(h^2\Big) \\ &= \frac{1}{2ih}\Big[f(ih) - f(ih)^*\Big] + O\Big(h^2\Big) \\ &= \frac{1}{h}\mathrm{Im} f(ih) + O\Big(h^2\Big). \end{aligned}$$

The approximation $f'(0) \approx \frac{1}{h}\mathrm{Im} f(ih)$ not only eliminates the roundoff error jeopardizing the central difference approximation, but it also requires one less function evaluation. Of course, the latter advantage is partially offset by the necessity of using complex arithmetic.

As an example, consider the problem of differentiating the analytic function $f(x) = e^x/(\sin^3 x + \cos^3 x)$ at $x = 1.5$. Table 18.1 reproduces a single-precision numerical experiment from reference [14] and shows the lethal effects of roundoff in the central difference formula. The formula $f'(x) \approx \frac{1}{h}\mathrm{Im} f(x + ih)$ approximates the true value $f'(1.5) = 3.62203$ extremely well and is stable even for small values of h. This stability makes it is possible to circumvent the delicate question of finding the right h to balance truncation and roundoff errors. ■

TABLE 18.1. Numerical Derivatives of $f(x) = e^x/(\sin^3 x + \cos^3 x)$

h	$\frac{1}{2h}[f(x+h) - f(x-h)]$	$\frac{1}{h}\text{Im} f(x+ih)$
10^{-2}	3.62298	3.62109
10^{-3}	3.62229	3.62202
10^{-4}	3.62158	3.62203
10^{-5}	3.60012	3.62203
10^{-6}	3.57628	3.62203
10^{-7}	4.76837	3.62203
10^{-8}	0.00000	3.62203
10^{-9}	0.00000	3.62203

18.5 Convolution

Proposition 18.2.1 suggests a fast method of computing the convolution of two sequences c_j and d_j of period n; namely, compute the transforms $\hat{c}_k$ and $\hat{d}_k$ via the fast Fourier transform, multiply pointwise to form the product transform $n\hat{c}_k\hat{d}_k$, and then invert the product transform via the fast inverse Fourier transform. This procedure requires on the order of $O(n \ln n)$ operations, whereas the naive evaluation of a convolution requires on the order of n^2 operations unless one of the sequences consists mostly of zeros. Here are some examples where fast convolution is useful.

Example 18.5.1 (*Repeated Differencing*)**.** The classical finite difference $\Delta c_j = c_{j+1} - c_j$ corresponds to convolution against the sequence

$$d_j = \begin{cases} 1 & j = -1 \bmod n \\ -1 & j = 0 \bmod n \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the sequence $\Delta^r c_j$ is sent into the sequence $(u_n^k - 1)^r \hat{c}_k$ under the finite Fourier transform. ■

Example 18.5.2 (*Data Smoothing*)**.** In many statistical applications, observations $x_0, \ldots, x_{n-1}$ are smoothed by a linear filter w_j. Smoothing creates a new sequence y_j according to the recipe

$$y_j = w_r x_{j-r} + w_{r-1} x_{j-r+1} + \cdots + w_{-r+1} x_{j+r-1} + w_{-r} x_{j+r}.$$

For instance, $y_j = \frac{1}{3}(x_{j-1} + x_j + x_{j+1})$ replaces x_j by a moving average of x_j and its two nearest neighbors. For the convolution paradigm to make sense, we must extend x_j and w_j to be periodic sequences of period n and pad w_j with zeros so that $w_{r+1} = \cdots = w_{n-r-1} = 0$. In many situations it is natural to require the weights to satisfy $w_j \geq 0$ and $\sum_{j=-r}^{r} w_j = 1$. Problem 3 provides the finite Fourier transforms of two popular smoothing sequences. ■

Example 18.5.3 (*Multiplication of Generating Functions*). One can write the generating function $R(s)$ of the sum $X + Y$ of two independent, nonnegative, integer-valued random variables X and Y as the product $R(s) = P(s)Q(s)$ of the generating function $P(s) = \sum_{j=0}^{\infty} p_j s^j$ of X and the generating function $Q(s) = \sum_{j=0}^{\infty} q_j s^j$ of Y. The coefficients of $R(s)$ are given by the convolution formula

$$r_k = \sum_{j=0}^{k} p_j q_{k-j}.$$

Assuming that the p_j and q_j are 0 or negligible for $j \geq m$, we can view the two sequences as having period $n = 2m$ provided we set $p_j = q_j = 0$ for $j = m, \ldots, n-1$. Introducing these extra zeros makes it possible to write

$$r_k = \sum_{j=0}^{n-1} p_j q_{k-j} \tag{9}$$

without embarrassment. The r_j returned by the suggested procedure are correct in the range $0 \leq j \leq m-1$. Clearly, the same process works if $P(s)$ and $Q(s)$ are arbitrary polynomials of degree $m-1$ or less. ■

Example 18.5.4 (*Multiplication of Large Integers*). If p and q are large integers, then we can express them in base b as

$$\begin{aligned} p &= p_0 + p_1 b + \cdots + p_{m-1} b^{m-1} \\ q &= q_0 + q_1 b + \cdots + q_{m-1} b^{m-1}, \end{aligned}$$

where each $0 \leq p_j \leq b-1$ and $0 \leq q_j \leq b-1$. We can represent the product $r = pq$ as $r = \sum_{k=0}^{n-1} r_k b^k$ with the r_k given by equation (9) and $n = 2m$. Although a given r_k may not satisfy the constraint $r_k \leq b-1$, once we replace it by its representation in base b and add and carry appropriately, we quickly recover the base b representation of r. For very large integers, computing r via the fast Fourier transform represents a large savings. ■

Example 18.5.5 (*Fast Solution of a Renewal Equation*). The discrete renewal equation

$$u_n = a_n + \sum_{m=0}^{n} f_m u_{n-m} \tag{10}$$

arises in many applications of probability theory [7]. Here f_n is a known discrete probability density with $f_0 = 0$, and a_n is a known sequence with partial sums converging absolutely to $\sum_{n=0}^{\infty} a_n = a$. Beginning with the initial value $u_0 = a_0$, it takes on the order of n^2 operations to compute $u_0, \ldots, u_n$ recursively via the convolution equation (10).

If we multiply both sides of (10) by s^n and sum on n, then we get the equation

$$U(s) = A(s) + F(s)U(s), \tag{11}$$

involving the generating functions

$$U(s) = \sum_{n=0}^{\infty} u_n s^n, \quad A(s) = \sum_{n=0}^{\infty} a_n s^n, \quad F(s) = \sum_{n=0}^{\infty} f_n s^n.$$

The solution

$$U(s) = \frac{A(s)}{1 - F(s)}$$

of equation (11) has a singularity at $s = 1$. This phenomenon is merely a reflection of the fact that the u_n do not tend to 0 as n tends to ∞. Indeed, under a mild hypothesis on the coefficients f_n, one can show that $\lim_{n\to\infty} u_n = a/\mu$, where $\mu = \sum_{n=0}^{\infty} n f_n$ [7]. The required hypothesis on the f_n says that the set $\{n: f_n > 0\}$ has greatest common divisor 1. Equivalently, the only complex number s satisfying both $F(s) = 1$ and $|s| = 1$ is $s = 1$. (See Problem 12.)

These observations suggest that it would be better to estimate the coefficients $v_n = u_n - a/\mu$ of the generating function

$$\begin{aligned} V(s) &= U(s) - \frac{a}{\mu(1-s)} \\ &= \frac{A(s)\mu(1-s) - a[1-F(s)]}{[1-F(s)]\mu(1-s)}. \end{aligned}$$

A double application of l'Hopital's rule implies that

$$\lim_{s\to 1} V(s) = \frac{aF''(1)}{2\mu^2} - \frac{A'(1)}{\mu}.$$

In other words, we have removed the singularity of $U(s)$ in forming $V(s)$. Provided $F(s)$ satisfies the greatest common divisor hypothesis, we can now recover the coefficients v_n by the approximate Fourier series method of Section 18.4. The advantage of this oblique attack on the problem is that it takes on the order of only $n \ln n$ operations to compute $u_0, \ldots, u_{n-1}$.

As a concrete illustration of the proposed method, consider the classical problem of computing the probability u_n of a new run of r heads ending at trial n in a sequence of coin-tossing trials. If p and $q = 1-p$ are the head and tail probabilities per trial, respectively, then in this case the appropriate

TABLE 18.2. Renewal Probabilities in a Coin Tossing Example

n	u_n	n	u_n	n	u_n
0	1.0000	5	0.1563	10	0.1670
1	0.0000	6	0.1719	11	0.1665
2	0.2500	7	0.1641	12	0.1667
3	0.1250	8	0.1680	13	0.1666
4	0.1875	9	0.1660	∞	0.1667

renewal equation has $A(s) = 1$ and

$$F(s) = \frac{p^r s^r (1 - ps)}{1 - s + qp^r s^{r+1}}. \tag{12}$$

(See reference [7] or Problem 13.) A brief but tedious calculation shows that $F(s)$ has mean and variance

$$\mu = \frac{1 - p^r}{qp^r}, \quad \sigma^2 = \frac{1}{(qp^r)^2} - \frac{2r + 1}{qp^r} - \frac{p}{q^2},$$

which may be combined to give $F''(1) = \sigma^2 + \mu^2 - \mu$. Fourier transforming $n = 32$ values of $V(s)$ on the boundary of the unit circle when $r = 2$ and $p = 1/2$ yields the renewal probabilities displayed in Table 18.2. In this example, convergence to the limiting value occurs so rapidly that the value of introducing the finite Fourier transform is debatable. Other renewal equations exhibit less rapid convergence. ■

18.6 Time Series

The canonical example of a time series is a stationary sequence $Z_0, Z_1, \ldots$ of real, square-integrable random variables. The sample average $\frac{1}{n}\sum_{j=0}^{n-1} Z_j$ over the n data points collected is the natural estimator of the common theoretical mean μ of the Z_j. Of considerably more interest is the autocovariance sequence

$$c_k = \text{Cov}(Z_j, Z_{j+k}) = c_{-k}.$$

Since we can subtract from each Z_j the sample mean, let us assume that each Z_j has mean 0. Given this simplification, the natural estimator of c_k is

$$d_k = \frac{1}{n} \sum_{j=0}^{n-k-1} Z_j Z_{j+k}.$$

If the finite sequence $Z_0, \ldots, Z_{n-1}$ is padded with n extra zeros and extended to a periodic sequence e_j of period $2n$, then

$$\begin{aligned} d_k &= \frac{2}{2n} \sum_{j=0}^{2n-1} e_j e_{j+k} \\ &= \frac{2}{2n} \sum_{j=0}^{2n-1} e_{j-k} e_j. \end{aligned}$$

According to properties (a) and (c) of Proposition 18.2.1, $d_0, \ldots, d_{n-1}$ can be quickly computed by inverting the finite Fourier transform $2|\hat{e}_k|^2$.

If the terms c_k of the autocovariance sequence decline sufficiently rapidly, then $\sum_k |c_k| < \infty$, and the Fourier series $\sum_k c_k e^{2\pi i k x}$ converges absolutely

to a continuous function $f(x)$ called the spectral density of the time series. One of the goals of times series analysis is to estimate the periodic function $f(x)$. The periodogram

$$I_n(x) = \frac{1}{n}\Big|\sum_{j=0}^{n-1} Z_j e^{-2\pi ijx}\Big|^2$$

provides an asymptotically unbiased estimator of $f(x)$. Indeed, the dominated convergence theorem and the premise $\sum_k |c_k| < \infty$ together imply

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E}[I_n(x)] &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=0}^{n-1} \mathrm{E}(Z_j Z_k) e^{2\pi i(k-j)x} \\
&= \lim_{n\to\infty} \sum_{m=-n+1}^{n-1} \Big(1 - \frac{|m|}{n}\Big) c_m e^{2\pi imx} \\
&= \sum_m c_m e^{2\pi imx} \\
&= f(x).
\end{aligned}$$

As a byproduct of this convergence proof, we see that $f(x) \geq 0$. In view of the fact that the c_k are even, Problem 5 of Chapter 15 indicates that $f(x)$ is also even around both 0 and $1/2$.

Unfortunately, the sequence of periodogram estimators $I_n(x)$ is not consistent. Suppose we take two sequences l_n and m_n with $\lim_{n\to\infty} l_n/n = x$ and $\lim_{n\to\infty} m_n/n = y$. Then one can show that $\lim_{n\to\infty} \mathrm{Var}[I_n(l_n/n)]$ is proportional to $f(x)^2$ and that $\lim_{n\to\infty} \mathrm{Cov}[I_n(l_n/n), I_n(m_n/n)] = 0$ for $x \pm y \neq 0 \bmod 1$ [11]. The inconsistency of the periodogram has prompted statisticians to replace $I_n(k/n)$ by the smoothed estimator

$$\sum_{j=-r}^{r} w_j I_n\Big(\frac{j+k}{n}\Big)$$

with positive weights w_j satisfying $w_{-j} = w_j$ and $\sum_{j=-r}^{r} w_j = 1$. The smoothed periodogram decreases mean square error at the expense of increasing bias slightly. This kind of compromise occurs throughout statistics. Of course, the value of the fast Fourier transform in computing the finite Fourier transforms $\frac{1}{n}\sum_{j=0}^{n-1} Z_j e^{-2\pi ijk/n}$ and smoothing the periodogram should be obvious. Here, as elsewhere, speed of computation dictates much of statistical practice.

18.7 Problems

1. Explicitly calculate the finite Fourier transforms of the four sequences $c_j = 1$, $c_j = 1_{\{0\}}$, $c_j = (-1)^j$, and $c_j = 1_{\{0,1,\ldots,n/2-1\}}$ defined on $\{0, 1, \ldots, n-1\}$. For the last two sequences assume that n is even.

2. Show that the sequence $c_j = j$ on $\{0, 1, \ldots, n-1\}$ has finite Fourier transform

$$\hat{c}_k = \begin{cases} \frac{n-1}{2} & k = 0 \\ -\frac{1}{2} + \frac{i}{2}\cot\frac{k\pi}{n} & k \neq 0. \end{cases}$$

3. For $0 \leq r < n/2$, define the rectangular and triangular smoothing sequences

$$c_j = \frac{1}{2r+1} 1_{\{-r \leq j \leq r\}}$$
$$d_j = \frac{1}{r} 1_{\{-r \leq j \leq r\}} \left(1 - \frac{|j|}{r}\right)$$

and extend them to have period n. Show that

$$\hat{c}_k = \frac{1}{n(2r+1)} \frac{\sin\frac{(2r+1)k\pi}{n}}{\sin\frac{k\pi}{n}}$$
$$\hat{d}_k = \frac{1}{nr^2} \left(\frac{\sin\frac{rk\pi}{n}}{\sin\frac{k\pi}{n}}\right)^2.$$

4. Prove parts (a) through (c) of Proposition 18.2.1.

5. From a periodic sequence c_k with period n, form the circulant matrix

$$C = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ \vdots & \vdots & \vdots & & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}.$$

For $u_n = e^{2\pi i/n}$ and m satisfying $0 \leq m \leq n-1$, show that the vector $(u_n^{0m}, u_n^{1m}, \ldots, u_n^{(n-1)m})^t$ is an eigenvector of C with eigenvalue $n\hat{c}_m$. From this fact deduce that the circulant matrix C can be written in the diagonal form $C = UDU^*$, where D is the diagonal matrix with kth diagonal entry $n\hat{c}_{k-1}$, U is the unitary matrix with entry $u_n^{(j-1)(k-1)}/\sqrt{n}$ in row j and column k, and U^* is the conjugate transpose of U.

6. For $0 \leq m \leq n-1$ and a periodic function $f(x)$ on [0,1], define the sequence $b_m = f(m/n)$. If $\hat{b}_k$ is the finite Fourier transform of the sequence b_m, then we can approximate $f(x)$ by $\sum_{k=-\lfloor n/2 \rfloor}^{\lfloor n/2 \rfloor} \hat{b}_k e^{2\pi ikx}$. Show that this approximation is exact when $f(x)$ is equal to $e^{2\pi ijx}$, $\cos(2\pi jx)$, or $\sin(2\pi jx)$ for j satisfying $0 \leq |j| < \lfloor n/2 \rfloor$.

7. Continuing Problem 6, let c_k be the kth Fourier series coefficient of a general periodic function $f(x)$. If $|c_k| \leq ar^{|k|}$ for constants $a \geq 0$ and

$0 \le r < 1$, then verify using equation (6) that

$$|\hat{b}_k - c_k| \le a r^n \frac{r^k + r^{-k}}{1 - r^n}$$

for $|k| < n$. Functions analytic around 0 automatically possess Fourier coefficients satisfying the bound $|c_k| \le a r^{|k|}$.

8. Continuing Problems 6 and 7, suppose a constant $a \ge 0$ and positive integer p exist such that

$$|c_k| \le \frac{a}{|k|^{p+1}}$$

for all $k \neq 0$. (As Problem 2 of Chapter 15 shows, this criterion holds if $f^{(p+1)}(x)$ is piecewise continuous.) Verify the inequality

$$|\hat{b}_k - c_k| \le \frac{a}{n^{p+1}} \sum_{j=1}^{\infty} \Big[\frac{1}{\Big(j + \frac{k}{n}\Big)^{p+1}} + \frac{1}{\Big(j - \frac{k}{n}\Big)^{p+1}} \Big]$$

when $|k| < n/2$. To simplify this inequality, demonstrate that

$$\begin{aligned} \sum_{j=1}^{\infty} \frac{1}{(j+\alpha)^{p+1}} &< \int_{\frac{1}{2}}^{\infty} (x+\alpha)^{-p-1} dx \\ &= \frac{1}{p\Big(\frac{1}{2} + \alpha\Big)^p} \end{aligned}$$

for $\alpha > -1/2$. Finally, conclude that

$$|\hat{b}_k - c_k| < \frac{a}{p n^{p+1}} \Big[\frac{1}{\Big(\frac{1}{2} + \frac{k}{n}\Big)^p} + \frac{1}{\Big(\frac{1}{2} - \frac{k}{n}\Big)^p} \Big].$$

9. For a complex number c with $|c| > 1$, show that the periodic function $f(x) = (c - e^{2\pi i x})^{-1}$ has the simple Fourier series coefficients $c_k = c^{-k-1} 1_{\{k \ge 0\}}$. Argue from equation (6) that the finite Fourier transform approximation $\hat{b}_k$ to c_k is

$$\hat{b}_k = \begin{cases} c^{-k-1} \frac{1}{1 - c^{-n}} & 0 \le k \le \frac{n}{2} - 1 \\ c^{-n-k-1} \frac{1}{1 - c^{-n}} & -\frac{n}{2} \le k \le 0. \end{cases}$$

10. For some purposes it is preferable to have a purely real transform. If $c_1, \ldots, c_{n-1}$ is a finite sequence of real numbers, then we define its finite sine transform by

$$\hat{c}_k = \frac{2}{n} \sum_{j=1}^{n-1} c_j \sin\Big(\frac{\pi k j}{n}\Big).$$

Show that this transform has inverse

$$\check{d}_j = \sum_{k=1}^{n-1} d_k \sin\Big(\frac{\pi kj}{n}\Big).$$

(Hint: It is helpful to consider $c_1, \ldots, c_{n-1}$ as part of a sequence of period $2n$ that is odd about n.)

11. From a real sequence c_k of period $2n$ we can concoct a complex sequence of period n according to the recipe $d_k = c_{2k} + ic_{2k+1}$. Because it is quicker to take the finite Fourier transform of the sequence d_k than of the sequence c_k, it is desirable to have a simple method of constructing $\hat{c}_k$ from $\hat{d}_k$. Show that

$$\hat{c}_k = \frac{1}{4}\Big(\hat{d}_k + \hat{d}^*_{n-k}\Big) - \frac{i}{4}\Big(\hat{d}_k - \hat{d}^*_{n-k}\Big)e^{-\frac{\pi ik}{n}}.$$

12. Let $F(s) = \sum_{n=1}^{\infty} f_n s^n$ be a probability generating function. Show that the equation $F(s) = 1$ has only the solution $s = 1$ on $|s| = 1$ if and only if the set $\{n: f_n > 0\}$ has greatest common divisor 1.

13. Let W be the waiting time until the first run of r heads in a coin-tossing experiment. If heads occur with probability p, and tails occur with probability $q = 1 - p$ per trial, then show that W has the generating function displayed in equation (12). (Hint: Argue that either $W = r$ or $W = k + 1 + W_k$, where $0 \le k \le r - 1$ is the initial number of heads and W_k is a probabilistic replica of W.)

14. Consider a power series $f(x) = \sum_{m=0}^{\infty} c_m x^m$ with radius of convergence $r > 0$. Prove that

$$\sum_{m=k \bmod n}^{\infty} c_m x^m = \frac{1}{n}\sum_{j=0}^{n-1} u_n^{-jk} f(u_n^j x)$$

for any x with $|x| < r$. As a special case, verify the identity

$$\sum_{m=k \bmod n}^{\infty} \binom{p}{m} = \frac{2^p}{n}\sum_{j=0}^{n-1} \cos\Big[\frac{(p-2k)j\pi}{n}\Big]\cos^p\Big[\frac{j\pi}{n}\Big]$$

for any positive integer p.

15. For a fixed positive integer n, we define the segmental functions ${}_n\alpha_j(x)$ of x as the finite Fourier transform coefficients

$${}_n\alpha_j(x) = \frac{1}{n}\sum_{k=0}^{n-1} e^{xu_n^k} u_n^{-jk}.$$

These functions generalize the hyperbolic trig functions $\cosh(x)$ and $\sinh(x)$. Prove the following assertions:

(a) ${}_n\alpha_j(x) = {}_n\alpha_{j+n}(x)$.
(b) ${}_n\alpha_j(x+y) = \sum_{k=0}^{n-1} {}_n\alpha_k(x)\,{}_n\alpha_{j-k}(y)$.
(c) ${}_n\alpha_j(x) = \sum_{k=0}^{\infty} x^{j+kn}/(j+kn)!$ for $0 \le j \le n-1$.

(d) $\frac{d}{dx}\left[{}_n\alpha_j(x)\right] = {}_n\alpha_{j-1}(x)$.

(e) Consider the differential equation $\frac{d^n}{dx^n}f(x) = kf(x)$ with initial conditions $\frac{d^j}{dx^j}f(0) = c_j$ for $0 \le j \le n-1$, where k and the c_j are constants. Show that

$$f(x) = \sum_{j=0}^{n-1} c_j k^{-\frac{j}{n}}\, {}_n\alpha_j(k^{\frac{1}{n}}x).$$

(f) The differential equation $\frac{d^n}{dx^n}f(x) = kf(x) + g(x)$ with initial conditions $\frac{d^j}{dx^j}f(0) = c_j$ for $0 \le j \le n-1$ has solution

$$f(x) = \int_0^x k^{-\frac{n-1}{n}}\, {}_n\alpha_{n-1}[k^{\frac{1}{n}}(x-y)]g(y)dy + \sum_{j=0}^{n-1} c_j k^{-\frac{j}{n}}\, {}_n\alpha_j(k^{\frac{1}{n}}x).$$

(g) $\lim_{x\to\infty} e^{-x}\, {}_n\alpha_j(x) = 1/n$.

(h) In a Poisson process of intensity 1, $e^{-x}\, {}_n\alpha_j(x)$ is the probability that the number of random points on $[0, x]$ equals j modulo n.

(i) Relative to this Poisson process, let N_x count every nth random point on $[0, x]$. Then N_x has probability generating function

$$P(s) = e^{-x} \sum_{j=0}^{n-1} s^{-\frac{j}{n}}\, {}_n\alpha_j(s^{\frac{1}{n}}x).$$

(j) Furthermore, N_x has mean

$$\mathrm{E}(N_x) = \frac{x}{n} - \frac{e^{-x}}{n} \sum_{j=0}^{n-1} j\, {}_n\alpha_j(x).$$

(k) $\lim_{x\to\infty}\left[\mathrm{E}(N_x) - x/n\right] = -(n-1)/(2n)$.

References

[1] Blackman RB, Tukey JW (1959) *The Measurement of Power Spectra.* Dover, New York

[2] Bloomfield P (1976) *Fourier Analysis of Time Series: An Introduction.* Wiley, New York

[3] Brigham EO (1974) *The Fast Fourier Transform.* Prentice-Hall, Englewood Cliffs, NJ

[4] Brillinger D (1975) *Time Series: Data Analysis and Theory.* Holt, Rinehart, and Winston, New York

[5] Cooley JW, Lewis PAW, Welch PD (1969) The finite Fourier transform. *IEEE Trans Audio Electroacoustics* AU-17:77–85

[6] Dym H, McKean HP (1972) *Fourier Series and Integrals.* Academic Press, New York
[7] Feller W (1968) *An Introduction to Probability Theory and Its Applications, Vol 1*, 3rd ed. Wiley, New York
[8] Henrici P (1979) Fast Fourier transform methods in computational complex analysis. *SIAM Review* 21:481–527
[9] Henrici P (1982) *Essentials of Numerical Analysis with Pocket Calculator Demonstrations.* Wiley, New York
[10] Karlin S, Taylor HM (1975) *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York
[11] Koopmans LH (1974) *The Spectral Analysis of Time Series.* Academic Press, New York
[12] Körner TW (1988) *Fourier Analysis.* Cambridge University Press, Cambridge
[13] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[14] Squire W, Trapp G (1998) Using complex variables to estimate derivatives of real functions. *SIAM Review* 40:110–112
[15] Wilf HS (1986) *Algorithms and Complexity.* Prentice-Hall, Englewood Cliffs, NJ

19
Wavelets

19.1 Introduction

Wavelets are just beginning to enter statistical theory and practice [2, 5, 7, 10, 12]. The pace of discovery is still swift, and except for orthogonal wavelets, the theory has yet to mature. However, the advantages of wavelets are already obvious in application areas such as image compression. Wavelets are more localized in space than the competing sines and cosines of Fourier series. They also use fewer coefficients in representing images, and they pick up edge effects better. The secret behind these successes is the capacity of wavelets to account for image variation on many different scales.

In this chapter we develop a small fraction of the relevant theory and briefly describe applications of wavelets to density estimation and image compression. For motivational purposes, we begin with the discontinuous wavelets of Haar. These wavelets are easy to understand but have limited utility. The recent continuous wavelets of Daubechies are both more subtle and more practical. Daubechies' wavelets fortunately lend themselves to fast computation. By analogy to the fast Fourier transform, there is even a fast wavelet transform [9]. The challenge to applied mathematicians, computer scientists, engineers, and statisticians is to find new applications that exploit wavelets. The edited volume [1] and the articles [6, 3] describe some opening moves by statisticians.

19.2 Haar's Wavelets

Orthonormal bases are not unique. For example, ordinary Fourier series and the beta distribution polynomials $\phi_n^{(1,1)}$ studied in Chapter 14 both provide bases for the space $\mathrm{L}^2[0,1]$ of square-integrable functions relative to the uniform distribution on $[0,1]$. Shortly after the turn of the twentieth century, Haar introduced yet another orthonormal basis for $\mathrm{L}^2[0,1]$. His construction anticipated much of the modern development of wavelets.

We commence our discussion of Haar's contribution with the indicator function $h_0(x) = 1_{[0,1)}(x)$ of the unit interval. This function satisfies the identities $\int h_0(x)dx = \int h_0(x)^2 dx = 1$. It can also be rescaled and translated to give the indicator function $h_0(2^j x - k)$ of the interval $[k/2^j, (k+1)/2^j)$. If we want to stay within the unit interval $[0,1]$, then we restrict j to be a nonnegative integer and k to be an integer between 0 and $2^j - 1$. If we prefer to range over the whole real line, then k can be any integer. For the sake of simplicity, let us focus on $[0,1]$. Since step functions are dense in $\mathrm{L}^2[0,1]$, we can approximate any square-integrable function by a linear combination of the $h_0(2^j x - k)$. Within a fixed level j, two different translates $h_0(2^j x - k)$ and $h_0(2^j x - l)$ are orthogonal, but across levels orthogonality fails. Thus, the normalized functions $2^{j/2} h_0(2^j x - k)$ do not provide an orthonormal basis.

Haar turned the scaling identity

$$h_0(x) = h_0(2x) + h_0(2x-1) \tag{1}$$

around to construct a second function

$$w(x) = h_0(2x) - h_0(2x-1), \tag{2}$$

which is 1 on $[0, 1/2)$ and -1 on $[1/2, 1)$. In modern terminology, $h_0(x)$ is called the scaling function and $w(x)$ the mother wavelet. We subject $w(x)$ to dilation and translation and construct a sequence of functions

$$h_n(x) = 2^{\frac{j}{2}} w(2^j x - k)$$

to supplement $h_0(x)$. Here $n > 0$ and j and k are uniquely determined by writing $n = 2^j + k$ subject to the constraint $0 \le k < 2^j$. As with the corresponding dilated and translated version of $h_0(x)$, the function $h_n(x)$ has support on the interval $[k/2^j, (k+1)/2^j) \subset [0,1)$. We claim that the sequence $\{h_n(x)\}_{n=0}^{\infty}$ constitutes an orthonormal basis of $\mathrm{L}^2[0,1]$.

To prove the claim, first note that

$$\begin{aligned}
\int_0^1 h_n^2(x)dx &= \int_0^1 \left[2^{\frac{j}{2}} w(2^j x - k)\right]^2 dx \\
&= \int_0^1 w(y)^2 dy \\
&= 1.
\end{aligned}$$

Second, observe that

$$\begin{aligned}\int_0^1 h_0(x)h_n(x)dx &= \int_0^1 h_n(x)dx \\ &= 0\end{aligned}$$

for any $n \geq 1$ because of the balancing positive and negative parts of $h_n(x)$. If $0 < m = 2^r + s < n$ for $0 \leq s < 2^r$, then

$$\begin{aligned}\int_0^1 h_m(x)h_n(x)dx &= 2^{\frac{r}{2}}2^{\frac{j}{2}}\int_0^1 w(2^r x - s)w(2^j x - k)dx \\ &= 2^{\frac{j-r}{2}}\int w(y - s)w(2^{j-r}y - k)dy. \qquad (3)\end{aligned}$$

If $r = j$ in the integral (3), then the support $[k, k+1)$ of the right integrand is disjoint from the support $[s, s+1)$ of the left integrand, and the integral is trivially 0. If $r < j$, then the support $[k/2^{j-r}, (k+1)/2^{j-r})$ of the right integrand is disjoint from the interval $[s, s+1)$ or wholly contained within either $[s, s+1/2)$ or $[s+1/2, s+1)$. If the two supports are disjoint, then again the integral is trivially 0. If they intersect, then the positive and negative contributions of the integral exactly cancel. This proves that the Haar functions $\{h_n(x)\}_{n=0}^{\infty}$ form an orthonormal sequence.

To verify completeness, it suffices to show that the indicator function $h_0(2^j x - k)$ of an arbitrary dyadic interval $[k/2^j, (k+1)/2^j) \subset [0, 1)$ can be written as a finite linear combination $\sum_n c_n h_n(x)$. For example,

$$1_{[0,\frac{1}{2})}(x) = h_0(2x) = \frac{1}{2}[h_0(x) + w(x)]$$

$$1_{[\frac{1}{2},1)}(x) = h_0(2x - 1) = \frac{1}{2}[h_0(x) - w(x)]$$

are immediate consequences of equations (1) and (2). The general case follows by induction from the analogous identities

$$h_0(2^j x - 2k) = \frac{1}{2}[h_0(2^{j-1}x - k) + w(2^{j-1}x - k)]$$

$$h_0(2^j x - 2k - 1) = \frac{1}{2}[h_0(2^{j-1}x - k) - w(2^{j-1}x - k)].$$

Obvious extensions of the above arguments show that we can construct an orthonormal basis for $\mathrm{L}^2(-\infty, \infty)$ from the functions $h_0(x - k)$ and $2^{j/2}w(2^j x - k)$, where j ranges over the nonnegative integers and k over all integers. In this basis, it is always possible to express the indicator function $h_0(2^r x - s)$ of an interval $[s/2^r, (s+1)/2^r)$ as a finite linear combination of the $h_0(x - k)$ and the $2^{j/2}w(2^j x - k)$ for $0 \leq j < r$.

19.3 Histogram Estimators

One application of the Haar functions is in estimating the common density function $f(x)$ of an i.i.d. sequence $X_1, \ldots, X_n$ of random variables. For j large and fixed, we can approximate $f(x)$ accurately in $\mathrm{L}^2(-\infty, \infty)$ by a linear combination of the orthonormal functions $g_k(x) = 2^{j/2} h_0(2^j x - k)$. The best choice of the coefficient c_k in the approximate expansion

$$f(x) \approx \sum_k c_k g_k(x)$$

is $c_k = \int g_k(z) f(z) dz = \mathrm{E}[g_k(X_1)]$. This suggests that we replace the expectation c_k by the sample average

$$\bar{c}_k = \frac{1}{n} \sum_{i=1}^{n} g_k(X_i).$$

The resulting estimator $\sum_k \bar{c}_k g_k(x)$ of $f(x)$ is called a histogram estimator.

If we let

$$a_{jk} = 2^j \int_{\frac{k}{2^j}}^{\frac{k+1}{2^j}} f(z) dz,$$

then we can express the expectation of the histogram estimator as

$$\begin{aligned} \mathrm{E}\Big[\sum_k \bar{c}_k g_k(x)\Big] &= \sum_k c_k g_k(x) \\ &= \sum_k a_{jk} 1_{[\frac{k}{2^j}, \frac{k+1}{2^j})}(x). \end{aligned}$$

If $f(x)$ is continuous, this sum tends to $f(x)$ as j tends to ∞. Since $g_k(x) g_l(x) = 0$ for $k \neq l$, the variance of the histogram estimator amounts to

$$\begin{aligned} \frac{1}{n} \operatorname{Var}\Big[\sum_k g_k(X_1) g_k(x)\Big] &= \frac{1}{n} \sum_k \operatorname{Var}\Big[g_k(X_1)\Big] g_k(x)^2 \\ &= \frac{2^j}{n} \sum_k a_{jk} \Big(1 - \frac{1}{2^j} a_{jk}\Big) 1_{[\frac{k}{2^j}, \frac{k+1}{2^j})}(x), \end{aligned}$$

which is small when the ratio $2^j/n$ is small. We can minimize the mean square error of the histogram estimator by taking some intermediate value for j and balancing bias against variance. Clearly, this general procedure extends to density estimators based on other orthonormal sequences [12].

19.4 Daubechies' Wavelets

Our point of departure in developing Daubechies' lovely generalization of the Haar functions is the scaling equation [2, 8, 11]

$$\psi(x) = \sum_{k=0}^{n-1} c_k \psi(2x-k). \tag{4}$$

When $n = 2$ and $c_0 = c_1 = 1$, the indicator function of the unit interval solves equation (4); this solution generates the Haar functions. Now we look for a continuous solution of (4) that leads to an orthogonal wavelet sequence on the real line instead of the unit interval. For the sake of simplicity, we limit our search to the special value $n = 4$. In fact, there exists a solution to (4) for every even $n > 0$. These higher-order Daubechies' scaling functions generate progressively smoother and less localized wavelet sequences.

In addition to continuity, we require that the scaling function $\psi(x)$ have bounded support. If the support of $\psi(x)$ is the interval $[a, b]$, then the support of $\psi(2x-k)$ is $[(a+k)/2, (b+k)/2]$. Thus, the right-hand side of equation (4) implies that $\psi(x)$ has support between $a/2$ and $(b+n-1)/2$. Equating these to a and b yields $a = 0$ and $b = n-1$. Because $\psi(x)$ vanishes outside $[0, n-1]$, continuity dictates that $\psi(0) = 0$ and $\psi(n-1) = 0$. Therefore, when $n = 4$, the only integers k permitting $\psi(k) \neq 0$ are $k = 1$ and 2. The scaling equation (4) determines the ratio $\psi(1)/\psi(2)$ through the eigenvector equation

$$\begin{pmatrix} \psi(1) \\ \psi(2) \end{pmatrix} = \begin{pmatrix} c_1 & c_0 \\ c_3 & c_2 \end{pmatrix} \begin{pmatrix} \psi(1) \\ \psi(2) \end{pmatrix}. \tag{5}$$

If we take $\psi(1) > 0$, then $\psi(1)$ and $\psi(2)$ are uniquely determined either by the convention $\int \psi(x)dx = 1$ or by the convention $\int |\psi(x)|^2 dx = 1$. As we will see later, these constraints can be simultaneously met. Once we have determined $\psi(x)$ for all integer values k, then these values determine $\psi(x)$ for all half-integer values $k/2$ through the scaling equation (4). The half-integer values $k/2$ determine the quarter-integer values $k/4$, and so forth. Since any real number is a limit of dyadic rationals $k/2^j$, the postulated continuity of $\psi(x)$ completely determines all values of $\psi(x)$. The scaling equation truly is a potent device.

Only certain values of the coefficients c_k are compatible with our objective of constructing an orthonormal basis for $\mathrm{L}^2(-\infty, \infty)$. To determine these values, we first note that the scaling equation (4) implies

$$\begin{aligned} 2\int \psi(x)dx &= 2\sum_k c_k \int \psi(2x-k)dx \\ &= \sum_k c_k \int \psi(z)dz. \end{aligned}$$

(Here and in the following, we omit limits of summation by defining $c_k = 0$ for k outside $0, \ldots, n-1$.) Assuming that $\int \psi(z)dz \neq 0$, we find

$$2 = \sum_k c_k. \tag{6}$$

If we impose the orthogonality constraints

$$1_{\{m=0\}} = \int \psi(x)\psi(x-m)^* dx$$

on the integer translates of the current unknown scaling function $\psi(x)$, then the scaling equation (4) implies

$$\begin{aligned} 1_{\{m=0\}} &= \int \psi(x)\psi(x-m)^* dx \\ &= \sum_k \sum_l c_k c_l^* \int \psi(2x-k)\psi(2x-2m-l)^* dx \\ &= \frac{1}{2} \sum_k \sum_l c_k c_l^* \int \psi(z)\psi(z+k-2m-l)^* dz \\ &= \frac{1}{2} \sum_k c_k c_{k-2m}^*. \end{aligned} \tag{7}$$

For reasons that will soon be apparent, we now define the mother wavelet

$$w(x) = \sum_k (-1)^k c_{1-k} \psi(2x-k). \tag{8}$$

In the case of the Haar functions, $w(x)$ satisfies $\int w(x)dx = 0$. In view of definition (8), imposing this constraint yields

$$\begin{aligned} 0 &= \sum_k (-1)^k c_{1-k} \int \psi(2x-k)dx \\ &= \frac{1}{2} \sum_k (-1)^k c_{1-k}. \end{aligned} \tag{9}$$

We can restate this result by taking the Fourier transform of equation (8). This gives

$$\hat{w}(y) = Q(\frac{y}{2})\hat{\psi}(\frac{y}{2}),$$

where

$$Q(y) = \frac{1}{2} \sum_k (-1)^k c_{1-k} e^{iky}.$$

From the identity $\int w(x)dx = \hat{w}(0) = 0$, we deduce that $Q(0) = 0$. Finally, the constraint $\int xw(x)dx = \frac{d}{idy}\hat{w}(0) = 0$, which is false for the Haar mother

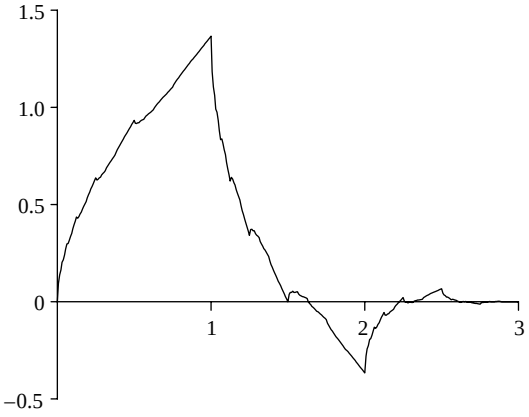

FIGURE 19.1. Plot of Daubechies' $\psi(x)$

wavelet, ensures a limited amount of symmetry in the current $w(x)$. This final constraint on the coefficients c_k amounts to

$$\begin{aligned} 0 &= \frac{d}{dy}\hat{w}(0) \\ &= \frac{1}{2}\Big[\frac{d}{dy}Q(0)\Big]\hat{\psi}(0) + \frac{1}{2}Q(0)\frac{d}{dy}\hat{\psi}(0) \\ &= \frac{1}{2}\Big[\frac{d}{dy}Q(0)\Big]\int \psi(x)dx \\ &= \frac{1}{4}\sum_k (-1)^k ikc_{1-k}. \end{aligned} \tag{10}$$

Our findings (6), (7), (9), and (10) can be summarized for $n = 4$ by the system of equations

$$\begin{aligned} c_0 + c_1 + c_2 + c_3 &= 2 \\ |c_0|^2 + |c_1|^2 + |c_2|^2 + |c_3|^2 &= 2 \\ c_0c_2^* + c_1c_3^* &= 0 \\ -c_0 + c_1 - c_2 + c_3 &= 0 \\ -c_0 + c_2 - 2c_3 &= 0. \end{aligned} \tag{11}$$

The first four of these equations are redundant and have the general solution

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \frac{1}{c^2+1}\begin{pmatrix} c(c-1) \\ 1-c \\ c+1 \\ c(c+1) \end{pmatrix}$$

for some real constant c. The last equation determines $c = \pm 1/\sqrt{3}$. With the coefficients c_k in hand, we return to equation (5) and identify the

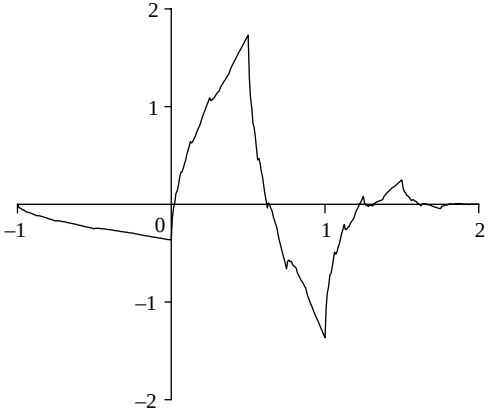

FIGURE 19.2. Plot of Daubechies' $w(x)$

eigenvector $\begin{pmatrix} c-1 \\ c+1 \end{pmatrix}$ determining the ratio of $\psi(1)$ to $\psi(2)$. By virtue of the fact that the coefficients c_k are real, choosing either $\psi(1)$ or $\psi(2)$ to be real forces $\psi(x)$ to be real for all x. This in turn compels $w(x)$ to be real. It follows that we can safely omit complex conjugate signs in calculating inner products.

Figures 19.1 and 19.2 plot Daubechies' $\psi(x)$ and $w(x)$ when $n = 4$ for the choices $c = -1/\sqrt{3}$, $\psi(1) = (1+\sqrt{3})/2$, and $\psi(2) = (1-\sqrt{3})/2$, which incidentally give the correct ratio $\psi(1)/\psi(2) = (c-1)/(c+1)$. The functions $\psi(x)$ and $w(x)$ are like no other special functions of applied mathematics. Despite our inability to express $\psi(x)$ explicitly, the scaling equation offers an effective means of computing its values on the dyadic rationals. Continuity fills in the holes.

The choices $c = -1/\sqrt{3}$, $\psi(1) = (1+\sqrt{3})/2$, and $\psi(2) = (1-\sqrt{3})/2$ also yield the partition of unity property

$$\sum_l \psi(x-l) = 1 \tag{12}$$

at any integer x. To prove that this property extends to all real numbers, let $\text{ev}(x) = \sum_m \psi(x-2m)$, and consider a half-integer x. The scaling relation (4), induction, and the first and fourth equations in (11) imply

$$\begin{aligned}
\sum_l \psi(x-l) &= \sum_l \sum_k c_k \psi(2x-2l-k) \\
&= c_0 \text{ev}(2x) + c_1[1-\text{ev}(2x)] + c_2 \text{ev}(2x) + c_3[1-\text{ev}(2x)] \\
&= c_1 + c_3 \\
&= 1.
\end{aligned}$$

Induction extends the partition-of-unity property (12) beyond half integers to all dyadic rationals, and continuity extends it from there to all real numbers.

At first glance it is not obvious that the choices $\psi(1) = (1+\sqrt{3})/2$ and $\psi(2) = (1-\sqrt{3})/2$ are compatible with the conventions $\int \psi(x)dx = 1$ and $\int \psi(x)^2 dx = 1$. Since $\psi(x)$ has bounded support and satisfies the partition-of-unity property, the first convention follows from

$$\begin{aligned} 1 &= \lim_{n\to\infty} \frac{1}{2n+1} \int \sum_{k=-n}^{n} \psi(x-k)dx \\ &= \int \psi(x)dx. \end{aligned}$$

Here we use the fact that $\sum_{k=-n}^{n} \psi(x-k) = 1_{[-n,n]}(x)$ except for small intervals around $-n$ and n. The orthogonality of the different $\psi(x-k)$ and the partition-of-unity property now justify the calculation

$$\begin{aligned} 1 &= \int \psi(x)dx \\ &= \int \psi(x) \sum_k \psi(x-k)dx \\ &= \int \psi(x)^2 dx. \end{aligned}$$

The scaling function $\psi(x)$ and the mother wavelet $w(x)$ together generate a wavelet basis for $\mathrm{L}^2(-\infty,\infty)$ consisting of all translates $\psi(x-m)$ of $\psi(x)$ plus all translated dilates $w_{jk}(x) = 2^{j/2}w(2^j x - k)$ of $w(x)$. Here m and k are arbitrary integers while j is an arbitrary nonnegative integer. The orthonormality of the translates $\psi(x-m)$ is built into the definition of $\psi(x)$. By the same reasoning that led to equation (7), we deduce that

$$\int w(x)\psi(x-m)dx = \frac{1}{2}\sum_k (-1)^k c_{1-k}c_{k-2m}.$$

This sum vanishes because the term $(-1)^k c_{1-k}c_{k-2m}$ exactly cancels the term $(-1)^{1-k+2m}c_{1-(1-k+2m)}c_{1-k+2m-2m}$ in which $1-k+2m$ replaces k. Now we see the purpose of the strange definition of $w(x)$. Orthonormality of the translates of $w(x)$ comes down to the constraints

$$\begin{aligned} 1_{\{m=0\}} &= \int w(x)w(x-m)dx \\ &= \frac{1}{2}\sum_k (-1)^k c_{1-k}(-1)^{k-2m}c_{1-k+2m} \int \psi(z)^2 dz \\ &= \frac{1}{2}\sum_l c_l c_{l+2m} \end{aligned}$$

already imposed on the c_k in equation (7).

The coefficient $2^{j/2}$ is chosen to make $\int w_{jk}(x)^2 dx = 1$. The orthogonality conditions $\int \psi(x) w_{jk}(x) dx = 0$ and $\int w_{jk}(x) w_{lm}(x) dx = 0$ for pairs $(j,k) \neq (l,m)$ follow by induction. For instance, induction implies

$$\begin{aligned}\int \psi(x) w_{jk}(x) dx &= \sum_l c_l \int \psi(2x-l) 2^{\frac{j}{2}} w(2^j x - k) dx \\ &= \sum_l c_l \int \psi(z) 2^{\frac{j}{2}-1} w(2^{j-1} z + 2^{j-1} l - k) dz \\ &= 0\end{aligned}$$

for $j > 0$. Thus, the wavelet sequence is orthonormal.

We next demonstrate that $\psi(2x-m)$ can be written as a finite sum

$$\psi(2x-m) = \sum_k r_{mk} \psi(x-k) + \sum_k s_{mk} w(x-k) \tag{13}$$

for certain coefficients r_{mk} and s_{mk}. Because the functions on the right of the representation (13) are orthonormal, we calculate r_{mk} as

$$\begin{aligned}\int \psi(2x-m)\psi(x-k) dx &= \sum_j c_j \int \psi(2x-m)\psi(2x-2k-j) dx \\ &= \frac{1}{2} c_{m-2k}\end{aligned}$$

and s_{mk} as

$$\begin{aligned}\int \psi(2x-m) w(x-k) dx &= \sum_j (-1)^j c_{1-j} \int \psi(2x-m)\psi(2x-2k-j) dx \\ &= \frac{(-1)^{m-2k}}{2} c_{1-m+2k}.\end{aligned}$$

In light of the second identity in (11), we conclude that

$$\begin{aligned}\int \psi(2x-m)^2 dx &= \frac{1}{2} \\ &= \frac{1}{4} \sum_k c_{m-2k}^2 + \frac{1}{4} \sum_k c_{1-m+2k}^2 \\ &= \sum_k r_{mk}^2 + \sum_k s_{mk}^2.\end{aligned}$$

Bessel's equality (3) of Chapter 15 now yields equation (13). Induction and substitution of $2^{l-1}x$ for $2x$ in equation (13) demonstrate that every function $\psi(2^l x - m)$ can be written as a finite linear combination of the functions $\psi(x-k)$ and the $w_{jk}(x)$ with $j < l$.

Finally, we address the completeness of the orthonormal wavelet sequence. Observe first that the sequence of functions $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$

is orthonormal for j fixed. To prove completeness, consider the projection

$$P_j f(x) = \sum_k \psi_{jk}(x) \int f(y)\psi_{jk}(y)dy$$

of a square-integrable function $f(x)$ onto the subspace spanned by the sequence $\{\psi_{jk}(x)\}_k$. It suffices to show that these projections converge to $f(x)$ in $\mathrm{L}^2(-\infty,\infty)$. Due to Bessel's inequality (2) in Chapter 15, convergence in $\mathrm{L}^2(-\infty,\infty)$ is equivalent to $\lim_{j\to\infty} \|P_j f\|^2 = \|f\|^2$. Since step functions are dense in $\mathrm{L}^2(-\infty,\infty)$, we can further reduce the problem to the case where $f(x)$ is the indicator function $1_{[a,b]}(x)$ of an interval. Making the change of variables $y = 2^j x$, we calculate

$$\begin{aligned}
\|P_j 1_{[a,b]}\|^2 &= \sum_k \Big[\int 1_{[a,b]}(x) 2^{\frac{j}{2}} \psi(2^j x - k)dx \Big]^2 \\
&= \sum_k \Big[\int 1_{[2^j a, 2^j b]}(y)\psi(y-k)dy \Big]^2 2^{-j}.
\end{aligned}$$

For the vast majority of indices k when j is large, the support of $\psi(y-k)$ is wholly contained within or wholly disjoint from $[2^j a, 2^j b]$. Hence, the condition $\int \psi(y)dy = 1$ implies that

$$\begin{aligned}
\lim_{j\to\infty} \|P_j 1_{[a,b]}\|^2 &= \lim_{j\to\infty} 2^{-j} \#\{k : k \in [2^j a, 2^j b]\} \\
&= b - a \\
&= \int 1_{[a,b]}(y)^2 dy.
\end{aligned}$$

This proves completeness.

Doubtless the reader has noticed that we have never proved that $\psi(x)$ exists and is continuous for all real numbers. For the sake of brevity, we refer interested readers to Problems 10 and 11 for a sketch of one attack on these thorny questions [8]. Although continuity of $\psi(x)$ is assured, differentiability is not. It turns out that $\psi(x)$ is left differentiable, but not right differentiable, at each dyadic rational of $[0,3]$. Problem 9 makes a start on the issue of differentiability.

19.5 Multiresolution Analysis

It is now time to step back and look at the larger landscape. At the coarsest level of detail, we have a (closed) subspace V_0 of $\mathrm{L}^2(-\infty,\infty)$ spanned by the translates $\psi(x-k)$ of the scaling function. This is the first of a hierarchy of closed subspaces V_j constructed from the translates $\psi_{jk}(x)$ of the dilated functions $2^{j/2}\psi(2^j x)$. The scaling equation (4) tells us that $V_j \subset V_{j+1}$ for every j, and completeness tells us that $\mathrm{L}^2(-\infty,\infty)$ is the closed span of

the union $\cup_{j=0}^{\infty} V_j$. This favorable state of affairs is marred by the fact that the functions $\psi_{jk}(x)$ are only orthogonal within a level j and not across levels. The remedy is to introduce the wavelets $w_{jk}(x)$. These are designed so that V_j is spanned by a basis of V_{j-1} plus the translates $w_{j-1,k}(x)$ of $w_{j-1,0}(x)$. This fact follows from the obvious generalization

$$\psi_{jk}(x) = \frac{1}{\sqrt{2}} \sum_l c_{k-2l} \psi_{j-1,l}(x) + \frac{1}{\sqrt{2}} \sum_l (-1)^{k-2l} c_{1-k+2l} w_{j-1,l}(x) \quad (14)$$

of equation (13). Representation (14) permits us to express the fine distinctions of V_j partially in terms of the coarser distinctions of V_{j-1}. The restatements

$$\begin{aligned} \psi_{jk}(x) &= \frac{1}{\sqrt{2}} \sum_l c_l \psi_{j+1,2k+l}(x) \\ w_{jk}(x) &= \frac{1}{\sqrt{2}} \sum_l (-1)^l c_{1-l} \psi_{j+1,2k+l}(x) \end{aligned} \quad (15)$$

of the scaling equation and the definition of $w(x)$ allow us to move in the opposite direction from coarse to fine.

19.6 Image Compression and the Fast Wavelet Transform

One of the major successes of wavelets is image compression. For the sake of simplicity, we will discuss the compression of one-dimensional images. Our remarks are immediately pertinent to acoustic recordings [13] and, with minor changes, to visual images. The fact that most images are finite in extent suggests that we should be using periodic wavelets rather than ordinary wavelets. It is possible to periodize wavelets by defining

$$\overline{w}_{jk}(x) = \sum_l w_{jk}(x-l) = \sum_l 2^{\frac{j}{2}} w(2^j x - k - l2^j) \quad (16)$$

for $j \geq 0$ and $0 \leq k < 2^j$. The reader is asked in Problem 12 to show that these functions of period 1 together with the constant 1 form an orthonormal basis for the space $\mathrm{L}^2[0,1]$. The constant 1 enters because $1 = \sum_l \psi(x-l)$.

To the subspace V_j in $\mathrm{L}^2(-\infty,\infty)$ corresponds the subspace $\overline{V}_j$ in $\mathrm{L}^2[0,1]$ spanned by the 2^j periodic functions

$$\overline{\psi}_{jk}(x) = \sum_l \psi_{jk}(x-l) = \sum_l 2^{\frac{j}{2}} \psi(2^j x - k - l2^j).$$

It is possible to pass between the $\overline{\psi}_{jk}(x)$ at level j and the basis functions $\overline{w}_{lm}(x)$ for $0 \le l < j$ via the analogs

$$\begin{aligned}
\overline{\psi}_{jk}(x) &= \frac{1}{\sqrt{2}} \sum_l c_{k-2l} \overline{\psi}_{j-1,l}(x) + \frac{1}{\sqrt{2}} \sum_l (-1)^{k-2l} c_{1-k+2l} \overline{w}_{j-1,l}(x) \\
\overline{\psi}_{jk}(x) &= \frac{1}{\sqrt{2}} \sum_l c_l \overline{\psi}_{j+1,2k+l}(x) \\
\overline{w}_{jk}(x) &= \frac{1}{\sqrt{2}} \sum_l (-1)^l c_{1-l} \overline{\psi}_{j+1,2k+l}(x)
\end{aligned} \tag{17}$$

of equations (14) and (15).

If we start with a linear approximation

$$f(x) \approx \sum_{k=0}^{2^j-1} r_k \overline{\psi}_{jk}(x) \tag{18}$$

to a given function $f(x)$ by the basis functions of $\overline{V}_j$, it is clear that the first of the recurrences in (17) permits us to replace this approximation by an equivalent linear approximation

$$f(x) \approx \sum_{k=0}^{2^{j-1}-1} s_k \overline{\psi}_{j-1,k}(x) + \sum_{k=0}^{2^{j-1}-1} t_k \overline{w}_{j-1,k}(x) \tag{19}$$

involving the basis functions of $\overline{V}_{j-1}$. We can then substitute

$$\sum_{k=0}^{2^{j-1}-1} s_k \overline{\psi}_{j-1,k}(x) = \sum_{k=0}^{2^{j-2}-1} u_k \overline{\psi}_{j-2,k}(x) + \sum_{k=0}^{2^{j-2}-1} v_k \overline{w}_{j-2,k}(x)$$

in equation (19), and so forth. This recursive procedure constitutes the fast wavelet transform. It is efficient because in the case of Daubechies' wavelets, only four coefficients c_k are involved, and because only half of the basis functions must be replaced at each level.

In image compression a function $f(x)$ is observed on an interval $[a, b]$. Extending the function slightly, we can easily make it periodic. We can also arrange that $[a, b] = [0, 1]$. If we choose 2^j sufficiently large, then the approximation (18) will be good provided each coefficient r_k satisfies $r_k = \int_0^1 f(x)\overline{\psi}_{jk}(x)dx$. We omit the practical details of how these integrations are done. Once the linear approximation (18) is computed, we can apply the fast wavelet transform to reduce the approximation to one involving the $\overline{w}_{l,k}(x)$ of order $l \le j-1$ and the constant 1. Image compression is achieved by throwing away any terms in the final expansion having coefficients smaller in absolute value than some threshold $\epsilon > 0$. If we store the image as the list of remaining coefficients, then we can readily reconstruct the image by forming the appropriate linear combination of basis functions.

If we want to return to the basis of $\overline{V}_j$ furnished by the $\overline{\psi}_{jk}(x)$, then the second and third recurrences of (17) make this possible.

As mentioned in the introduction to this chapter, the value of wavelet expansions derives from their ability to capture data at many different scales. A periodic wavelet $\overline{w}_{jk}(x)$ is quite localized if j is even moderately large. In regions of an image where there is little variation, the coefficients of the pertinent higher-order wavelets $\overline{w}_{jk}(x)$ are practically zero because $\int w(x)dx = 0$. Where edges or rapid oscillations occur, the higher-order wavelets $\overline{w}_{jk}(x)$ are retained.

19.7 Problems

1. Let $X_1, \ldots, X_n$ be a random sample from a well-behaved density $f(x)$. If $\{g_k(x)\}_{k=1}^{\infty}$ is a real, orthonormal basis for $\mathrm{L}^2(-\infty, \infty)$, then a natural estimator of $f(x)$ is furnished by

$$\bar{f}(x) = \sum_{k=1}^{\infty} \bar{c}_k g_k(x)$$
$$\bar{c}_k = \frac{1}{n} \sum_{i=1}^{n} g_k(X_i).$$

Show formally that

$$\mathrm{E}[\bar{f}(x)] = f(x)$$
$$\mathrm{Var}[\bar{f}(x)] = \frac{1}{n} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \int g_k(z) g_l(z) f(z) dz g_k(x) g_l(x) - \frac{1}{n} f(x)^2,$$

provided the orthogonal expansion of $f(x)$ converges pointwise to $f(x)$.

2. Let $C_1(x)$ be the uniform density on $[0, 1)$. The cardinal B-spline $C_m(x)$ of order m is the m-fold convolution of $C_1(x)$ with itself. Prove that this function satisfies the scaling equation

$$C_m(x) = \frac{1}{2^{m-1}} \sum_{k=0}^{m} \binom{m}{k} C_m(2x - k).$$

(Hint: Show that both sides have the same Fourier transform.)

3. For the choice $c = -1/\sqrt{3}$, show that $\begin{pmatrix} c-1 \\ c+1 \end{pmatrix}$ is the eigenvector sought in equation (5).

4. Write software to evaluate and graph Daubechies' scaling function and mother wavelet for $n = 4$ and $c = -1/\sqrt{3}$.

5. Suppose $\psi(x)$ is a continuous function with bounded support that satisfies the scaling equation (4) and the condition $\int \psi(x)dx = 1$. If

the coefficients satisfy $\sum_k c_k = 2$, then show that

$$2\int x\psi(x)dx = \sum_k kc_k.$$

6. Show that the Fourier transform of Daubechies' scaling function satisfies $\hat{\psi}(y) = P(y/2)\hat{\psi}(y/2)$, where $P(y) = (\sum_k c_k e^{iky})/2$. Conclude that $\hat{\psi}(y) = \prod_{k=1}^{\infty} P(y/2^k)$ holds.

7. Verify the identity

$$1 = \sum_k |\hat{\psi}(2\pi y + 2\pi k)|^2$$

satisfied by Daubechies' scaling function. (Hint: Apply the Parseval–Plancherel identity to the first line of (7), interpret the result as a Fourier series, and use the fact that the Fourier series of an integrable function determines the function almost everywhere.)

8. Demonstrate the identities

$$\begin{aligned} 1 &= |P(\pi y)|^2 + |P(\pi y + \pi)|^2 \\ 1 &= |Q(\pi y)|^2 + |Q(\pi y + \pi)|^2 \\ 0 &= P(\pi y)Q(\pi y)^* + P(\pi y + \pi)Q(\pi y + \pi)^* \end{aligned}$$

involving Daubechies' functions

$$\begin{aligned} P(y) &= \frac{1}{2}\sum_k c_k e^{iky} \\ Q(y) &= \frac{1}{2}\sum_k (-1)^k c_{1-k} e^{iky}. \end{aligned}$$

(Hint: For the first identity, see Problems 6 and 7.)

9. Show that Daubechies' scaling function $\psi(x)$ is left differentiable at $x = 3$ but not right differentiable at $x = 0$ when $n = 4$ and $c = -1/\sqrt{3}$. (Hint: Take difference quotients, and invoke the scaling equation (4).)

10. This problem and the next deal with Pollen's [8] proof of the existence of a unique continuous solution to the scaling equation (4). Readers are assumed to be familiar with some results from functional analysis [4]. Let $a = (1 + \sqrt{3})/4$ and $\overline{a} = (1 - \sqrt{3})/4$. If $f(x)$ is a function defined on $[0, 3]$, then we map it to a new function $M(f)(x)$ defined on $[0, 3]$ according to the piecewise formulas

$$\begin{aligned} M(f)\Big(\frac{0+x}{2}\Big) &= af(x) \\ M(f)\Big(\frac{1+x}{2}\Big) &= \overline{a}f(x) + ax + \frac{2+\sqrt{3}}{4} \\ M(f)\Big(\frac{2+x}{2}\Big) &= af(1+x) + \overline{a}x + \frac{\sqrt{3}}{4} \end{aligned}$$

$$M(f)\Big(\frac{3+x}{2}\Big) = \overline{a}f(1+x) - ax + \frac{1}{4}$$
$$M(f)\Big(\frac{4+x}{2}\Big) = af(2+x) - \overline{a}x + \frac{3-2\sqrt{3}}{4}$$
$$M(f)\Big(\frac{5+x}{2}\Big) = \overline{a}f(2+x)$$

for $x \in [0,1]$. To ensure that the transformation $M(f)(x)$ is well defined at the half-integers, we postulate that $f(x)$ takes the values $f(0) = f(3) = 0$, $f(1) = 2a$, and $f(2) = 2\overline{a}$. Show first that $M(f)(x) = f(x)$ at these particular points. Now consider the functional identities

$$2f(x) + f(1+x) = x + \frac{1+\sqrt{3}}{2}$$
$$2f(2+x) + f(1+x) = -x + \frac{3-\sqrt{3}}{2}$$

for $x \in [0,1]$. If $f(x)$ satisfies these two identities, then show that $M(f)(x)$ does as well. The set S of continuous functions $f(x)$ that have the values 0, $2a$, $2\overline{a}$, and 0 at 0, 1, 2, and 3 and that satisfy the two functional identities is nonempty. Indeed, prove that S contains the function that takes the required values and is linear between successive integers on $[0,3]$. Also show that S is a closed, convex subset of the Banach space (complete normed linear space) of continuous functions on $[0,3]$ under the norm $\|f\| = \sup_{x\in[0,3]} |f(x)|$. Given this fact, prove that $M(f)$ is a contraction mapping on S and therefore has a unique fixed point $\psi(x)$ [4]. Here it is helpful to note that $|\overline{a}| \leq |a| < 1$.

11. Continuing Problem 10, suppose we extend the continuous, fixed-point function $\psi(x)$ of the contraction map M to the entire real line by setting $\psi(x) = 0$ for $x \notin [0,3]$. Show that $\psi(x)$ satisfies the scaling equation (4). (Hint: You will have to use the two functional identities imposed on S as well as the functional identities implied by the fixed-point property of M.)

12. Demonstrate that the constant 1 plus the periodic wavelets defined by equation (16) constitute an orthonormal basis for $\mathrm{L}^2[0,1]$.

References

[1] Antoniadis A, Oppenheim G (editors) (1995) *Wavelets and Statistics.* Springer-Verlag, New York

[2] Daubechies I (1992) *Ten Lectures on Wavelets.* SIAM, Philadelphia

[3] Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. *J Amer Stat Assoc* 90:1200–1224

[4] Hoffman K (1975) *Analysis in Euclidean Space.* Prentice-Hall, Englewood Cliffs, NJ

[5] Jawerth B, Sweldens W (1994) An overview of wavelet based multiresolution analysis. *SIAM Review* 36:377–412
[6] Kolaczyk ED (1996) A wavelet shrinkage approach to tomographic image reconstruction. *J Amer Stat Assoc* 91:1079–1090
[7] Meyer Y (1993) *Wavelets: Algorithms and Applications.* Ryan RD, translator, SIAM, Philadelphia
[8] Pollen D (1992) Daubechies's scaling function on [0,3]. *Wavelets: A Tutorial in Theory and Applications,* Chui CK, editor, Academic Press, New York, pp 3–13
[9] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[10] Strang G (1989) Wavelets and dilation equations: A brief introduction. *SIAM Review* 31:614–627
[11] Strichartz RS (1993) How to make wavelets. *Amer Math Monthly* 100:539–556
[12] Walter GG (1994) *Wavelets and Other Orthogonal Systems with Applications.* CRC Press, Boca Raton, FL
[13] Wickerhauser MV (1992) Acoustic signal compression with wavelet packets. *Wavelets: A Tutorial in Theory and Applications,* Chui CK, editor, Academic Press, New York, pp 679–700

20
Generating Random Deviates

20.1 Introduction

Statisticians rely on a combination of mathematical theory and statistical simulation to develop new methods. Because simulations are often conducted on a massive scale, it is crucial that they be efficiently executed. In the current chapter, we investigate techniques for producing random samples from univariate and multivariate distributions. These techniques stand behind every successful simulation and play a critical role in Monte Carlo integration. Exceptionally fast code for simulations almost always depends on using a lower-level computer language such as C or Fortran. This limitation forces the statistician to write custom software. Mastering techniques for generating random variables (or deviates in this context) is accordingly a useful survival skill.

Almost all lower-level computer languages fortunately have facilities for computing a random sample from the uniform distribution on $[0, 1]$. Although there are important philosophical, mathematical, and statistical issues involved in whether and to what extent a deterministic computer can deliver independent uniform deviates [12, 16], we take the relaxed attitude that this problem has been solved for all practical purposes. Our focus here is on fast methods for turning uniform deviates into more complicated random samples [4, 2, 9, 10, 12, 14, 16, 17]. Because statisticians must constantly strike a balance between programming costs and machine efficiency, we stress methods that are straightforward to implement.

20.2 The Inverse Method

The inverse method embodied in the next proposition is one of the simplest and most natural methods of generating random variables [1].

Proposition 20.2.1. *Let X be a random variable with distribution function $F(x)$.*

(a) *If $F(x)$ is continuous, then $U = F(X)$ is uniformly distributed on $[0, 1]$.*
(b) *Even if $F(x)$ is not continuous, the inequality $\Pr[F(X) \leq t] \leq t$ is still true for all $t \in [0, 1]$.*
(c) *If $F^{[-1]}(y) = \inf\{x : F(x) \geq y\}$ for any $0 < y < 1$, and if U is uniform on $[0, 1]$, then $F^{[-1]}(U)$ has distribution function $F(x)$.*

Proof. Let us first demonstrate that

$$\Pr[F(X) \leq F(t)] = F(t). \tag{1}$$

To prove this assertion, note that $\{X > t\} \cap \{F(X) < F(t)\} = \emptyset$ and $\{X \leq t\} \cap \{F(X) > F(t)\} = \emptyset$ together entail

$$\{F(X) \leq F(t)\} = \{X \leq t\} \cup \{F(X) = F(t),\ X > t\}.$$

However, the event $\{F(X) = F(t),\ X > t\}$ maps under X to an interval of constancy of $F(x)$ and therefore has probability 0. Equation (1) follows immediately.

For part (a) let $u \in (0, 1)$. Because $F(x)$ is continuous, there exists t with $F(t) = u$. In view of equation (1),

$$\Pr[F(X) \leq u] = \Pr[F(X) \leq F(t)] = u.$$

Part (c) follows if we can show that the events $u \leq F(t)$ and $F^{[-1]}(u) \leq t$ are identical for u and $F(t)$ both in $(0, 1)$. Assume that $F^{[-1]}(u) \leq t$. Because $F(x)$ is increasing and right continuous, the set $\{x : u \leq F(x)\}$ is an interval containing its left endpoint. Hence, $u \leq F(t)$. Conversely, if $u \leq F(t)$, then $F^{[-1]}(u) \leq t$ by definition. Finally for part (b), apply part (c) and write $X = F^{[-1]}(U)$ for U uniform on $[0, 1]$. Then the inequality $U \leq F(X)$ implies

$$\Pr[F(X) \leq t] \leq \Pr(U \leq t) = t.$$

This completes the proof. □

Example 20.2.1 (*Exponential Distribution*). If X exponentially distributed with mean 1, then $F(x) = 1 - e^{-x}$, and $F^{[-1]}(u) = -\ln(1 - u)$. Because $1 - U$ is uniform on $[0, 1]$ when U is uniform on $[0, 1]$, the random variable $-\ln U$ is distributed as X. The positive multiple $Y = -\mu \ln U$ is exponentially distributed with mean μ. ■

Example 20.2.2 (*Cauchy Distribution*). The distribution function $F(x) = 1/2 + \arctan(x)/\pi$ of a standard Cauchy random variable X has inverse $F^{[-1]}(u) = \tan[\pi(u - 1/2)]$. To generate a Cauchy random variable

$Y = \sigma X + \mu$ with location and scale parameters μ and σ, simply take $Y = \sigma \tan[\pi(U - 1/2)] + \mu$ for U uniform on $[0, 1]$. ■

Example 20.2.3 (*Probability Plots*). If $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics of a random sample from a continuous distribution function $F(x)$, then taking $U_i = F(X_i)$ generates the order statistics $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ of a random sample from the uniform distribution. The fact that $\mathrm{E}[U_{(i)}] = i/(n+1)$ suggests that a plot of the points $(i/[n+1], F[X_{(i)}])$ should fall approximately on a straight line. This is the motivation for the diagnostic tool of probability plotting [15]. ■

Example 20.2.4 (*Discrete Uniform*). One can sample a number uniformly from the set $\{1, 2, \ldots, n\}$ by taking $\lfloor nU \rfloor + 1$, where U is uniform on $[0, 1]$ and $\lfloor r \rfloor$ denotes the greatest integer less than or equal to the real number r. ■

Example 20.2.5 (*Geometric*). In a Bernoulli sampling scheme with success probability p, the number of trials N until the first success follows a geometric distribution. If we choose λ so that $q = 1 - p = e^{-\lambda}$, then N can be represented as $N = \lfloor X \rfloor + 1$, where X is exponentially distributed with intensity λ. Indeed,

$$\begin{aligned} \Pr(N = k + 1) &= \Pr(k \leq X < k + 1) \\ &= e^{-\lambda k} - e^{-\lambda(k+1)} \\ &= q^k - q^{k+1}. \end{aligned}$$

In light of Example 20.2.1, $N = \lfloor -\ln(U)/\lambda \rfloor + 1$, where U is uniform on $[0, 1]$. For the geometric that counts total failures until success rather than total trials, we replace N by $N - 1$. ■

20.3 Normal Random Deviates

Although in principle normal random deviates can be generated by the inverse method, the two preferred methods involve substantially less computation. Both the Box and Muller and the Marsaglia methods generate two independent, standard normal deviates X and Y at a time starting from two independent, uniform deviates U and V. The Box and Muller method transforms the random Cartesian coordinates (X, Y) in the plane to random polar coordinates (Θ, R). It is clear from their joint density $e^{-\frac{r^2}{2}} r/(2\pi)$ that Θ and R are independent, with Θ uniformly distributed on $[0, 2\pi]$ and R^2 exponentially distributed with mean 2. Example 20.2.1 says we can generate Θ and R^2 by taking $\Theta = 2\pi U$ and $R^2 = -2 \ln V$. Transforming from polar coordinates back to Cartesian coordinates, we define $X = R \cos \Theta$ and $Y = R \sin \Theta$.

In Marsaglia's polar method, a random point (U, V) in the unit square is transformed into a random point (S, T) in the square $[-1, 1] \times [-1, 1]$ by

taking $S = 2U - 1$ and $T = 2V - 1$. If $W^2 = S^2 + T^2 > 1$, then the random point (S, T) falls outside the unit circle. When this occurs, the current U and V are discarded and resampled. If $W^2 = S^2 + T^2 \leq 1$, then the point (S, T) generates a uniformly distributed angle Θ with $\cos\Theta = S/W$ and $\sin\Theta = T/W$. Furthermore, the distribution of the random variable $Z = -2\ln W^2$ is

$$\begin{aligned}\Pr(Z \leq z) &= \Pr(W \geq e^{-\frac{z}{4}}) \\ &= 1 - \frac{\pi(e^{-\frac{z}{4}})^2}{\pi} \\ &= 1 - e^{-\frac{z}{2}},\end{aligned}$$

which implies that Z is distributed as R^2 in the Box and Muller method. Thus, we need only set

$$\begin{aligned}X &= \sqrt{-2\ln W^2}\frac{S}{W} \\ Y &= \sqrt{-2\ln W^2}\frac{T}{W}\end{aligned}$$

to recover the normally distributed pair (X, Y).

The polar method avoids the trigonometric function evaluations of the Box and Muller method but uses $4/\pi$ as many random pairs (U, V) on average. Both methods generate normal deviates with mean μ and variance σ^2 by replacing X and Y by $\sigma X + \mu$ and $\sigma Y + \mu$.

20.4 Acceptance–Rejection Method

The acceptance–rejection method is predicated on the notion of majorization [7]. Suppose we want to sample from a complicated probability density $f(x)$ that is majorized by a simple probability density $g(x)$ in the sense that $f(x) \leq cg(x) = h(x)$ for all x and some constant $c > 1$. If we sample a deviate X distributed according to $g(x)$, then we can accept or reject X as a representative of $f(x)$. John von Neumann suggested making this decision based on sampling a uniform deviate U and accepting X if and only if $U \leq f(X)/h(X)$. This procedure gives the probability of generating an accepted value in the interval $(x, x + dx)$ as proportional to

$$g(x)dx\frac{f(x)}{h(x)} = \frac{1}{c}f(x)dx.$$

In other words, the density function of the accepted deviates is precisely $f(x)$. The fraction of sampled deviates accepted is $1/c$.

As we have seen in Example 20.2.1, generating exponential deviates is computationally quick. This fact suggests exploiting exponential curves as majorizing functions in the acceptance–rejection method [2]. On a log scale,

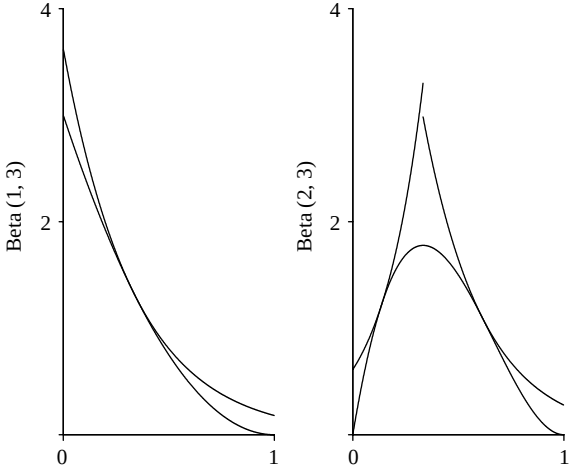

FIGURE 20.1. Exponential Envelopes for Two Beta Densities

an exponential curve is a straight line. If a density $f(x)$ is log-concave, then any line tangent to $\ln f(x)$ will lie above $\ln f(x)$. Thus, log-concave densities are ideally suited to acceptance–rejection sampling with piecewise exponential envelopes. Commonly encountered log-concave densities include the normal, the gamma with shape parameter $\alpha \geq 1$, the beta with parameters α and $\beta \geq 1$, the exponential power density, and Fisher's z density. The reader can easily check log concavity in each of these examples and in the three additional examples mentioned in Problem 5 by showing that $\frac{d^2}{dx^2} \ln f(x) \leq 0$ on the support of $f(x)$.

A strictly log-concave density $f(x)$ defined on an interval is unimodal. The mode m of $f(x)$ may occur at either endpoint or on the interior of the interval. In the former case, we suggest using a single exponential envelope; in the latter case, two exponential envelopes oriented in opposite directions from the mode m. Figure 20.1 depicts the two situations. With different left and right envelopes, the appropriate majorizing function is

$$h(x) = \begin{cases} c_l \lambda_l e^{-\lambda_l (m-x)} & x < m \\ c_r \lambda_r e^{-\lambda_r (x-m)} & x \geq m. \end{cases}$$

Note that $h(x)$ has total mass $c = c_l + c_r$. To minimize the rejection rate and maximize the efficiency of sampling, we minimize the mass constants c_l and c_r. Geometrically this is accomplished by choosing optimal tangent points x_l and x_r. The tangency condition for the right envelope amounts to

$$\begin{aligned} f(x_r) &= c_r \lambda_r e^{-\lambda_r (x_r - m)} \\ f'(x_r) &= -c_r \lambda_r^2 e^{-\lambda_r (x_r - m)}. \end{aligned} \tag{2}$$

These equations allow us to solve for λ_r as $-f'(x_r)/f(x_r)$ and then for c_r as

$$c_r(x_r) = -\frac{f(x_r)^2}{f'(x_r)} e^{-\frac{f'(x_r)}{f(x_r)}(x_r - m)}.$$

Finding x_r to minimize c_r is now a matter of calculus. A similar calculation for the left envelope shows that $c_l(x_l) = -c_r(x_l)$.

Example 20.4.1 (*Exponential Power Density*). The exponential power density

$$f(x) = \frac{e^{-|x|^\alpha}}{2\Gamma(1 + \frac{1}{\alpha})}, \qquad \alpha \geq 1,$$

has mode $m = 0$. For $x_r \geq 0$ we have

$$\lambda_r = \alpha x_r^{\alpha - 1}$$
$$c_r(x_r) = \frac{e^{(\alpha-1)x_r^\alpha}}{2\Gamma(1 + \frac{1}{\alpha})\alpha x_r^{\alpha-1}}.$$

The equation $\frac{d}{dx} c_r(x) = 0$ has solution $-x_l = x_r = \alpha^{-1/\alpha}$. This allows us to calculate the acceptance probability

$$\frac{1}{2c_r(x_r)} = \Gamma\Big(1 + \frac{1}{\alpha}\Big)\alpha^{\frac{1}{\alpha}} e^{\frac{1}{\alpha} - 1},$$

which ranges from 1 at $\alpha = 1$ (the double or bilateral exponential distribution) to $e^{-1} = .368$ as α tends to ∞. For a normal density ($\alpha = 2$), the acceptance probability reduces to $\sqrt{\pi/2e} \approx .76$. In practical implementations, the acceptance–rejection method for normal deviates is slightly less efficient than the polar method. ■

A completely analogous development holds for a discrete density $f(x)$ defined and positive on an interval of integers. Now, however, we substitute the easily generated geometric distribution for the exponential distribution [3, 8]. In extending the notion of log concavity to a discrete density $f(x)$, we linearly interpolate $\ln f(x)$ between supporting integers as shown in Figure 20.2. If the linearly interpolated function is concave, then $f(x)$ is said to be log concave. Analytically, log concavity of $f(x)$ is equivalent to the inequality

$$\ln f(x) \geq \frac{1}{2}[\ln f(x-1) + \ln f(x+1)]$$

for all supporting integers x. This inequality is in turn equivalent to the inequality $f(x)^2 \geq f(x-1)f(x+1)$ for all integers x.

For a discrete density with an interior mode m, the majorizing function

$$h(x) = \begin{cases} c_l(1 - q_l)q_l^{m-1-x} & x < m \\ c_r(1 - q_r)q_r^{x-m} & x \geq m \end{cases}$$

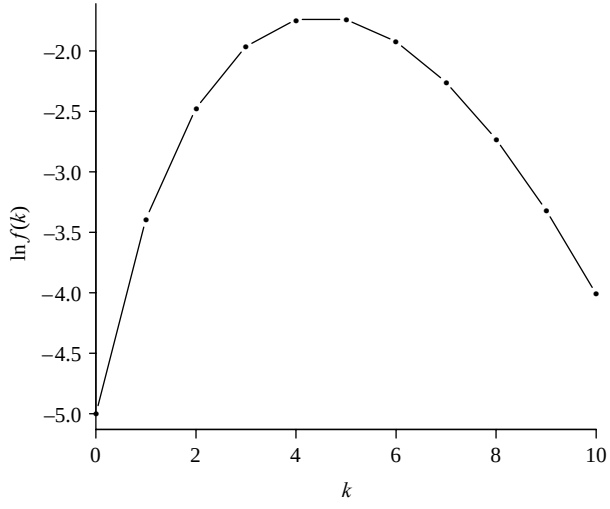

FIGURE 20.2. Linearly Interpolated Log Poisson Density

consists of two geometric envelopes oriented in opposite directions from the mode m. The analog of the tangency condition (2) is

$$\begin{aligned} f(x_r) &= c_r(1-q_r)q_r^{x_r-m} \\ f(x_r+1) &= c_r(1-q_r)q_r^{x_r+1-m}. \end{aligned}$$

Solving these two equations gives $q_r = f(x_r+1)/f(x_r)$ and

$$c_r(x_r) = \frac{f(x_r)}{1-\frac{f(x_r+1)}{f(x_r)}}\left[\frac{f(x_r+1)}{f(x_r)}\right]^{m-x_r}.$$

We now minimize the mass constant $c_r(x_r)$ by adjusting x_r. To the left of the mode, a similar calculation yields

$$\begin{aligned} c_l(x_l) &= \frac{f(x_l)}{1-\frac{f(x_l)}{f(x_l+1)}}\left[\frac{f(x_l)}{f(x_l+1)}\right]^{x_l+1-m} \\ &= -c_r(x_l). \end{aligned}$$

Example 20.4.2 (*Poisson Distribution*). For the Poisson density $f(x) = \lambda^x e^{-\lambda}/x!$, the mode $m = \lfloor \lambda \rfloor$ because $f(x+1)/f(x) = \lambda/(x+1)$. It follows that the mass constant

$$c_r(x_r) = \frac{\frac{\lambda^{x_r}e^{-\lambda}}{x_r!}\left(\frac{\lambda}{x_r+1}\right)^{m-x_r}}{1-\frac{\lambda}{x_r+1}}.$$

To minimize $c_r(x_r)$, we treat x_r as a continuous variable and invoke Stirling's asymptotic approximation in the form

$$\ln x! = \ln(x+1)! - \ln(x+1)$$

$$\begin{aligned}
&= \left(x + \frac{3}{2}\right) \ln(x+1) - (x+1) + \ln\sqrt{2\pi} - \ln(x+1) \\
&= \left(x + \frac{1}{2}\right) \ln(x+1) - (x+1) + \ln\sqrt{2\pi}.
\end{aligned}$$

Substitution of this expression in the expansion of $\ln c_r(x_r)$ produces

$$\begin{aligned}
\ln c_r(x_r) = {} & x_r \ln\lambda - \lambda - \left(x_r + \frac{1}{2}\right) \ln(x_r+1) + (x_r+1) - \ln\sqrt{2\pi} \\
& + (m - x_r) \ln\lambda - (m - x_r) \ln(x_r+1) - \ln\left(1 - \frac{\lambda}{x_r+1}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{d}{dx_r} \ln c_r(x_r) &= \ln\lambda - \ln(x_r+1) - \frac{x_r + \frac{1}{2}}{x_r+1} + 1 - \ln\lambda \\
&\quad + \ln(x_r+1) - \frac{m - x_r}{x_r+1} - \frac{\frac{\lambda}{(x_r+1)^2}}{1 - \frac{\lambda}{x_r+1}} \\
&= \frac{x_r^2 + \left(\frac{3}{2} - m - \lambda\right) x_r + \frac{1}{2} - m - \frac{3}{2}\lambda + m\lambda}{(x_r+1)(x_r+1-\lambda)}.
\end{aligned}$$

Setting this derivative equal to 0 identifies x_r and x_l as the two roots of the quadratic equation $x^2 + (3/2 - m - \lambda)x + 1/2 - m - 3\lambda/2 + m\lambda = 0$. ■

The efficiency of the acceptance–rejection method depends heavily on programming details. For instance, the initial choice of the left or right envelope in a two-envelope problem involves comparing a uniform random variable U to the ratio $r = c_l/(c_l + c_r)$. Once this choice is made, U can be reused to generate the appropriate exponential deviate. Indeed, given $U < r$, the random variable $V = U/r$ is uniformly distributed on $[0, 1]$ and independent of U. Similarly, given the complementary event $U \geq r$, the random variable $W = (U - r)/(1 - r)$ is also uniformly distributed on $[0, 1]$ and independent of U. In the continuous case, the acceptance step is based on the ratio $f(x)/h(x)$. Usually it is more efficient to base acceptance on the log ratio

$$\ln \frac{f(x)}{h(x)} = \begin{cases} \ln f(x) - \ln f(x_l) + \lambda_l(x_l - x) & x < m \\ \ln f(x) - \ln f(x_r) + \lambda_r(x - x_r) & x \geq m \end{cases},$$

from which one cancels as many common terms as possible. Finally, we stress that piecewise exponential envelopes are not the only majorizing functions possible. Problems 11, 12, and 13 provide some alternative examples.

20.5 Ratio Method

The ratio method is a kind of generalization of the polar method. Suppose that $f(x)$ is a probability density and $h(x) = cf(x)$ for $c > 0$. Consider the set $S_h = \{(u, v) : 0 < u \leq \sqrt{h(v/u)}\}$ in the plane. If this set is bounded, then we can enclose it in a well-behaved set such as a rectangle and sample uniformly from the enclosing set. The next proposition shows how this leads to a method for sampling from $f(x)$.

Proposition 20.5.1. *Suppose $k_u = \sup_x \sqrt{h(x)}$ and $k_v = \sup_x |x|\sqrt{h(x)}$ are finite. Then the rectangle $[0, k_u] \times [-k_v, k_v]$ encloses S_h. If $h(x) = 0$ for $x < 0$, then the rectangle $[0, k_u] \times [0, k_v]$ encloses S_h. Finally, if the point (U, V) sampled uniformly from the enclosing set falls within S_h, then the ratio $X = V/U$ is distributed according to $f(x)$.*

Proof. From the definition of S_h it is clear that the permitted u lie in $[0, k_u]$. Multiplying the inequality $|v|u/|v| \leq \sqrt{h(v/u)}$ by $|v|/u$ implies that $|v| \leq k_v$. If $h(x) = 0$ for $x < 0$, then no $v < 0$ yields a pair (u, v) in S_h. Finally, note that the transformation $(u, v) \to (u, v/u)$ has Jacobian u^{-1}. Hence,

$$\begin{aligned}\int\int 1_{\{\frac{v}{u} \leq x_0\}} 1_{\{0<u\leq\sqrt{h(\frac{v}{u})}\}} du\, dv &= \int\int 1_{\{x \leq x_0\}} 1_{\{0<u\leq\sqrt{h(x)}\}} u du\, dx \\ &= \int_{-\infty}^{x_0} \frac{1}{2} h(x) dx\end{aligned}$$

is the distribution function of the accepted X up to a normalizing constant. □

Example 20.5.1 (*Gamma with Shape Parameter $\alpha > 1$*)**.** Here we take $h(x) = x^{\alpha-1}e^{-x}1_{(0,\infty)}(x)$. The maximum of $\sqrt{h(x)}$ occurs at $x = \alpha - 1$ and equals $k_u = [(\alpha - 1)/e]^{(\alpha-1)/2}$. Likewise, the maximum of $x\sqrt{h(x)}$ occurs at $x = \alpha + 1$ and equals $k_v = [(\alpha + 1)/e]^{(\alpha+1)2}$. To carry out the ratio method, we sample uniformly from the rectangular region $[0, k_u] \times [0, k_v]$ by multiplying two independent, uniform deviates U and V by k_u and k_v, respectively. The ratio $X = k_v V/(k_u U)$ is accepted as a random deviate from $f(x)$ if and only if

$$k_u U \leq X^{\frac{\alpha-1}{2}} e^{-\frac{X}{2}},$$

which simplifies to

$$\frac{2}{\alpha - 1} \ln U - 1 - \ln W + W \leq 0$$

for $W = X/(\alpha - 1)$. ■

20.6 Deviates by Definition

In many cases we can generate a random variable by exploiting its definition in terms of simpler random variables or familiar stochastic processes. Here are some examples.

Example 20.6.1 (*Binomial*). For a small number of trials n, a binomial deviate S_n can be quickly generated by taking n independent, uniform deviates $U_1, \ldots, U_n$ and setting $S_n = \sum_{i=1}^{n} 1_{\{U_i \leq p\}}$, where p is the success probability per trial. ■

Example 20.6.2 (*Negative Binomial*). Consider a Bernoulli sampling process with success probability p. The number of trials S_n until the nth success follows a negative binomial distribution. When $n = 1$ we recover the geometric distribution. Adding n independent geometric deviates gives the negative binomial. In view of the ease with which we can generate geometric deviates (Example 20.2.5), we can sample S_n quickly for n small. ■

Example 20.6.3 (*Poisson*). To generate a Poisson deviate X with mean λ, consider a Poisson process with unit intensity. The number of random points falling on the interval $[0, \lambda]$ follows a Poisson distribution with mean λ. Furthermore, the waiting times between successive random points are independent, exponentially distributed random variables with common mean 1. If we generate a sequence $Z_1, Z_2, \ldots$ of independent, exponential deviates and stop with Z_j satisfying $\sum_{i=1}^{j-1} Z_i \leq \lambda < \sum_{i=1}^{j} Z_i$, then $X = j - 1$. Rephrasing the stopping condition as $\prod_{i=1}^{j-1} e^{-Z_i} \geq e^{-\lambda} > \prod_{i=1}^{j} e^{-Z_i}$ allows us to use uniform deviates $U_1, \ldots, U_j$ since Example 20.2.1 implies that U_i and e^{-Z_i} are identically distributed. This procedure is more efficient for small λ than the acceptance–rejection method discussed in Example 20.4.2. ■

Example 20.6.4 (*Lognormal*). If a random variable X can be represented as a function $f(Y)$ of another random variable that is easy to generate, then obviously we should sample Y and compute $f(Y)$ to generate X. For example, if X is a standard normal deviate, then $e^{\sigma X + \mu}$ is a lognormal deviate for all choices of the mean μ and standard deviation σ of the normal deviate $\sigma X + \mu$. ■

Example 20.6.5 (*Chi-Square*). A chi-square distribution with n degrees of freedom is a gamma distribution with shape parameter $\alpha = n/2$ and scale parameter $\beta = 2$. The acceptance–rejection method sketched in Problem 7 or the ratio method discussed in Example 20.5.1 delivers a gamma deviate with shape parameter α and scale parameter 1. Doubling their output when $\alpha = n/2$ gives a χ^2_n deviate. Alternatively for small n, we can exploit the definition of χ^2_n as a sum of squares of n independent, standard normal deviates. Once we have generated a χ^2_n deviate, we can compute derived

deviates such as the inverse chi-square, the inverse chi, and the log chi-square by forming $1/\chi_n^2$, $1/\chi_n$, and $\ln \chi_n^2$, respectively. ■

Example 20.6.6 (*F Distribution*). If χ_m^2 and χ_n^2 are independent chi-square random variables with m and n degrees of freedom, respectively, then the ratio

$$F_{mn} = \frac{\frac{1}{m}\chi_m^2}{\frac{1}{n}\chi_n^2}$$

follows an F distribution. Since we can readily generate chi-square deviates, this definition provides a convenient method for generating F deviates. ■

Example 20.6.7 (*Student's t Distribution*). If X is a standard normal deviate and χ_n^2 is an independent chi-square deviate, then the definition $T_n = X/\sqrt{\frac{1}{n}\chi_n^2}$ gives a convenient method of generating t deviates. ■

Example 20.6.8 (*Beta*). If X_α and X_β are independent gamma deviates with shape parameters α and β and scale parameter 1, then the ratio $X_\alpha/(X_\alpha + X_\beta)$ is by definition a beta deviate with parameter pair (α, β). ■

20.7 Multivariate Deviates

We make no attempt to be systematic in presenting the following examples.

Example 20.7.1 (*Multivariate Normal*). In R^n the simplest multivariate normal random vector X has n independent, standard normal components. To generate a multivariate normal deviate Y with mean vector μ and covariance matrix Ω, we first generate X and then form $Y = \Omega^{1/2}X + \mu$. Any square root $\Omega^{1/2}$ will do; for instance, we can use the Cholesky decomposition of Ω. ■

Example 20.7.2 (*Multivariate t*). Let X be a multivariate normal deviate with mean vector $\mathbf{0}$ and covariance matrix Ω, and let χ_ν^2 be an independent chi-square deviate with possibly noninteger degrees of freedom ν. The translated ratio

$$T_\nu = \frac{X}{\sqrt{\chi_\nu^2/\nu}} + \mu \tag{3}$$

follows a multivariate t distribution with location vector μ, scale matrix Ω, and degree of freedom ν [5]. For ν small, the t distribution has fatter tails than the normal distribution and offers the opportunity of estimating location and scale parameters robustly [13]. As ν tends to ∞, the t distribution tends to the normal with mean vector μ and covariance matrix Ω. Again, the most natural way of generating T_ν is via its definition (3). ■

Example 20.7.3 (*Multivariate Uniform*). In R^n there are many sets of finite Lebesgue measure from which we might want sample uniformly. The rectangle $[a,b] = \prod_{i=1}^{n}[a_i,b_i]$ is the simplest. In this case we take n independent uniform deviates $U_1,\ldots,U_n$ and construct the vector V with ith component $V_i = (b_i - a_i)U_i + a_i$. To sample uniformly from the unit sphere $S_n = \{x : \|x\|_2 \leq 1\}$, we sample a standard, multivariate normal random vector X and note that $V = X/\|X\|_2$ is uniformly distributed on the surface of S_n. We then choose an independent radius $R \leq 1$ and form the contracted point RV within S_n. Since the volume of a sphere depends on the nth power of its radius, we construct R by the inverse method employing the distribution function $F(r) = r^n$ on $[0,1]$. More complicated sets can be accommodated by enclosing them within a rectangle or sphere and using a rejection procedure. ■

Example 20.7.4 (*Dirichlet*). The Dirichlet density is the natural generalization of the beta density [11]. To generate a Dirichlet deviate $Y = (Y_1,\ldots,Y_n)^t$, we take n independent gamma deviates $X_1,\ldots,X_n$ with shape parameters α_i and scale parameters 1 and form the ratios

$$Y_i = \frac{X_i}{\sum_{i=1}^{n} X_i}.$$

The random vector Y lives on the simplex

$$\Delta_n = \{(y_1,\ldots,y_n)^t : y_1 > 0\ ,\ldots,y_n > 0,\ \sum_{i=1}^{n} y_i = 1\}$$

and has density

$$f(y) = \frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n} y^{\alpha_i - 1}$$

there relative to the uniform measure. When each $\alpha_i = 1$, the deviate Y is uniformly distributed on Δ_n, and the reduced deviate $(Y_1,\ldots,Y_{n-1})^t$ derived by deleting Y_n is uniformly distributed over the reduced simplex

$$\left\{(y_1,\ldots,y_{n-1})^t : y_1 > 0,\ldots,y_{n-1} > 0, \sum_{i=1}^{n-1} y_i \leq 1\right\}.$$

■

Example 20.7.5 (*Order Statistics*). At first glance, generating order statistics $Y_{(1)},\ldots,Y_{(n)}$ from a distribution $F(y)$ appears easy. However, if n is large and we are interested only in a few order statistics at the beginning or end of the sequence, we can do better than generate all n deviates Y_i and order them in $O(n\ln n)$ steps. Consider the special case of exponential deviates X_i with mean 1. From the calculation

$$\Pr(X_{(1)} \geq x) = \prod_{i=1}^{n}\Pr(X_i \geq x) = e^{-nx},$$

we find that $X_{(1)}$ is exponentially distributed with intensity n. Because of the lack of memory property of the exponential, the $n-1$ random points to the right of $X_{(1)}$ provide an exponentially distributed sample of size $n-1$ starting at $X_{(1)}$. Duplicating our argument for $X_{(1)}$, we find that the difference $X_{(2)} - X_{(1)}$ is independent of $X_{(1)}$ and exponentially distributed with intensity $n-1$. Arguing inductively we now see that $Z_1 = X_{(1)}$ and that the differences $Z_{i+1} = X_{(i+1)} - X_{(i)}$ are independent and that Z_i is exponentially distributed with intensity $n-i+1$. This result, which is proved more rigorously in [6], suggests that we can sample the Z_i and add them to get the $X_{(i)}$. If we are interested only in the $X_{(i)}$ for $i \leq j$, then we omit generating $Z_{j+1}, \ldots, Z_n$.

To capitalize on this special case, note that Example 20.2.1 permits us to generate the n-order statistics from the uniform distribution by defining $U_{(i)} = e^{-X_{(n-i+1)}}$. The order statistics are reversed here because e^{-x} is strictly decreasing. Given the fact that $1-U$ is uniform when U is uniform, we can equally well define $U_{(i)} = 1 - e^{-X_{(i)}}$. If we desire the order statistics $Y_{(i)}$ from a general, continuous distribution function $F(y)$ with inverse $F^{[-1]}(u)$, then we apply the inverse method and set

$$Y_{(i)} = F^{[-1]}(e^{-X_{(n-i+1)}})$$

or

$$Y_{(i)} = F^{[-1]}(1 - e^{-X_{(i)}}).$$

The first construction is more convenient if we want only the last j-order statistics, and the second construction is more convenient if we want only the first j-order statistics. In both cases we generate only $X_{(1)}, \ldots, X_{(j)}$. ■

20.8 Problems

1. Discuss how you would use the inverse method to generate a random variable with (a) the continuous logistic density

$$f(x|\mu,\sigma) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma[1+e^{-\frac{x-\mu}{\sigma}}]^2},$$

(b) the Pareto density

$$f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}} 1_{(\alpha,\infty)}(x),$$

and (c) the Weibull density

$$f(x|\delta,\gamma) = \frac{\gamma}{\delta} x^{\gamma-1} e^{-\frac{x^\gamma}{\delta}} 1_{(0,\infty)}(x),$$

where α, β, γ, δ, and σ are taken positive.

2. Continuing Problem 1, discuss how the inverse method applies to (d) the Gumbel density

$$f(x) = e^{-x}e^{-e^{-x}},$$

(e) the arc sine density

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}1_{(0,1)}(x),$$

and (f) the slash density

$$f(x) = \alpha x^{\alpha-1}1_{(0,1)}(x),$$

where $\alpha > 0$.

3. Demonstrate how Examples 20.2.4 and 20.2.5 follow from Proposition 20.2.1.

4. Show that the normal distribution, the gamma distribution with shape parameter $\alpha \geq 1$, the beta distribution with parameters α and $\beta \geq 1$, the exponential power distribution with parameter $\alpha \geq 1$, and Fisher's z distribution of Problem 8 all have log-concave densities.

5. Verify that the logistic and Weibull ($\gamma \geq 1$) densities of Problem 1 and the Gumbel density of Problem 2 are log concave. Prove that the Cauchy density is not log concave.

6. Check that the Poisson, binomial, negative binomial, and hypergeometric distributions have discrete log-concave densities.

7. Verify that the gamma distribution with shape parameter $\alpha > 1$ and scale parameter $\beta = 1$ has mode $m = \alpha - 1$ and that the beta distribution with parameters $\alpha > 1$ and $\beta > 1$ has mode

$$m = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

Demonstrate that the corresponding optimal tangency points of the acceptance–rejection method of Section 20.4 are the roots of the respective quadratics

$$\begin{aligned} m(m-x)^2 - x &= (\alpha-1)(\alpha-1-x)^2 - x \\ &\quad (\alpha+\beta-1)x^2 + (2m - \alpha - \alpha m - \beta m)x + \alpha m - m \\ &= (\alpha+\beta-1)x^2 + (1-2\alpha)x + \frac{(1-\alpha)^2}{\alpha+\beta-2}. \end{aligned}$$

(Hints: You may want to use a computer algebra program such as Maple. The beta distribution involves a quartic polynomial, one of whose quadratic factors has imaginary roots.)

8. If X has an F distribution with m and n degrees of freedom, then $\ln(X)/2$ has Fisher's Z distribution. Show that $\ln X$ has density

$$f(x) = \frac{m^{\frac{m}{2}}n^{\frac{n}{2}}\Gamma(\frac{m}{2}+\frac{n}{2})e^{\frac{my}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(n+me^y)^{\frac{m}{2}+\frac{n}{2}}}.$$

Prove that $f(x)$ has mode $m = 0$ and that the corresponding optimal tangency points x_l and x_r of the acceptance–rejection method of Section 20.4 are the roots of the transcendental equation

$$m(nx - 2)e^x = 2n + mnx.$$

(Hint: You may want to use a computer algebra program such as Maple.)

9. Demonstrate that the binomial density $f(x) = \binom{n}{x}p^x q^{n-x}$ has mode $\lfloor (n+1)p \rfloor$ and that the negative binomial density $f(x) = \binom{x+n-1}{n-1}p^n q^x$ has mode $\lfloor (n-1)q/p \rfloor$. Verify that the corresponding optimal tangency points x_l and x_r of the acceptance–rejection method of Section 20.4 are the roots of the respective quadratic equations

$$2x^2 + (3 - 2m - 2np)x + 1 - 2m - p + 2mp - 3np + 2mnp = 0,$$
$$2px^2 + (3p - 2mp - 2nq)x + 1 - 2m - 3nq + 2mnq = 0.$$

(Hints: Use Stirling's formula and a computer algebra program such as Maple.)

10. Check that the hypergeometric density

$$f(x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$$

has mode $m = \lfloor (R+1)(n+1)/(2+N) \rfloor$. Prove that the corresponding optimal tangency points x_l and x_r of the acceptance–rejection method of Section 20.4 are the roots of the quadratic

$$0 = (2N + 2)x^2 + (3N - 2mN - 4m + 4 - 2nR)x$$
$$+ \ (2m - 1)(R - N + n - 1 + nR) - 2nR.$$

(Hints: Use Stirling's formula and a computer algebra program such as Maple. This quadratic is one of two possible quadratics. Why can you discard the other one?)

11. The von Mises distribution for a random angle Θ has density

$$f(\theta) = \frac{e^{\kappa \cos\theta}}{I_0(\kappa)},$$

where $\theta \in [0, 2\pi]$, $\kappa > 0$, and $I_0(\kappa)$ is a Bessel function. Devise an acceptance–rejection method for generating random deviates from $f(\theta)$.

12. Devise an acceptance–rejection method for generating beta deviates based on the inequality $x^{\alpha-1}(1-x)^{\beta-1} \leq x^{\alpha-1} + (1-x)^{\beta-1}$.

13. When $\alpha < 1$, show that the gamma density

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}e^{-x}1_{(0,\infty)}(x)$$

with scale 1 is majorized by the mixture density

$$g(x) = \frac{e}{e+\alpha}\alpha x^{\alpha-1}1_{(0,1)}(x) + \frac{\alpha}{e+\alpha}e^{1-x}1_{[1,\infty)}(x)$$

with mass constant $c = (e+\alpha)/[e\Gamma(\alpha)\alpha]$. Give a detailed algorithm for implementing the acceptance–rejection method employing the majorizing function $h(x) = cg(x)$.

14. Let S and T be independent deviates sampled from the uniform distribution on $[-1, 1]$. Show that conditional on the event $S^2 + T^2 \leq 1$ the ratio S/T is Cauchy.

15. Specify the ratio method for generating normal deviates starting from the multiple $h(x) = e^{-x^2/2}$ of the standard normal density. Show that the smallest enclosing rectangle is defined by the inequalities $0 \leq u \leq 1$ and $v^2 \leq 2/e$.

16. Describe and implement in computer code an algorithm for sampling from the hypergeometric distribution. Use the "deviate by definition" method of Section 20.6.

17. Describe how to generate deviates from the noncentral chi-square, noncentral F, and noncentral t distributions. Implement one of these algorithms in computer code.

18. Suppose the n-dimensional random deviate X is uniformly distributed within the unit sphere $S_n = \{x : \|x\| \leq 1\}$. If Ω is a covariance matrix with square root $\Omega^{1/2}$, then show that $Y = \Omega^{1/2}X$ is uniformly distributed within the ellipsoid $\{y : y^t\Omega^{-1}y \leq 1\}$.

References

[1] Angus J (1994) The probability integral transform and related results. *SIAM Review* 36:652–654

[2] Devroye L (1986) *Non-Uniform Random Variate Generation.* Springer-Verlag, New York

[3] Devroye L (1987) A simple generator for discrete log-concave distributions. *Computing* 39:87–91

[4] Dagpunar J (1988) *Principles of Random Variate Generation.* Oxford University Press, Oxford

[5] Fang KT, Kotz S, Ng KW (1990) *Symmetric Multivariate and Related Distributions.* Chapman & Hall, London

[6] Feller W (1971) *An Introduction to Probability Theory and its Applications, Vol 2*, 2nd ed. Wiley, New York

[7] Flury BD (1990) Acceptance-rejection sampling made easy. *SIAM Review* 32:474–476

[8] Hörmann W (1994) A universal generator for discrete log-concave distributions. *Computing* 52:89–96

[9] Kalos MH, Whitlock PA (1986) *Monte Carlo Methods, Vol 1: Basics.* Wiley, New York

[10] Kennedy WJ Jr, Gentle JE (1980) *Statistical Computing.* Marcel Dekker, New York
[11] Kingman JFC (1993) *Poisson Processes.* Oxford University Press, Oxford
[12] Knuth D (1981) *The Art of Computer Programming, 2: Seminumerical Algorithms*, 2nd ed. Addison-Wesley, Reading MA
[13] Lange KL, Little RJA, Taylor JMG (1989) Robust modeling using the t distribution. *J Amer Stat Assoc* 84:881–896
[14] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[15] Rice JA (1995) *Mathematical Statistics and Data Analysis*, 2nd ed. Duxbury Press, Belmont, CA
[16] Ripley BD (1983) Computer Generation of Random Variables. *International Stat Review* 51:301–319
[17] Rubinstein RY (1981) *Simulation and the Monte Carlo Method.* Wiley, New York

21

Independent Monte Carlo

21.1 Introduction

Monte Carlo integration is a rough and ready technique for calculating high-dimensional integrals and dealing with nonsmooth integrands [3, 4, 7, 9, 10]. Although quadrature methods can be extended to multiple dimensions, these deterministic techniques are almost invariably defeated by the curse of dimensionality. For example, if a quadrature method relies on n quadrature points in one dimension, then its product extension to d dimensions relies on n^d quadrature points. Even in one dimension, quadrature methods perform best for smooth functions. Both Romberg acceleration and Gaussian quadrature certainly exploit smoothness.

Monte Carlo techniques ignore smoothness and substitute random points for fixed quadrature points. If we wish to approximate the integral

$$\mathrm{E}[f(X)] = \int f(x) d\mu(x)$$

of an arbitrary integrand $f(x)$ against a probability measure μ, then we can take an i.i.d. sample $X_1, \ldots, X_n$ from μ and estimate $\int f(x) d\mu(x)$ by the sample average $\frac{1}{n} \sum_{i=1}^{n} f(X_i)$. The law of large numbers implies that these Monte Carlo estimates converge to $\mathrm{E}[f(X)]$ as n tends to ∞. If $f(x)$ is square integrable, then the central limit theorem allows us to refine this conclusion by asserting that the estimator $\frac{1}{n} \sum_{i=1}^{n} f(X_i)$ is approximately normally distributed about $\mathrm{E}[f(X)]$ with standard deviation $\sqrt{\mathrm{Var}[f(X)]/n}$. In practice, we estimate the order of the Monte Carlo error

as $\sqrt{v/n}$, where

$$v = \frac{1}{n-1}\sum_{i=1}^{n}\Big[f(X_i) - \frac{1}{n}\sum_{j=1}^{n} f(X_j)\Big]^2$$

is the usual unbiased estimator of $\mathrm{Var}[f(X)]$.

The central limit theorem perspective also forces on us two conclusions. First, the error estimate does not depend directly on the dimensionality of the underlying space. This happy conclusion is balanced by the disappointing realization that the error in estimating $\mathrm{E}[f(X)]$ declines at the slow rate $n^{-1/2}$. In contrast, the errors encountered in quadrature formulas with n quadrature points typically vary as $O(n^{-k})$ for k at least 2. Rather than bemoan the $n^{-1/2}$ rate of convergence in Monte Carlo integration, practitioners now attempt to reduce the $\mathrm{Var}[f(X)]$ part of the standard error formula $\sqrt{\mathrm{Var}[f(X)]/n}$.

The diverse applications of the Monte Carlo method almost defy summary. We conclude this chapter with a nontrivial construction of a permutation test for testing independence in large, sparse contingency tables. This example features discrete simulation and serves as an antidote to our emphasis on numerical integration.

21.2 Importance Sampling

Importance sampling is one technique for variance reduction. Suppose that the probability measure μ is determined by a density $g(x)$ relative to a measure ν such as Lebesgue measure or counting measure. If $h(x)$ is another density relative to ν with $h(x) > 0$ when $f(x)g(x) \neq 0$, then we can write

$$\int f(x)g(x)d\nu(x) = \int \frac{f(x)g(x)}{h(x)}h(x)d\nu(x).$$

Thus, if $Y_1, \ldots, Y_n$ is an i.i.d. sample from $h(x)$, then the sample average $\frac{1}{n}\sum_{i=1}^{n} f(Y_i)g(Y_i)/h(Y_i)$ offers an alternative estimator of $\int f(x)g(x)d\nu(x)$. This estimator has smaller variance than $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ if and only if

$$\int \Big[\frac{f(x)g(x)}{h(x)}\Big]^2 h(x)d\nu(x) \leq \int f(x)^2 g(x)d\nu(x).$$

If we choose $h(x) = |f(x)|g(x)/\int |f(z)|g(z)d\nu(z)$, then the Cauchy–Schwarz inequality implies

$$\begin{aligned}\int \Big[\frac{f(x)g(x)}{h(x)}\Big]^2 h(x)d\nu(x) &= \Big[\int |f(x)|g(x)d\nu(x)\Big]^2 \\ &\leq \int f(x)^2 g(x)d\nu(x)\int g(x)d\nu(x) \\ &= \int f(x)^2 g(x)d\nu(x).\end{aligned}$$

Equality occurs if and only if $|f(x)|$ is constant with probability 1 relative to $g(x)d\nu(x)$. If $f(x)$ is nonnegative and $h(x)$ is chosen according to the above recipe, then the variance of the estimator $\frac{1}{n}\sum_{i=1}^{n} f(Y_i)g(Y_i)/h(Y_i)$ reduces to 0.

This elegant result is slightly irrelevant since $\int |f(x)|g(x)d\nu(x)$ is unknown. However, it suggests that the variance of the sample average estimator will be reduced if $h(x)$ resembles $|f(x)|g(x)$. When this is the case, random points tend to be sampled where they are most needed to achieve accuracy.

Example 21.2.1 (*Expected Returns to the Origin*)**.** In a three-dimensional, symmetric random walk on the integer lattice [2], the expected number of returns to the origin equals

$$\frac{1}{2^3}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\frac{3}{3-\cos(\pi x_1)-\cos(\pi x_2)-\cos(\pi x_3)}dx_1dx_2dx_3.$$

Detailed analytic calculations too lengthy to present here show that this integral approximately equals 1.516. A crude Monte Carlo estimate based on 10,000 uniform deviates from the cube $[-1,1]^3$ is 1.478 ± 0.036. The singularity of the integrand at the origin $\mathbf{0}$ explains the inaccuracy and implies that the estimator has infinite variance. Thus, the standard error 0.036 attached to the estimate 1.478 is bogus.

We can improve the estimate by importance sampling. Let

$$S_3 = \{(x_1, x_2, x_3) : r \le 1\}$$

be the unit sphere in R^3, where $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$. Since the singularity near the origin behaves like a multiple of r^{-2}, we decompose the integral as

$$\begin{aligned}
&\frac{1}{2^3}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\frac{3}{3-\cos(\pi x_1)-\cos(\pi x_2)-\cos(\pi x_3)}dx_1dx_2dx_3\\
&\quad=\frac{1}{(2^3-\frac{4\pi}{3})}\int_{[-1,1]^3\setminus S_3}\frac{3(2^3-\frac{4\pi}{3})/2^3}{3-\cos(\pi x_1)-\cos(\pi x_2)-\cos(\pi x_3)}dx_1dx_2dx_3\\
&\qquad+\int_{S_3}\frac{3(4\pi r^2)/2^3}{3-\cos(\pi x_1)-\cos(\pi x_2)-\cos(\pi x_3)}\frac{1}{4\pi r^2}dx_1dx_2dx_3. \qquad (1)
\end{aligned}$$

Sampling from the density $1/(4\pi r^2)$ on S_3 concentrates random points near the origin. As the brief calculation

$$\Pr(R^* \le r) = \frac{1}{4\pi}\int_0^r\int_0^{2\pi}\int_0^{\pi}\frac{1}{s^2}s^2\sin\phi\,d\phi\,d\theta\,ds = r$$

with spherical coordinates shows, the radius R^* under such sampling is uniformly distributed on $[0,1]$. This contrasts with the nonuniform

distribution

$$\Pr(R \le r) = \frac{3}{4\pi} \int_0^r \int_0^{2\pi} \int_0^{\pi} s^2 \sin\phi \, d\phi \, d\theta \, ds = r^3$$

of the radius R under uniform sampling on S_3. These formulas demonstrate that we can generate R^* by taking a uniform sample from S_3 and setting $R^* = R^3$.

A strategy for computing the expected number of returns is now clear. We sample a point (x_1, x_2, x_3) uniformly from the cube $[-1, 1]^3$. If $r > 1$, then the point is uniform in $[-1, 1]^3 \setminus S_3$, and we use it to compute a Monte Carlo estimate of the first integral on the right of (1). If $r \le 1$, then we have a uniform point in S_3. If we replace r by r^3 and adjust (x_1, x_2, x_3) accordingly, then we use the revised point in S_3 to compute a Monte Carlo estimate of the second integral on the right of (1) based on the density $1/(4\pi r^2)$. Carrying out this procedure with 10,000 random points from $[-1, 1]^3$ and adding the two Monte Carlo estimates produces the improved estimate 1.513 ± 0.030 of the expected number of returns. ■

21.3 Stratified Sampling

In stratified sampling, we partition the domain of integration S of an expectation $\mathrm{E}[f(X)] = \int_S f(x) d\mu(x)$ into $m > 1$ disjoint subsets S_i and sample a fixed number of points $X_{i1}, \ldots, X_{in_i}$ from each S_i according to the conditional probability measure $\mu(A \mid S_i)$. If we estimate the conditional expectation $\mathrm{E}[f(X) \mid X \in S_i]$ by $\frac{1}{n_i} \sum_{j=1}^{n_i} f(X_{ij})$, then the weighted estimator

$$\sum_{i=1}^{m} \mu(S_i) \frac{1}{n_i} \sum_{j=1}^{n_i} f(X_{ij}) \tag{2}$$

is unbiased for $\mathrm{E}[f(X)]$. If the n_i are chosen carefully, then the variance $\sum_{i=1}^{m} \mu(S_i)^2 \operatorname{Var}[f(X) \mid X \in S_i]/n_i$ of this estimator will be smaller than the variance $\operatorname{Var}[f(X)]/n$ of the sample average Monte Carlo estimator with the same number of points $n = \sum_{i=1}^{m} n_i$ drawn randomly from S.

For instance, if we take $n_i = n\mu(S_i)$, then the variance of the stratified estimator (2) reduces to

$$\frac{1}{n} \sum_{i=1}^{m} \mu(S_i) \operatorname{Var}[f(X) \mid X \in S_i] = \frac{1}{n} \mathrm{E}\{\operatorname{Var}[f(X) \mid Z]\},$$

where Z is a random variable satisfying $Z = i$ when the single random point X drawn from μ falls in S_i. Since we can write

$$\operatorname{Var}[f(X)] = \mathrm{E}\{\operatorname{Var}[f(X) \mid Z]\} + \operatorname{Var}\{\mathrm{E}[f(X) \mid Z]\},$$

it is clear that the stratified estimator has smaller variance than the sample average estimator.

In principle, one can improve on proportional sampling. To minimize the variance of the stratified estimator, we treat the n_i as continuous variables, introduce a Lagrange multiplier λ, and look for a stationary point of the Lagrangian

$$\sum_{i=1}^{m} \mu(S_i)^2 \frac{1}{n_i} \operatorname{Var}[f(X) \mid X \in S_i] + \lambda\Big(n - \sum_{i=1}^{m} n_i\Big).$$

Equating its partial derivative with respect to n_i to zero and taking into account the constraint $\sum_{i=1}^{m} n_i = n$ yields

$$n_i = n \frac{\mu(S_i)\sqrt{\operatorname{Var}[f(X) \mid X \in S_i]}}{\sum_{k=1}^{m} \mu(S_k)\sqrt{\operatorname{Var}[f(X) \mid X \in S_k]}}.$$

Although the exact values of the conditional variances $\operatorname{Var}[f(X) \mid X \in S_i]$ are inaccessible in practice, we can estimate them using a small pilot sample of points from each S_i. Once this is done, we can collect a more intelligent, final stratified sample that puts the most points where $f(x)$ shows the most variation. Obviously, it is harder to give general advice about how to choose the strata S_i and compute their probabilities $\mu(S_i)$ in the first place.

21.4 Antithetic Variates

In the method of antithetic variates, we look for unbiased estimators V and W of an integral that are negatively correlated rather than independent. The average $(V + W)/2$ is also unbiased, and its variance

$$\operatorname{Var}\Big(\frac{V+W}{2}\Big) = \frac{1}{4}\operatorname{Var}(V) + \frac{1}{4}\operatorname{Var}(W) + \frac{1}{2}\operatorname{Cov}(V, W)$$

is reduced compared to what it would be if V and W were independent. The next proposition provides a sufficient condition for achieving negative correlation. Its proof exploits coupled random variables; by definition these reside on the same probability space [6].

Proposition 21.4.1. *Suppose X is a random variable and the functions $f(x)$ and $g(x)$ are both increasing or both decreasing. If the random variables $f(X)$ and $g(X)$ have finite second moments, then*

$$\operatorname{Cov}[f(X), g(X)] \geq 0.$$

If $f(x)$ is increasing and $g(x)$ is decreasing, or vice versa, then the reverse inequality holds.

Proof. Consider a second random variable Y independent of X but having the same distribution. If $f(x)$ and $g(x)$ are both increasing or both

decreasing, then the product $[f(X) - f(Y)][g(X) - g(Y)] \geq 0$. Hence,

$$\begin{aligned} 0 &\leq \mathrm{E}\{[f(X) - f(Y)][g(X) - g(Y)]\} \\ &= \mathrm{E}[f(X)g(X)] + \mathrm{E}[f(Y)g(Y)] - \mathrm{E}[f(X)]\,\mathrm{E}[g(Y)] - \mathrm{E}[f(Y)]\,\mathrm{E}[g(X)] \\ &= 2\,\mathrm{Cov}[f(X), g(X)]. \end{aligned}$$

The same proof with obvious modifications holds when one of the two functions is increasing and the other is decreasing. □

Example 21.4.1 (*Antithetic Uniform Estimators*). Consider the integral $\int f(x)g(x)dx$, where $f(x)$ is increasing and the density $g(x)$ has distribution function $G(x)$. If $U_1, \ldots, U_n$ is an i.i.d. sample from the uniform distribution, then $f[G^{[-1]}(U_i)]$ and $f[G^{[-1]}(1-U_i)]$ are both unbiased estimators of $\int f(x)g(x)dx$. According to Proposition 21.4.1, they are negatively correlated, and the estimator

$$\frac{1}{2n}\sum_{i=1}^{n}\{f[G^{[-1]}(U_i)] + f[G^{[-1]}(1 - U_i)]\}$$

has smaller variance than $\frac{1}{2n}\sum_{i=1}^{2n} f[G^{[-1]}(U_i)]$. ■

21.5 Control Variates

In computing $\mathrm{E}[f(X)]$, suppose that we can calculate exactly the expectation $\mathrm{E}[g(X)]$ for a function $g(x)$ close to $f(x)$. Then it makes sense to write

$$\mathrm{E}[f(X)] = \mathrm{E}[f(X) - g(X)] + \mathrm{E}[g(X)]$$

and approximate $\mathrm{E}[f(X) - g(X)]$ by a Monte Carlo estimate rather than $\mathrm{E}[f(X)]$. Example 21.2.1 provides a test case of this tactic. Near the origin the integrand

$$\begin{aligned} f(x) &= \frac{3}{3 - \cos(\pi x_1) - \cos(\pi x_2) - \cos(\pi x_3)} \\ &\approx \frac{6}{\pi^2 r^2} \\ &= g(x). \end{aligned}$$

By transforming to spherical coordinates it is straightforward to calculate

$$\frac{1}{\frac{4\pi}{3}}\int_{S_3} g(x)dx_1dx_2dx_3 = \frac{18}{\pi^2}.$$

In Example 21.2.1 we can avoid importance sampling by forming a Monte Carlo estimate of the conditional expectation $\mathrm{E}[f(X) - g(X) \,|\, X \in S_3]$ and adding it to the exact conditional expectation

$$\mathrm{E}[g(X) \,|\, X \in S_3] = \frac{18}{\pi^2}.$$

Making this change but proceeding otherwise precisely as sketched in Example 21.2.1 yields the impressive approximation 1.517 ± 0.015 for the expected number of returns to the origin. In this case, the method of control variates operates by subtracting off the singularity of the integrand and performs better than importance sampling.

21.6 Rao–Blackwellization

The Rao–Blackwell theorem in statistics takes an unbiased estimator and replaces it by its conditional expectation given a sufficient statistic. A similar variance-reduction procedure is possible for random samples generated by the acceptance–rejection method [1]. Recall that in the acceptance–rejection method we sample from an envelope density $h(x)$ that satisfies an inequality $g(x) \leq ch(x)$ relative to a target density $g(x)$. We accept a sampled point X drawn from the density $h(x)$ based on an independent uniform deviate U. If $U \leq W = g(X)/[ch(X)]$, then we accept X; otherwise, we reject X. The accepted points conform to the target density $g(x)$.

In computing an expectation $\mathrm{E}[f(X)] = \int f(x)g(x)dx$ by the Monte Carlo method, suppose we generate n points $X_1, \ldots, X_n$ according to $h(x)$ and accept precisely m of them, including the last, using uniform deviates $U_1, \ldots, U_n$. The Monte Carlo estimator of $\mathrm{E}[f(X)]$ is

$$\frac{1}{m} \sum_{i=1}^{n} 1_{\{U_i \leq W_i\}} f(X_i).$$

The conditional expectation

$$\begin{aligned} &\frac{1}{m} \mathrm{E}\Big[\sum_{i=1}^{n} 1_{\{U_i \leq W_i\}} f(X_i) \mid n, X_1, \ldots, X_n \Big] \\ &\quad = \frac{1}{m} \sum_{i=1}^{n} \mathrm{E}\Big[1_{\{U_i \leq W_i\}} \mid n, W_1, \ldots, W_n \Big] f(X_i) \end{aligned}$$

retains the unbiased character of the Monte Carlo estimator while reducing its variance. This Rao–Blackwellized estimator uses both the rejected and the accepted points with appropriate weights for each.

To make this scheme viable, we must compute the conditional probability p_i that the ith deviate X_i is accepted, given its success probability $W_i = w_i$ and the fact that there are m successes in n trials, with the last trial ending in a success. Let q_{jk} be the probability of j successes among k Bernoulli trials with respective success probabilities $w_1, \ldots, w_k$, and let $q_{jk}^{(i)}$ be the probability of the intersection of this same event with the event that trial i is a success. In view of the fact that the last trial is a success, we calculate $p_n = 1$ and $p_i = q_{m-1,n-1}^{(i)}/q_{m-1,n-1}$ for $1 \leq i < n$. Computation of the

Poisson-binomial probabilities q_{jk} can be carried out by the algorithm given in Section 1.7 of Chapter 1. The joint probabilities $q_{jk}^{(i)}$ can be computed from the initial conditions $q_{ji}^{(i)} = w_i q_{j-1,i-1}$ and the recurrence

$$q_{jk}^{(i)} = w_k q_{j-1,k-1}^{(i)} + (1 - w_k) q_{j,k-1}^{(i)}$$

for $k > i$. If m and n are both large, then the approximation $p_i = w_i$ is probably adequate for practical purposes, and computation of the q_{jk} and $q_{jk}^{(i)}$ can be avoided.

21.7 Exact Tests of Independence in Contingency Tables

As an illustration of the operation of the Monte Carlo method in a discrete setting, we now turn to the problem of testing independence in large, sparse contingency tables [5]. Consider a table with n multivariate observations scored on m factors. We will denote a typical cell of the table by a multi-index $\mathbf{i} = (i_1, \ldots, i_m)$, where i_j represents the level of factor j in the cell. If the probability associated with level k of factor j is p_{jk}, then under the assumption of independence of the various factors, a multivariate observation falls in cell $\mathbf{i} = (i_1, \ldots, i_m)$ with probability

$$p_{\mathbf{i}} = \prod_{j=1}^{m} p_{ji_j}.$$

Furthermore, the cell counts $\{n_{\mathbf{i}}\}$ from the sample follow a multinomial distribution with parameters $(n, \{p_{\mathbf{i}}\})$.

For the purposes of testing independence, the probabilities p_{jk} are nuisance parameters. In exact inference, one conditions on the marginal counts $\{n_{jk}\}_k$ of each factor j. These, of course, follow a multinomial distribution with parameters $(n, \{p_{jk}\}_k)$. Because under the null hypothesis of independence, marginal counts are independent from factor to factor, the conditional distribution of the cell counts is

$$\begin{aligned} \Pr(\{n_{\mathbf{i}}\} \mid \{n_{jk}\}) &= \frac{\binom{n}{\{n_{\mathbf{i}}\}} \prod_{\mathbf{i}} p_{\mathbf{i}}^{n_{\mathbf{i}}}}{\prod_{j=1}^{m} \binom{n}{\{n_{jk}\}_k} \prod_k (p_{jk})^{n_{jk}}} \\ &= \frac{\binom{n}{\{n_{\mathbf{i}}\}}}{\prod_{j=1}^{m} \binom{n}{\{n_{jk}\}_k}}. \end{aligned} \tag{3}$$

One of the pleasant facts of exact inference is that the multivariate Fisher–Yates distribution (3) does not depend on the marginal probabilities p_{jk}. Problem 9 indicates how to compute the moments of (3).

We can also derive the Fisher–Yates distribution by a counting argument involving a related but different sample space. Consider an $m \times n$ matrix whose rows correspond to factors and whose columns correspond to the multivariate observations attributed to the different cells. At factor j there are n_{jk} observations representing level k. If we uniquely label each of these $n = \sum_k n_{jk}$ observations, then there are $n!$ distinguishable permutations of the level labels in row j. The uniform sample space consists of the $(n!)^m$ matrices derived from the $n!$ permutations of each of the m rows. Each such matrix is assigned probability $1/(n!)^m$. For instance, if we distinguish duplicate labels by a superscript $*$, then the 3×4 matrix

$$\begin{pmatrix} a_1 & a_2 & a_1^* & a_2^* \\ b_3 & b_1 & b_1^* & b_2 \\ c_2 & c_1 & c_3 & c_2^* \end{pmatrix} \tag{4}$$

for $m = 3$ factors and $n = 4$ multivariate observations represents one out of $(4!)^3$ equally likely matrices and yields the nonzero cell counts

$$\begin{aligned} n_{a_1 b_3 c_2} &= 1 \\ n_{a_2 b_1 c_1} &= 1 \\ n_{a_1 b_1 c_3} &= 1 \\ n_{a_2 b_2 c_2} &= 1. \end{aligned}$$

To count the number of matrices consistent with a cell count vector $\{n_{\mathbf{i}}\}$, note that the n multivariate observations can be assigned to the columns of a typical matrix from the uniform space in $\binom{n}{\{n_{\mathbf{i}}\}}$ ways. Within each such assignment, there are $\prod_k n_{jk}!$ consistent permutations of the labels at level j; over all levels, there are $\prod_{j=1}^m \prod_k n_{jk}!$ consistent permutations. It follows that the cell count vector $\{n_{\mathbf{i}}\}$ has probability

$$\begin{aligned} \Pr(\{n_{\mathbf{i}}\}) &= \frac{\binom{n}{\{n_{\mathbf{i}}\}} \prod_{j=1}^m \prod_k n_{jk}!}{(n!)^m} \\ &= \frac{\binom{n}{\{n_{\mathbf{i}}\}}}{\prod_{j=1}^m \binom{n}{\{n_{jk}\}_k}}. \end{aligned}$$

In other words, we recover the Fisher–Yates distribution.

The uniform sample space also suggests a device for random sampling from the Fisher–Yates distribution [5]. If we arrange our n multivariate observations in an $m \times n$ matrix as described above and randomly permute the entries within each row, then we get a new matrix whose cell counts are drawn from the Fisher–Yates distribution. For example, appropriate permutations within each row of the matrix (4) produce the matrix

$$\begin{pmatrix} a_1 & a_1^* & a_2 & a_2^* \\ b_1 & b_1^* & b_2 & b_3 \\ c_2 & c_2^* & c_1 & c_3 \end{pmatrix}$$

with nonzero cell counts

$$\begin{aligned} n_{a_1b_1c_2} &= 2 \\ n_{a_2b_2c_1} &= 1 \\ n_{a_2b_3c_3} &= 1. \end{aligned}$$

Problem 10 asks the reader to devise an efficient algorithm for generating random permutations.

Iterating this permutation procedure r times generates an independent, random sample $Z_1, \ldots, Z_r$ of tables from the Fisher–Yates distribution. In practice, it suffices to permute all rows except the bottom row m because cell counts do not depend on the order of the columns in a matrix such as (4). Given the observed value T_{obs} of a test statistic T for independence, we estimate the corresponding p-value by the sample average $\frac{1}{r}\sum_{l=1}^{r} 1_{\{T(Z_l)\geq T_{\text{obs}}\}}$.

In Fisher's exact test, the statistic T is the negative of the Fisher–Yates probability (3). Thus, the null hypothesis of independence is rejected if the observed Fisher–Yates probability is too low. The chi-square statistic $\sum_{\mathbf{i}}[n_{\mathbf{i}}-\mathrm{E}(n_{\mathbf{i}})]^2/\mathrm{E}(n_{\mathbf{i}})$ is also reasonable for testing independence, provided we estimate its p-value by random sampling and do not foolishly rely on the standard chi-square approximation. As noted in Problem 9, the expectation $\mathrm{E}(n_{\mathbf{i}}) = n\prod_{j=1}^{m}(n_{ji_j}/n)$.

Lazzeroni and Lange [5] apply this Monte Carlo method to test for linkage equilibrium (independence) at six linked genetic markers on chromosome 11. Here each marker locus is considered a factor, and each allele of a marker is considered a level. An observation of a chromosome with a particular sequence of alleles at the marker loci is called a haplotype. With only random 180 haplotypes and 2, 2, 10, 5, 3, and 2 alleles at the six loci, the resulting contingency table qualifies as sparse. Large sample chi-square tests strongly suggest linkage disequilibrium (dependence). Monte Carlo exact tests do not reach significance at the 0.1 level and correct this misimpression.

21.8 Problems

1. Consider the integral $\int_0^1 \cos(\pi x/2)dx = 2/\pi$ [4]. Interpreting this as an expectation relative to the uniform distribution on $[0,1]$, show that

$$\mathrm{Var}\left[\cos\left(\frac{\pi X}{2}\right)\right] \;=\; \frac{1}{2} - \left(\frac{2}{\pi}\right)^2 \;\approx\; 0.095.$$

The importance density $h(x) = 3(1-x^2)/2$ roughly resembles the integrand $\cos(\pi x/2)$. Demonstrate numerically that

$$\text{Var}\left[\frac{2\cos\left(\frac{\pi Y}{2}\right)}{3(1-Y^2)}\right] \approx 0.00099$$

when Y is sampled from $h(y)$. Thus, importance sampling reduces the variance of the corresponding Monte Carlo estimator by almost a factor of 100.

2. Continuing Problem 1, devise a ratio method for sampling from the density $h(x) = 3(1-x^2)/2$ on $[0,1]$. Show that $[0,\sqrt{3/2}] \times [0,\sqrt{3/8}]$ is an enclosing rectangle.

3. Write a program to carry out the naive Monte Carlo method, the importance sampling method, and the control variate method for estimating the random walk integral discussed in Example 21.2.1 and Section 21.5. Show analytically that about 52.4 percent of the points generated in the cube $[-1,1]^3$ fall within S_3.

4. Proposition 21.4.1 demonstrates what a powerful tool coupling of random variables is in proving inequalities. Here is another example involving the monotonicity of power functions in hypothesis testing. Let X and Y follow binomial distributions with n trials and success probabilities $p < q$, respectively. We can realize X and Y on the same probability space by scattering n points randomly on the unit interval. Interpret X as the number of points less than or equal to the cutoff p, and interpret Y as the number of points less than or equal to the cutoff q. Show that this interpretation leads to an immediate proof of the inequality $\Pr(X \geq k) \leq \Pr(Y \geq k)$ for all k.

5. Continuing Problem 4, suppose that X follows the hypergeometric distribution

$$\Pr(X = i) = \frac{\binom{r}{i}\binom{n-r}{m-i}}{\binom{n}{m}}.$$

Let Y follow the same hypergeometric distribution except that $r+1$ replaces r. Prove that $\Pr(X \geq k) \leq \Pr(Y \geq k)$ for all k. (Hint: Consider an urn with r red balls, 1 white ball, and $n-r-1$ black balls. If we draw m balls from the urn without replacement, then X is the number of red balls drawn, and Y is the number of red or white balls drawn.)

6. Suppose you compute by stratified sampling from the two subintervals $[0,\sqrt{3}/2)$ and $[\sqrt{3}/2,1]$ a Monte Carlo estimate of the integral $\int_0^1 \sqrt{1-x^2}dx$. If you employ n points overall, how many points should you allot to each subinterval to achieve the greatest variance reduction [7]?

7. In the stratified sampling method, suppose the domain of integration splits into two subsets S_1 and S_2 such that

$$\begin{aligned}\Pr(X \in S_1) &= \Pr(X \in S_2)\\ \mathrm{E}[f(X) \mid X \in S_1] &= \mathrm{E}[f(X) \mid X \in S_2]\\ \mathrm{Var}[f(X) \mid X \in S_1] &= \mathrm{Var}[f(X) \mid X \in S_2].\end{aligned}$$

Show that the stratified estimate of $\mathrm{E}[f(X)]$ with $3n/4$ points drawn randomly from S_1 and $n/4$ points drawn randomly from S_2 is actually worse than the sample mean estimate with all n points drawn randomly from $S_1 \cup S_2$.

8. The method of control variates can be used to estimate the moments of the sample median $X_{(n)}$ from a random sample of size $2n-1$ from a symmetric distribution. Because we expect the difference $X_{(n)} - \bar{X}$ between the sample median and the sample mean to be small, the moments of $\bar{X}$ serve as a first approximation to the moments of $X_{(n)}$. Put this insight into practice by writing a Monte Carlo program to compute $\mathrm{Var}(X_{(n)})$ for a sample from the standard normal distribution.

9. To compute moments under the Fisher–Yates distribution (3), let

$$u^{\underline{r}} = u(u-1)\cdots(u-r+1)$$

be a falling factorial power, and let $\{l_{\mathbf{i}}\}$ be a collection of nonnegative integers indexed by the cell labels $\mathbf{i} = (i_1 \ldots, i_m)$. Setting $l = \sum_{\mathbf{i}} l_{\mathbf{i}}$ and $l_{jk} = \sum_{\mathbf{i}} 1_{\{i_j = k\}} l_{\mathbf{i}}$, show that

$$\mathrm{E}\Big(\prod_{\mathbf{i}} n_{\mathbf{i}}^{\underline{l_{\mathbf{i}}}}\Big) = \frac{\prod_{j=1}^{m}\prod_{k}(n_{jk})^{\underline{l_{jk}}}}{(n^{\underline{l}})^{m-1}}.$$

In particular, verify that $\mathrm{E}(n_{\mathbf{i}}) = n\prod_{j=1}^{m}(n_{ji_j}/n)$.

10. Devise an efficient algorithm to generate random permutations of a given length [8]. Describe how you could use this algorithm to generate random strings containing exactly m 0's and n 1's.

References

[1] Casella G, Robert CP (1996) Rao–Blackwellisation of sampling schemes. *Biometrika* 83:81–94

[2] Doyle PG, Snell JL (1984) *Random Walks and Electrical Networks.* The Mathematical Association of America

[3] Hammersley JM, Handscomb DC (1964) *Monte Carlo Methods.* Methuen, London

[4] Kalos MH, Whitlock PA (1986) *Monte Carlo Methods, Vol 1: Basics.* Wiley, New York

[5] Lazzeroni LC, Lange K (1997) Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Ann Stat* 25:138–168

[6] Lindvall T (1992) *Lectures on the Coupling Method.* Wiley, New York
[7] Morgan BJT (1984) *Elements of Simulation.* Chapman & Hall, London
[8] Nijenhuis A, Wilf HS (1978) *Combinatorial Algorithms for Computers and Calculators*, 2nd ed. Academic Press, New York
[9] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[10] Rubinstein RY (1981) *Simulation and the Monte Carlo Method.* Wiley, New York

22
Bootstrap Calculations

22.1 Introduction

For most statisticians, observations take precedence over theory. Reliance on normality assumptions or large sample theory is particularly suspect in small samples. Although the bootstrap perfectly embodies this skepticism, it initially met with intellectual resistance. The notion that one can learn something useful about the properties of estimators, confidence intervals, and hypothesis tests by resampling data was alien to most statisticians of the past. The computational demands of the bootstrap alone made it unthinkable. These intellectual and practical objections began to crumble with the advent of modern computing. The introduction of the jackknife by Quenouille [15] and Tukey [18] demonstrated some of the virtues of data resampling. In the last decade, Efron's bootstrap has largely supplanted the jackknife [7].

Like most good ideas, the principle behind the bootstrap is simple. Suppose for the sake of argument that we draw an i.i.d. sample $\mathbf{x} = (x_1, \ldots, x_n)$ from some unknown probability distribution $F(x)$. If we want to understand the sampling properties of a complicated estimator $T(\mathbf{x})$ of a parameter $t(F)$ of $F(x)$, then we study the properties of the corresponding estimator $T(\mathbf{x}^*)$ on the space of i.i.d. samples $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ drawn from a data-based approximation $F_n^*(x)$ to $F(x)$. In the case of the nonparametric bootstrap, $F_n^*(x)$ is the empirical distribution function, putting weight $1/n$ on each of the n observed points $x_1, \ldots, x_n$. In the case of the parametric bootstrap, we assume a parametric form for $F_\alpha(x)$, estimate the

parameter α from the data by $\hat{\alpha}$, and then sample from $F_n^*(x) = F_{\hat{\alpha}}(x)$. The bootstrap correspondence principle suggests that not only do $T(\mathbf{x})$ and $T(\mathbf{x}^*)$ have similar distributions, but equally important in practice, that $T(\mathbf{x}) - t(F)$ and $T(\mathbf{x}^*) - t(F_n^*)$ have similar distributions. In many examples, the identity $t(F_n^*) = T(\mathbf{x})$ holds.

These insights are helpful, but finding the theoretical sampling distribution of $T(\mathbf{x}^*)$ is usually impossible. Undeterred by this fact, Efron [7] suggested that we approximate the distribution and moments of $T(\mathbf{x}^*)$ by independent Monte Carlo sampling, in effect substituting computing brawn for mathematical weakness. As we have seen in Chapter 20, there are inefficient and efficient ways of carrying out Monte Carlo estimation. The theme of this chapter is efficient Monte Carlo estimation for the nonparametric bootstrap, the more interesting and widely applied version of the bootstrap. Limitations of space prevent us from delving more deeply into the theoretical justifications of the bootstrap. The underlying theory involves contiguity of probability measures and Edgeworth expansions. The papers [1, 2] and books [12, 17] are good places to start for mathematically sophisticated readers. Efron and Tibshirani [9] and Davison and Hinkley [5] provide practical guides to the bootstrap that go well beyond the abbreviated account presented here.

22.2 Range of Applications

In the bootstrap literature, the scope of the word "parameter" is broadened to include any function $t(F)$ of a probability distribution $F(x)$. The moments and central moments

$$\mu_k(F) = \int x^k dF(x)$$
$$\omega_k(F) = \int [x - \mu_1(F)]^k dF(x)$$

are typical parameters whenever they exist. The pth quantile

$$\xi_p(F) = \inf\{x: F(x) \geq p\}$$

is another commonly encountered parameter. If we eschew explicit parametric models, then the natural estimators of these parameters are the corresponding sample statistics

$$\hat{\mu}_k(\mathbf{x}) = \mu_k(F_n^*) = \frac{1}{n}\sum_{i=1}^n x_i^k$$
$$\hat{\omega}_k(\mathbf{x}) = \omega_k(F_n^*) = \frac{1}{n}\sum_{i=1}^n [x_k - \mu_1(F_n^*)]^k$$
$$\hat{\xi}_p(\mathbf{x}) = \xi_p(F_n^*) = \inf\{x: F_n^*(x) \geq p\}.$$

By construction, these estimators obey the rule $T(\mathbf{x}) = t(F_n^*)$ and consequently are called "plug-in" estimators by Efron and Tibshirani [9]. The unbiased version of the sample variance,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\omega_2(F_n^*),$$

fails to qualify as a plug-in estimator. In this chapter we consider only plug-in estimators.

Efron [7] discusses calculation of the variance of the sample median $\hat{\xi}_{1/2}(\mathbf{x}) = \xi_{1/2}(F_n^*)$ as a motivating example for the bootstrap. Classical large sample theory implies that

$$\mathrm{Var}[\xi_{\frac{1}{2}}(F_n^*)] \asymp \frac{1}{4nf(\xi_{\frac{1}{2}})^2},$$

where $f(x) = F'(x)$ is the density of $F(x)$ [16]. However, $f(x)$ is hard to estimate accurately even in large samples. The bootstrap provides a way out of this dilemma that avoids direct estimation of $f(\xi_{1/2})$.

In general, the bootstrap correspondence principle suggests that we estimate the variance of an estimator $T(\mathbf{x})$ by the variance of the corresponding estimator $T(\mathbf{x}^*)$ on the bootstrap sample space. Because the variance $\mathrm{Var}(T)^*$ of $T(\mathbf{x}^*)$ is usually difficult to calculate exactly, we take independent bootstrap samples $\mathbf{x}_b^* = (x_{b1}^*, \ldots, x_{bn}^*)$ for $b = 1, \ldots, B$ and approximate $\mathrm{Var}(T)^*$ by

$$\widehat{\mathrm{Var}}(T)^* = \frac{1}{B-1}\sum_{b=1}^{B}[T(\mathbf{x}_b^*) - \hat{\mathrm{E}}(T)^*]^2,$$

where

$$\hat{\mathrm{E}}(T)^* = \frac{1}{B}\sum_{b=1}^{B}T(\mathbf{x}_b^*).$$

If we can calculate $\mathrm{E}(T)^*$ exactly, then in the approximation $\widehat{\mathrm{Var}}(T)^*$ we substitute $\mathrm{E}(T)^*$ for the sample mean $\hat{\mathrm{E}}(T)^*$ and replace the divisor $B-1$ by B. For many purposes, it is adequate to choose B in the range 25 to 100; in hypothesis testing based on bootstrap pivots, Booth and Sarkar [3] argue for the much higher value $B = 800$.

Bias reduction is another valuable application of the bootstrap provided users bear in mind that bias reduction can increase the variability of an estimator. As an example, we follow Hall [12] and consider the problem of estimating the third power μ_1^3 of the mean μ_1 of a distribution function $F(x)$. The natural nonparametric estimator is $\bar{x}^3 = (\frac{1}{n}\sum_{i=1}^{n}x_i)^3$. Proposition proposition4.1 or a straightforward calculation shows that this

estimator has expectation

$$\begin{aligned}\mathrm{E}(\bar{x}^3) &= \mathrm{E}\left[\mu_1 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)\right]^3 \\ &= \mu_1^3 + \frac{3\mu_1\omega_2}{n} + \frac{\omega_3}{n^2},\end{aligned}$$

with a bias of order $O(n^{-1})$. The corresponding bootstrap estimator $\bar{x^*}^3$ has expectation

$$\mathrm{E}(\bar{x^*}^3) = \hat{\mu}_1^3 + \frac{3\hat{\mu}_1\hat{\omega}_2}{n} + \frac{\hat{\omega}_3}{n^2}.$$

Note in this case that we can calculate bootstrap moments exactly without resort to Monte Carlo simulation. To form an estimator with less bias, the bootstrap correspondence principle suggests that we subtract the bias of the bootstrap estimator from the original estimator.

In this concrete case, we can show that the revised estimator

$$\bar{x}^3 - \frac{3\hat{\mu}_1\hat{\omega}_2}{n} - \frac{\hat{\omega}_3}{n^2}$$

has less bias than the original estimator $\bar{x}^3$. If we pass to random variables $y_i = x_i - \mu_1$ with zero means, then

$$\begin{aligned}\mathrm{E}(\hat{\mu}_1\hat{\omega}_2) &= \mu_1\,\mathrm{E}(\hat{\omega}_2) + \mathrm{E}\left[\bar{y}\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2\right] \\ &= \frac{(n-1)\mu_1\omega_2}{n} + \mathrm{E}\left[\bar{y}\frac{1}{n}\sum_{j=1}^{n}y_j^2 - \bar{y}^3\right] \\ &= \frac{(n-1)\mu_1\omega_2}{n} + \frac{\omega_3}{n} - \frac{\omega_3}{n^2}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{E}(\hat{\omega}_3) &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^3 - 3y_i^2\bar{y} + 3y_i\bar{y}^2 - \bar{y}^3)\right] \\ &= \omega_3 - \frac{3\omega_3}{n} + \frac{3\omega_3}{n^2} - \frac{\omega_3}{n^2}.\end{aligned}$$

Thus, the expectation of the revised estimator is

$$\begin{aligned}&\mathrm{E}\left(\bar{x}^3 - \frac{3\hat{\mu}_1\hat{\omega}_2}{n} - \frac{\hat{\omega}_3}{n^2}\right) \\ &\quad= \mu_1^3 + \frac{3\mu_1\omega_2}{n} + \frac{\omega_3}{n^2} - \frac{3}{n}\left[\frac{(n-1)\mu_1\omega_2}{n} + \frac{\omega_3}{n} - \frac{\omega_3}{n^2}\right] \\ &\qquad- \frac{1}{n^2}\left(\omega_3 - \frac{3\omega_3}{n} + \frac{3\omega_3}{n^2} - \frac{\omega_3}{n^2}\right) \\ &\quad= \mu_1^3 + \frac{3(\mu_1\omega_2 - \omega_3)}{n^2} + \frac{6\omega_3}{n^3} - \frac{2\omega_3}{n^4}.\end{aligned}$$

The crucial thing to note here is that the bias is now of order $O(n^{-2})$ rather than $O(n^{-1})$.

In more complicated examples, we approximate the bootstrap bias

$$\text{bias}^* = \text{E}[T(\mathbf{x}^*)] - t(F_n^*)$$

by the Monte Carlo average

$$\widehat{\text{bias}}_B^* = \frac{1}{B}\sum_{b=1}^{B} T(\mathbf{x}_b^*) - t(F_n^*).$$

In accord with the bootstrap correspondence principle, the revised estimator $T(\mathbf{x}) - \widehat{\text{bias}}_B^*$ usually has much less bias than $T(\mathbf{x})$.

Our third application involves confidence intervals and is the focus of much recent research. Recall that a set $C(\mathbf{x}) = C(x_1, \ldots, x_n)$ is a $1-\alpha$ level confidence set for a parameter $t(F)$ if $\Pr[t(F) \in C(\mathbf{x})] \geq 1-\alpha$. A good share of theoretical statistics is devoted to the construction and interpretation of confidence intervals. The bootstrap correspondence principle cuts through this thicket of complexity. Suppose we have a plug-in estimator $T(\mathbf{x})$ of $t(F)$ with an attached estimator $V(\mathbf{x})$ of its variance. For instance, $T(\mathbf{x})$ could be the sample mean $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $V(\mathbf{x})$ could be $1/n$ times the sample variance $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2$.

In the bootstrap-t method, we consider the studentized variable

$$R(\mathbf{x}) = \frac{T(\mathbf{x}) - t(F)}{\sqrt{V(\mathbf{x})}}$$

and its bootstrap analog

$$R(\mathbf{x}^*) = \frac{T(\mathbf{x}^*) - T(\mathbf{x})}{\sqrt{V(\mathbf{x}^*)}}.$$

If G_n and G_n^* denote the distribution functions of $R(\mathbf{x})$ and $R(\mathbf{x}^*)$, respectively, then the bootstrap correspondence principle suggests that the percentiles $\xi_\alpha(G_n)$ and $\xi_\alpha(G_n^*)$ are approximately equal for all $\alpha \in (0,1)$. It follows that

$$\begin{aligned}\Pr\left[t(F) \geq T(\mathbf{x}) - \sqrt{V(\mathbf{x})}\xi_{1-\alpha}(G_n^*)\right] &= \Pr\left[R(\mathbf{x}) \leq \xi_{1-\alpha}(G_n^*)\right] \\ &\approx \Pr\left[R(\mathbf{x}^*) \leq \xi_{1-\alpha}(G_n^*)\right] \\ &\geq 1-\alpha.\end{aligned}$$

In other words, $T(\mathbf{x}) - \sqrt{V(\mathbf{x})}\xi_{1-\alpha}(G_n^*)$ is an approximate level $1-\alpha$ lower confidence bound for $t(F)$. Efron and Tibshirani [9] recommend 1000 Monte Carlo bootstrap resamples to approximate the percentile point $\xi_{1-\alpha}(G_n^*)$.

Example 22.2.1 (*Bootstrapping Residuals in Linear Regression*). Bootstrap-t confidence intervals turn out to be useful in the linear regression model $y = X\beta + u$. In regression we can resample cases (y_i, x_i) or residuals

$r_i = y_i - \hat{y}_i$. In bootstrapping residuals, we sample $\mathbf{r}^* = (r_1^*, \ldots, r_n^*)$ from $\mathbf{r} = (r_1, \ldots, r_n)$ with replacement and construct the new vector of dependent (or response) variables $y^* = X\hat{\beta} + \mathbf{r}^*$. Recall that the least squares estimator $\hat{\beta} = (X^t X)^{-1} X^t y$ is unbiased for β provided the errors u_i satisfy $\mathrm{E}(u_i) = 0$. If, in addition, the errors are uncorrelated with common variance $\mathrm{Var}(u_i) = \sigma^2$, then $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$. The predicted value of y is obviously defined as $\hat{y} = X\hat{\beta}$.

For the least squares estimator to qualify as a plug-in estimator in bootstrapping residuals, the condition $\mathrm{E}(\hat{\beta}^*) = \hat{\beta}$ must hold. Problem 6 asks the reader to verify this condition for the intercept model $X = (\mathbf{1}, Z)$ under the above hypotheses. Bootstrap-t confidence intervals for the components of β are based on the studentized variables $(\hat{\beta}_i - \beta_i)/(\hat{\sigma}\sqrt{w_{ii}})$ and their bootstrap analogs, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2$ and w_{ii} is the ith diagonal entry of $(X^t X)^{-1}$. ■

The bootstrap percentile method of constructing confidence bounds is predicated on the existence of a continuous, symmetric distribution function $H(z) = 1 - H(-z)$ and a continuous, strictly increasing transformation $\psi(z)$ such that

$$\Pr\{\psi[T(\mathbf{x}^*)] - \psi[T(\mathbf{x})] \le z\} = H(z).$$

For example, $H(z)$ might be the standard normal distribution function and $\psi(z)$ a normalizing and variance-stabilizing transformation. If $\xi_\alpha(H)$ is the α-percentile of $H(z)$, then the bootstrap correspondence principle implies

$$\begin{aligned}
&\Pr\{\psi^{-1}[\psi(T(\mathbf{x})) + \xi_\alpha(H)] \le t(F)\} \\
&\quad = \Pr\{\psi[T(\mathbf{x})] - \psi[t(F)] \le -\xi_\alpha(H)\} \\
&\quad \approx \Pr\{\psi[T(\mathbf{x}^*)] - \psi[T(\mathbf{x})] \le -\xi_\alpha(H)\} \\
&\quad = H[-\xi_\alpha(H)] \\
&\quad = 1 - \alpha.
\end{aligned}$$

One of the pleasant features of the percentile method is that the lower confidence bound $\psi^{-1}[\psi(T(\mathbf{x})) + \xi_\alpha(H)]$ of $t(F)$ coincides with the α-percentile of $T(\mathbf{x}^*)$. This fact follows from the trivial rearrangement

$$\begin{aligned}
&\Pr\{T(\mathbf{x}^*) \le \psi^{-1}[\psi(T(\mathbf{x})) + \xi_\alpha(H)]\} \\
&\quad = \Pr\{\psi[T(\mathbf{x}^*)] - \psi[T(\mathbf{x})] \le \xi_\alpha(H)\} \\
&\quad = \alpha.
\end{aligned}$$

Thus, neither the transformation $\psi(x)$ nor the distribution function $H(x)$ need ever be explicitly calculated. In practice, one takes the bootstrap Monte Carlo estimate of the α-percentile of $T(\mathbf{x}^*)$ as the lower confidence bound of $t(F)$.

Unfortunately, neither the bootstrap-t nor the bootstrap percentile transformation is a panacea. To use the bootstrap-t, we need a good estimator

$V(\mathbf{x})$ of the variance of $T(\mathbf{x})$. Such estimators are not always available. Best results are obtained when $R(\mathbf{x}^*) = [T(\mathbf{x}^*) - T(\mathbf{x})]/\sqrt{V(\mathbf{x}^*)}$ is nearly pivotal. The simplicity of the bootstrap percentile method and its invariance under monotone transformations recommend it, but it seldom performs well unless n is large. Interested readers can pursue some of the more elaborate and potentially better bootstrap schemes for constructing confidence intervals in the books [9, 12, 17].

22.3 Balanced Bootstrap Simulations

We now turn to methods of reducing Monte Carlo sampling error. The first of these methods, balanced bootstrap resampling [6], is best illustrated by an example where simulation is unnecessary. According to the bootstrap correspondence principle, we estimate the bias of the sample mean $\bar{x}$ of n i.i.d. observations $\mathbf{x} = (x_1, \ldots, x_n)$ by the Monte Carlo difference

$$\frac{1}{B}\sum_{b=1}^{B} \bar{x}_b^* - \bar{x}.$$

By chance alone, this difference often is nonzero. We can eliminate this artificial bias by adopting nonindependent Monte Carlo sampling. Balanced bootstrap resampling retains as much randomness in the bootstrap resamples $\mathbf{x}_b$ as possible while forcing each original observation x_i to appear exactly B times in the bootstrap resamples. The naive implementation of the balanced bootstrap involves concatenating B copies of the data $(x_1, \ldots, x_n)$, randomly permuting the resulting data vector v of length nB, and then taking successive blocks of size n from v for the B bootstrap resamples.

A drawback of this permutation method of producing balanced bootstrap resamples is that it requires storage of v. Gleason [10] proposed an acceptance–rejection algorithm that sequentially creates only one block of v at a time and therefore minimizes computer storage. We can visualize Gleason's algorithm by imagining n urns, with urn i initially filled with B replicates of observation x_i. In the first step of the algorithm, we simply choose an urn at random and extract one of its replicate observations. This forms the first observation of the first bootstrap resample. In filling out the first resample and constructing subsequent ones, we must adjust our sampling procedure to reflect the fact that the urns may contain different numbers of observations. Let c_i be the number of observations currently left in urn i, and set $c = \max_i c_i$. If there are k nonempty urns, we choose one of these k urns at random, say the ith, and propose extracting a replicate x_i. Our decision to accept or reject the replicate is determined by selecting a random uniform deviate U. If $U \leq c_i/c$, then we accept the replicate; otherwise, we reject the replicate. When a replicate x_i is accepted, c_i is

reduced by 1, and the maximum c is recomputed. In practice, recomputation of c is the most time-consuming step of the algorithm. Gleason [10] suggests some remedies that require less frequent updating of c.

In the special case of bootstrap bias estimation for a statistic $T(\bar{x})$ based on the sample mean, there are two alternatives to the balanced bootstrap. Both involve adding corrections to the usual bias approximation

$$\frac{1}{B}\sum_{b=1}^{B} T(\bar{x}_b^*) - T(\bar{x}) \tag{1}$$

under independent Monte Carlo sampling. Here $\bar{x}_b^*$ is the sample mean of the bth bootstrap resample $\mathbf{x}_b^* = (x_{b1}^*, \ldots, x_{bn}^*)$. If dT is the differential of T, and $\bar{x^*}$ is the grand mean of the B bootstrap resamples, then the linear bias approximation is

$$\frac{1}{B}\sum_{b=1}^{B} T(\bar{x}_b^*) - T(\bar{x}) - dT(\bar{x})(\bar{x^*} - \bar{x}),$$

and the centered bias approximation [8] is

$$\frac{1}{B}\sum_{b=1}^{B} T(\bar{x}_b^*) - T(\bar{x}) - [T(\bar{x^*}) - T(\bar{x})] = \frac{1}{B}\sum_{b=1}^{B} T(\bar{x}_b^*) - T(\bar{x^*}).$$

Under balanced resampling $\bar{x^*} = \bar{x}$, and both the linear and centered bias approximations coincide with the standard bias approximation (1).

The balanced bootstrap applies to a wide range of problems, including estimation of distribution functions and quantiles. In practice, it works best for statistics that are nearly linear in the resampling vector. One can impose additional constraints on balanced resampling such as the requirement that every pair of observations occur the same number of times in the B bootstrap resamples. These higher-order balanced bootstraps are harder to generate, and limited evidence suggests that they are less effective than other methods of reducing Monte Carlo error [9].

22.4 Antithetic Bootstrap Simulations

In certain situations, the method of antithetic resampling discussed in Chapter 21 is well suited to the bootstrap. Hall [11, 12] suggests implementing antithetic resampling by replacing scalar i.i.d. observations $(x_1, \ldots, x_n)$ by their order statistics $x_{(1)} \le \cdots \le x_{(n)}$. He then defines the permutation $\pi(i) = n - i + 1$ that reverses the order statistics; in other words, $x_{(\pi[1])} \ge \cdots \ge x_{(\pi[n])}$. For any bootstrap resample $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$, we can construct a corresponding antithetic bootstrap resample $\mathbf{x}^{**}$ by substituting $x_{(\pi[i])}$ for every appearance of $x_{(i)}$ in $\mathbf{x}^*$. Often it is intuitively clear that the identically distributed statistics $T^* = T(\mathbf{x}^*)$ and $T^{**} = T(\mathbf{x}^{**})$

are negatively correlated. Negative correlation makes it advantageous to approximate the mean $\mathrm{E}(T^*)$ by taking sample averages of the statistic $(T^* + T^{**})/2$.

If $T(\mathbf{x})$ is a monotone function of the sample mean $\bar{x}$, then negative correlation is likely to occur. For example, when $T^* = T(\bar{x}^*)$ is symmetric about its mean, the residuals $T^* - \mathrm{E}(T^*)$ and $T^{**} - \mathrm{E}(T^*)$ are always of opposite sign, and a large positive value of one residual is matched by a large negative value of the other residual. This favorable state of affairs for the antithetic bootstrap persists for statistics $T^* = T(\bar{x}^*)$ having low skewness.

For a vector sample $(x_1, \ldots, x_n)$, the order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$ no longer exist. However, for a statistic $T(\bar{x})$ depending only on the sample mean, we can order the observations by their deflection of $T(\bar{x}^*)$ from $T(\bar{x})$. To a first approximation, this deflection is measured by

$$T(x_i) - T(\bar{x}) \approx dT(\bar{x})(x_i - \bar{x}).$$

Hall [11, 12] recommends implementing antithetic resampling after ordering the x_i by their approximate deflections $dT(\bar{x})(x_i - \bar{x})$.

22.5 Importance Resampling

In standard bootstrap resampling, each observation x_i is resampled uniformly with probability $1/n$. In some applications it is helpful to implement importance sampling by assigning different resampling probabilities p_i to the different observations x_i [4, 13]. For instance, with univariate observations $(x_1, \ldots, x_n)$, we may want to emphasize one of the tails of the empirical distribution. If we elect to resample nonuniformly according to the multinomial distribution with proportions $p = (p_1, \ldots, p_n)^t$, then the equality

$$\begin{aligned}
\mathrm{E}[T(\mathbf{x}^*)] &= \mathrm{E}_p \left[T(\mathbf{x}^*) \frac{\binom{n}{m_1^* \cdots m_n^*} \left(\frac{1}{n}\right)^n}{\binom{n}{m_1^* \cdots m_n^*} \prod_{i=1}^n p_i^{m_i^*}} \right] \\
&= \mathrm{E}_p \left[T(\mathbf{x}^*) \prod_{i=1}^n (np_i)^{-m_i^*} \right]
\end{aligned}$$

connects the uniform expectation and the importance expectation on the bootstrap resampling space. Here m_i^* represents the number of times sample point x_i appears in $\mathbf{x}^*$. Thus, we can approximate the mean $\mathrm{E}[T(\mathbf{x}^*)]$ by taking a bootstrap average

$$\frac{1}{B} \sum_{b=1}^B T(\mathbf{x}_b^*) \prod_{i=1}^n (np_i)^{-m_{bi}^*} \tag{2}$$

with multinomial sampling relative to p. This Monte Carlo approximation has variance

$$\frac{1}{B}\Big\{ \mathrm{E}_p\left[T(\mathbf{x}^*)^2 \prod_{i=1}^n (np_i)^{-2m_{bi}^*}\right] - \mathrm{E}\left[T(\mathbf{x}^*)\right]^2\Big\},$$

which we can minimize with respect to p by minimizing the second moment $\mathrm{E}_p\left[T(\mathbf{x}^*)^2 \prod_{i=1}^n (np_i)^{-2m_{bi}^*}\right]$.

Hall [12] suggests approximately minimizing the second moment by taking a preliminary uniform bootstrap sample of size B_1. Based on the preliminary resample, we approximate $\mathrm{E}_p\left[T(\mathbf{x}^*)^2 \prod_{i=1}^n (np_i)^{-2m_{bi}^*}\right]$ by the Monte Carlo average

$$\begin{aligned} s(p) &= \frac{1}{B_1}\sum_{b=1}^{B_1} T(\mathbf{x}_b^*)^2 \prod_{i=1}^n (np_i)^{-2m_{bi}^*} \prod_{i=1}^n (np_i)^{m_{bi}^*} \\ &= \frac{1}{B_1}\sum_{b=1}^{B_1} T(\mathbf{x}_b^*)^2 \prod_{i=1}^n (np_i)^{-m_{bi}^*}. \end{aligned} \tag{3}$$

The function $s(p)$ serves as a surrogate for $\mathrm{E}_p\left[T(\mathbf{x}^*)^2 \prod_{i=1}^n (np_i)^{-2m_{bi}^*}\right]$. It is possible to minimize $s(p)$ on the open simplex

$$U = \{p : p_i > 0,\ i = 1, \ldots, n,\ \sum_{i=1}^n p_i = 1\}$$

by standard methods.

For instance, we can apply the adaptive barrier method sketched in Chapter 14 when n is not too large. The method of geometric programming offers another approach [14]. Both methods are facilitated by the convexity of $s(p)$ on U. Convexity is evident from the form

$$\begin{aligned} d^2 s(p) = \frac{1}{B_1}\sum_{b=1}^{B_1} T(\mathbf{x}_b^*)^2 \prod_{i=1}^n (np_i)^{-m_{bi}^*} \\ \times \left\{ \begin{pmatrix} \frac{m_{b1}^*}{p_1} \\ \vdots \\ \frac{m_{bn}^*}{p_n} \end{pmatrix} \begin{pmatrix} \frac{m_{b1}^*}{p_1} & \cdots & \frac{m_{bn}^*}{p_n} \end{pmatrix} + \begin{pmatrix} \frac{m_{b1}^*}{p_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{m_{bn}^*}{p_n^2} \end{pmatrix} \right\}. \end{aligned} \tag{4}$$

of the second differential. Problems 9 and 10 ask the reader to verify this formula and check that $s(p)$ is strictly convex and attains its minimum on U provided there exists for each index i an index b such that $T(\mathbf{x}_b^*) \neq 0$ and $m_{bi}^* \neq 0$. When this condition fails, one or more optimal p_i are estimated as 0. This makes it impossible to resample some observations and suggests that the constraint $p_i \geq \epsilon > 0$ be imposed for all i and some small ϵ.

Once the optimal vector p_{opt} is calculated, we take a second bootstrap resample of size B_2 with multinomial probabilities p_{opt} and compute the

sample average (2) with B_2 replacing B and p_{opt} replacing p. Call the result A_2. Given the outcomes of the first bootstrap, we can also easily compute a sample average A_1 approximating $\mathrm{E}[T(\mathbf{x}^*)]$ under uniform resampling. Each of these sample averages has an attached sample variance V_i. The convex combination

$$\frac{V_2}{V_1+V_2}A_1+\frac{V_1}{V_1+V_2}A_2$$

is unbiased for $\mathrm{E}[T(\mathbf{x}^*)]$ and should have nearly minimal variance. (See Problem 11.)

The obvious strengths of Hall's strategy of importance sampling are its generality, its adaptive nature, and its use of stage-one resampling for both approximating the importance density and the sought-after expectation. In practice the strategy has at least two drawbacks. First, it entails solving a nontrivial optimization problem as an intermediate step. Second, Hall [12] argues theoretically that the method offers little advantage in approximating certain expectations such as central moments. However, Hall's method appears to yield substantial dividends in approximating distribution functions and tail probabilities.

Example 22.5.1 (*Hormone Patch Data*). Efron and Tibshirani [9] discuss estimation of the ratio $\mathrm{E}(Y)/\mathrm{E}(Z)$ in the hormone patch data appearing in Table 22.1. The natural estimator of $\mathrm{E}(Y)/\mathrm{E}(Z)$ is the ratio of sample means $\bar{y}/\bar{z}$. Sampling the bootstrap analog $\bar{y}^*/\bar{z}^*$ gives a feel for the distribution of $\bar{y}/\bar{z}$. To illustrate the value of importance sampling, we now consider Monte Carlo approximation of the right-tail probability $\Pr(\bar{y}^*/\bar{z}^* > .2)$. In $B = 500$ ordinary bootstrap replicates, this event occurred only 6 times. Thus, we are fairly far out in the right tail. Minimizing the importance criterion (3) for these 500 replicates yields the optimal importance probability vector

$$p_{\text{opt}} = (0.0357, 0.3928, 0.0414, 0.2266, 0.0788, 0.1475, 0.0416, 0.0357)^t.$$

Figure 22.1 plots running Monte Carlo averages approximating the right-tail probability $\Pr(\bar{y}^*/\bar{z}^* > .2)$ based on the ordinary bootstrap and the importance bootstrap with a total of $B = 100{,}000$ bootstrap resamples for each method. Averages are plotted in increments of 500 bootstrap resamples. Clearly, importance sampling converges more quickly. ■

TABLE 22.1. Hormone Patch Data

Subject	y	z	**Subject**	y	z
1	8406	−1200	5	4795	−1290
2	2342	2601	6	3516	351
3	8187	−2705	7	4796	−638
4	8459	1982	8	10238	−2719

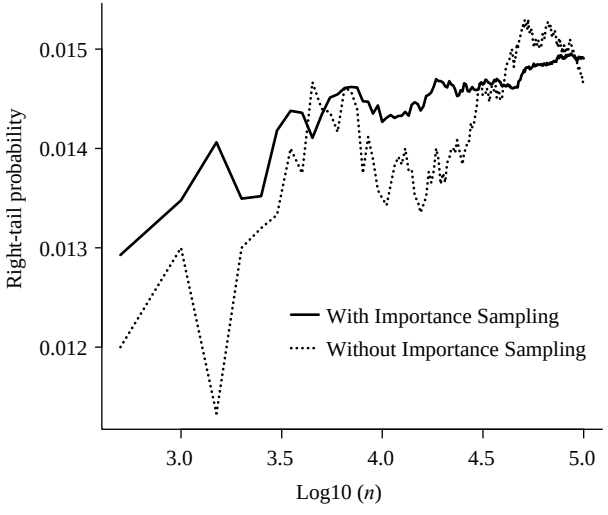

FIGURE 22.1. Running Monte Carlo Averages for $\Pr(\bar{y}^*/\bar{z}^* > .2)$

22.6 Problems

1. If the n observations $\{x_1, \ldots, x_n\}$ are distinct, then we can represent a bootstrap resample by an n-tuple $(m_1, \ldots, m_n)$ of nonnegative integers, where m_i indicates the number of repetitions of x_i in the resample. Show that there are $\binom{2n-1}{n}$ such n-tuples. If the n-tuple $(m_1, \ldots, m_n)$ is assigned multinomial probability $\binom{n}{m_1 \cdots m_n} n^{-n}$, then further demonstrate that the most probable n-tuple is $(1, \ldots, 1)$. Apply Stirling's formula and show that

$$\begin{aligned} \binom{2n-1}{n} &\asymp \frac{2^{2n-1}}{\sqrt{n\pi}} \\ \binom{n}{1 \cdots 1} \left(\frac{1}{n}\right)^n &\asymp \sqrt{2n\pi} e^{-n}. \end{aligned}$$

Finally, prove that a given x_i appears in a bootstrap resample with approximate probability $1 - e^{-1} \approx 0.632$.

2. Problem 1 indicates that the most likely bootstrap resample from n distinct observations $\{x_1, \ldots, x_n\}$ has probability $p_n \asymp \sqrt{2n\pi} e^{-n}$. Show that b bootstrap samples are all distinct with probability

$$\begin{aligned} q_{nb} &\geq (1 - p_n)(1 - 2p_n) \cdots (1 - [b-1]p_n) \\ &\geq 1 - \frac{1}{2} b(b-1) p_n. \end{aligned}$$

For $n = 20$ and $b = 2000$, compute the bound $q_{nb} \geq 0.954$ [12].

3. The bias of the estimator $\bar{y}/\bar{z}$ in the hormone patch data of Example 22.5.1 can be corrected by the bootstrap. Plot the bias of the bootstrap

TABLE 22.2. Mouse Survival Data

Group	Survival Time	Group	Survival Time
Treatment	94	Control	10
Treatment	38	Control	40
Treatment	23	Control	104
Treatment	197	Control	51
Treatment	99	Control	27
Treatment	16	Control	146
Treatment	141	Control	30
Control	52	Control	46

analog $\bar{y}^*/\bar{z}^*$ as calculated by the ordinary bootstrap, the centered bootstrap, and the balanced bootstrap as a function of the number of bootstrap replicates B. Which bootstrap bias approximation converges most rapidly to the true bias of $\bar{y}^*/\bar{z}^*$?

4. Suppose $X = (Y, Z)$ is a bivariate random vector with independent components Y and Z. If you are given n independent realizations $x_1, \ldots, x_n$ of X, what alternative distribution would be a reasonable substitute for the empirical distribution in bootstrap resampling?

5. In a certain medical experiment summarized in Table 22.2, mice were assigned to either a treatment or control group, and their survival times recorded in days [9]. Compute a 90 percent bootstrap-t confidence interval for the mean of each group.

6. In the linear regression model of Example 22.2.1, prove that the condition $\mathrm{E}(\hat{\beta}^*) = \hat{\beta}$ holds for the intercept model $X = (\mathbf{1}, Z)$ under bootstrapping residuals. (Hint: Show that $1^t\hat{\mathbf{r}} = 0$, where $\hat{\mathbf{r}}$ is the residual vector.)

7. If $n = 2m - 1$ is odd, and the observations $\{x_1, \ldots, x_n\}$ are distinct, then the sample median $x_{(m)}$ is uniquely defined. Prove that a bootstrap resample $\{x_1^*, \ldots, x_n^*\}$ has median $x_{(m)}^*$ with distribution function

$$G_n^*(x_{(k)}) = \Pr(x_{(m)}^* \leq x_{(k)}) = \sum_{j=m}^{n} \binom{n}{j} \frac{k^j (n-k)^{n-j}}{n^n}.$$

By definition, the quantile $\xi_\alpha(G_n^*)$ of $x_{(m)}^*$ is the smallest order statistic $x_{(k)}$ satisfying $G_n^*(x_{(k)}) \geq \alpha$. The bootstrap percentile method gives $\xi_\alpha(G_n^*)$ as an approximate $1 - \alpha$ lower confidence bound for the true median.

8. Suppose that $T(\mathbf{x})$ is a plug-in estimator of $t(F)$. If the bootstrap analog $T(\mathbf{x}^*)$ of $T(\mathbf{x})$ has distribution function G_n^*, then derive the approximate $1-\alpha$ upper and lower confidence bounds $2T(\mathbf{x}) - \xi_\alpha(G_n^*)$ and $2T(\mathbf{x}) - \xi_{1-\alpha}(G_n^*)$ of $t(F)$. Here $\xi_\alpha(G_n^*)$ denotes the α-percentile of G_n^*.

9. Calculate the second differential (4) of the function $s(p)$ defined in equation (3).

10. Show that the function $s(p)$ in equation (3) is convex on the open simplex $U = \{p : p_i > 0,\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = 1\}$. Prove in addition that $s(p)$ is strictly convex and attains its minimum on U provided there exists for each index i an index b such that $T(\mathbf{x}_b^*) \neq 0$ and $m_{bi}^* \neq 0$.

11. Suppose two independent random variables Y_1 and Y_2 have the same mean but different variances v_1 and v_2. Demonstrate that the convex combination $\beta Y_1 + (1 - \beta)Y_2$ with minimal variance is achieved by taking $\beta = v_2/(v_1 + v_2)$.

References

[1] Babu GJ, Singh K (1984) On a one term Edgeworth correction for Efron's bootstrap. *Sankhyā A* 46:219–232

[2] Bickel PJ, Freedman DA (1981) Some asymptotics for the bootstrap. *Ann Stat* 9:1196–1217

[3] Booth JG, Sarkar S (1998) Monte Carlo approximation of bootstrap variances. *The American Statistician* (in press)

[4] Davison AC (1988) Discussion of papers by DV Hinkley and TJ DiCiccio and JP Romano. *J Roy Stat Soc B* 50:356–357

[5] Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge

[6] Davison AC, Hinkley DV, Schechtman E (1986) Efficient bootstrap simulation. *Biometrika* 73:555–566

[7] Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7:1–26

[8] Efron B (1990) More efficient bootstrap computations. *J Amer Stat Assoc* 85:79–89

[9] Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap.* Chapman & Hall, New York

[10] Gleason JR (1988) Algorithms for balanced bootstrap simulations. *Amer Statistician* 42:263–266

[11] Hall P (1989) Antithetic resampling for the bootstrap. *Biometrika* 76:713–724

[12] Hall P (1992) *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York

[13] Johns MV Jr (1988) Importance sampling for bootstrap confidence intervals. *J Amer Stat Assoc* 83:709–714

[14] Peressini AL, Sullivan FE, Uhl JJ Jr (1988) *The Mathematics of Nonlinear Programming.* Springer-Verlag, New York

[15] Quenouille M (1949) Approximate tests of correlation in time series. *J Roy Stat Soc Ser B* 11:18–44

[16] Serfling RJ (1980) *Approximation Theorems in Mathematical Statistics.* Wiley, New York

[17] Shao J, Tu D (1995) *The Jackknife and Bootstrap*. Springer-Verlag, New York
[18] Tukey JW (1958) Bias and confidence in not quite large samples. (Abstract) *Ann Math Stat* 29:614

23
Finite-State Markov Chains

23.1 Introduction

Applied probability and statistics thrive on models. Markov chains are one of the richest sources of good models for capturing dynamical behavior with a large stochastic component [2, 3, 5, 6, 8, 10, 11, 13]. Certainly, every research statistician should be comfortable formulating and manipulating Markov chains. In this chapter we give a quick overview of some of the relevant theory of Markov chains in the simple context of finite-state chains. We cover both discrete-time and continuous-time chains in what we hope is a lively blend of applied probability, graph theory, linear algebra, and differential equations. Since this may be a first account for many readers, we stress intuitive explanations and computational techniques rather than mathematical rigor.

To convince readers of the statistical utility of Markov chains, we introduce the topic of hidden Markov chains [1, 4, 17]. This brings in Baum's forward and backward algorithms and inhomogeneous chains. Limitations of space prevent us from considering specific applications. Interested readers can consult our listed references on speech recognition [17], physiological models of single-ion channels [7], gene mapping by radiation hybrids [14], and alignment of multiple DNA sequences [19].

23.2 Discrete-Time Markov Chains

For the sake of simplicity, we will only consider chains with a finite-state space [2, 6, 8, 10, 11]. The movement of such a chain from epoch to epoch (equivalently, generation to generation) is governed by its transition probability matrix $P = (p_{ij})$. If Z_n denotes the state of the chain at epoch n, then $p_{ij} = \Pr(Z_n = j \mid Z_{n-1} = i)$. As a consequence, every entry of P satisfies $p_{ij} \geq 0$, and every row of P satisfies $\sum_j p_{ij} = 1$. Implicit in the definition of p_{ij} is the fact that the future of the chain is determined by its present regardless of its past. This Markovian property is expressed formally by the equation

$$\Pr(Z_n = i_n \mid Z_{n-1} = i_{n-1}, \ldots, Z_0 = i_0) = \Pr(Z_n = i_n \mid Z_{n-1} = i_{n-1}).$$

The n-step transition probability $p_{ij}^{(n)} = \Pr(Z_n = j \mid Z_0 = i)$ is given by the entry in row i and column j of the matrix power P^n. This follows because the decomposition

$$p_{ij}^{(n)} = \sum_{i_1} \cdots \sum_{i_{n-1}} p_{ii_1} \cdots p_{i_{n-1}j}$$

over all paths $i \to i_1 \to \cdots \to i_{n-1} \to j$ corresponds to matrix multiplication. A question of fundamental theoretical importance is whether the matrix powers P^n converge. If the chain eventually forgets its starting state, then the limit should have identical rows. Denoting the common limiting row by π, we deduce that $\pi = \pi P$ from the calculation

$$\begin{aligned} \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix} &= \lim_{n \to \infty} P^{n+1} \\ &= \left(\lim_{n \to \infty} P^n \right) P \\ &= \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix} P. \end{aligned}$$

Any probability distribution π on the states of the chain satisfying the condition $\pi = \pi P$ is termed an equilibrium (or stationary) distribution of the chain. For finite-state chains, equilibrium distributions always exist [6, 8]. The real issue is uniqueness.

Mathematicians have attacked the uniqueness problem by defining appropriate ergodic conditions. For finite-state Markov chains, two ergodic assumptions are invoked. The first is aperiodicity; this means that the greatest common divisor of the set $\{n \geq 1 : p_{ii}^{(n)} > 0\}$ is 1 for every state i. Aperiodicity trivially holds when $p_{ii} > 0$ for all i. The second ergodic assumption is irreducibility; this means that for every pair of states (i, j), there exists a positive integer n_{ij} such that $p_{ij}^{(n_{ij})} > 0$. In other words,

every state is reachable from every other state. Said yet another way, all states communicate. For an irreducible chain, Problem 1 states that the integer n_{ij} can be chosen independently of the particular pair (i, j) if and only if the chain is also aperiodic. Thus, we can merge the two ergodic assumptions into the single assumption that some power P^n has all entries positive. Under this single ergodic condition, we showed in Chapter 6 that a unique equilibrium distribution π exists and that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$. Because all states communicate, the entries of π are necessarily positive.

Equally important is the ergodic theorem [6, 8]. This theorem permits one to run a chain and approximate theoretical means by sample means. More precisely, let $f(z)$ be some function defined on the states of an ergodic chain. Then $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(Z_i)$ exists and equals the theoretical mean $\mathrm{E}_\pi[f(Z)] = \sum_z \pi_z f(z)$ of $f(Z)$ under the equilibrium distribution π. This result generalizes the law of large numbers for independent sampling.

The equilibrium condition $\pi = \pi P$ can be restated as the system of equations

$$\pi_j = \sum_i \pi_i p_{ij} \tag{1}$$

for all j. In many Markov chain models, the stronger condition

$$\pi_j p_{ji} = \pi_i p_{ij} \tag{2}$$

holds for all pairs (i, j). If this is the case, then the probability distribution π is said to satisfy detailed balance. Summing equation (2) over i yields the equilibrium condition (1). An irreducible Markov chain with equilibrium distribution π satisfying detailed balance is said to be reversible. Irreducibility is imposed to guarantee that π is unique and has positive entries.

If $i_1, \ldots, i_m$ is any sequence of states in a reversible chain, then detailed balance implies

$$\begin{aligned}
\pi_{i_1} p_{i_1 i_2} &= \pi_{i_2} p_{i_2 i_1} \\
\pi_{i_2} p_{i_2 i_3} &= \pi_{i_3} p_{i_3 i_2} \\
&\vdots \\
\pi_{i_{m-1}} p_{i_{m-1} i_m} &= \pi_{i_m} p_{i_m i_{m-1}} \\
\pi_{i_m} p_{i_m i_1} &= \pi_{i_1} p_{i_1 i_m}.
\end{aligned}$$

Multiplying these equations together and cancelling the common positive factor $\pi_{i_1} \cdots \pi_{i_m}$ from both sides of the resulting equality give Kolmogorov's circulation criterion [12]

$$p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{m-1} i_m} p_{i_m i_1} = p_{i_1 i_m} p_{i_m i_{m-1}} \cdots p_{i_3 i_2} p_{i_2 i_1}. \tag{3}$$

Conversely, if an irreducible Markov chain satisfies Kolmogorov's criterion, then the chain is reversible. This fact can be demonstrated by

explicitly constructing the equilibrium distribution and showing that it satisfies detailed balance. The idea behind the construction is to choose some arbitrary reference state i and to pretend that π_i is given. If j is another state, let $i \to i_1 \to \cdots \to i_m \to j$ be any path leading from i to j. Then the formula

$$\pi_j = \pi_i \frac{p_{ii_1}p_{i_1i_2}\cdots p_{i_mj}}{p_{ji_m}p_{i_mi_{m-1}}\cdots p_{i_1i}} \tag{4}$$

defines π_j. A straightforward application of Kolmogorov's criterion (3) shows that the definition (4) does not depend on the particular path chosen from i to j. To validate detailed balance, suppose that k is adjacent to j. Then $i \to i_1 \to \cdots \to i_m \to j \to k$ furnishes a path from i to k through j. It follows from (4) that $\pi_k = \pi_j p_{jk}/p_{kj}$, which is obviously equivalent to detailed balance. In general, the value of π_i is not known beforehand. Setting $\pi_i = 1$ produces the equilibrium distribution up to a normalizing constant.

Example 23.2.1 (*Random Walk on a Graph*)**.** Consider a connected graph with vertex set V and edge set E. The number of edges $d(v)$ incident on a given vertex v is called the degree of v. Due to the connectedness assumption, $d(v) > 0$ for all $v \in V$. Now define the transition probability matrix $P = (p_{uv})$ by

$$p_{uv} = \begin{cases} \frac{1}{d(u)} & \text{for } \{u, v\} \in E \\ 0 & \text{for } \{u, v\} \notin E. \end{cases}$$

This Markov chain is irreducible because of the connectedness assumption; it is also aperiodic unless the graph is bipartite. (A graph is said to be bipartite if we can partition its vertex set into two disjoint subsets F and M, say females and males, such that each edge has one vertex in F and the other vertex in M.) If V has m edges, then the equilibrium distribution π of the chain has components $\pi_v = d(v)/(2m)$. It is trivial to show that this choice of π satisfies detailed balance. ■

Example 23.2.2 (*Wright's Model of Genetic Drift*)**.** Consider a population of m organisms from some animal or plant species. Each member of this population carries two genes at some genetic locus, and these genes assume two forms (or alleles) labeled a_1 and a_2. At each generation, the population reproduces itself by sampling $2m$ genes with replacement from the current pool of $2m$ genes. If Z_n denotes the number of a_1 alleles at generation n, then it is clear that the Z_n constitute a Markov chain with transition probability matrix

$$p_{jk} = \binom{2m}{k}\left(\frac{j}{2m}\right)^k\left(1 - \frac{j}{2m}\right)^{2m-k}.$$

This chain is reducible because once one of the states 0 or $2m$ is reached, then the corresponding allele is fixed in the population, and no further

variation is possible. An infinite number of equilibrium distributions exist determined by $\pi_0 = \alpha$ and $\pi_{2m} = 1 - \alpha$ for $\alpha \in [0, 1]$. ■

Example 23.2.3 (*Ehrenfest's Model of Diffusion*). Consider a box with m gas molecules. Suppose the box is divided in half by a rigid partition with a very small hole. Molecules drift aimlessly around each half until one molecule encounters the hole and passes through. Let Z_n be the number of molecules in the left half of the box at epoch n. If epochs are timed to coincide with molecular passages, then the transition matrix of the chain is

$$p_{jk} = \begin{cases} 1 - \frac{j}{m} & \text{for } k = j + 1 \\ \frac{j}{m} & \text{for } k = j - 1 \\ 0 & \text{otherwise.} \end{cases}$$

This chain is periodic with period 2, irreducible, and reversible with equilibrium distribution $\pi_j = \binom{m}{j} 2^{-m}$. ■

23.3 Hidden Markov Chains

Hidden Markov chains incorporate both observed data and missing data. The missing data are the sequence of states visited by a Markov chain; the observed data provide partial information about this sequence of states. Denote the sequence of visited states by $Z_1, \ldots, Z_n$ and the observation taken at epoch i when the chain is in state Z_i by $Y_i = y_i$. Baum's algorithms [1, 4, 17] recursively compute the likelihood of the observed data

$$P = \Pr(Y_1 = y_1, \ldots, Y_n = y_n) \tag{5}$$

in a way that avoids enumerating all possible realizations $Z_1, \ldots, Z_n$.

The likelihood (5) is constructed from three ingredients: (a) the initial distribution π at the first epoch of the chain, (b) the epoch-dependent transition probabilities $p_{ijk} = \Pr(Z_{i+1} = k \mid Z_i = j)$, and (c) the conditional densities $\phi_i(y_i \mid j) = \Pr(Y_i = y_i \mid Z_i = j)$. The dependence of the transition probability p_{ijk} on i makes the chain inhomogeneous over time and allows greater flexibility in modeling. If the chain is homogeneous, then π is often taken as the equilibrium distribution. Implicit in the definition of $\phi_i(y_i \mid j)$ are the assumptions that $Y_1, \ldots, Y_n$ are independent given $Z_1, \ldots, Z_n$ and that Y_i depends only on Z_i. Finally, with obvious changes in notation, the observed data can be continuously rather than discretely distributed.

Baum's forward algorithm is based on recursively evaluating the joint probabilities

$$\alpha_i(j) = \Pr(Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Z_i = j).$$

At the first epoch, $\alpha_1(j) = \pi_j$ by definition; the obvious update to $\alpha_i(j)$ is

$$\alpha_{i+1}(k) = \sum_j \alpha_i(j)\phi_i(y_i \mid j)p_{ijk}. \tag{6}$$

The likelihood (5) can be recovered by computing $\sum_j \alpha_n(j)\phi_n(y_n \mid j)$ at the final epoch n.

In Baum's backward algorithm, we recursively evaluate the conditional probabilities

$$\beta_i(k) = \Pr(Y_{i+1} = y_{i+1}, \ldots, Y_n = y_n \mid Z_i = k),$$

starting by convention at $\beta_n(k) = 1$ for all k. The required update is clearly

$$\beta_i(j) = \sum_k p_{ijk}\phi_{i+1}(y_{i+1} \mid k)\beta_{i+1}(k). \tag{7}$$

In this instance, the likelihood is recovered at the first epoch by forming the sum $\sum_j \pi_j\phi_1(y_1 \mid j)\beta_1(j)$.

Baum's algorithms (6) and (7) are extremely efficient, particularly if the observations y_i strongly limit the number of compatible states at each epoch i. In statistical practice, maximization of the likelihood with respect to model parameters is usually an issue. Most maximum likelihood algorithms require the score in addition to the likelihood. These partial derivatives can often be computed quickly in parallel with other quantities in Baum's forward and backward algorithms. For example, suppose that a parameter θ impacts only the transition probabilities p_{ijk} for a specific epoch i [14]. Since we can write the likelihood as

$$P = \sum_j \sum_k \alpha_i(j)\phi_i(y_i \mid j)p_{ijk}\phi_{i+1}(y_{i+1} \mid k)\beta_{i+1}(k),$$

it follows that

$$\frac{\partial}{\partial\theta}P = \sum_j \sum_k \alpha_i(j)\phi_i(y_i \mid j)\Big[\frac{\partial}{\partial\theta}p_{ijk}\Big]\phi_{i+1}(y_{i+1} \mid k)\beta_{i+1}(k). \tag{8}$$

Similarly, if θ only enters the conditional density $\phi_i(y_i \mid j)$ for a given i, then the representation

$$P = \sum_j \alpha_i(j)\phi_i(y_i \mid j)\beta_i(j)$$

leads to the partial derivative formula

$$\frac{\partial}{\partial\theta}P = \sum_j \alpha_i(j)\Big[\frac{\partial}{\partial\theta}\phi_i(y_i \mid j)\Big]\beta_i(j). \tag{9}$$

Finally, if θ enters only into the initial distribution π, then

$$\frac{\partial}{\partial\theta}P = \sum_j \Big[\frac{\partial}{\partial\theta}\pi_j\Big]\phi_1(y_1 \mid j)\beta_1(j). \tag{10}$$

These formulas suggest that an efficient evaluation of P and its partial derivatives can be orchestrated by carrying out the backward algorithm first, saving all resulting $\beta_i(j)$, and then carrying out the forward algorithm while simultaneously computing all partial derivatives. Note that if a parameter θ enters into several of the factors defining P, then by virtue of the product rule of differentiation, we can express $\frac{\partial}{\partial\theta}P$ as a sum of the corresponding right-hand sides of equations (8), (9), and (10). Given a partial derivative $\frac{\partial}{\partial\theta}P$ of the likelihood P, we compute the corresponding entry in the score vector by taking the quotient $\frac{\partial}{\partial\theta}P/P$.

Besides evaluating and maximizing the likelihood, statisticians are often interested in finding a most probable sequence of states of the hidden Markov chain given the observed data. The Viterbi algorithm solves this problem by dynamic programming [7]. We proceed by solving the intermediate problems

$$\gamma_k(z_k) = \max_{z_1,\ldots,z_{k-1}} \Pr(Z_1 = z_1, \ldots, Z_k = z_k, Y_1 = y_1, \ldots, Y_k = y_k)$$

for each $k = 1, \ldots, n$, beginning with $\gamma_1(z_1) = \pi_{z_1}\phi_1(y_1 \mid z_1)$. When we reach $k = n$, then $\max_{z_n} \gamma_n(z_n)$ yields the largest joint probability

$$\Pr(Z_1 = z_1, \ldots, Z_n = z_n, Y_1 = y_1, \ldots, Y_n = y_n)$$

and consequently the largest conditional probability

$$\Pr(Z_1 = z_1, \ldots, Z_n = z_n \mid Y_1 = y_1, \ldots, Y_n = y_n)$$

as well. If we have kept track of one solution sequence $z_1(z_n), \ldots, z_{n-1}(z_n)$ for each $\gamma_n(z_n)$, then obviously we can construct a best overall sequence by taking the best z_n and appending to it $z_1(z_n), \ldots, z_{n-1}(z_n)$. To understand better the recursive phase of the algorithm, let

$$\delta_k(z_1, \ldots, z_k) = \Pr(Z_1 = z_1, \ldots, Z_k = z_k, Y_1 = y_1, \ldots, Y_k = y_k).$$

In this notation, we express $\gamma_{k+1}(z_{k+1})$ as

$$\begin{aligned}
\gamma_{k+1}(z_{k+1}) &= \max_{z_1,\ldots,z_k} \delta_{k+1}(z_1, \ldots, z_{k+1}) \\
&= \max_{z_1,\ldots,z_k} \delta_k(z_1, \ldots, z_k) p_{k,z_k,z_{k+1}} \phi_{k+1}(y_{k+1} \mid z_{k+1}) \\
&= \max_{z_k} p_{k,z_k,z_{k+1}} \phi_{k+1}(y_{k+1} \mid z_{k+1}) \max_{z_1,\ldots,z_{k-1}} \delta_k(z_1, \ldots, z_k) \\
&= \max_{z_k} p_{k,z_k,z_{k+1}} \phi_{k+1}(y_{k+1} \mid z_{k+1}) \gamma_k(z_k)
\end{aligned}$$

and create a maximizing sequence $z_1(z_{k+1}), \ldots, z_k(z_{k+1})$ for each z_{k+1} from the corresponding best z_k and its recorded sequence $z_1(z_k), \ldots, z_{k-1}(z_k)$.

23.4 Continuous-Time Markov Chains

Continuous-time Markov chains are often more realistic than discrete-time Markov chains. Just as in the discrete case, the behavior of a chain is described by an indexed family Z_t of random variables giving the state occupied by the chain at each time t. However, now the index t ranges over real numbers rather than integers. Of fundamental theoretical importance are the probabilities $p_{ij}(t) = \Pr(Z_t = j \mid Z_0 = i)$. For a chain having a finite number of states, these probabilities can be found by solving a matrix differential equation. To derive this equation, we use the short-time approximation

$$p_{ij}(t) = \lambda_{ij} t + o(t) \tag{11}$$

for $i \neq j$, where λ_{ij} is the transition rate (or infinitesimal transition probability) from state i to state j. Equation (11) implies the further short-time approximation

$$p_{ii}(t) = 1 - \lambda_i t + o(t), \tag{12}$$

where $\lambda_i = \sum_{j \neq i} \lambda_{ij}$.

The alternative perspective of competing risks sharpens our intuitive understanding of equations (11) and (12). Imagine that a particle executes the Markov chain by moving from state to state. If the particle is currently in state i, then each neighboring state independently beckons the particle to switch positions. The intensity of the temptation exerted by state j is the constant λ_{ij}. In the absence of competing temptations, the particle waits an exponential length of time T_{ij} with intensity λ_{ij} before moving to state j. Taking into account the competing temptations, the particle moves at the moment $\min_j T_{ij}$, which is exponentially distributed with intensity λ_i. Once the particle decides to move, it moves to state j with probability λ_{ij}/λ_i. Equations (11) and (12) now follow from the approximations

$$\begin{aligned} \left(1 - e^{-\lambda_i t}\right) \frac{\lambda_{ij}}{\lambda_i} &= \lambda_{ij} t + o(t) \\ e^{-\lambda_i t} &= 1 - \lambda_i t + o(t). \end{aligned}$$

Next consider the Chapman–Kolmogorov relation

$$p_{ij}(t+h) = p_{ij}(t) p_{jj}(h) + \sum_{k \neq j} p_{ik}(t) p_{kj}(h), \tag{13}$$

which simply says the chain must pass through some intermediate state k at time t enroute to state j at time $t+h$. Substituting the approximations (11) and (12) in (13) yields

$$p_{ij}(t+h) = p_{ij}(t)(1 - \lambda_j h) + \sum_{k \neq j} p_{ik}(t) \lambda_{kj} h + o(h).$$

Sending h to 0 in the difference quotient

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = -p_{ij}(t)\lambda_j + \sum_{k \neq j} p_{ik}(t)\lambda_{kj} + \frac{o(h)}{h}$$

produces the forward differential equation

$$p'_{ij}(t) = -p_{ij}(t)\lambda_j + \sum_{k \neq j} p_{ik}(t)\lambda_{kj}. \tag{14}$$

The system of differential equations (14) can be summarized in matrix notation by introducing the matrices $P(t) = [p_{ij}(t)]$ and $\Lambda = (\Lambda_{ij})$, where $\Lambda_{ij} = \lambda_{ij}$ for $i \neq j$ and $\Lambda_{ii} = -\lambda_i$. The forward equations in this notation become

$$\begin{aligned} P'(t) &= P(t)\Lambda \\ P(0) &= I, \end{aligned} \tag{15}$$

where I is the identity matrix. It is easy to check that the solution of the initial-value problem (15) is furnished by the matrix exponential [9, 13]

$$\begin{aligned} P(t) &= e^{t\Lambda} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}(t\Lambda)^k. \end{aligned} \tag{16}$$

Probabilists call Λ the infinitesimal generator or infinitesimal transition matrix of the process. The infinite series (16) converges because its partial sums form a Cauchy sequence. This fact follows directly from the inequality

$$\left\| \sum_{k=0}^{m} \frac{1}{k!}(t\Lambda)^k - \sum_{k=0}^{m+n} \frac{1}{k!}(t\Lambda)^k \right\| \leq \sum_{k=m+1}^{m+n} \frac{1}{k!}|t|^k \|\Lambda\|^k.$$

A probability distribution $\pi = (\pi_i)$ on the states of a continuous-time Markov chain is a row vector whose components satisfy $\pi_i \geq 0$ for all i and $\sum_i \pi_i = 1$. If

$$\pi P(t) = \pi \tag{17}$$

holds for all $t \geq 0$, then π is said to be an equilibrium distribution for the chain. Written in components, the eigenvector equation (17) reduces to $\sum_i \pi_i p_{ij}(t) = \pi_j$. Again this is completely analogous to the discrete-time theory. For small t equation (17) can be rewritten as

$$\pi(I + t\Lambda) + o(t) = \pi.$$

This approximate form makes it obvious that $\pi\Lambda = \mathbf{0}$ is a necessary condition for π to be an equilibrium distribution. Multiplying (16) on the left by π shows that $\pi\Lambda = \mathbf{0}$ is also a sufficient condition for π to be an equilibrium distribution. In components this necessary and sufficient condition

amounts to

$$\sum_{j \neq i} \pi_j \lambda_{ji} = \pi_i \sum_{j \neq i} \lambda_{ij} \tag{18}$$

for all i. If all the states of a Markov chain communicate, then there is one and only one equilibrium distribution π. Furthermore, each of the rows of $P(t)$ approaches π as t approaches ∞. Lamperti [13] provides a clear exposition of these facts.

Fortunately, the annoying feature of periodicity present in discrete-time theory disappears in the continuous-time theory. The definition and properties of reversible chains carry over directly from discrete time to continuous time provided we substitute infinitesimal transition probabilities for transition probabilities. For instance, the detailed balance condition becomes

$$\pi_i \lambda_{ij} = \pi_j \lambda_{ji}$$

for all pairs $i \neq j$. Kolmogorov's circulation criterion for reversibility continues to hold; when it is true, the equilibrium distribution is constructed from the infinitesimal transition probabilities exactly as in discrete time.

Example 23.4.1 (*Oxygen Attachment to Hemoglobin*). A hemoglobin molecule has four possible sites to which oxygen (O_2) can attach. If the surrounding concentration s_o of O_2 is sufficiently high, then we can model the number of sites occupied on a hemoglobin molecule as a continuous-time Markov chain [18]. Figure 23.1 depicts the model; in the figure, each arc is labeled by an infinitesimal transition probability and each state by the number of O_2 molecules attached to the hemoglobin molecule. The forward rates $s_o k_{+j}$ incorporate the concentration of O_2. Because this chain is reversible, we can calculate its equilibrium distribution starting from the reference state 0 as $\pi_i = \pi_0 s_o^i \prod_{j=1}^{i} k_{+j}/k_{-j}$. ■

Example 23.4.2 (*Continuous-Time Multi-Type Branching Processes*). Matrix exponentials also appear in the theory of continuous-time branching processes. In such a process one follows a finite number of independently acting particles that reproduce and die. In a multi-type branching process, each particle is classified in one of n possible categories. A type i particle lives an exponentially distributed length of time with a death intensity of λ_i. At the end of its life, a type i particle reproduces both particles of its own type and particles of other types. Suppose that on average it produces f_{ij} particles of type j. We would like to calculate the average number of

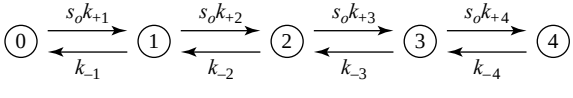

FIGURE 23.1. A Markov Chain Model for Oxygen Attachment to Hemoglobin

particles $m_{ij}(t)$ of type j at time $t \geq 0$ starting with a single particle of type i at time 0. Since particles of type j at time $t+h$ either arise from particles of type j at time t which do not die during $(t, t+h)$ or from particles of type k which die during $(t, t+h)$ and reproduce particles of type j, we find that

$$m_{ij}(t+h) = m_{ij}(t)(1-\lambda_j h) + \sum_k m_{ik}(t)\lambda_k f_{kj} h + o(h).$$

Forming the corresponding difference quotients and sending h to 0 yield the differential equations

$$m'_{ij}(t) = \sum_k m_{ik}(t)\lambda_k(f_{kj} - 1_{\{k=j\}}),$$

which we summarize as the matrix differential equation $M'(t) = M(t)\Omega$ for the $n \times n$ matrices $M(t) = [m_{ij}(t)]$ and $\Omega = [\lambda_i(f_{ij} - 1_{\{i=j\}})]$. Again the solution is provided by the matrix exponential $M(t) = e^{t\Omega}$ subject to the initial condition $M(0) = I$. ■

23.5 Calculation of Matrix Exponentials

From the definition of the matrix exponential e^A, it is easy to deduce that it is continuous in A and satisfies $e^{A+B} = e^A e^B$ whenever $AB = BA$. It is also straightforward to check the differentiability condition

$$\frac{d}{dt}e^{tA} = Ae^{tA} = e^{tA}A.$$

Of more practical importance is how one actually calculates e^{tA} [16]. In some cases it is possible to do so analytically. For instance, if u and v are column vectors with the same number of components, then

$$e^{suv^t} = \begin{cases} I + suv^t & \text{if } v^t u = 0 \\ I + \frac{e^{sv^t u} - 1}{v^t u} uv^t & \text{otherwise.} \end{cases}$$

This follows from the formula $(uv^t)^i = (v^t u)^{i-1} uv^t$. The special case having $u = (-\alpha, \beta)^t$ and $v = (1, -1)^t$ permits us to calculate the finite-time transition matrix

$$P(s) = \exp\left[s \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}\right]$$

for a two-state Markov chain.

If A is a diagonalizable $n \times n$ matrix, then we can write $A = TDT^{-1}$ for D a diagonal matrix with ith diagonal entry ρ_i. In this situation note that $A^2 = TDT^{-1}TDT^{-1} = TD^2T^{-1}$ and in general that $A^i = TD^iT^{-1}$.

Hence,

$$\begin{aligned} e^{tA} &= \sum_{i=0}^{\infty} \frac{1}{i!}(tA)^i \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} T(tD)^i T^{-1} \\ &= T e^{tD} T^{-1}, \end{aligned}$$

where

$$e^{tD} = \begin{pmatrix} e^{\rho_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\rho_n t} \end{pmatrix}.$$

Even if we cannot diagonalize A explicitly, we can usually do so numerically.

When $t > 0$ is small, another method is to truncate the series expansion for e^{tA} to $\sum_{i=0}^{n}(tA)^i/i!$ for n small. For larger t such truncation can lead to serious errors. If the truncated expansion is accurate for all $t \leq c$, then for arbitrary t one can exploit the property $e^{(s+t)A} = e^{sA}e^{tA}$ of the matrix exponential. Thus, if $t > c$, take the smallest positive integer k such that $2^{-k}t \leq c$ and approximate $e^{2^{-k}tA}$ by the truncated series. Applying the multiplicative property we can compute e^{tA} by squaring $e^{2^{-k}tA}$, squaring the result $e^{2^{-k+1}tA}$, squaring the result of this, and so forth, a total of k times. Problems 11 and 12 explore how the errors encountered in this procedure can be controlled.

23.6 Problems

1. Demonstrate that a finite-state Markov chain is ergodic (irreducible and aperiodic) if and only if some power P^n of the transition matrix P has all entries positive. (Hints: For sufficiency, show that if some power P^n has all entries positive, then P^{n+1} has all entries positive. For necessity, note that $p_{ij}^{(r+s+t)} \geq p_{ik}^{(r)} p_{kk}^{(s)} p_{kj}^{(t)}$, and use the number theoretic fact that the set $\{s : p_{kk}^{(s)} > 0\}$ contains all sufficiently large positive integers s if k is aperiodic. See the appendix of [3] for the requisite number theory.)

2. Show that Kolmogorov's criterion (3) implies that definition (4) does not depend on the particular path chosen from i to j.

3. In the Bernoulli–Laplace model, we imagine two boxes with m particles each. Among the $2m$ particles there are b black particles and w white particles, where $b + w = 2m$ and $b \leq w$. At each epoch one particle is randomly selected from each box, and the two particles are exchanged. Let Z_n be the number of black particles in the first box.

Is the corresponding chain irreducible, aperiodic, or reversible? Show that its equilibrium distribution is hypergeometric.

4. In Example 23.2.1, show that the chain is aperiodic if and only if the underlying graph is not bipartite.

5. Let P be the transition matrix and π the equilibrium distribution of a reversible Markov chain with n states. Define an inner product $\langle u, v\rangle_\pi$ on complex column vectors u and v with n components by

$$\langle u, v\rangle_\pi = \sum_i u_i \pi_i v_i^*.$$

Verify that P satisfies the self-adjointness condition

$$\langle Pu, v\rangle_\pi = \langle u, Pv\rangle_\pi,$$

and conclude by standard arguments that P has only real eigenvalues. Formulate a similar result for a reversible continuous-time chain.

6. Let $Z_0, Z_1, Z_2, \ldots$ be a realization of an ergodic chain. If we sample every kth epoch, then show (a) that the sampled chain $Z_0, Z_k, Z_{2k}, \ldots$ is ergodic, (b) that it possesses the same equilibrium distribution as the original chain, and (c) that it is reversible if the original chain is. Thus, we can estimate theoretical means by sample averages using only every kth epoch of the original chain.

7. A Markov chain is said to be embedded in a base Markov chain if there exists a map $f : C^* \to C$ from the state space C^* of the base chain onto the state space C of the embedded chain [15]. This map partitions the states of C^* into equivalence classes under the equivalence relation $x \sim y$ when $f(x) = f(y)$. If $Q = (q_{uv})$ denotes the matrix of transition probabilities of the base chain, then it is natural to define the transition probabilities of the embedded chain by

$$p_{f(u)f(v)} = \sum_{w \sim v} q_{uw}.$$

For the embedding to be probabilistically consistent, it is necessary that

$$\sum_{w \sim v} q_{uw} = \sum_{w \sim v} q_{xw} \tag{19}$$

for all $x \sim u$. A distribution ν on the base chain induces a distribution μ on the embedded chain according to

$$\mu_{f(u)} = \sum_{w \sim u} \nu_w. \tag{20}$$

Mindful of these conventions, show that the embedded Markov chain is irreducible if the base Markov chain is irreducible and is aperiodic if the base chain is aperiodic. If the base chain is reversible with stationary distribution ν, then show that the embedded chain is reversible with induced stationary distribution μ given by (20).

8. Suppose that Λ is the infinitesimal transition matrix of a continuous-time Markov chain, and let $\mu \geq \max_i \lambda_i$. If $R = I + \mu^{-1}\Lambda$, then prove that R has nonnegative entries and that

$$S(t) = \sum_{i=0}^{\infty} e^{-\mu t} \frac{(\mu t)^i}{i!} R^i$$

coincides with $P(t)$. (Hint: Verify that $S(t)$ satisfies the same defining differential equation and the same initial condition as $P(t)$.)

9. Consider a continuous-time Markov chain with infinitesimal transition matrix Λ and equilibrium distribution π. If the chain is at equilibrium at time 0, then show that it experiences $t\sum_i \pi_i \lambda_i$ transitions on average during the time interval $[0, t]$, where $\lambda_i = \sum_{j \neq i} \lambda_{ij}$ and λ_{ij} denotes a typical off-diagonal entry of Λ.

10. Verify the inequalities

$$\|e^{tA}\| \leq e^{|t| \cdot \|A\|}$$
$$\|e^{-tA}\| \geq e^{-|t| \cdot \|A\|}$$

for any square matrix A and matrix norm induced by a vector norm.

11. Derive the error estimate

$$\|e^{tA} - \sum_{i=0}^{n} \frac{1}{i!}(tA)^i\| \leq \frac{|t|^{n+1}\|A\|^{n+1}}{(n+1)!} \frac{1}{1 - \frac{|t| \cdot \|A\|}{n+2}}$$

for any square matrix A and matrix norm induced by a vector norm.

12. Consider the partial sums $S_n = \sum_{i=0}^{n} B^i/i!$ for some square matrix B. Show that $BS_n = S_n B$ and that for any $\epsilon > 0$

$$\|e^B - S_n\| < \epsilon$$
$$\|S_n\| \leq \|e^B\|\Big(1 + \frac{\epsilon}{\|e^B\|}\Big),$$

provided n is large enough and the indicated matrix norm is induced by a vector norm. If n is chosen to satisfy the last two inequalities, then show that

$$\|e^{2^k B} - S_n^{2^k}\| \leq \epsilon \|e^B\|^{2^k - 1}\Big(2 + \frac{\epsilon}{\|e^B\|}\Big)^{2^k - 1}$$

for any positive integer k. In conjunction with Problem 11, conclude that we can approximate e^{tA} arbitrarily closely by $S_n^{2^k}$ for $B = 2^{-k}tA$ and n sufficiently large. Hint:

$$\|e^{2^k B} - S_n^{2^k}\| \leq \|e^{2^{k-1}B} + S_n^{2^{k-1}}\| \cdot \|e^{2^{k-1}B} - S_n^{2^{k-1}}\|.$$

13. Let A and B be the 2×2 real matrices

$$A = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \qquad B = \begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix}.$$

Show that

$$e^A = e^a \begin{pmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{pmatrix}, \qquad e^B = e^\lambda \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

(Hints: Note that 2×2 matrices of the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ are isomorphic to the complex numbers under the correspondence $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \leftrightarrow a+bi$. For the second case write $B = \lambda I + C$.)

14. Prove that $\det(e^A) = e^{\mathrm{tr}(A)}$, where tr is the trace function. (Hint: Since the diagonalizable matrices are dense in the set of matrices [9], by continuity you may assume that A is diagonalizable.)

References

[1] Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* 3:1–8

[2] Bhattacharya RN, Waymire EC (1990) *Stochastic Processes with Applications.* Wiley, New York

[3] Billingsley P (1986) *Probability and Measure,* 2nd ed. Wiley, New York

[4] Devijver PA (1985) Baum's forward-backward algorithm revisited. *Pattern Recognition Letters* 3:369–373

[5] Doyle PG, Snell JL (1984) *Random Walks and Electrical Networks.* The Mathematical Association of America

[6] Feller W (1968) *An Introduction to Probability Theory and its Applications, Vol 1*, 3rd ed. Wiley, New York

[7] Fredkin DR, Rice JA (1992) Bayesian restoration of single-channel patch clamp recordings. *Biometrics* 48:427–448

[8] Grimmett GR, Stirzaker DR (1992) *Probability and Random Processes*, 2nd ed. Oxford University Press, Oxford

[9] Hirsch MW, Smale S (1974) *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic Press, New York

[10] Karlin S, Taylor HM (1975) *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York

[11] Karlin S, Taylor HM (1981) *A Second Course in Stochastic Processes.* Academic Press, New York

[12] Kelly FP (1979) *Reversibility and Stochastic Networks.* Wiley, New York

[13] Lamperti J (1977) *Stochastic Processes. A Survey of the Mathematical Theory.* Springer-Verlag, New York

[14] Lange K, Boehnke M, Cox DR, Lunetta KL (1995) Statistical methods for polyploid radiation hybrid mapping. *Genome Res* 5:136–150

[15] Lazzeroni LC, Lange K (1997) Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Ann Stat* 25:138–168

[16] Moler C, Van Loan C (1978) Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review* 20:801–836

[17] Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77:257–285
[18] Rubinow SI (1975) *Introduction to Mathematical Biology.* Wiley, New York
[19] Waterman MS (1995) *Introduction to Computational Biology: Maps, Sequences, and Genomes.* Chapman & Hall, London

24

Markov Chain Monte Carlo

24.1 Introduction

The Markov chain Monte Carlo (MCMC) revolution sweeping statistics is drastically changing how statisticians perform integration and summation. In particular, the Metropolis algorithm and Gibbs sampling make it straightforward to construct a Markov chain that samples from a complicated conditional distribution. Once a sample is available, then any conditional expectation can be approximated by forming its corresponding sample average. The implications of this insight are profound for both classical and Bayesian statistics. As a bonus, trivial changes to the Metropolis algorithm yield simulated annealing, a general-purpose algorithm for solving difficult combinatorial optimization problems.

Our limited goal in this chapter is to introduce a few of the major MCMC themes. In describing the theory we will use the notation of discrete-time Markov chains. Readers should bear in mind that most of the theory carries over to chains with continuous-state spaces; our examples exploit this fact. One issue of paramount importance is how rapidly the underlying chains reach equilibrium. This is the Achilles heel of the whole business and not just a mathematical nicety. Unfortunately, probing this delicate issue is scarcely possible in the confines of a single chapter. We analyze one example to give readers a feel for the power of coupling arguments.

Readers interested in pursuing MCMC methods and simulated annealing further will enjoy the pioneering articles [5, 8, 12, 14, 16]. The elementary surveys [2, 3] of Gibbs sampling and the Metropolis algorithm are quite

readable, as are the books [6, 11, 20]. The mathematical surveys [10, 19] provide good starting points in understanding the rate of convergence of MCMC methods. Computer implementation of simulated annealing can be found in [18].

24.2 The Hastings–Metropolis Algorithm

The Hastings–Metropolis algorithm is a device for constructing a Markov chain with a prescribed equilibrium distribution π on a given state space [12, 16]. Each step of the chain is broken into two stages, a proposal stage and an acceptance stage. If the chain is currently in state i, then in the proposal stage a new destination state j is proposed according to a probability density $q_{ij} = q(j \mid i)$. In the subsequent acceptance stage, a random number is drawn uniformly from $[0, 1]$ to determine whether the proposed step is actually taken. If this number is less than the Hastings–Metropolis acceptance probability

$$a_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}, \tag{1}$$

then the proposed step is taken. Otherwise, the proposed step is declined, and the chain remains in place.

A few comments about this strange procedure are in order. First, the resemblance of the Hastings–Metropolis algorithm to acceptance–rejection sampling should make the reader more comfortable. Second, like most good ideas, the algorithm has gone through successive stages of abstraction and generalization. For instance, Metropolis et al. [16] considered only symmetric proposal densities with $q_{ij} = q_{ji}$. In this case the acceptance probability reduces to

$$a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\}. \tag{2}$$

In this simpler setting it is clear that any proposed destination j with $\pi_j > \pi_i$ is automatically accepted. Finally, in applying either formula (1) or formula (2), it is noteworthy that the π_i need only be known up to a multiplicative constant.

To prove that π is the equilibrium distribution of the chain constructed from the Hastings–Metropolis scheme (1), it suffices to check that detailed balance holds. If π puts positive weight on all points of the state space, it is clear that we must impose the requirement that the inequalities $q_{ij} > 0$ and $q_{ji} > 0$ are simultaneously true or simultaneously false. This requirement is also implicit in definition (1). Now suppose without loss of generality that the fraction

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \leq 1$$

for some $j \neq i$. Then detailed balance follows immediately from

$$\begin{aligned}\pi_i q_{ij} a_{ij} &= \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \\ &= \pi_j q_{ji} \\ &= \pi_j q_{ji} a_{ji}.\end{aligned}$$

Besides checking that π is the equilibrium distribution, we should also be concerned about whether the Hastings–Metropolis chain is irreducible and aperiodic. Aperiodicity is the rule because the acceptance–rejection step allows the chain to remain in place. Problem 2 states a precise result and a counterexample. Irreducibility holds provided the entries of π are positive and the proposal matrix $Q = (q_{ij})$ is irreducible.

24.3 Gibbs Sampling

The Gibbs sampler is a special case of the Hastings–Metropolis algorithm for Cartesian product state spaces [5, 8, 11]. Suppose that each sample point $i = (i_1, \ldots, i_m)$ has m components. The Gibbs sampler updates one component of i at a time. If the component is chosen randomly and resampled conditional on the remaining components, then the acceptance probability is 1. To prove this assertion, let i_c be the uniformly chosen component, and denote the remaining components by $i_{-c} = (i_1, \ldots, i_{c-1}, i_{c+1}, \ldots, i_m)$. If j is a neighbor of i reachable by changing only component i_c, then $j_{-c} = i_{-c}$. For such a neighbor j, the proposal probability

$$q_{ij} = \frac{1}{m} \cdot \frac{\pi_j}{\sum_{\{k:k_{-c}=i_{-c}\}} \pi_k}$$

satisfies $\pi_i q_{ij} = \pi_j q_{ji}$, and the ratio appearing in the acceptance probability (1) is 1.

In contrast to random sampling of components, we can repeatedly cycle through the components in some fixed order, say $1, 2, \ldots, m$. If the transition matrix for changing component c while leaving other components unaltered is $P^{(c)}$, then the transition matrices for random sampling and sequential (or cyclic) sampling are $R = \frac{1}{m}\sum_c P^{(c)}$ and $S = P^{(1)} \cdots P^{(m)}$, respectively. Because each $P^{(c)}$ satisfies $\pi P^{(c)} = \pi$, we have $\pi R = \pi$ and $\pi S = \pi$ as well. Thus, π is the unique equilibrium distribution for R or S if either is irreducible. However, as pointed out in Problem 3, R satisfies detailed balance while S ordinarily does not.

Example 24.3.1 (*Ising Model*). Consider m elementary particles equally spaced around the boundary of the unit circle. Each particle c can be in one of two magnetic states—spin up with $i_c = 1$ or spin down with $i_c = -1$. The Gibbs distribution

$$\pi_i \propto e^{\beta \sum_d i_d i_{d+1}} \tag{3}$$

takes into account nearest-neighbor interactions in the sense that states like $(1, 1, 1, \ldots, 1, 1, 1)$ are favored and states like $(1, -1, 1, \ldots, 1, -1, 1)$ are shunned for $\beta > 0$. (Note that in (3) the index $m+1$ of i_{m+1} is interpreted as 1.) Specification of the normalizing constant (or partition function)

$$Z = \sum_i e^{\beta \sum_d i_d i_{d+1}}$$

is unnecessary to carry out Gibbs sampling. If we elect to resample component c, then the choices $j_c = -i_c$ and $j_c = i_c$ are made with respective probabilities

$$\frac{e^{\beta(-i_{c-1}i_c - i_c i_{c+1})}}{e^{\beta(i_{c-1}i_c + i_c i_{c+1})} + e^{\beta(-i_{c-1}i_c - i_c i_{c+1})}} = \frac{1}{e^{2\beta(i_{c-1}i_c + i_c i_{c+1})} + 1}$$
$$\frac{e^{\beta(i_{c-1}i_c + i_c i_{c+1})}}{e^{\beta(i_{c-1}i_c + i_c i_{c+1})} + e^{\beta(-i_{c-1}i_c - i_c i_{c+1})}} = \frac{1}{1 + e^{-2\beta(i_{c-1}i_c + i_c i_{c+1})}}.$$

When the number of particles m is even, the odd-numbered particles are independent given the even-numbered particles, and vice versa. This fact suggests alternating between resampling all odd-numbered particles and resampling all even-numbered particles. Such multi-particle updates take longer to execute but create more radical rearrangements than single-particle updates. ■

Example 24.3.2 (*A Normal Random Sample with Conjugate Priors***).** Consider a random sample $y = (y_1, \ldots, y_n)$ from a normal density with mean μ and variance τ^{-1}. Suppose that μ is subject to a normal prior with mean 0 and variance ω^{-1} and τ is subject to a gamma prior with shape parameter α and scale parameter β. Given that the two priors are independent, the joint density of data and parameters is

$$(2\pi)^{-\frac{n+1}{2}} \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{i=1}^n (y_i - \mu)^2} \omega^{\frac{1}{2}} e^{-\frac{\omega}{2}\mu^2} \frac{\tau^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-\frac{\tau}{\beta}}.$$

Gibbs sampling from the joint posterior distribution of μ and τ requires the conditional density of μ given y and τ and the conditional density of τ given y and μ. The first of these two conditional densities is normally distributed with mean $n\tau\bar{y}/(\omega + n\tau)$ and variance $1/(\omega + n\tau)$, where $\bar{y}$ is the sample mean $\frac{1}{n}\sum_{i=1}^n y_i$. The second is gamma distributed with shape parameter $n/2 + \alpha$ and scale parameter $1/(ns_n^2/2 + 1/\beta)$, where s_n^2 is the sample variance $\frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2$. Choosing conjugate priors here eases the analysis as it does throughout Bayesian statistics. ■

Example 24.3.3 (*Data Augmentation and Allele Frequency Estimation***).** Data augmentation uses missing data in a slightly different fashion than the EM algorithm [20, 21]. In data augmentation we sample from the joint conditional distribution of the missing data and the parameters given the observed data. For example, consider the ABO allele frequency estimation problem of Chapter 10. If we put a Dirichlet prior with parameters

$(\alpha_A, \alpha_B, \alpha_O)$ on the allele frequencies, then the joint density of the observed data and the parameters amounts to

$$\binom{n}{n_A\ n_B\ n_{AB}\ n_O}(p_A^2 + 2p_Ap_O)^{n_A}(p_B^2 + 2p_Bp_O)^{n_B}(2p_Ap_B)^{n_{AB}}(p_O^2)^{n_O} \times \frac{\Gamma(\alpha_A + \alpha_B + \alpha_O)}{\Gamma(\alpha_A)\Gamma(\alpha_B)\Gamma(\alpha_O)} p_A^{\alpha_A - 1} p_B^{\alpha_B - 1} p_O^{\alpha_O - 1}.$$

Extracting the posterior density of the allele frequencies p_A, p_B, and p_O appears formidable. However, sampling from the posterior distribution becomes straightforward if we augment the observed data by specifying the underlying counts $n_{A/A}$, $n_{A/O}$, $n_{B/B}$, and $n_{B/O}$ of individuals with genotypes A/A, A/O, B/B, and B/O, respectively.

Sequential Gibbs sampling alternates between sampling the complete data $(n_{A/A}, n_{A/O}, n_{B/B}, n_{B/O}, n_{AB}, n_O)$ conditional on the observed data (n_A, n_B, n_{AB}, n_O) and the parameters (p_A, p_B, p_O) and sampling the parameters (p_A, p_B, p_O) conditional on the complete data. The marginal distribution of the parameters from a sequential Gibbs sample coincides with the posterior distribution. To sample $n_{A/A}$, we simply draw from a binomial distribution with n_A trials and success probability $p_A^2/(p_A^2+2p_Ap_O)$. The complementary variable $n_{A/O}$ is determined by the linear constraint $n_{A/A} + n_{A/O} = n_A$. Sampling $n_{B/B}$ and $n_{B/O}$ is done similarly. Sampling the parameters (p_A, p_B, p_O) conditional on the complete data is accomplished by sampling from a Dirichlet distribution with parameters

$$\begin{aligned} \alpha_A + 2n_{A/A} + n_{A/O} + n_{AB}, \\ \alpha_B + 2n_{B/B} + n_{B/O} + n_{AB}, \\ \alpha_O + n_{A/O} + n_{B/O} + 2n_O. \end{aligned}$$

Again choosing a conjugate prior is critical in keeping the sampling process simple. ■

24.4 Other Examples of Hastings–Metropolis Sampling

Although constructing good proposal densities is an art, some general techniques are worth mentioning [3, 22]. In the following list, our use of discrete notation should not obscure the fact that the methods have wider applicability.

Example 24.4.1 (*Compatible Proposal Densities*). In randomized Gibbs sampling the proposal density q_{ij} already satisfies detailed balance relative to the equilibrium distribution π. Therefore, no Metropolis adjustment is necessary. This enviable situation is hardly unique to Gibbs sampling. For

instance, consider the $n!$ different permutations

$$\sigma = (\sigma_1, \ldots, \sigma_n)$$

of the set $\{1, \ldots, n\}$ equipped with the uniform distribution $\pi_\sigma = 1/n!$ [4]. Declare a permutation ω to be a neighbor of σ if there exist two indices $i \neq j$ such that $\omega_i = \sigma_j$, $\omega_j = \sigma_i$, and $\omega_k = \sigma_k$ for $k \notin \{i, j\}$. Evidently, each permutation has $\binom{n}{2}$ neighbors. If we put $q_{\sigma\omega} = 1/\binom{n}{2}$ for each neighbor ω of σ, then $q_{\sigma\omega}$ satisfies detailed balance. Thus, randomly choosing a pair of indices $i \neq j$ and switching σ_i with σ_j produces a Markov chain on the set of permutations. This chain has period 2. It can be made aperiodic by allowing the randomly chosen indices i and j to coincide.

Hastings [12] suggests a continuous generalization of this chain that generates random $n \times n$ orthogonal matrices R with $\det(R) = 1$. These multidimensional rotations form a compact subgroup of the set of all $n \times n$ matrices and consequently possess an invariant Haar probability measure [17]. The proposal stage of Hastings' algorithm consists of choosing at random two indices $i \neq j$ and an angle $\theta \in [-\pi, \pi]$. The proposed replacement for the current rotation matrix R is then $S = E_{ij}(\theta)R$, where the matrix $E_{ij}(\theta)$ is a rotation in the (i, j) plane through angle θ. Note that $E_{ij}(\theta) = (e_{kl})$ coincides with the identity matrix except for entries $e_{ii} = e_{jj} = \cos\theta$, $e_{ij} = \sin\theta$, and $e_{ji} = -\sin\theta$. By virtue of the fact $E_{ij}(\theta)^{-1} = E_{ij}(-\theta)$, the proposal density is symmetric, and the Markov chain induced on the set of multidimensional rotations is reversible with respect to Haar measure. Thus, a proposed S is always accepted by the Metropolis mechanism. Problem 8 shows that this Markov chain is irreducible. ■

Example 24.4.2 (*Random Walk*)**.** Random walk sampling occurs when the proposal density $q_{ij} = q_{j-i}$ for some density q_k. This construction requires that the sample space be closed under subtraction. If $q_k = q_{-k}$, then the Metropolis acceptance probability (2) applies. ■

Example 24.4.3 (*Autoregressive Chains*)**.** This method is intended for continuous-state spaces. Given the current point x, a candidate point y is chosen by taking $y = a + B(x - a) + z$, where a is a fixed vector, B is a compatible fixed matrix, and z is sampled from some density $q(z)$. When B is the identity matrix I, we recover the random walk method. ■

Example 24.4.4 (*Independence Sampler*)**.** For the particular proposal density $q_{ij} = q_j$, candidate points are drawn independently of the current point. Ideally q_i should be close to π_i for most i. Furthermore, the ratio q_i/π_i should not be allowed to become too small. Indeed, if it is exceptionally small for a given state i, then it is exceptionally hard to exit i. ■

Example 24.4.5 (*Acceptance–Rejection with a Pseudo-Dominating Density*)**.** In the acceptance–rejection method of random sampling, it is easy to sample from a dominating density $g(y)$ and hard to sample from a tar-

get density $f(y)$. Suppose we have reason to believe that $f(y) \leq cg(y)$ with high probability for some constant c. We do not exclude the possibility of $f(y) > cg(y)$ happening with small probability, but we do require that $f(y)$ and $g(y)$ have the same support, and we accept a point y sampled from $g(y)$ with probability $\min\{f(y)/[cg(y)], 1\}$. Tierney [22] has noted that subjecting the accepted y to a second round of acceptance–rejection with the correct acceptance probability yields a reversible Markov chain with $f(x)$ as its equilibrium distribution. Both rounds of rejection can be summarized by the single Hastings–Metropolis acceptance probability

$$a(x,y) = \min\left\{\frac{f(y)}{cg(y)}, 1\right\} \times \min\left\{\min\left\{\frac{cg(x)}{f(x)}, 1\right\} \times \max\left\{\frac{f(y)}{cg(y)}, 1\right\}, 1\right\}$$

for points y sampled from $g(y)$ when the chain is currently in state x.

To check the detailed balance condition

$$f(x)g(y)a(x,y) = f(y)g(x)a(y,x) \tag{4}$$

for all $x \neq y$, we consider four different cases. In case 1, $f(x) \leq cg(x)$ and $f(y) \leq cg(y)$. In this instance, both sides of equation (4) reduce to $f(x)f(y)/c$. In case 2, $f(x) > cg(x)$ and $f(y) \leq cg(y)$, and both sides of equation (4) become $f(y)g(x)$. Case 3 with the opposite inequalities $f(x) \leq cg(x)$ and $f(y) > cg(y)$ is handled similarly. Finally in case 4, $f(x) > cg(x)$ and $f(y) > cg(y)$, and equation (4) reduces to

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

which is obviously true. ■

24.5 Some Practical Advice

Virtually no commercial software exists for carrying out MCMC methods. Crafting good algorithms and writing software to implement them is time consuming and error prone. Our ignorance about the rates of convergence of most chains is also worrisome. Nonetheless, the range of problems that can be solved by MCMC methods is so impressive that most statisticians are willing to invest the programming time and take the risks. Given the difficulty and potential for abuse of MCMC methods, we offer the following advice:

(a) Every chain must start somewhere. In sampling from posterior densities there are three obvious possibilities. We can set initial parameter values equal to frequentist estimates, to sampled values from their corresponding priors, or to means or medians of their corresponding priors.

(b) In choosing priors for sequential Gibbs sampling, it is clearly advantageous to select independent, conjugate priors. This makes sampling from the various conditional densities straightforward.

(c) As Example 24.3.3 shows, data augmentation can render sampling much easier. Whenever the EM algorithm works well in the frequentist version of a problem, data augmentation is apt to help in sampling from the posterior density in a Bayesian version of the same problem.

(d) Calculation of sample averages from a Markov chain should not commence immediately. Every chain needs a burn-in period to reach equilibrium.

(e) Not every epoch need be taken into account in forming a sample average of a complicated statistic. This is particularly true if the chain reaches equilibrium slowly; in such circumstances, values from neighboring states are typically highly correlated. Problem 6 of Chapter 23 validates the procedure of sampling a statistic at every kth epoch of a chain.

(f) Just as with independent sampling, we can achieve variance reduction by replacing a sampled statistic by its conditional expectation. Suppose, for instance, in Example 24.3.3 that we want to find the posterior mean of the number $n_{A/A}$ of people of genotype A/A. If we run the chain m epochs after burn-in, then we can estimate the posterior mean by the sample average $\frac{1}{m}\sum_{i=1}^{m} n_{A/A}^{(i)}$. However, the estimator

$$\frac{1}{m}\sum_{i=1}^{m} n_A \frac{p_A^{(i)\,2}}{p_A^{(i)\,2} + 2p_A^{(i)}p_O^{(i)}}$$

is apt to have smaller variance since we have eliminated the noise introduced by sampling $n_{A/A}^{(i)}$ at epoch i.

(g) Undiagnosed slow convergence can lead to grievous errors in statistical inference. Gelman and Rubin suggest some monitoring techniques that are applicable if multiple, independent realizations of a chain are available [7]. Widely dispersed starting states for the independent chains are particularly helpful in revealing slow convergence. If a chain is known to converge rapidly, multiple independent runs are no better than a single long chain in computing expectations.

(h) For chains that converge slowly, importance sampling and running parallel, coupled chains offer speedups. Problems 10 and 11 briefly explain these techniques.

24.6 Convergence of the Independence Sampler

For the independence sampler, it is possible to give a coupling bound on the rate of convergence to equilibrium [15]. Suppose that $X_0, X_1, \ldots$ represents

the sequence of states visited by the independence sampler starting from $X_0 = x_0$. We couple this Markov chain to a second independence sampler $Y_0, Y_1, \ldots$ starting from the equilibrium distribution π. By definition, each Y_k has distribution π. The two chains are coupled by a common proposal stage and a common uniform deviate U sampled in deciding whether to accept the common proposed point. Once $X_k = Y_k$ for some k, then $X_l = Y_l$ for all $l \geq k$. Let T denote the random epoch when X_k first meets Y_k and the X chain attains equilibrium.

The importance ratios $w_j = \pi_j/q_j$ determine what proposed points are accepted. Without loss of generality, assume that the states of the chain are numbered $1, \ldots, n$ and that $w_1 \geq w_j$ for all j. If $X_k = x \neq y = Y_k$, then according to equation (1) the next proposed point is accepted by both chains with probability

$$\begin{aligned}\sum_{j=1}^n q_j \min\left\{\frac{w_j}{w_x}, \frac{w_j}{w_y}, 1\right\} &= \sum_{j=1}^n \pi_j \min\left\{\frac{1}{w_x}, \frac{1}{w_y}, \frac{1}{w_j}\right\} \\ &\geq \frac{1}{w_1}.\end{aligned}$$

In other words, at each trial the two chains meet with at least probability $1/w_1$. This translates into the tail probability bound

$$\Pr(T > k) \leq \left(1 - \frac{1}{w_1}\right)^k. \tag{5}$$

Now let $\pi^{(k)}$ denote the distribution of X_k. To exploit the bound (5), we introduce the total variation norm [4]

$$\begin{aligned}\|\pi^{(k)} - \pi\|_{TV} &= \sup_{A \subset \{1,\ldots,n\}} |\Pr(X_k \in A) - \pi(A)| \\ &= \frac{1}{2}\sum_{i=1}^n |\Pr(X_k = i) - \pi_i|.\end{aligned} \tag{6}$$

In view of the fact that X_k is at equilibrium when $T \leq k$, we have

$$\begin{aligned}&\pi^{(k)}(A) \\ &\quad = \Pr(X_k \in A | T \leq k) \Pr(T \leq k) + \Pr(X_k \in A | T > k) \Pr(T > k) \\ &\quad = \pi(A) \Pr(T \leq k) + \Pr(X_k \in A | T > k) \Pr(T > k).\end{aligned}$$

This implies the bound

$$\begin{aligned}\|\pi^{(k)} - \pi\|_{TV} &\leq \Pr(T > k) \\ &\leq \left(1 - \frac{1}{w_1}\right)^k\end{aligned}$$

on the total variation distance of X_k from equilibrium.

24.7 Simulated Annealing

In simulated annealing we are interested in finding the most probable state of a Markov chain [14, 18]. If this state is k, then $\pi_k > \pi_i$ for all $i \neq k$. To accentuate the weight given to state k, we can replace the equilibrium distribution π by a distribution putting probability

$$\pi_i^{(\tau)} = \frac{\pi_i^{\frac{1}{\tau}}}{\sum_j \pi_j^{\frac{1}{\tau}}}$$

on state i. Here τ is a small, positive parameter traditionally called temperature. With a symmetric proposal density, the distribution $\pi_i^{(\tau)}$ can be attained by running a chain with Metropolis acceptance probability

$$a_{ij} = \min\left\{\left(\frac{\pi_j}{\pi_i}\right)^{\frac{1}{\tau}}, 1\right\}. \tag{7}$$

In fact, what is done in simulated annealing is that the chain is run with τ gradually decreasing to 0. If τ starts out large, then in the early steps of simulated annealing, almost all proposed steps are accepted, and the chain broadly samples the state space. As τ declines, fewer unfavorable steps are taken, and the chain eventually settles on some nearly optimal state. With luck, this state is k or a state equivalent to k if several states are optimal. Simulated annealing is designed to mimic the gradual freezing of a substance into a crystalline state of perfect symmetry and hence minimum energy.

Example 24.7.1 (*The Traveling Salesman Problem*). A salesman must visit n towns, starting and ending in his hometown. Given the distance d_{ij} between every pair of towns i and j, in what order should he visit the towns to minimize the length of his circuit? This problem belongs to the class of NP-complete problems; these have deterministic solutions that are conjectured to increase in complexity at an exponential rate in n.

In the simulated annealing approach to the traveling salesman problem, we assign to each permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$ a cost $c_\sigma = \sum_{i=1}^n d_{\sigma_i, \sigma_{i+1}}$, where $\sigma_{n+1} = \sigma_1$. Defining $\pi_\sigma \propto e^{-c_\sigma}$ turns the problem of minimizing the cost into one of finding the most probable permutation σ. In the proposal stage of simulated annealing, we randomly select two indices $i \neq j$ and reverse the block of integers beginning at σ_i and ending at σ_j in the current permutation $(\sigma_1, \ldots, \sigma_n)$. This proposal is accepted with probability (7) depending on the temperature τ. In *Numerical Recipes'* [18] simulated annealing algorithm for the traveling salesman problem, τ is lowered in multiplicative decrements of 10 percent after every $100n$ epochs or every $10n$ accepted steps, whichever comes first. ■

24.8 Problems

1. An acceptance function $a : [0, \infty] \mapsto [0, 1]$ satisfies the functional identity $a(x) = xa(1/x)$. Prove that the detailed balance condition

$$\pi_i q_{ij} a_{ij} = \pi_j q_{ji} a_{ji}$$

holds if the acceptance probability a_{ij} is defined by

$$a_{ij} = a\Big(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}\Big)$$

in terms of an acceptance function $a(x)$. Check that the Barker function $a(x) = x/(1 + x)$ qualifies as an acceptance function and that any acceptance function is dominated by the Metropolis acceptance function in the sense that $a(x) \leq \min\{x, 1\}$ for all x.

2. The Metropolis acceptance mechanism (2) ordinarily implies aperiodicity of the underlying Markov chain. Show that if the proposal distribution is symmetric and if some state i has a neighboring state j such that $\pi_i > \pi_j$, then the period of state i is 1, and the chain, if irreducible, is aperiodic. For a counterexample, assign probability $\pi_i = 1/4$ to each vertex i of a square. If the two vertices adjacent to a given vertex i are each proposed with probability $1/2$, then show that all proposed steps are accepted by the Metropolis criterion and that the chain is periodic with period 2.

3. Consider the Cartesian product space $\{0, 1\} \times \{0, 1\}$ equipped with the probability distribution

$$(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = \Big(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\Big).$$

Demonstrate that sequential Gibbs sampling does not satisfy detailed balance by showing that $\pi_{00} s_{00,11} \neq \pi_{11} s_{11,00}$, where $s_{00,11}$ and $s_{11,00}$ are entries of the matrix S for first resampling component one and then resampling component two.

4. In a Bayesian random effects model, m_i observations y_{ij} are drawn from a normal density with mean η_i and variance τ^{-1} for $i = 1, \ldots, n$. Impose a normal prior on η_i with mean μ and variance ω^{-1}, a normal prior on μ with mean 0 and variance δ^{-1}, and gamma priors on τ and ω with shape parameters α_τ and α_ω and scale parameters β_τ and β_ω, respectively. Assume independence among the y_{ij} given all model parameters, independence of the η_i given the hyperparameters μ, τ, and ω, and independence among the hyperparameters themselves. Calculate the conditional densities necessary to conduct a Gibbs sample from the posterior distribution of the parameter vector $(\eta_1, \ldots, \eta_n, \mu, \tau, \omega)$.

5. Carry out the Gibbs sampling procedure for the ABO allele frequency model described in Example 24.3.3 using the duodenal ulcer data of Chapter 10. Estimate the posterior medians, means, variances, and covariances of the allele frequency parameters.

6. Let $x_1, \ldots, x_n$ be observations from independent Poisson random variables with means $\lambda_1 t_1, \ldots, \lambda_n t_n$, where the t_i are known times and the λ_i are unknown intensities. Suppose that λ_i has gamma prior with shape parameter α and scale parameter β and that β has inverse gamma prior $\delta^\gamma e^{-\delta/\beta}/[\beta^{\gamma+1}\Gamma(\gamma)]$. Here α, δ, and γ are given, while the λ_i and β are parameters. If the priors on the λ_i are independent, what is the joint density of the data and the parameters? Design a sequential Gibbs scheme to sample from the posterior distribution of the parameters given the data.

7. If the component updated in the randomly sampled component version of Gibbs sampling depends probabilistically on the current state of the chain, how must the Hastings–Metropolis acceptance probability be modified to preserve detailed balance? Under the appropriate modification, the acceptance probability is no longer always 1.

8. Show that Hastings' Markov chain on multidimensional rotations is irreducible. (Hint: Prove that every multidimensional rotation R can be written as a finite product of matrices of the form $E_{ij}(\theta)$. Using a variation of Jacobi's method discussed in Chapter 8, argue inductively that you can zero out the off-diagonal entries of R in a finite number of multiplications by appropriate two-dimensional rotations $E_{ij}(\theta)$. The remaining diagonal entries all equal ± 1. There are an even number of -1 diagonal entries, and these can be converted to $+1$ diagonal entries in pairs.)

9. Write a program implementing the sampling method of Example 24.4.1. Use the program to estimate $\mathrm{E}(\sum_{i=1}^n R_{ii}^2)$ for a random rotation R. The exact value of this integral is 1.

10. Importance sampling is one remedy when the states of a Markov chain communicate poorly [12]. Suppose that π is the equilibrium distribution of the chain. If we sample from a chain whose distribution is ν, then we can recover approximate expectations with respect to π by taking weighted averages. In this scheme, the state z is given weight $w_z = \pi_z/\nu_z$. If $Z_0, Z_1, Z_2 \ldots$ is a run from the chain with equilibrium distribution ν, then under the appropriate ergodic assumptions prove that

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} w_{Z_i} f(Z_i)}{\sum_{i=0}^{n-1} w_{Z_i}} = \mathrm{E}_\pi[f(X)].$$

The choice $\nu_z \propto \pi_z^{1/\tau}$ for $\tau > 1$ lowers the peaks and raises the valleys of π [13]. Unfortunately, if ν differs too much from π, then the estimator

$$\frac{\sum_{i=0}^{n-1} w_{Z_i} f(Z_i)}{\sum_{i=0}^{n-1} w_{Z_i}}$$

of the expectation $\mathrm{E}_\pi[f(X)]$ will have a large variance for n of moderate size.

11. Another device to improve mixing of a Markov chain is to run several parallel chains on the same state space and occasionally swap their states [9]. If π is the distribution of the chain we wish to sample from, then let $\pi^{(1)} = \pi$, and define $m-1$ additional distributions $\pi^{(2)}, \ldots, \pi^{(m)}$. For instance, incremental heating can be achieved by taking

$$\pi_z^{(k)} \propto \pi_z^{\frac{1}{1+(k-1)\tau}}$$

for $\tau > 0$. At epoch n, we sample for each chain k a state Z_{nk} given the chain's previous state $Z_{n-1,k}$. We then randomly select chain i with probability $1/m$ and consider swapping states between it and chain $j = i+1$. (When $i = m$, no swap is performed.) Under appropriate ergodic assumptions on the m participating chains, show that if the acceptance probability for the proposed swap is

$$\min\left\{\frac{\pi_{Z_{nj}}^{(i)}\pi_{Z_{ni}}^{(j)}}{\pi_{Z_{ni}}^{(i)}\pi_{Z_{nj}}^{(j)}}, 1\right\},$$

then the product chain is ergodic with equilibrium distribution given by the product distribution $\pi^{(1)} \otimes \pi^{(2)} \otimes \cdots \otimes \pi^{(m)}$. The marginal distribution of this distribution for chain 1 is just π. Therefore, we can throw away the outcomes of chains 2 through m and estimate expectations with respect to π by forming sample averages from the embedded run of chain 1. (Hint: The fact that no swap is possible at each step allows the chains to run independently for an arbitrary number of steps.)

12. Demonstrate equality (6) for the total variation norm.

13. It is known that every planar graph can be colored by four colors [1]. Design, program, and test a simulated annealing algorithm to find a four coloring of any planar graph. (Suggestions: Represent the graph by a list of nodes and a list of edges. Assign to each node a color represented by a number between 1 and 4. The cost of a coloring is the number of edges with incident nodes of the same color. In the proposal stage of the simulated annealing solution, randomly choose a node, randomly reassign its color, and recalculate the cost. If successful, simulated annealing will find a coloring with the minimum cost of 0.)

References

[1] Brualdi RA (1977) *Introductory Combinatorics.* North-Holland, New York

[2] Casella G, George EI (1992) Explaining the Gibbs sampler. *Amer Statistician* 46:167–174

[3] Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. *Amer Statistician* 49:327–335
[4] Diaconis, P (1988) *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics, Hayward, CA
[5] Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *J Amer Stat Assoc* 85:398–409
[6] Gelman A, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis.* Chapman & Hall, London
[7] Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences (with discussion). *Stat Sci* 7:457–511
[8] Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Machine Intell* 6:721–741
[9] Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* Keramidas EM, editor, Interface Foundation, Fairfax, VA pp 156–163
[10] Gidas B (1995) Metropolis-type Monte Carlo simulation and simulated annealing. *Topics in Contemporary Probability and its Applications*, Snell JL, editor, CRC Press, Boca Raton, FL, pp 159–232
[11] Gilks WR, Richardson S, Spiegelhalter DJ (editors) (1996) *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London
[12] Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109
[13] Jennison C (1993) Discussion on the meeting of the Gibbs sampler and other Markov chain Monte Carlo methods. *J Roy Stat Soc B* 55:54–56
[14] Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
[15] Liu JS (1996) Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat and Computing* 6:113–119
[16] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Physics* 21:1087–1092
[17] Nachbin L (1965) *The Haar Integral,* Van Nostrand, Princeton, NJ
[18] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge
[19] Rosenthal JS (1995) Convergence rates of Markov chains. *SIAM Review* 37:387–405
[20] Tanner MA (1993) *Tools for Statistical Inference: Methods for Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed. Springer-Verlag, New York
[21] Tanner M, Wong W (1987) The calculation of posterior distributions (with discussion). *J Amer Stat Assoc* 82:528–550
[22] Tierney L (1994) Markov chains for exploring posterior distributions (with discussion). *Ann Stat* 22:1701–1762

Index