Dirk P. Kroese · Joshua C.C. Chan

Statistical Modeling and Computation

Springer

Statistical Modeling and Computation

Dirk P. Kroese • Joshua C.C. Chan

Statistical Modeling and Computation

Springer

Dirk P. Kroese
The University of Queensland
School of Mathematics and Physics
Brisbane, Australia

Joshua C.C. Chan
Department of Economics
Australian National University
Canberra, Australia

Printed on acid-free paper

In memory of Reuven Rubinstein, my Friend and Mentor

<div align="right">

Dirk Kroese

</div>

To Raquel

<div align="right">

Joshua Chan

</div>

Preface

Statistics provides one of the few principled means to extract information from random data and has perhaps more interdisciplinary connections than any other field of science. However, for a beginning student of statistics, the abundance of mathematical concepts, statistical philosophies, and numerical techniques can seem overwhelming. The purpose of this book is to provide a comprehensive and accessible introduction to modern statistics, illuminating its many facets, from both classical (frequentist) and Bayesian points of view. The book offers an integrated treatment of mathematical statistics and modern statistical computation.

The book is aimed at beginning students of statistics and practitioners who would like to fully understand the theory and key numerical techniques of statistics. It is based on a progression of undergraduate statistics courses at The University of Queensland and the Australian National University. Parts of the book have also been successfully tested at the University of New South Wales. Emphasis is laid on the mathematical and computational aspects of statistics. No prior knowledge of statistics is required, but we assume that the reader has a basic knowledge of mathematics, which forms an essential basis for the development of the statistical theory. Starting from scratch, the book gradually builds up to an advanced undergraduate level, providing a solid basis for possible postgraduate research. Throughout the text we illustrate the theory by providing working code in Matlab, rather than relying on black-box statistical packages. We make frequent use of the symbol ☞ in the margin to facilitate cross-referencing between related pages. The book is accompanied by the web site www.statmodcomp.org from which the Matlab code and data files can be downloaded. In addition, we provide an R equivalent for each Matlab program.

The book is structured into three parts. In Part I we introduce the fundamentals of probability theory. We discuss models for random experiments, conditional probability and independence, random variables, and probability distributions. Moreover, we explain how to carry out random experiments on a computer.

In Part II we introduce the general framework for statistical modeling and inference, from both classical and Bayesian perspectives. We discuss a variety of common models for data, such as independent random samples, linear regression,

and ANOVA models. Once a model for the data is determined one can carry out a mathematical analysis of the model on the basis of the available data. We discuss a wide range of concepts and techniques for statistical inference, including likelihood-based estimation and hypothesis testing, sufficiency, confidence intervals, and kernel density estimation. We encompass both classical and Bayesian approaches and also highlight popular Monte Carlo sampling techniques.

In Part III we address the statistical analysis and computation of a variety of advanced models, such as generalized linear models, autoregressive and moving average models, Gaussian models, and state space models. Particular attention is paid to fast numerical techniques for classical and Bayesian inference on these models. Throughout the book our leading principle is that the mathematical formulation of a statistical model goes hand in hand with the specification of its simulation counterpart.

The book contains a large number of illustrative examples and problem sets (with solutions). To keep the book fully self-contained, we include the more technical proofs and mathematical theory in Appendix B. Appendix A features a concise introduction to MATLAB.

Brisbane, Australia   Dirk Kroese
Canberra, Australia   Joshua Chan

Acknowledgements

Contents

Abbreviations and Acronyms

ANOVA   Analysis of variance
AR   Autoregressive
ARMA   Autoregressive-moving average
cdf   Cumulative distribution function
EM   Expectation–maximization
iid   Independent and identically distributed
pdf   Probability density function (discrete or continuous)
PGF   Probability generating function
KDE   Kernel density estimate/estimator
MA   Moving average
MCMC   Markov chain Monte Carlo
MGF   Moment generating function
ML(E)   Maximum likelihood (estimate/estimator)
PRESS   Predicted residual sum of squares

Mathematical Notation

Throughout this book we use notation in which different fonts and letter cases signify different types of mathematical objects. For example, vectors $\mathbf{a}, \mathbf{b}, \mathbf{x}, \ldots$ are written in lowercase boldface font and matrices A, B, X in uppercase normal font. Sans serif fonts indicate probability distributions, such as N, Exp, and Bin. Probability and expectation symbols are written in blackboard bold font: $\mathbb{P}$ and $\mathbb{E}$. MATLAB code and functions will always be written in `typewriter` font.

Traditionally, classical and Bayesian statistics use a *different* notation system for random variables and their probability density functions. In classical statistics and probability theory random variables usually are denoted by uppercase letters $X, Y, Z, \ldots$ and their outcomes by lowercase letters $x, y, z, \ldots$. Bayesian statisticians typically use lowercase letters for both. More importantly, in the Bayesian notation system, it is common to use the *same* letter f (or p) for different probability densities, as in $f(x, y) = f(x) f(y)$. Classical statisticians and probabilists would prefer a different symbol for each function, as in $f(x, y) = f_X(x) f_Y(y)$. We will predominantly use the classical notation, especially in the first part of the book. However, when dealing with Bayesian models and inference, such as in Chaps. 8 and 11, it will be convenient to switch to the Bayesian notation system. Here is a list of frequently used symbols:

$\approx$	Is approximately
$\propto$	Is proportional to
∞	Infinity
$\otimes$	Kronecker product
$\overset{\text{def}}{=}$	Is defined as
$\sim$	Is distributed as
$\overset{\text{iid}}{\sim}, \sim_{\text{iid}}$	Are independent and identically distributed as
$\overset{\text{approx.}}{\sim}$	Is approximately distributed as
$\mapsto$	Maps to
$A \cup B$	Union of sets A and B
$A \cap B$	Intersection of sets A and B
A^c	Complement of set A
$A \subset B$	A is a subset of B
$\emptyset$	Empty set
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
∇f	Gradient of f
$\nabla^2 f$	Hessian of f
$A^\top, \mathbf{x}^\top$	Transpose of matrix A or vector $\mathbf{x}$
$\text{diag}(\mathbf{a})$	Diagonal matrix with diagonal entries defined by $\mathbf{a}$
$\text{tr}(A)$	Trace of matrix A
$\det(A)$	Determinant of matrix A

$\mid A \mid$	Absolute value of the determinant of matrix A. Also, number of elements in set A or absolute value of real number A
argmax	argmax $f(x)$ is a value x^* for which $f(x^*) \geqslant f(x)$ for all x
d	Differential symbol
$\mathbb{E}$	Expectation
e	Euler's constant $\lim_{n \to \infty}(1 + 1/n)^n = 2.71828\ldots$
$I_A, I\{A\}$	Indicator function: equal to 1 if the condition/event A holds and 0 otherwise.
ln	(Natural) logarithm
$\mathbb{N}$	Set of natural numbers $\{0, 1, \ldots\}$
φ	Pdf of the standard normal distribution
Φ	Cdf of the standard normal distribution
$\mathbb{P}$	Probability measure
$\mathcal{O}$	Big-O order symbol: $f(x) = \mathcal{O}(g(x))$ if $\mid f(x) \mid \leqslant \alpha g(x)$ for some constant α as $x \to a$
o	Little-o order symbol: $f(x) = o(g(x))$ if $f(x)/g(x) \to 0$ as $x \to a$
$\mathbb{R}$	The real line = one-dimensional Euclidean space
$\mathbb{R}_+$	Positive real line: $[0, \infty)$
$\mathbb{R}^n$	n-Dimensional Euclidean space
$\widehat{\boldsymbol{\theta}}$	Estimate/estimator
$\mathbf{x}, \mathbf{y}$	Vectors
$\mathbf{X}, \mathbf{Y}$	Random vectors
$\mathbb{Z}$	Set of integers $\{\ldots, -1, 0, 1, \ldots\}$

Probability Distributions

Ber	Bernoulli distribution
Beta	Beta distribution
Bin	Binomial distribution
Cauchy	Cauchy distribution
χ^2	Chi-squared distribution
Dirichlet	Dirichlet distribution
DU	Discrete uniform distribution
Exp	Exponential distribution
F	F distribution
Gamma	Gamma distribution
Geom	Geometric distribution
InvGamma	Inverse-gamma distribution
Mnom	Multinomial distribution
N	Normal or Gaussian distribution
Poi	Poisson distribution
t	Student's t distribution
TN	Truncated normal distribution
U	Uniform distribution
Weib	Weibull distribution

Part I
Fundamentals of Probability

In Part I of the book we consider the *probability* side of statistics. In particular, we will consider how random experiments can be modeled mathematically and how such modeling enables us to compute various properties of interest for those experiments.

Chapter 1
Probability Models

1.1 Random Experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but which is nevertheless subject to analysis. Examples of random experiments are:

1. Tossing a die and observing its face value.
2. Measuring the amount of monthly rainfall in a certain location.
3. Counting the number of calls arriving at a telephone exchange during a fixed time period.
4. Selecting at random fifty people and observing the number of left-handers.
5. Choosing at random ten people and measuring their heights.

The goal of *probability* is to understand the behavior of random experiments by analyzing the corresponding *mathematical models*. Given a mathematical model for a random experiment one can calculate quantities of interest such as probabilities and expectations. Moreover, such mathematical models can typically be implemented on a computer, so that it becomes possible to *simulate* the experiment. Conversely, any computer implementation of a random experiment implicitly defines a mathematical model. Mathematical models for random experiments are also the basis of *statistics*, where the objective is to infer which of several competing models best fits the observed data. This often involves the estimation of model parameters from the data.

Example 1.1 (Coin Tossing). One of the most fundamental random experiments is the one where a coin is tossed a number of times. Indeed, much of probability theory can be based on this simple experiment. To better understand how this coin toss experiment behaves, we can carry it out on a computer, using programs such as MATLAB. The following simple MATLAB program simulates a sequence of 100 tosses with a fair coin (i.e., Heads and Tails are equally likely) and plots the results in a bar chart.

```
x = (rand(1,100) < 0.5)    % generate the coin tosses
bar(x)                     % plot the results in a bar chart
```

The function `rand` draws uniform random numbers from the interval $[0, 1]$—
in this case a 1×100 vector of such numbers. By testing whether the uni-
form numbers are less than 0.5, we obtain a vector x of 1s and 0s, indicating
Heads and Tails, say. Typical outcomes for three such experiments are given in
Fig. 1.1.
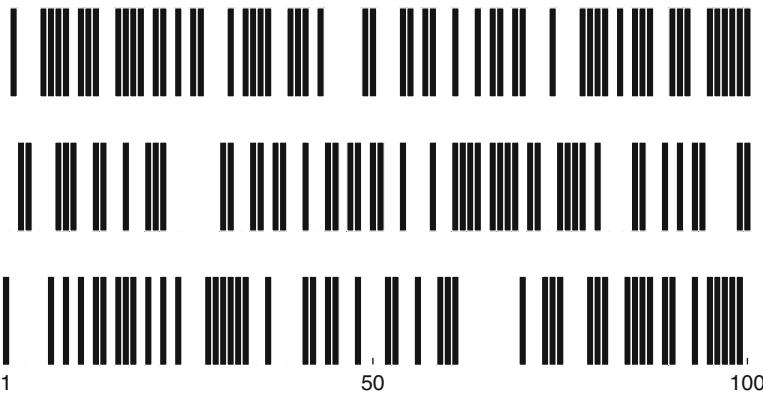
Fig. 1.1 Three experiments where a fair coin is tossed 100 times. The *dark bars* indicate when
"Heads" ($= 1$) appears

We can also plot the average number of Heads against the number of tosses. In
the same MATLAB program, this is accomplished by adding two lines of code:

```
y = cumsum(x)./[1:100] % calculate the cumulative sum and
              % divide this elementwise by the vector [1:100]
plot(y)       % plot the result in a line graph
```

The result of three such experiments is depicted in Fig. 1.2. Notice that the
average number of Heads seems to converge to 0.5, but there is a lot of random
fluctuation.

Similar results can be obtained for the case where the coin is *biased*,
with a probability of Heads of p, say. Here are some typical *probability*
questions.

- What is the probability of x Heads in 100 tosses?
- What is the expected number of Heads?
- How long does one have to wait until the first Head is tossed?
- How fast does the average number of Heads converge to p?
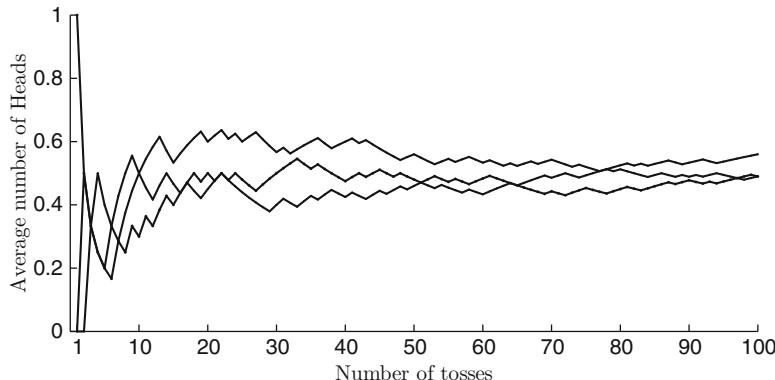
Fig. 1.2 The average number of Heads in n tosses, where $n = 1, \ldots, 100$

A statistical analysis would start from observed data of the experiment—for example, all the outcomes of 100 tosses are known. Suppose the probability of Heads p is not known. Typical *statistics* questions are:

- Is the coin fair?
- How can p be best estimated from the data?
- How accurate/reliable would such an estimate be?

The mathematical models that are used to describe random experiments consist of three building blocks: a *sample space*, a set of *events*, and a *probability*. We will now describe each of these objects.

1.2 Sample Space

Although we cannot predict the outcome of a random experiment with certainty, we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

Definition 1.1. (Sample Space). The **sample space** Ω of a random experiment is the set of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively and observe their face values:

$$\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), \ldots, (6, 6)\} .$$

2. Measure the lifetime of a machine in days:

$$\Omega = \mathbb{R}_+ = \{\text{ positive real numbers }\}.$$

3. Count the number of arriving calls at an exchange during a specified time interval:

$$\Omega = \{0, 1, \ldots\}.$$

4. Measure the heights of 10 people:

$$\Omega = \{(x_1, \ldots, x_{10}) : x_i \geqslant 0, i = 1, \ldots, 10\} = \mathbb{R}_+^{10}.$$

Here $(x_1, \ldots, x_{10})$ represents the outcome that the height of the first selected person is x_1, the height of the second person is x_2, and so on.

Notice that for modeling purposes it is often easier to take the sample space larger than is strictly necessary. For example, the actual lifetime of a machine would in reality not span the entire positive real axis, and the heights of the ten selected people would not exceed 9 ft.

1.3   Events

Often we are not interested in a single outcome but in whether or not one in a *group* of outcomes occurs.

> **Definition 1.2. (Event).** An **event** is a subset of the sample space Ω to which a probability can be assigned.

Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events are:

1. The event that the sum of two dice is 10 or more:

$$A = \{(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)\}.$$

2. The event that a machine is functioning for less than 1000 days:

$$A = [0, 1000).$$

3. The event that out of a group of 50 people 5 are left-handed:

$$A = \{5\}.$$

Example 1.2 (Coin Tossing). Suppose that a coin is tossed 3 times and that we record either Heads or Tails at every toss. The sample space can then be written as

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\},$$

where, for instance, HTH means that the first toss is Heads, the second Tails, and the third Heads. An alternative (but equivalent) sample space is the set $\{0, 1\}^3$ of binary vectors of length 3; for example, HTH corresponds to (1,0,1) and THH to (0,1,1).

The event A that the third toss is Heads is

$$A = \{\text{HHH, HTH, THH, TTH}\}.$$

Since events are sets, we can apply the usual set operations to them, as illustrated in the *Venn diagrams* in Fig. 1.3.

1. The set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur.
2. The set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur.
3. The event A^c (A **complement**) is the event that A does *not* occur.
4. If $B \subset A$ (B is a **subset** of A), then event B is said to *imply* event A.

| $A \cap B$ | $A \cup B$ | A^c | $B \subset A$ |
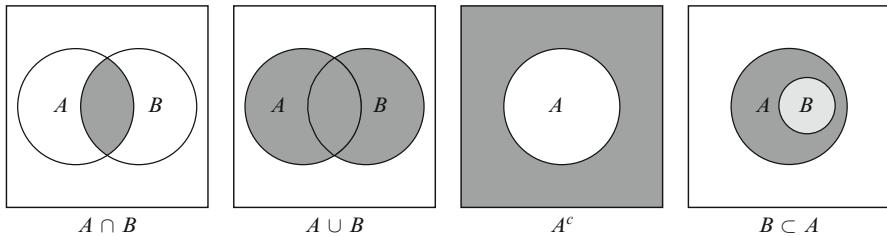
Fig. 1.3 Venn diagrams of set operations. Each *square* represents the sample space Ω

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$ (empty set), are called **disjoint** events.

Example 1.3 (Casting Two Dice). Suppose we cast two dice consecutively. The sample space is $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$. Let $A = \{(6,1),\ldots,(6,6)\}$ be the event that the first die is 6, and let $B = \{(1,6),\ldots,(6,6)\}$ be the event that the second die is 6. Then $A \cap B = \{(6,1),\ldots,(6,6)\} \cap \{(1,6),\ldots,(6,6)\} = \{(6,6)\}$ is the event that both dice are 6.

Example 1.4 (System Reliability). In Fig. 1.4 three systems are depicted, each consisting of three unreliable components. The *series* system works if all components work; the *parallel* system works if at least one of the components works; and the *2-out-of-3 system* works if at least 2 out of 3 components work.
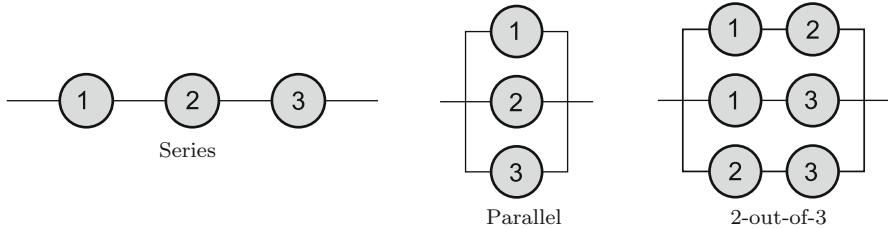
Fig. 1.4 Three unreliable systems

Let A_i be the event that the ith component is functioning, $i = 1, 2, 3$; and let D_a, D_b, D_c be the events that, respectively, the series, parallel, and 2-out-of-3 system are functioning. Then, $D_a = A_1 \cap A_2 \cap A_3$ and $D_b = A_1 \cup A_2 \cup A_3$. Also,

$$D_c = (A_1 \cap A_2 \cap A_3) \cup (A_1^c \cap A_2 \cap A_3) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c)$$

$$= (A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3) \,.$$

Two useful results in the theory of sets are the following, due to De Morgan:

Theorem 1.1. (De Morgan's Laws). If $\{A_i\}$ is a collection of sets, then

$$\left(\bigcup_i A_i \right)^c = \bigcap_i A_i^c \tag{1.1}$$

and

$$\left(\bigcap_i A_i \right)^c = \bigcup_i A_i^c \,. \tag{1.2}$$

Proof. If we interpret A_i as the event that component i works in Example 1.4, then the left-hand side of (1.1) is the event that the parallel system is not working. The right-hand side of (1.1) is the event that all components are not working. Clearly these two events are identical. The proof for (1.2) follows from a similar reasoning;
☞ 19   see also Problem 1.2.   □

1.4   Probability

The third ingredient in the model for a random experiment is the specification of
the probability of the events. It tells us how *likely* it is that a particular event will
occur.

Definition 1.3. (Probability). A **probability** $\mathbb{P}$ is a function which assigns
a number between 0 and 1 to each event and which satisfies the following
rules:

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$.
2. $\mathbb{P}(\Omega) = 1$.
3. For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

$$\textbf{Sum Rule:} \qquad \mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i) . \qquad (1.3)$$

The crucial property (1.3) is called the **sum rule** of probability. It simply states
that if an event can happen in several distinct ways (expressed as a union of events,
none of which are overlapping), then the probability that at least one of these events
happens (i.e., the probability of the union) is simply the sum of the probabilities of
the individual events. Figure 1.5 illustrates that the probability $\mathbb{P}$ has the properties
of a *measure*. However, instead of measuring lengths, areas, or volumes, $\mathbb{P}(A)$
measures the likelihood or probability of an event A as a number between 0 and 1.
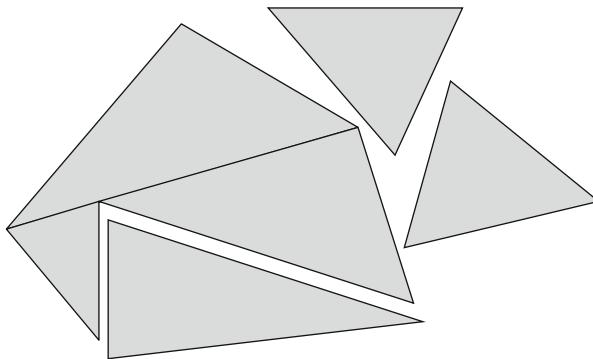
Fig. 1.5 A probability rule $\mathbb{P}$ has exactly the same properties as an area measure. For example,
the total area of the union of the nonoverlapping triangles is equal to the sum of the areas of the
individual triangles

The following theorem lists some important properties of a probability measure. These properties are direct consequences of the three rules defining a probability measure.

Theorem 1.2. (Properties of a Probability). Let A and B be events and $\mathbb{P}$ a probability. Then,

1. $\mathbb{P}(\emptyset) = 0$,
2. if $A \subset B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$,
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

1. Since $\Omega = \Omega \cup \emptyset$ and $\Omega \cap \emptyset = \emptyset$, it follows from the sum rule that $\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset)$. Therefore, by Rule 2 of Definition 1.3, we have $1 = 1 + \mathbb{P}(\emptyset)$, from which it follows that $\mathbb{P}(\emptyset) = 0$.
2. If $A \subset B$, then $B = A \cup (B \cap A^c)$, where A and $B \cap A^c$ are disjoint. Hence, by the sum rule, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, which (by Rule 1) is greater than or equal to $\mathbb{P}(A)$.
3. $\Omega = A \cup A^c$, where A and A^c are disjoint. Hence, by the sum rule and Rule 2: $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$, and thus $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
4. Write $A \cup B$ as the disjoint union of A and $B \cap A^c$. Then, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Also, $B = (A \cap B) \cup (B \cap A^c)$, so that $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c)$. Combining these two equations gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. □

We have now completed our general model for a random experiment. Of course for any *specific* model we must carefully specify the sample space Ω and probability $\mathbb{P}$ that best describe the random experiment.

Example 1.5 (Casting a Die). Consider the experiment where a fair die is cast. How should we specify Ω and $\mathbb{P}$? Obviously, $\Omega = \{1, 2, \ldots, 6\}$; and common sense dictates that we should define $\mathbb{P}$ by

$$\mathbb{P}(A) = \frac{|A|}{6}, \quad A \subset \Omega,$$

where $|A|$ denotes the number of elements in set A. For example, the probability of getting an even number is $\mathbb{P}(\{2, 4, 6\}) = 3/6 = 1/2$.

In many applications the sample space is *countable*: $\Omega = \{a_1, a_2, \ldots, a_n\}$ or $\Omega = \{a_1, a_2, \ldots\}$. Such a sample space is said to be **discrete**. The easiest way to specify a probability $\mathbb{P}$ on a discrete sample space is to first assign a probability p_i to each **elementary event** $\{a_i\}$ and then to define

$$\mathbb{P}(A) = \sum_{i : a_i \in A} p_i \text{ for all } A \subset \Omega.$$

Fig. 1.6 A discrete sample space
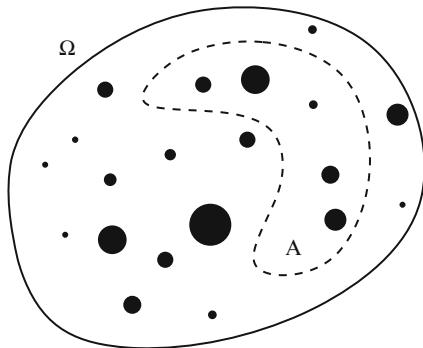
This idea is graphically represented in Fig. 1.6. Each element a_i in the sample space is assigned a probability weight p_i represented by a black dot. To find the probability of an event A we have to sum up the weights of all the elements in the set A.

Again, it is up to the modeler to properly specify these probabilities. Fortunately, in many applications, all elementary events are *equally likely*, and thus the probability of each elementary event is equal to 1 divided by the total number of elements in Ω. In such case the probability of an event $A \subset \Omega$ is simply

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega} \ ,$$

provided that the total number of elements in Ω is finite. The calculation of such probabilities thus reduces to *counting*; see Problem 1.6.   ☞ 19

When the sample space is not countable, for example, $\Omega = \mathbb{R}_+$, it is said to be **continuous**.

Example 1.6 (Drawing a Random Point in the Unit Interval). We draw at random a point in the interval $[0, 1]$ such that each point is equally likely to be drawn. How do we specify the model for this experiment?

The sample space is obviously $\Omega = [0, 1]$, which is a continuous sample space. We cannot define $\mathbb{P}$ via the elementary events $\{x\}$, $x \in [0, 1]$ because each of these events has probability 0. However, we can define $\mathbb{P}$ as follows. For each $0 \leqslant a \leqslant b \leqslant 1$, let

$$\mathbb{P}([a, b]) = b - a \ .$$

This completely defines $\mathbb{P}$. In particular, the probability that a point will fall into any (sufficiently nice) set A is equal to the *length* of that set.

Describing a random experiment by specifying explicitly the sample space and the probability measure is not always straightforward or necessary. Sometimes it is useful to model only certain *observations* on the experiment. This is where *random variables* come into play, and we will discuss these in Chap. 2.   ☞ 23

1.5   Conditional Probability and Independence

How do probabilities change when we know that some event $B \subset \Omega$ has occurred? Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$, which is called the *conditional probability* of A given B. The situation is illustrated in Fig. 1.7.

Fig. 1.7 What is the probability that A occurs given that the outcome is known to lie in B?
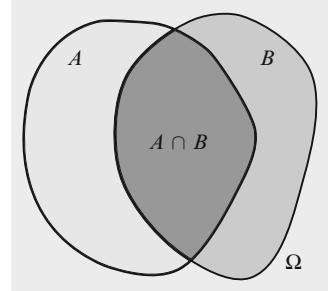
Definition 1.4. (Conditional Probability). The **conditional probability** of A given B (with $\mathbb{P}(B) \neq 0$) is defined as:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \ . \tag{1.4}$$

Example 1.7 (Casting Two Dice). We cast two fair dice consecutively. Given that the sum of the dice is 10, what is the probability that one 6 is cast? Let B be the event that the sum is 10:

$$B = \{(4, 6), (5, 5), (6, 4)\} \ .$$

Let A be the event that one 6 is cast:

$$A = \{(1, 6), \ldots, (5, 6), (6, 1), \ldots, (6, 5)\} \ .$$

Then, $A \cap B = \{(4, 6), (6, 4)\}$. And, since for this experiment all elementary events are equally likely, we have

$$\mathbb{P}(A \mid B) = \frac{2/36}{3/36} = \frac{2}{3} \ .$$

Example 1.8 (Monty Hall Problem). Consider a quiz in which the final contestant
is to choose a prize which is hidden behind one of three curtains (A, B, or C).
Suppose without loss of generality that the contestant chooses curtain A. Now the
quiz master (Monty) always opens one of the other curtains: if the prize is behind B,
Monty opens C; if the prize is behind C, Monty opens B; and if the prize is behind
A, Monty opens B or C with equal probability, e.g., by tossing a coin (of course the
contestant does not see Monty tossing the coin!) (Fig. 1.8).
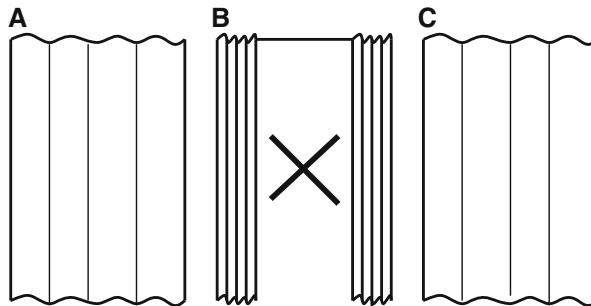
Fig. 1.8 Given that Monty opens curtain (B), should the contestant stay with his/her original
choice (A) or switch to the other unopened curtain (C)?

Suppose, again without loss of generality, that Monty opens curtain B. The con-
testant is now offered the opportunity to switch to curtain C. Should the contestant
stay with his/her original choice (A) or switch to the other unopened curtain (C)?

Notice that the sample space here consists of four possible outcomes: Ac, the
prize is behind A and Monty opens C; Ab, the prize is behind A and Monty opens
B; Bc, the prize is behind B and Monty opens C; and Cb, the prize is behind C
and Monty opens B. Let A, B, C be the events that the prize is behind A, B, and C,
respectively. Note that $A = \{Ac, Ab\}$, $B = \{Bc\}$, and $C = \{Cb\}$; see Fig. 1.9.

Fig. 1.9 The sample space
for the Monty Hall problem
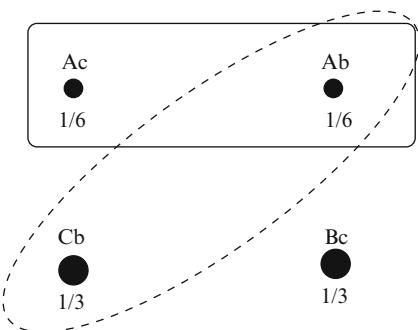
Now, obviously $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C)$, and since Ac and Ab are equally likely, we have $\mathbb{P}(\{Ab\}) = \mathbb{P}(\{Ac\}) = 1/6$. Monty opening curtain B means that we have information that event $\{Ab, Cb\}$ has occurred. The probability that the prize is behind A given this event is therefore

$$\mathbb{P}(A \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Ac, Ab\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Ab\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{1}{3} \, .$$

This is what is to be expected: the fact that Monty opens a curtain does not give any extra information that the prize is behind A. Obviously, $\mathbb{P}(B \,|\, \text{B is opened}) = 0$. It follows then that $\mathbb{P}(C \,|\, \text{B is opened})$ must be 2/3, since the conditional probabilities must sum up to 1. Indeed,

$$\mathbb{P}(C \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Cb\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\frac{1}{3}}{\frac{1}{6} + \frac{1}{3}} = \frac{2}{3} \, .$$

Hence, given the information that B is opened, it is twice as likely that the prize is behind C than behind A. Thus, the contestant should switch!

1.5.1 Product Rule

By the definition of conditional probability (1.4) we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \, \mathbb{P}(B \,|\, A) \, .$$

It is not difficult to generalize this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the **product rule** of probability. We leave the proof as an exercise; see Problem 1.11.

☞ 20

> **Theorem 1.3. (Product Rule).** Let $A_1, \ldots, A_n$ be a sequence of events with $\mathbb{P}(A_1 \cdots A_{n-1}) > 0$. Then,
>
> $$\mathbb{P}(A_1 \cdots A_n) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1.5)$$
> $$\mathbb{P}(A_1) \, \mathbb{P}(A_2 \,|\, A_1) \, \mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) \, .$$

Example 1.9 (Urn Problem). We draw consecutively three balls from an urn with 5 white and 5 black balls, without putting them back. What is the probability that all drawn balls will be black?

Let A_i be the event that the ith ball is black. We wish to find the probability of $A_1 A_2 A_3$, which by the product rule (1.5) is

$$\mathbb{P}(A_1)\,\mathbb{P}(A_2\mid A_1)\,\mathbb{P}(A_3\mid A_1 A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} \approx 0.083\ .$$

Example 1.10 (Birthday Problem). What is the probability that in a group of n people all have different birthdays? We can use the product rule. Let A_i be the event that the first i people have different birthdays, $i = 1, 2, \ldots$. Note that $\cdots \subset A_3 \subset A_2 \subset A_1$. Therefore, $A_n = A_1 \cap A_2 \cap \cdots \cap A_n$, and thus by the product rule

$$\mathbb{P}(A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2\mid A_1)\,\mathbb{P}(A_3\mid A_2)\cdots\mathbb{P}(A_n\mid A_{n-1})\ .$$

Now $\mathbb{P}(A_k\mid A_{k-1}) = (365 - k + 1)/365$, because given that the first $k - 1$ people have different birthdays, there are no duplicate birthdays among the first k people if and only if the birthday of the kth person is chosen from the $365 - (k-1)$ remaining birthdays. Thus, we obtain

$$\mathbb{P}(A_n) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - n + 1}{365}, \quad n \geqslant 1\ . \tag{1.6}$$

A graph of $\mathbb{P}(A_n)$ against n is given in Fig. 1.10. Note that the probability $\mathbb{P}(A_n)$ rapidly decreases to zero. For $n = 23$ the probability of having no duplicate birthdays is already less than 1/2.
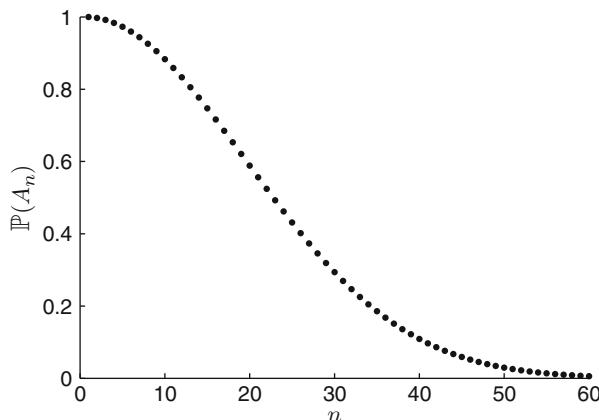
Fig. 1.10   The probability of having no duplicate birthday in a group of n people against n

1.5.2   Law of Total Probability and Bayes' Rule

Suppose that $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω; see Fig. 1.11.
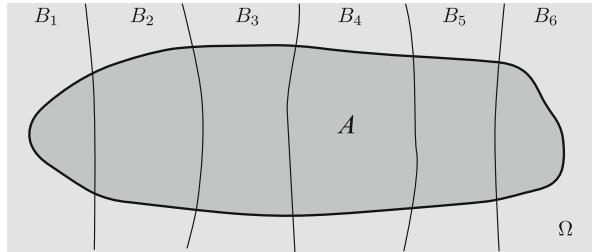
Fig. 1.11   A partition $B_1, \ldots, B_6$ of the sample space Ω. Event A is partitioned into events $A \cap B_1$, $\ldots, A \cap B_6$

A partitioning of the state space can sometimes make it easier to calculate probabilities via the following theorem.

Theorem 1.4. (Law of Total Probability). Let A be an event and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i) \, . \tag{1.7}$$

Proof. The sum rule gives $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$, and by the product rule we have $\mathbb{P}(A \cap B_i) = \mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i)$. $\square$

Combining the law of total probability with the definition of conditional probability gives **Bayes' Rule**:

Theorem 1.5. (Bayes Rule). Let A be an event with $\mathbb{P}(A) > 0$ and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j) \, \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i) \, \mathbb{P}(B_i)} \, . \tag{1.8}$$

Proof. By definition, $\mathbb{P}(B_j \mid A) = \mathbb{P}(A \cap B_j)/\mathbb{P}(A) = \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)/\mathbb{P}(A)$. Now apply the law of total probability to $\mathbb{P}(A)$. □

Example 1.11 (Quality Control Problem). A company has three factories (1, 2, and 3) that produce the same chip, each producing 15 %, 35 %, and 50 % of the total production. The probability of a faulty chip at factory 1, 2, and 3 is 0.01, 0.05, and 0.02, respectively. Suppose we select randomly a chip from the total production and this chip turns out to be faulty. What is the conditional probability that this chip has been produced in factory 1?

Let B_i denote the event that the chip has been produced in factory i. The $\{B_i\}$ form a partition of Ω. Let A denote the event that the chip is faulty. We are given the information that $\mathbb{P}(B_1) = 0.15$, $\mathbb{P}(B_2) = 0.35$, $\mathbb{P}(B_3) = 0.5$ as well as $\mathbb{P}(A \mid B_1) = 0.01$, $\mathbb{P}(A \mid B_2) = 0.05$, $\mathbb{P}(A \mid B_3) = 0.02$.

We wish to find $\mathbb{P}(B_1 \mid A)$, which by Bayes' rule is given by

$$\mathbb{P}(B_1 \mid A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 \ .$$

1.5.3 Independence

Independence is a very important concept in probability and statistics. Loosely speaking it models the *lack of information* between events. We say events A and B are *independent* if the knowledge that B has occurred does not change the probability that A occurs. More precisely, A and B are said to be independent if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Since $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, an alternative definition of independence is: A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. This definition covers the case where $B = \emptyset$.

We can extend the definition to arbitrarily many events [compare with the product rule (1.5)]:

Definition 1.5. (Independence). The events $A_1, A_2, \ldots,$ are said to be **independent** if for any k and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \ . \qquad (1.9)$$

Remark 1.1. In most cases independence of events is a *model assumption*. That is, $\mathbb{P}$ is chosen such that certain events are independent.

Example 1.12 (Coin Tossing and the Binomial Law). We toss a coin n times. The sample space can be written as the set of binary n-tuples:

$$\Omega = \{(\underbrace{0,\ldots,0}_{n \text{ times}}),\ldots,(1,\ldots,1)\}\,.$$

Here, 0 represents Tails and 1 represents Heads. For example, the outcome $(0,1,0,1,\ldots)$ means that the first time Tails is thrown, the second time Heads, the third time Tails, the fourth time Heads, etc.

How should we define $\mathbb{P}$? Let A_i denote the event of Heads at the ith throw, $i = 1,\ldots,n$. Then, $\mathbb{P}$ should be such that the following holds:

- The events $A_1,\ldots,A_n$ should be *independent* under $\mathbb{P}$.
- $\mathbb{P}(A_i)$ should be the same for all i. Call this known or unknown probability p ($0 \leqslant p \leqslant 1$).

These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are Heads and the last $n-k$ are Tails is

$$\mathbb{P}(\{(\underbrace{1,1,\ldots,1}_{k \text{ times}},\underbrace{0,0,\ldots,0}_{n-k \text{ times}})\}) = \mathbb{P}(A_1 \cap \cdots \cap A_k \cap A_{k+1}^c \cap \cdots \cap A_n^c)$$

$$= \mathbb{P}(A_1)\cdots\mathbb{P}(A_k)\,\mathbb{P}(A_{k+1}^c)\cdots\mathbb{P}(A_n^c) = p^k(1-p)^{n-k}\,.$$

Note that if A_i and A_j are independent, then so are A_i and A_j^c; see Problem 1.12.

Let B_k be the event that k Heads are thrown in total. The probability of this event is the sum of the probabilities of elementary events $\{(x_1,\ldots,x_n)\}$ for which $x_1 + \cdots + x_n = k$. Each of these events has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ of these. We thus obtain the **binomial law**:

$$\mathbb{P}(B_k) = \binom{n}{k}\,p^k(1-p)^{n-k}, \quad k = 0,1,\ldots,n\,. \tag{1.10}$$

Example 1.13 (Geometric Law). There is another important law associated with the coin toss experiment. Let C_k be the event that Heads appears for the first time at the kth toss, $k = 1,2,\ldots$. Then, using the same events $\{A_i\}$ as in the previous example, we can write

$$C_k = A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k\,.$$

Using the independence of $A_1^c,\ldots,A_{k-1}^c, A_k$, we obtain the **geometric law**:

$$\mathbb{P}(C_k) = \mathbb{P}(A_1^c)\cdots\mathbb{P}(A_{k-1}^c)\,\mathbb{P}(A_k)$$

$$= \underbrace{(1-p)\cdots(1-p)}_{k-1 \text{ times}}\,p = (1-p)^{k-1}\,p\,.$$

1.6 Problems

1.1. For each of the five random experiments at the beginning of Sect. 1.1 define a convenient sample space.

1.2. Interpret De Morgan's rule (1.2) in terms of an unreliable series system.

1.3. Let $\mathbb{P}(A) = 0.9$ and $\mathbb{P}(B) = 0.8$. Show that $\mathbb{P}(A \cap B) \geqslant 0.7$.

1.4. Throw two fair dice one after the other.

(a) What is the probability that the second die is 3, given that the sum of the dice is 6?
(b) What is the probability that the first die is 3 and the second is not 3?

1.5. An "expert" wine taster has to try to match 6 glasses of wine to 6 wine labels. Each label can only be chosen once.

(a) Formulate a sample space Ω for this experiment.
(b) Assuming the wine taster is a complete fraud, define an appropriate probability $\mathbb{P}$ on the sample space.
(c) What is the probability that the wine taster guesses 4 labels correctly, assuming he/she guesses them randomly?

1.6. Many counting problems can be cast into the framework of drawing k balls from an urn with n balls, numbered $1, \ldots, n$; see Fig. 1.12.

Fig. 1.12 Draw k balls from an urn with $n = 10$ numbered balls
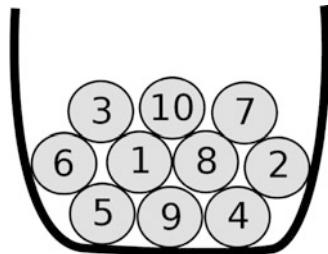
The drawing can be done in several ways. Firstly, the k balls could be drawn one-by-one or all at the same time. In the first case the **order** in which the balls are drawn can be noted. In the second case we can still assume that the balls are drawn one-by-one, but we do not note the order. Secondly, once a ball is drawn, it can either be put back into the urn or be left out. This is called drawing with and without **replacement**, respectively. There are thus four possible random experiments. Prove that for each of these experiments the total number of possible outcomes is the following:

1. Ordered, with replacement: n^k.
2. Ordered, without replacement: $^n P_k = n(n-1) \cdots (n-k+1)$.

3. Unordered, without replacement: $^nC_k = \binom{n}{k} = \frac{^nP_k}{k!} = \frac{n!}{(n-k)!\,k!}$.

4. Unordered, with replacement: $\binom{n+k-1}{k}$.

Provide a sample space for each of these experiments. Hint: it is important to use a notation that clearly shows whether the arrangements of numbers are ordered or not. Denote ordered arrangements by *vectors*, e.g., (1, 1, 2), and unordered arrangements by *sets*, e.g., $\{1, 2, 3\}$ or *multisets*, e.g., $\{1, 1, 2\}$.

1.7. Formulate the birthday problem in terms of an urn experiment, as in Problem 1.6, and derive the probability (1.6) by counting.

1.8. Three cards are drawn from a full deck of cards, noting the order. The cards may be numbered from 1 to 52.

(a) Give the sample space. Is each elementary event equally likely?
(b) What is the probability that we draw three Aces?
(c) What is the probability that we draw one Ace, one King, and one Queen (not necessarily in that order)?
(d) What is the probability that we draw no pictures (no A, K, Q, or J)?

1.9. In a group of 20 people there are three brothers. The group is separated at random into two groups of 10. What is the probability that the brothers are in the same group?

1.10. Two fair dice are thrown.

(a) Find the probability that both dice show the same face.
(b) Find the same probability, using the extra information, that the sum of the dice is not greater than 4.

1.11. Prove the product rule (1.5). Hint: first show it for the case of three events:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A \cap B) .$$

1.12. If A and B are independent events, then A and B^c are also independent. Prove this.

1.13. Select at random 3 people from a large population. What is the probability that they all have the same birthday?

1.14. In a large population 40 % votes for A and 60 % for B. Suppose we select at random 10 people. What is the probability that in this group exactly 4 people will vote for A?

1.15. A certain AIDS test has a 0.98 probability of giving a positive result when the blood is infected and a 0.07 probability of giving a positive result when the blood is not infected (a so-called false-positive). Suppose 1 % of the population carries the HIV virus.

1. Using the law of total probability, what is the probability that the test is positive for a randomly selected person?
2. What is the probability that a person is indeed infected, *given* that the test yields a positive result?

1.16. A box has three identical-looking coins. However, the probability of success (Heads) is different for each coin: coin 1 is fair, coin 2 has a success probability of 0.4, and coin 3 has a success probability of 0.6. We pick one coin at random and throw it 100 times. Suppose 43 Heads come up. Using this information, assess the probability that coin 1, 2, or 3 was chosen.

1.17. In a binary communication channel, 0s and 1s are transmitted with equal probability. The probability that a 0 is correctly received (as a 0) is 0.95. The probability that a 1 is correctly received (as a 1) is 0.99. Suppose we receive a 0, what is the probability that, in fact, a 1 was sent?

1.18. A fair coin is tossed 20 times.

1. What is the probability of exactly 10 Heads?
2. What is the probability of 15 or more Heads?

1.19. Two fair dice are cast (at the same time) until their sum is 12.

1. What is the probability that we have to wait exactly 10 tosses?
2. What is the probability that we do not have to wait more than 100 tosses?

1.20. Independently throw 10 balls into one of three boxes, numbered 1, 2, and 3, with probabilities 1/4, 1/2, and 1/4, respectively.

1. What is the probability that box 1 has 2 balls, box 2 has 5 balls, and box 3 has 3 balls?
2. What is the probability that box 1 remains empty?

1.21. Implement a MATLAB program that performs 100 tosses with a fair die. Hint: use the `rand` and `ceil` functions, where `ceil(x)` returns the smallest integer larger than or equal to `x`.

1.22. For each of the four urn experiments in Problem 1.6 implement a MATLAB program that simulates the experiment. Hint: in addition to the functions `rand` and `ceil`, you may wish to use the `sort` function.

1.23. Verify your answers for Problem 1.20 with a computer simulation, where the experiment is repeated many times.

Chapter 2
Random Variables and Probability Distributions

Specifying a model for a random experiment via a complete description of the sample space Ω and probability measure $\mathbb{P}$ may not always be necessary or convenient. In practice we are only interested in certain *numerical measurements* pertaining to the experiment. Such random measurements can be included into the model via the notion of a *random variable*.

2.1 Random Variables

> **Definition 2.1. (Random Variable).** A **random variable** is a *function* from the sample space Ω to $\mathbb{R}$.

Example 2.1 (Sum of Two Dice). We throw two fair dice and note the sum of their face values. If we throw the dice consecutively and observe both throws, the sample space is $\Omega = \{(1, 1), \ldots, (6, 6)\}$. The function X defined by $X(i, j) = i + j$ is a random variable which maps the outcome (i, j) to the sum $i + j$, as depicted in Fig. 2.1.

Note that five outcomes in the sample space are mapped to 8. A natural notation for the corresponding set of outcomes is $\{X = 8\}$. Since all outcomes in Ω are equally likely, we have

$$\mathbb{P}(\{X = 8\}) = \frac{5}{36} \ .$$

This notation is very suggestive and convenient. From a nonmathematical viewpoint we can interpret X as a "random" variable, that is, a variable that can take several values with certain probabilities. In particular, it is not difficult to check that

$$\mathbb{P}(\{X = x\}) = \frac{6 - |7 - x|}{36}, \quad x = 2, \ldots, 12 \ .$$
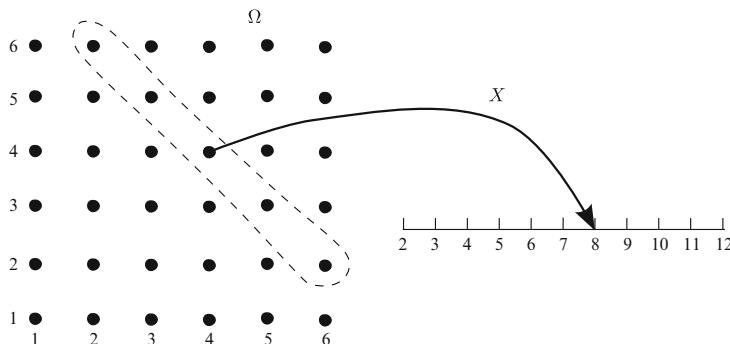
Fig. 2.1 Random variable X represents the sum of two dice

Although random variables are, mathematically speaking, *functions*, it is often convenient to view them as observations of a random experiment that has not yet taken place. In other words, a random variable is considered as a measurement that becomes available *tomorrow*, while all the thinking about the measurement can be carried out *today*. For example, we can specify today exactly the probabilities pertaining to the random variables.

We often denote random variables with *capital* letters from the last part of the alphabet, e.g., X, X_1, X_2, ..., Y, Z. Random variables allow us to use natural and intuitive notations for certain events, such as $\{X = 10\}$, $\{X > 1000\}$, and $\{\max(X, Y) \leqslant Z\}$.

☞ 18   **Example 2.2 (Coin Tossing).** In Example 1.12 we constructed a probability model for the random experiment where a biased coin is tossed n times. Suppose we are not interested in a specific outcome but only in the total number of Heads, X, say. In particular, we would like to know the probability that X takes certain values between 0 and n. Example 1.12 suggests that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n \,, \tag{2.1}$$

providing all the information about X that we could possibly wish to know. To justify (2.1) mathematically, we can reason as in Example 2.1. First, define X as the function that assigns to each outcome $\omega = (x_1, \ldots, x_n)$ the number $x_1 + \cdots + x_n$. Thus, X is a random variable in mathematical terms, that is, a function. Second, the event B_k that there are exactly k Heads in n throws can be written as

$$B_k = \{\omega \in \Omega : X(\omega) = k\} \,.$$

If we write this as $\{X = k\}$, and further abbreviate $\mathbb{P}(\{X = k\})$ to $\mathbb{P}(X = k)$, then we obtain (2.1) directly from (1.10).

We give some more examples of random variables without specifying the sample space:

1. The number of defective transistors out of 100 inspected ones.
2. The number of bugs in a computer program.
3. The amount of rain in a certain location in June.
4. The amount of time needed for an operation.

The set of all possible values that a random variable X can take is called the **range** of X. We further distinguish between discrete and continuous random variables:

- **Discrete** random variables can only take *countably many* values.
- **Continuous** random variables can take a continuous range of values, for example, any value on the positive real line $\mathbb{R}_+$.

2.2 Probability Distribution

Let X be a random variable. We would like to designate the probabilities of events such as $\{X = x\}$ and $\{a \leqslant X \leqslant b\}$. If we can specify all probabilities involving X, we say that we have determined the **probability distribution** of X. One way to specify the probability distribution is to give the probabilities of all events of the form $\{X \leqslant x\}$, $x \in \mathbb{R}$. This leads to the following definition.

> **Definition 2.2. (Cumulative Distribution Function).** The **cumulative distribution function** (cdf) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by
>
> $$F(x) = \mathbb{P}(X \leqslant x), \quad x \in \mathbb{R} .$$
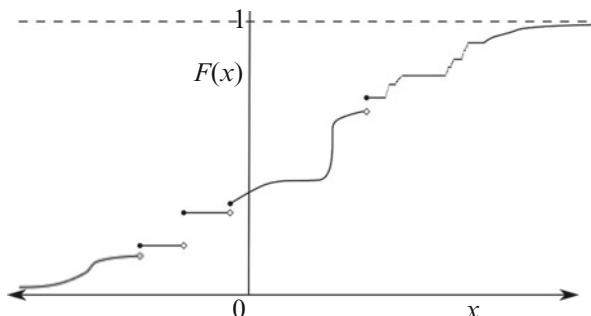
Fig. 2.2 A cumulative distribution function (*cdf*)

Note that we have used $\mathbb{P}(X \leqslant x)$ as a shorthand notation for $\mathbb{P}(\{X \leqslant x\})$. From now on we will use this type of abbreviation throughout the book. In Fig. 2.2 the graph of a general cdf is depicted.

Theorem 2.1. (Properties of Cdf). Let F be the cdf of a random variable X. Then,

1. F is bounded between 0 and 1: $0 \leqslant F(x) \leqslant 1$,

2. F is increasing: if $x \leqslant y$, then $F(x) \leqslant F(y)$,

3. F is right-continuous: $\lim_{h \downarrow 0} F(x + h) = F(x)$.

Proof.

☞ 9
1. Let $A = \{X \leqslant x\}$. By Rule 1 in Definition 1.3, $0 \leqslant \mathbb{P}(A) \leqslant 1$.
2. Suppose $x \leqslant y$. Define $A = \{X \leqslant x\}$ and $B = \{X \leqslant y\}$. Then, $A \subset B$ and, by
☞ 10   Theorem 1.2, $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
3. Take any sequence $x_1, x_2, \ldots$ decreasing to x. We have to show that $\lim_{n \to \infty} \mathbb{P}(X \leqslant x_n) = \mathbb{P}(X \leqslant x)$ or, equivalently, $\lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}(A)$, where $A_n = \{X > x_n\}$ and $A = \{X > x\}$. Let $B_n = \{x_{n-1} \geqslant X > x_n\}$, $n = 1, 2, \ldots$, with x_0 defined as ∞. Then, $A_n = \cup_{i=1}^{n} B_i$ and $A = \cup_{i=1}^{\infty} B_i$. Since the $\{B_i\}$ are disjoint, we have by the sum rule:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \stackrel{\text{def}}{=} \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n \to \infty} \mathbb{P}(A_n) \,,$$

as had to be shown.   □

Conversely, any function F with the above properties can be used to specify the distribution of a random variable X.

If X has cdf F, then the probability that X takes a value in the interval $(a, b]$ (excluding a, including b) is given by

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) \,.$$

To see this, note that $\mathbb{P}(X \leqslant b) = \mathbb{P}(\{X \leqslant a\} \cup \{a < X \leqslant b\})$, where the events $\{X \leqslant a\}$ and $\{a < X \leqslant b\}$ are disjoint. Thus, by the sum rule, $F(b) = F(a) + \mathbb{P}(a < X \leqslant b)$, which leads to the result above. Note however that

$$\mathbb{P}(a \leqslant X \leqslant b) = F(b) - F(a) + \mathbb{P}(X = a)$$
$$= F(b) - F(a) + F(a) - F(a-)$$
$$= F(b) - F(a-) \,,$$

where $F(a-)$ denotes the limit from below: $\lim_{x \uparrow a} F(x)$.

2.2.1 Discrete Distributions

> **Definition 2.3. (Discrete Distribution).** A random variable X is said to have a **discrete distribution** if $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ for some finite or countable set of values $x_1, x_2, \ldots$, such that $\sum_i \mathbb{P}(X = x_i) = 1$. The **discrete probability density function (pdf)** of X is the function f defined by $f(x) = \mathbb{P}(X = x)$.

We sometimes write f_X instead of f to stress that the discrete probability density function refers to the discrete random variable X. The easiest way to specify the distribution of a discrete random variable is to specify its pdf. Indeed, by the sum rule, if we know $f(x)$ for all x, then we can calculate all possible probabilities involving X. Namely,

☞ 9

$$\mathbb{P}(X \in B) = \sum_{x \in B} f(x) \tag{2.2}$$

for any subset B in the range of X, as illustrated in Fig. 2.3.
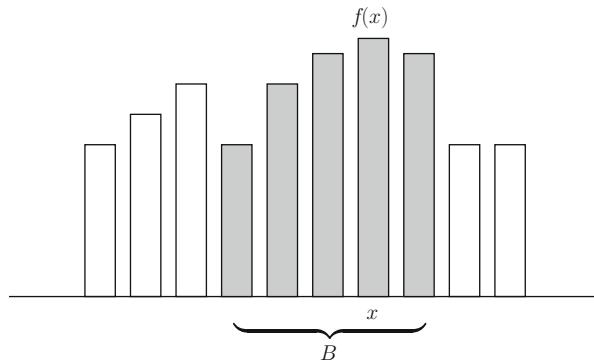
Fig. 2.3 Discrete probability density function

Example 2.3 (Sum of Two Dice, Continued). Toss two fair dice and let X be the sum of their face values. The discrete pdf is given in Table 2.1, which follows directly from Example 2.1.

Table 2.1 Discrete pdf of the sum of two fair dice

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

2.2.2   Continuous Distributions

Definition 2.4. (Continuous Distribution). A random variable X with cdf F is said to have a **continuous distribution** if there exists a positive function f with *total integral 1* such that for all $a < b$,

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) = \int_a^b f(u)\,du \ . \tag{2.3}$$

Function f is called the **probability density function (pdf)** of X.

Remark 2.1. Note that we use the *same* notation f for both the discrete and the continuous pdf, to stress the similarities between the discrete and continuous case. We will even drop the qualifier "discrete" or "continuous" when it is clear from the context with which case we are dealing. Henceforth we will use the notation $X \sim f$ and $X \sim F$ to indicate that X is distributed according to pdf f or cdf F.

In analogy to the discrete case (2.2), once we know the pdf, we can calculate any probability of interest by means of integration:

$$\mathbb{P}(X \in B) = \int_B f(x)\,dx \ , \tag{2.4}$$

as illustrated in Fig. 2.4.
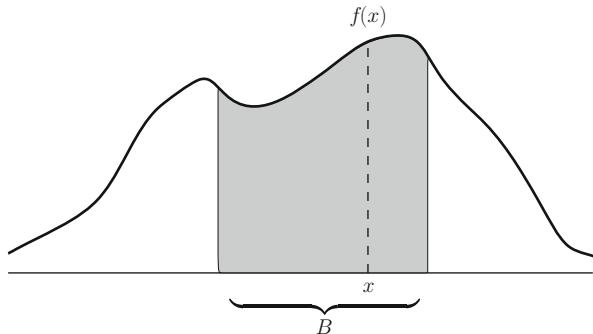
Fig. 2.4 Probability density function (*pdf*)

Suppose that f and F are the pdf and cdf of a continuous random variable X, as in Definition 2.4. Then F is simply a *primitive* (also called antiderivative) of f:

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^x f(u)\,du \ .$$

Conversely, f is the *derivative* of the cdf F:

$$f(x) = \frac{d}{dx} F(x) = F'(x) \,.$$

It is important to understand that in the continuous case $f(x)$ is not equal to the probability $\mathbb{P}(X = x)$, because the latter is 0 for all x. Instead, we interpret $f(x)$ as the *density* of the probability distribution at x, in the sense that for any small h,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_x^{x+h} f(u)\, du \approx h\, f(x) \,. \tag{2.5}$$

Note that $\mathbb{P}(x \leqslant X \leqslant x + h)$ is equal to $\mathbb{P}(x < X \leqslant x + h)$ in this case.

Example 2.4 (Random Point in an Interval). Draw a random number X from the interval of real numbers $[0, 2]$, where each number is equally likely to be drawn. What are the pdf f and cdf F of X? Using the same reasoning as in Example 1.6, we see that

$$\mathbb{P}(X \leqslant x) = F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/2 & \text{if } 0 \leqslant x \leqslant 2, \\ 1 & \text{if } x > 2. \end{cases}$$

By differentiating F we find

$$f(x) = \begin{cases} 1/2 & \text{if } 0 \leqslant x \leqslant 2, \\ 0 & \text{otherwise.} \end{cases}$$

Note that this density is *constant* on the interval $[0, 2]$ (and zero elsewhere), reflecting the fact that each point in $[0, 2]$ is equally likely to be drawn.

2.3   Expectation

Although all probability information about a random variable is contained in its cdf or pdf, it is often useful to consider various numerical characteristics of a random variable. One such number is the *expectation* of a random variable, which is a "weighted average" of the values that X can take. Here is a more precise definition.

Definition 2.5. (Expectation of a Discrete Random Variable). Let X be a *discrete* random variable with pdf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \sum_x x\, \mathbb{P}(X = x) = \sum_x x\, f(x) \,. \tag{2.6}$$

The expectation of X is sometimes written as μ_X. It is assumed that the sum in (2.6) is well defined—possibly ∞ or $-\infty$. One way to interpret the expectation is as a *long-run average payout*. Suppose in a game of dice the payout X (dollars) is the largest of the face values of two dice. To play the game a fee of d dollars must be paid. What would be a fair amount for d? The answer is

$$d = \mathbb{E}X = 1 \times \mathbb{P}(X = 1) + 2 \times \mathbb{P}(X = 2) + \cdots + 6 \times \mathbb{P}(X = 6)$$

$$= 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + 3 \times \frac{5}{36} + 4 \times \frac{7}{36} + 5 \times \frac{9}{36} + 6 \times \frac{11}{36} = \frac{161}{36} \approx 4.47 \ .$$

Namely, if the game is played many times, the long-run fraction of tosses in which the maximum face value is 1, 2,..., 6, is $\frac{1}{36}, \frac{3}{36}, \ldots, \frac{11}{36}$, respectively. Hence, the long-run average payout of the game is the weighted sum of $1, 2, \ldots, 6$, where the weights are the long-run fractions (probabilities). The game is "fair" if the long-run average profit $\mathbb{E}X - d$ is zero.

The expectation can also be interpreted as a *center of mass*. Imagine that point masses with weights $p_1, p_2, \ldots, p_n$ are placed at positions $x_1, x_2, \ldots, x_n$ on the real line; see Fig. 2.5.
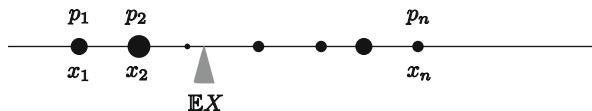
Fig. 2.5 The expectation as a center of mass

The center of mass—the place where the weights are balanced—is

$$\text{center of mass} = x_1 \, p_1 + \cdots + x_n \, p_n \ ,$$

which is exactly the expectation of the discrete variable X that takes values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. An obvious consequence of this interpretation is that for a *symmetric* pdf the expectation is equal to the symmetry point (provided that the expectation exists). In particular, suppose that $f(c + y) = f(c - y)$ for all y. Then,

$$\mathbb{E}X = c \, f(c) + \sum_{x>c} x f(x) + \sum_{x<c} x f(x)$$

$$= c \, f(c) + \sum_{y>0}(c + y) f(c + y) + \sum_{y>0}(c - y) f(c - y)$$

$$= c \, f(c) + \sum_{y>0} c \, f(c + y) + c \sum_{y>0} f(c - y) = c \sum_{x} f(x) = c \ .$$

For continuous random variables we can define the expectation in a similar way, replacing the sum with an integral.

Definition 2.6. (Expectation of a Continuous Random Variable). Let X be a *continuous* random variable with pdf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x\;. \tag{2.7}$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$, is also a random variable. The following theorem simply states that the expected value of a function of X is the weighted average of the values that this function can take.

Theorem 2.2. (Expectation of a Function of a Random Variable). If X is *discrete* with pdf f, then for any real-valued function g

$$\mathbb{E}\,g(X) = \sum_x g(x)\,f(x)\;.$$

Similarly, if X is *continuous* with pdf f, then

$$\mathbb{E}\,g(X) = \int_{-\infty}^{\infty} g(x)\,f(x)\,\mathrm{d}x\;.$$

Proof. The proof is given for the discrete case only, as the continuous case can be proven in a similar way. Let $Y = g(X)$, where X is a discrete random variable with pdf f_X and g is a function. Let f_Y be the (discrete) pdf of the random variable Y. It can be expressed in terms of f_X in the following way:

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_{x:g(x)=y} f_X(x)\;.$$

Thus, the expectation of Y is

$$\mathbb{E}Y = \sum_y y\, f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} y f_X(x)$$

$$= \sum_x g(x)\, f_X(x)\;. \qquad \square$$

Example 2.5 (Die Experiment and Expectation). Find $\mathbb{E}X^2$ if X is the outcome of the toss of a fair die. We have

$$\mathbb{E}X^2 = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = \frac{91}{6} .$$

An important consequence of Theorem 2.2 is that the expectation is "linear".

Theorem 2.3. (Properties of the Expectation). For any real numbers a and b, and functions g and h,

1. $\mathbb{E}[a\,X + b] = a\,\mathbb{E}X + b$,
2. $\mathbb{E}[g(X) + h(X)] = \mathbb{E}g(X) + \mathbb{E}h(X)$.

Proof. Suppose X has pdf f. The first statement follows (in the discrete case) from

$$\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x x\, f(x) + b \sum_x f(x) = a\,\mathbb{E}X + b .$$

Similarly, the second statement follows from

$$\mathbb{E}(g(X) + h(X)) = \sum_x (g(x) + h(x)) f(x) = \sum_x g(x) f(x) + \sum_x h(x) f(x)$$

$$= \mathbb{E}g(X) + \mathbb{E}h(X) .$$

The continuous case is proven analogously, simply by replacing sums with integrals.
$\square$

Another useful numerical characteristic of the distribution of X is the *variance* of X. This number, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X.

Definition 2.7. (Variance and Standard Deviation). The **variance** of a random variable X, denoted as $\mathrm{Var}(X)$, is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 . \qquad (2.8)$$

The square root of the variance is called the **standard deviation**. The number $\mathbb{E}X^r$ is called the rth **moment** of X.

> **Theorem 2.4. (Properties of the Variance).** For any random variable X the following properties hold for the variance:
>
> 1. $\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
> 2. $\text{Var}(a + bX) = b^2 \text{Var}(X)$.

Proof. Write $\mathbb{E}X = \mu$, so that $\text{Var}(X) = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. By the linearity of the expectation, the last expectation is equal to the sum $\mathbb{E}X^2 - 2\mu\mathbb{E}X + \mu^2 = \mathbb{E}X^2 - \mu^2$, which proves the first statement. To prove the second statement, note that the expectation of $a + bX$ is equal to $a + b\mu$. Consequently,

$$\text{Var}(a + bX) = \mathbb{E}(a + bX - (a + b\mu))^2 = \mathbb{E}(b^2(X - \mu)^2) = b^2\text{Var}(X) .$$

$\square$

Note that Property 1 in Theorem 2.4 implies that $\mathbb{E}X^2 \geqslant (\mathbb{E}X)^2$, because $\text{Var}(X) \geqslant 0$. This is a special case of a much more general result, regarding the expectation of convex functions. A real-valued function $h(x)$ is said to be **convex** if for each x_0, there exist constants a and b such that (1) $h(x) \geqslant ax + b$ for all x and (2) $h(x_0) = ax_0 + b$. Examples are the functions $x \mapsto x^2$, $x \mapsto e^x$, and $x \mapsto -\ln x$.

> **Theorem 2.5. (Jensen's Inequality).** Let $h(x)$ be a convex function and X a random variable. Then,
>
> $$\mathbb{E}h(X) \geqslant h(\mathbb{E}X) . \tag{2.9}$$

Proof. Let $x_0 = \mathbb{E}X$. Because h is convex, there exist constants a and b such that $h(X) \geqslant aX + b$ and $h(x_0) = ax_0 + b$. Hence, $\mathbb{E}h(X) \geqslant \mathbb{E}(aX + b) = ax_0 + b = h(x_0) = h(\mathbb{E}X)$. $\square$

2.4   Transforms

Many probability calculations—such as the evaluation of expectations and variances—are facilitated by the use of *transforms*. We discuss here a number of such transforms.

Definition 2.8. (Probability Generating Function). Let X be a *nonnegative* and *integer-valued* random variable with discrete pdf f. The **probability generating function** (PGF) of X is the function G defined by

$$G(z) = \mathbb{E}\, z^X = \sum_{x=0}^{\infty} z^x\, f(x)\,, \quad |z| < R\,,$$

where $R \geqslant 1$ is the **radius of convergence**.

Example 2.6 (Poisson Distribution). Let X have a discrete pdf f given by

$$f(x) = \mathrm{e}^{-\lambda}\, \frac{\lambda^x}{x!}\,, \quad x = 0, 1, 2, \dots\,.$$

X is said to have a **Poisson distribution**. The PGF of X is given by

$$\begin{aligned}
G(z) &= \sum_{x=0}^{\infty} z^x\, \mathrm{e}^{-\lambda}\, \frac{\lambda^x}{x!} \\
&= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} \\
&= \mathrm{e}^{-\lambda} \mathrm{e}^{z\lambda} = \mathrm{e}^{-\lambda(1-z)}\,.
\end{aligned}$$

As this is finite for every z, the radius of convergence is here $R = \infty$.

Theorem 2.6. (Derivatives of a PGF). The kth derivative of a PGF $\mathbb{E}\, z^X$ can be obtained by *differentiation under the expectation sign*:

$$\begin{aligned}
\frac{\mathrm{d}^k}{\mathrm{d}z^k} \mathbb{E}\, z^X &= \mathbb{E}\, \frac{\mathrm{d}^k}{\mathrm{d}z^k} z^X \\
&= \mathbb{E}\left[X(X-1)\cdots(X-k+1) z^{X-k} \right] \quad \text{for } |z| < R\,,
\end{aligned}$$

where $R \geqslant 1$ is the radius of convergence of the PGF.

Proof. The proof is deferred to Appendix B.2.   □

Let $G(z)$ be the PGF of a random variable X. Thus, $G(z) = z^0 \, \mathbb{P}(X = 0) + z^1\, \mathbb{P}(X = 1) + z^2\, \mathbb{P}(X = 2) + \cdots$. Substituting $z = 0$ gives $G(0) = \mathbb{P}(X = 0)$. By Theorem 2.6 the derivative of G is

$$G'(z) = \mathbb{P}(X = 1) + 2z\, \mathbb{P}(X = 2) + 3z^2\, \mathbb{P}(X = 3) + \cdots\,.$$

In particular, $G'(0) = \mathbb{P}(X = 1)$. By differentiating $G'(z)$, we see that the second derivative of G at 0 is $G''(0) = 2\,\mathbb{P}(X = 2)$. Repeating this procedure gives the following corollary to Theorem 2.6.

Corollary 2.1. (Probabilities from PGFs). Let X be a nonnegative integer-valued random variable with PGF $G(z)$. Then,

$$\mathbb{P}(X = k) = \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}G(0) \ .$$

The PGF thus uniquely determines the discrete pdf. Another consequence of Theorem 2.6 is that expectations, variances, and moments can be easily found from the PGF.

Corollary 2.2. (Moments from PGFs). Let X be a nonnegative integer-valued random variable with PGF $G(z)$ and kth derivative $G^{(k)}(z)$. Then,

$$\lim_{\substack{z \to 1 \\ |z| < 1}} \frac{\mathrm{d}^k}{\mathrm{d}z^k}G(z) = \mathbb{E}\left[X(X - 1)\cdots(X - k + 1)\right] \ . \tag{2.10}$$

In particular, if the expectation and variance of X are finite, then $\mathbb{E}X = G'(1)$ and $\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2$.

Proof. The proof is deferred to Appendix B.2.   □   ☞ 369

Definition 2.9. (Moment Generating Function). The **moment generating function** (MGF) of a random variable X is the function $M : \mathbb{R} \to [0, \infty]$ given by

$$M(s) = \mathbb{E}\,\mathrm{e}^{sX} \ .$$

In particular, for a discrete random variable with pdf f,

$$M(s) = \sum_x \mathrm{e}^{sx}\,f(x) \ ,$$

and for a continuous random variable with pdf f,

$$M(s) = \int_{-\infty}^{\infty} \mathrm{e}^{sx}\,f(x)\,\mathrm{d}x \ .$$

Note that $M(s)$ can be infinite for certain values of s. We sometimes write M_X to stress the role of X.

Similar to the PGF, the MGF has the **uniqueness property**: two MGFs are the same if and only if their corresponding cdfs are the same. In addition, the integer moments of X can be computed from the derivatives of M, as summarized in the next theorem. The proof is similar to that of Theorem 2.6 and Corollary 2.2 and is given in Appendix B.3.

Theorem 2.7. (Moments from MGFs). If the MGF is finite in an open interval containing 0, then all moments $\mathbb{E}X^n$, $n = 0, 1, \ldots$ are finite and satisfy

$$\mathbb{E}X^n = M^{(n)}(0) \,,$$

where $M^{(n)}(0)$ is the nth derivative of M evaluated at 0.

Note that under the conditions of Theorem 2.7, the variance of X can be obtained from the MGF as

$$\mathrm{Var}(X) = M''(0) - (M'(0))^2 \,.$$

2.5   Common Discrete Distributions

In this section we give a number of common discrete distributions and list some of their properties. Note that the discrete pdf of each of these distributions, denoted f, depends on one or more parameters; so in fact we are dealing with *families* of distributions.

2.5.1   *Bernoulli Distribution*

Definition 2.10.   (Bernoulli Distribution). A random variable X is said to have a **Bernoulli** distribution with success probability p if X can only assume the values 0 and 1, with probabilities

$$f(0) = \mathbb{P}(X = 0) = 1 - p \quad \text{and} \quad f(1) = \mathbb{P}(X = 1) = p \,.$$

We write $X \sim \mathsf{Ber}(p)$.

The Bernoulli distribution is the most fundamental of all probability distributions. It models a single coin toss experiment. Three important properties of the Bernoulli are summarized in the following theorem.

Theorem 2.8. (Properties of the Bernoulli Distribution). Let $X \sim \text{Ber}(p)$. Then,

1. $\mathbb{E}X = p$,
2. $\text{Var}(X) = p(1 - p)$,
3. the PGF is $G(z) = 1 - p + zp$.

Proof. The expectation and the variance of X can be obtained by direct computation:

$$\mathbb{E}X = 0 \times \mathbb{P}(X = 0) + 1 \times \mathbb{P}(X = 1) = 0 \times (1 - p) + 1 \times p = p$$

and

$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X - (\mathbb{E}X)^2 = p - p^2 = p(1 - p) ,$$

where we have used the fact that in this case $X^2 = X$. Finally, the PGF is given by $G(z) = z^0(1 - p) + z^1 p = 1 - p + zp$. $\qquad\square$

2.5.2 Binomial Distribution

Definition 2.11. (Binomial Distribution). A random variable X is said to have a **binomial** distribution with parameters n and p if X has pdf

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n . \qquad (2.11)$$

We write $X \sim \text{Bin}(n, p)$.

From Example 2.2 we see that X can be interpreted as the total number of Heads in n successive coin flip experiments, with probability of Heads equal to p. An example of the graph of the pdf is given in Fig. 2.6. Theorem 2.9 lists some important properties of the binomial distribution.

Fig. 2.6 The pdf of the
Bin$(10, 0.7)$ distribution
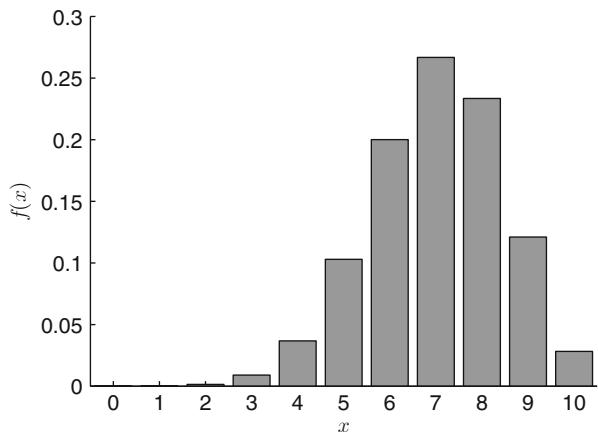
Theorem 2.9. (Properties of the Binomial Distribution). Let $X \sim$ Bin(n, p). Then,

1. $\mathbb{E}X = np$,
2. Var$(X) = np(1 - p)$,
3. the PGF is $G(z) = (1 - p + zp)^n$.

Proof. Using Newton's binomial formula:

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} ,$$

we see that

$$G(z) = \sum_{k=0}^{n} z^k \binom{n}{k} p^k (1 - p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (z\,p)^k (1 - p)^{n-k} = (1 - p + zp)^n .$$

From Corollary 2.2 we obtain the expectation and variance via $G'(1) = np$ and $G''(1) + G'(1) - (G'(1))^2 = (n-1)np^2 + np - n^2 p^2 = np(1 - p)$. $\qquad\square$

2.5.3   Geometric Distribution

Definition 2.12. (Geometric Distribution). A random variable X is said to have a **geometric** distribution with parameter p if X has pdf

$$f(x) = \mathbb{P}(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, 3, \ldots . \tag{2.12}$$

We write $X \sim$ Geom(p).

From Example 1.13 we see that X can be interpreted as the number of tosses
needed until the first Heads occurs in a sequence of coin tosses, with the probability
of Heads equal to p. An example of the graph of the pdf is given in Fig. 2.7.
Theorem 2.10 summarizes some properties of the geometric distribution.
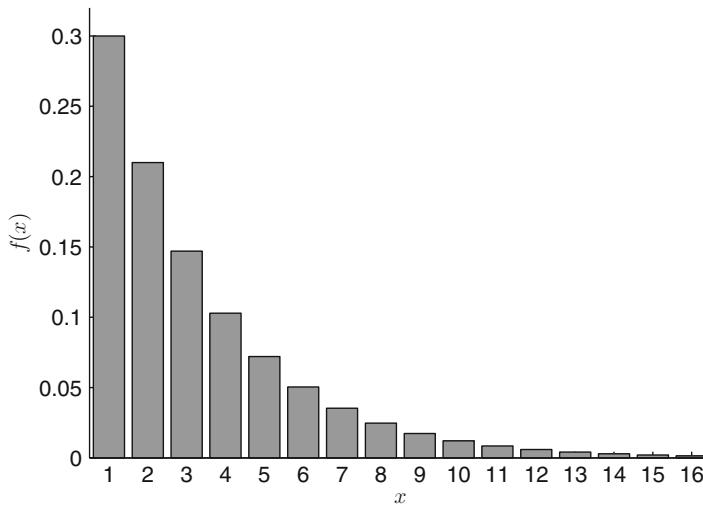
Fig. 2.7 The pdf of the Geom(0.3) distribution

Theorem 2.10. (Properties of the Geometric Distribution). Let $X \sim$
Geom(p). Then,

1. $\mathbb{E}X = 1/p$,
2. $\mathrm{Var}(X) = (1 - p)/p^2$,
3. the PGF is

$$G(z) = \frac{z\,p}{1 - z\,(1 - p)} , \quad |z| < \frac{1}{1 - p} . \tag{2.13}$$

Proof. The PGF of X follows from

$$G(z) = \sum_{x=1}^{\infty} z^x p (1 - p)^{x-1} = z\,p \sum_{k=0}^{\infty} (z(1 - p))^k = \frac{z\,p}{1 - z\,(1 - p)} ,$$

using the well-known result for *geometric sums*: $1 + a + a^2 + \cdots = (1 - a)^{-1}$, for
$|a| < 1$. By Corollary 2.2 the expectation is therefore

$$\mathbb{E}X = G'(1) = \frac{1}{p} \ .$$

By differentiating the PGF twice we find the variance:

$$\text{Var}(X) = G''(1) + G'(1) - (G''(1))^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2} \ . \qquad \square$$

One property of the geometric distribution that deserves extra attention is the **memoryless property**. Consider again the coin toss experiment. Suppose we have tossed the coin k times without a success (Heads). What is the probability that we need more than x additional tosses before getting a success? The answer is, obviously, the same as the probability that we require more than x tosses if we start from scratch, that is, $\mathbb{P}(X > x) = (1-p)^x$, irrespective of k. The fact that we have already had k failures does not make the event of getting a success in the next trial(s) any more likely. In other words, the coin does not have a memory of what happened—hence the name memoryless property.

> **Theorem 2.11. (Memoryless Property).** Let $X \sim \text{Geom}(p)$. Then for any $x, k = 1, 2, \ldots$,
>
> $$\mathbb{P}(X > k + x \mid X > k) = \mathbb{P}(X > x) \ .$$

☞ 12   *Proof.* By the definition of conditional probability,

$$\mathbb{P}(X > k + x \mid X > k) = \frac{\mathbb{P}(\{X > k + x\} \cap \{X > k\})}{\mathbb{P}(X > k)} \ .$$

The event $\{X > k + x\}$ is a subset of $\{X > k\}$; hence their intersection is $\{X > k + x\}$. Moreover, the probabilities of the events $\{X > k + x\}$ and $\{X > k\}$ are $(1-p)^{k+x}$ and $(1-p)^k$, respectively. Therefore,

$$\mathbb{P}(X > k + x \mid X > k) = \frac{(1-p)^{k+x}}{(1-p)^k} = (1-p)^x = \mathbb{P}(X > x) \ ,$$

as required. $\square$

2.5.4 Poisson Distribution

> **Definition 2.13. (Poisson Distribution).** A random variable X is said to have a **Poisson** distribution with parameter $\lambda > 0$ if X has pdf

(continued)

(continued)

$$f(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots. \tag{2.14}$$

We write $X \sim \mathsf{Poi}(\lambda)$.

The Poisson distribution may be viewed as the limit of the $\mathsf{Bin}(n, \lambda/n)$ distribution. Namely, if $X_n \sim \mathsf{Bin}(n, \lambda/n)$, then

$$
\begin{aligned}
\mathbb{P}(X_n = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \frac{n \times (n-1) \times \cdots \times (n-x+1)}{n \times n \times \cdots \times n} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}.
\end{aligned}
$$

As $n \to \infty$ the second and fourth factors converge to 1 and the third factor to $e^{-\lambda}$ (this is one of the defining properties of the exponential function). Hence, we have

$$\lim_{n \to \infty} \mathbb{P}(X_n = x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

An example of the graph of the Poisson pdf is given in Fig. 2.8. Theorem 2.12 summarizes some properties of the Poisson distribution.
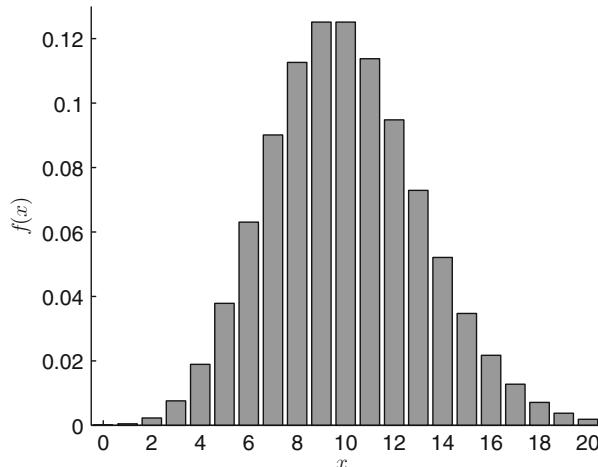
Fig. 2.8 The pdf of the $\mathsf{Poi}(10)$ distribution

Theorem 2.12. (Properties of the Poisson Distribution). Let $X \sim \mathsf{Poi}(\lambda)$. Then,

1. $\mathbb{E}X = \lambda$,
2. $\mathrm{Var}(X) = \lambda$,
3. the PGF is $G(z) = \mathrm{e}^{-\lambda(1-z)}$.

☞ 34   *Proof.* The PGF was derived in Example 2.6. It follows from Corollary 2.2 that $\mathbb{E}X = G'(1) = \lambda$ and

$$\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda .$$

Thus, the parameter λ can be interpreted as both the expectation and variance of X.

□

2.6   Common Continuous Distributions

In this section we give a number of common continuous distributions and list some of their properties. Note that the pdf of each of these distributions depends on one or more parameters; so, as in the previous section, we are dealing with *families* of distributions.

2.6.1   *Uniform Distribution*

Definition 2.14. (Uniform Distribution). A random variable X is said to have a **uniform** distribution on the interval $[a, b]$ if its pdf is given by

$$f(x) = \frac{1}{b - a}, \quad a \leqslant x \leqslant b .$$

We write $X \sim \mathsf{U}[a, b]$ (and $X \sim \mathsf{U}(a, b)$ for a uniform random variable on an open interval (a, b)).

Fig. 2.9 The pdf of the
uniform distribution on $[a, b]$
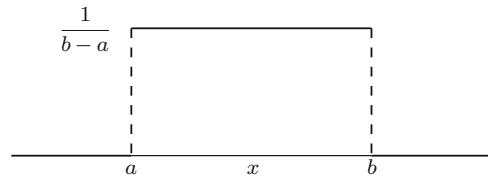
The random variable $X \sim \mathsf{U}[a, b]$ can model a randomly chosen point from the interval $[a, b]$, where each choice is equally likely. A graph of the pdf is given in Fig. 2.9.

Theorem 2.13. (Properties of the Uniform Distribution). Let $X \sim \mathsf{U}[a, b]$. Then,

1. $\mathbb{E}X = (a + b)/2$,
2. $\mathrm{Var}(X) = (b - a)^2/12$.

Proof. We have

$$\mathbb{E}X = \int_a^b \frac{x}{b - a}\, \mathrm{d}x = \frac{1}{b - a}\left[\frac{b^2 - a^2}{2}\right] = \frac{a + b}{2}$$

and

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_a^b \frac{x^2}{b - a}\, \mathrm{d}x - \left(\frac{a + b}{2}\right)^2$$

$$= \frac{b^3 - a^3}{3(b - a)} - \left(\frac{a + b}{2}\right)^2 = \frac{(b - a)^2}{12} \, .$$

$\square$

2.6.2   Exponential Distribution

Definition 2.15. (Exponential Distribution). A random variable X is said to have an **exponential** distribution with parameter λ if its pdf is given by

$$f(x) = \lambda\, \mathrm{e}^{-\lambda x}, \quad x \geqslant 0 \, . \qquad (2.15)$$

We write $X \sim \mathsf{Exp}(\lambda)$.

The exponential distribution can be viewed as a continuous version of the geometric distribution. Graphs of the pdf for various values of λ are given in Fig. 2.10. Theorem 2.14 summarizes some properties of the exponential distribution.
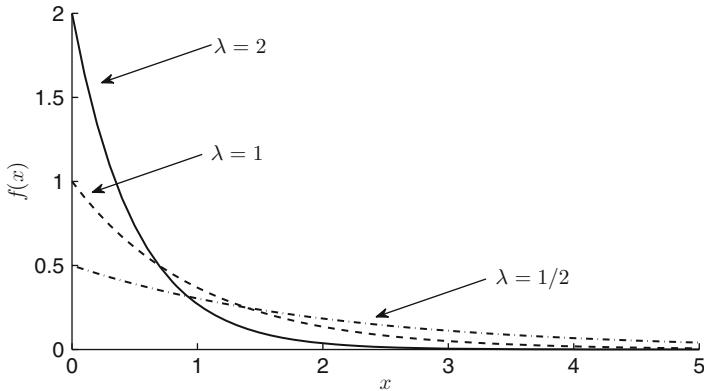
Fig. 2.10 The pdf of the $\text{Exp}(\lambda)$ distribution for various λ

Theorem 2.14. (Properties of the Exponential Distribution). Let $X \sim \text{Exp}(\lambda)$. Then,

1. $\mathbb{E}X = 1/\lambda$,
2. $\text{Var}(X) = 1/\lambda^2$,
3. the MGF of X is $M(s) = \lambda/(\lambda - s)$, $s < \lambda$,
4. the cdf of X is $F(x) = 1 - e^{-\lambda x}$, $x \geqslant 0$,
5. the **memoryless property** holds: for any $s, t > 0$,

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t) \, . \tag{2.16}$$

Proof.

3. The MGF is given by

$$M(s) = \int_0^\infty e^{sx} \lambda e^{-\lambda x} \mathrm{d}x = \lambda \int_0^\infty e^{-(\lambda - s)x} \, \mathrm{d}x = \lambda \left[\frac{-e^{-(\lambda - s)x}}{\lambda - s} \right]_0^\infty$$

$$= \frac{\lambda}{\lambda - s}, \quad s < \lambda \quad (\text{and } M(s) = \infty \text{ for } s \geqslant \lambda).$$

☞ 36   1. From Theorem 2.7, we obtain

$$\mathbb{E}X = M'(0) = \left. \frac{\lambda}{(\lambda - s)^2} \right|_{s=0} = \frac{1}{\lambda} \, .$$

2. Similarly, the second moment is $\mathbb{E}X^2 = M''(0) = \frac{2\lambda}{(\lambda-s)^3}\big|_{s=0} = 2/\lambda^2$, so that the variance is

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \;.$$

4. The cdf of X is given by

$$F(x) = \mathbb{P}(X \leqslant x) = \int_0^x \lambda e^{-\lambda u}\mathrm{d}u = \left[-e^{-\lambda u}\right]_0^x = 1 - e^{-\lambda x}, \quad x \geqslant 0 \;.$$

Note that the tail probability $\mathbb{P}(X > x)$ is exponentially decaying:

$$\mathbb{P}(X > x) = e^{-\lambda x}, \quad x \geqslant 0 \;.$$

5. Similar to the proof of the memoryless property for the geometric distribution (Theorem 2.11), we have

$$\mathbb{P}(X > s + t \mid X > s) = \frac{\mathbb{P}(X > s + t,\, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)}$$

$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t) \;. \qquad \square$$

The memoryless property can be interpreted as a "non-aging" property. For example, when X denotes the lifetime of a machine, then, given the fact that the machine is alive at time s, the remaining lifetime of the machine, $X - s$, has the same exponential distribution as a completely new machine. In other words, the machine has no memory of its age and does not deteriorate (although it will break down eventually).

2.6.3 Normal (Gaussian) Distribution

In this section we introduce the most important distribution in the study of statistics: the normal (or Gaussian) distribution. Additional properties of this distribution will be given in Sect. 3.6.

Definition 2.16. (Normal Distribution). A random variable X is said to have a **normal** distribution with parameters μ and σ^2 if its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R} \;. \qquad (2.17)$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

The parameters μ and σ^2 turn out to be the expectation and variance of the distribution, respectively. If $\mu = 0$ and $\sigma = 1$, then

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

and the distribution is known as the **standard normal** distribution. The cdf of the standard normal distribution is often denoted by Φ and its pdf by φ. In Fig. 2.11 the pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution is plotted for various μ and σ^2.
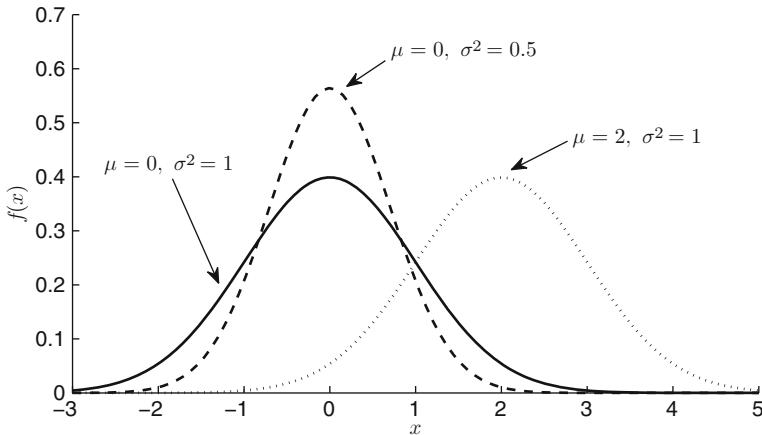
Fig. 2.11   The pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution for various μ and σ^2

We next consider some important properties of the normal distribution.

Theorem 2.15. (Standardization). Let $X \sim \mathsf{N}(\mu, \sigma^2)$ and define $Z = (X - \mu)/\sigma$. Then Z has a standard normal distribution.

Proof. The cdf of Z is given by

$$\mathbb{P}(Z \leqslant z) = \mathbb{P}((X - \mu)/\sigma \leqslant z) = \mathbb{P}(X \leqslant \mu + \sigma z)$$

$$= \int_{-\infty}^{\mu + \sigma z} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2} dy = \Phi(z) ,$$

where we make a change of variable $y = (x - \mu)/\sigma$ in the fourth equation. Hence, $Z \sim \mathsf{N}(0, 1)$.   □

The rescaling procedure in Theorem 2.15 is called **standardization**. It follows from Theorem 2.15 that any $X \sim \mathsf{N}(\mu, \sigma^2)$ can be written as

$$X = \mu + \sigma Z, \quad \text{where } Z \sim \mathsf{N}(0,1) .$$

In other words, any normal random variable can be viewed as an **affine transformation**—that is, a linear transformation plus a constant—of a standard normal random variable.

Next we prove the earlier claim that the parameters μ and σ^2 are, respectively, the expectation and variance of the distribution.

Theorem 2.16. (Expectation and Variance for the Normal Distribution).
If $X \sim \mathsf{N}(\mu, \sigma^2)$, then $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof. Since the pdf is symmetric around μ and $\mathbb{E}X < \infty$, it follows that $\mathbb{E}X = \mu$. To show that the variance of X is σ^2, we first write $X = \mu + \sigma Z$, where $Z \sim \mathsf{N}(0,1)$. Then, $\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2 \mathrm{Var}(Z)$. Hence, it suffices to show that $\mathrm{Var}(Z) = 1$. Now, since the expectation of Z is 0, we have

$$\mathrm{Var}(Z) = \mathbb{E}Z^2 = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2} \, \mathrm{d}z = \int_{-\infty}^{\infty} z \times \frac{z}{\sqrt{2\pi}} \, e^{-z^2/2} \, \mathrm{d}z .$$

We apply integration by parts to the last integral to find

$$\mathbb{E}Z^2 = \left[-\frac{z}{\sqrt{2\pi}} \, e^{-z^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2} \, \mathrm{d}z = 1 ,$$

since the last integrand is the pdf of the standard normal distribution. $\square$

Theorem 2.17. (MGF for the Normal Distribution). The MGF of $X \sim \mathsf{N}(\mu, \sigma^2)$ is

$$\mathbb{E}e^{sX} = e^{s\mu + s^2\sigma^2/2}, \quad s \in \mathbb{R} . \tag{2.18}$$

Proof. Write $X = \mu + \sigma Z$, where $Z \sim \mathsf{N}(0,1)$. We have

$$\mathbb{E}e^{sZ} = \int_{-\infty}^{\infty} e^{sz} \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2} \, \mathrm{d}z = e^{s^2/2} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} \, e^{-(z-s)^2/2}}_{\text{pdf of } \mathsf{N}(s,1)} \, \mathrm{d}z = e^{s^2/2} ,$$

so that $\mathbb{E}e^{sX} = \mathbb{E}e^{s(\mu + \sigma Z)} = e^{s\mu} \, \mathbb{E}e^{s\sigma Z} = e^{s\mu} e^{\sigma^2 s^2/2} = e^{s\mu + \sigma^2 s^2/2}$. $\square$

2.6.4 Gamma and χ^2 Distribution

Definition 2.17. (Gamma Distribution). A random variable X is said to have a **gamma** distribution with **shape** parameter $\alpha > 0$ and **scale** parameter $\lambda > 0$ if its pdf is given by

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geqslant 0 , \tag{2.19}$$

where Γ is the gamma function. We write $X \sim \mathsf{Gamma}(\alpha, \lambda)$.

The **gamma function** $\Gamma(\alpha)$ is an important special function in mathematics, defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \, e^{-u} \, du . \tag{2.20}$$

We mention a few properties of the Γ function:

1. $\Gamma(\alpha + 1) = \alpha \, \Gamma(\alpha)$ for $\alpha \in \mathbb{R}_+$.
2. $\Gamma(n) = (n - 1)!$ for $n = 1, 2, \ldots..$
3. $\Gamma(1/2) = \sqrt{\pi}$.

Two special cases of the $\mathsf{Gamma}(\alpha, \lambda)$ distribution are worth mentioning. Firstly, the $\mathsf{Gamma}(1, \lambda)$ distribution is simply the $\mathsf{Exp}(\lambda)$ distribution. Secondly, the $\mathsf{Gamma}(n/2, \ 1/2)$ distribution, where $n \in \{1, 2, \ldots\}$, is called the **chi-squared** distribution with n **degrees of freedom**. We write $X \sim \chi_n^2$. A graph of the pdf of the χ_n^2 distribution for various n is given in Fig. 2.12.
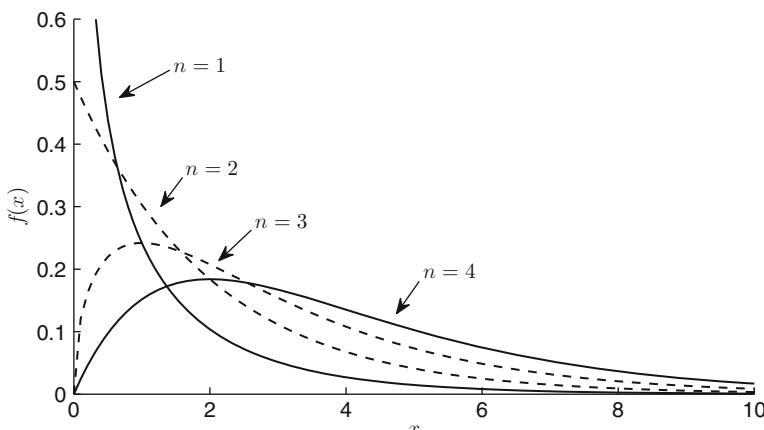
Fig. 2.12 The pdf of the χ_n^2 distribution for various degrees of freedom n

The following theorem summarizes some properties of the gamma distribution.

Theorem 2.18. (Properties of the Gamma Distribution). Let $X \sim$ Gamma(α, λ). Then,

1. $\mathbb{E}X = \alpha/\lambda$,
2. $\mathrm{Var}(X) = \alpha/\lambda^2$,
3. the MGF is $M(s) = [\lambda/(\lambda - s)]^\alpha$, $s < \lambda$ (and ∞ otherwise).

Proof.

3. For $s < \lambda$, the MGF of X at s is given by

$$M(s) = \mathbb{E}\, e^{sX} = \int_0^\infty \frac{e^{-\lambda x}\, \lambda^\alpha\, x^{\alpha-1}}{\Gamma(\alpha)}\, e^{sx}\, \mathrm{d}x$$

$$= \left(\frac{\lambda}{\lambda - s}\right)^\alpha \int_0^\infty \underbrace{\frac{e^{-(\lambda-s)x}\, (\lambda - s)^\alpha\, x^{\alpha-1}}{\Gamma(\alpha)}}_{\text{pdf of Gamma}(\alpha, \lambda - s)}\, \mathrm{d}x$$

$$= \left(\frac{\lambda}{\lambda - s}\right)^\alpha . \tag{2.21}$$

1. Consequently, by Theorem 2.7,   ☞ 36

$$\mathbb{E}X = M'(0) = \frac{\alpha}{\lambda} \left(\frac{\lambda}{\lambda - s}\right)^{\alpha+1}\Bigg|_{s=0} = \frac{\alpha}{\lambda}.$$

2. Similarly, $\mathrm{Var}(X) = M''(0) - (M'(0))^2 = (\alpha + 1)\alpha/\lambda^2 - (\alpha/\lambda)^2 = \alpha/\lambda^2$. □

2.6.5 F Distribution

Definition 2.18. (F Distribution). Let m and n be strictly positive integers. A random variable X is said to have an **F distribution** with **degrees of freedom** m and n if its pdf is given by

$$f(x) = \frac{\Gamma(\frac{m+n}{2})\, (m/n)^{m/2} x^{(m-2)/2}}{\Gamma(\frac{m}{2})\, \Gamma(\frac{n}{2})\, [1 + (m/n)x]^{(m+n)/2}}, \quad x \geqslant 0, \tag{2.22}$$

where Γ denotes the gamma function. We write $X \sim \mathsf{F}(m, n)$.

The F distribution plays an important role in classical statistics, through
Theorem 3.11. A graph of the pdf of the $\mathsf{F}(m, n)$ distribution for various m and n is given in Fig. 2.13.
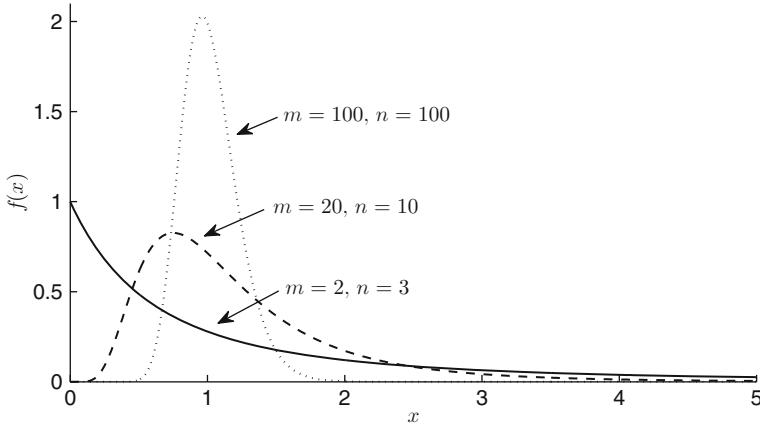
Fig. 2.13 The pdf of the $\mathsf{F}(m, n)$ distribution for various degrees of freedom m and n

2.6.6 Student's t Distribution

Definition 2.19. (Student's t Distribution). A random variable X is said to have a **Student's t** distribution with parameter $\nu > 0$ if its pdf is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}, \qquad (2.23)$$

where Γ denotes the gamma function. We write $X \sim \mathsf{t}_\nu$. For integer values the parameter ν is referred to as the **degrees of freedom** of the distribution.

A graph of the pdf of the t_ν distribution for various ν is given in Fig. 2.14. Note that the pdf is symmetric. Moreover, it can be shown that the pdf of the t_ν distribution converges to the pdf of the $\mathsf{N}(0, 1)$ distribution as $\nu \to \infty$. The t_1 distribution is called the **Cauchy distribution**.

For completeness we mention that if $X \sim \mathsf{t}_\nu$, then

$$\mathbb{E}X = 0 \quad (\nu > 1) \quad \text{and} \quad \mathrm{Var}(X) = \frac{\nu}{\nu - 2}, \quad (\nu > 2)\,.$$
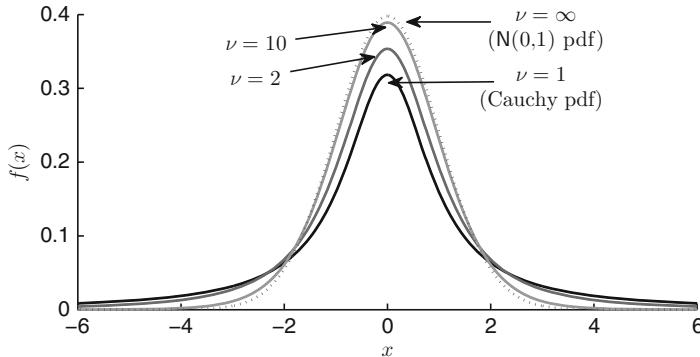
Fig. 2.14 The pdfs of t_1 (Cauchy), t_2, t_{10}, and t_∞ (N(0, 1)) distributions

The t and F distributions are related in the following way.

Theorem 2.19. (Relationship Between the t and F Distribution). For integer $n \geqslant 1$, if $X \sim t_n$, then $X^2 \sim F(1, n)$.

Proof. Let $Z = X^2$. We can express the cdf of Z in terms of the cdf of X. Namely, for every $z > 0$, we have

$$F_Z(z) = \mathbb{P}(X^2 \leqslant z) = \mathbb{P}(-\sqrt{z} \leqslant X \leqslant \sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z}) .$$

Differentiating with respect to z gives the following relation between the two pdfs:

$$f_Z(z) = f_X(\sqrt{z}) \frac{1}{2\sqrt{z}} + f_X(-\sqrt{z}) \frac{1}{2\sqrt{z}} = f_X(\sqrt{z}) \frac{1}{\sqrt{z}} ,$$

using the symmetry of the t distribution. Substituting (2.23) into the last equation yields

$$f_Z(z) = c(n) \frac{z^{-1/2}}{(1 + z/n)^{(n+1)/2}}, \quad z > 0$$

for some constant $c(n)$. The only pdf of this form is that of the $F(1, n)$ distribution. $\square$

2.7 Generating Random Variables

This section shows how to generate random variables on a computer. We first discuss a modern uniform random generator and then introduce two general methods

for drawing from an arbitrary one-dimensional distribution: the inverse-transform method and the acceptance–rejection method.

2.7.1　Generating Uniform Random Variables

The MATLAB `rand` function simulates the drawing of a uniform random number on the interval $(0, 1)$ by generating *pseudo*random numbers, that is, numbers that, although not actually random (because the computer is a deterministic device), behave for all intended purposes as truly random. The following algorithm (L'Ecuyer 1999) uses simple recurrences to produce high-quality pseudorandom numbers, in the sense that the numbers pass all currently known statistical tests for randomness and uniformity.

Algorithm 2.1. (Combined Multiple-Recursive Generator).

1. Suppose N random numbers are required. Define $m_1 = 2^{32} - 209$ and $m_2 = 2^{32} - 22853$.
2. Initialize a vector $(X_{-2}, X_{-1}, X_0) = (12345, 12345, 12345)$ and a vector $(Y_{-2}, Y_{-1}, Y_0) = (12345, 12345, 12345)$.
3. For $t = 1$ to N let

$$X_t = (1403580 \, X_{t-2} - 810728 \, X_{t-3}) \bmod m_1 \, ,$$

$$Y_t = (527612 \, Y_{t-1} - 1370589 \, Y_{t-3}) \bmod m_2 \, ,$$

and output the tth random number as

$$U_t = \begin{cases} \dfrac{X_t - Y_t + m_1}{m_1 + 1} & \text{if } X_t \leqslant Y_t \, , \\ \dfrac{X_t - Y_t}{m_1 + 1} & \text{if } X_t > Y_t \, . \end{cases}$$

Here, $x \bmod m$ means the remainder of x when divided by m. The initialization in Step 2 determines the initial state—the so-called **seed**—of the random number stream. Restarting the stream from the same seed produces the same sequence.

Algorithm 2.1 is implemented as a core MATLAB uniform random number generator from Version 7. Currently the default generator in MATLAB is the *Mersenne twister*, which also passes (most) statistical tests and tends to be a little faster. However, it is considerably more difficult to implement. A typical usage of MATLAB's uniform random number generator is as follows.

```
>>rng(1,'combRecursive') % use the CMRG with seed 1
>>rand(1,5)    % draw 5 random numbers

ans =
    0.4957    0.2243    0.2073    0.6823    0.6799

>>rng(1234)   % set the seed to 1234
>>rand(1,5)

ans =
    0.2830    0.2493    0.3600    0.9499    0.8071

>>rng(1234)   % reset the seed to 1234

>>rand(1,5)

ans =
    0.2830    0.2493    0.3600    0.9499    0.8071
```

2.7.2   Inverse-Transform Method

Once we have a method for drawing a uniform random number, we can, in principle, simulate a random variable X from *any* cdf F by using the following algorithm.

Algorithm 2.2. (Inverse-Transform Method).

1. Generate U from $\mathsf{U}(0, 1)$.
2. Return $X = F^{-1}(U)$, where F^{-1} is the inverse function of F.
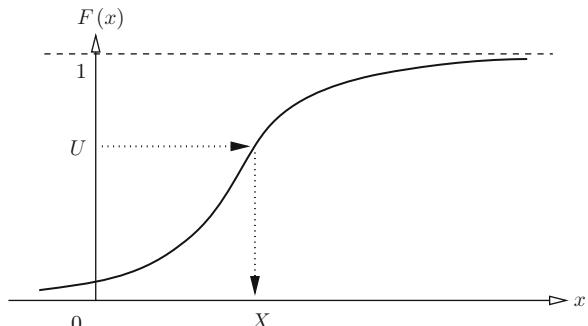
Fig. 2.15 The inverse-transform method

Figure 2.15 illustrates the inverse-transform method. We see that the random variable $X = F^{-1}(U)$ has cdf F, since

$$\mathbb{P}(X \leqslant x) = \mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x) \,. \qquad (2.24)$$

Example 2.7 (Generating Uniformly on a Unit Disk). Suppose we wish to draw a random point (X, Y) uniformly on the unit disk; see Fig. 2.16. In polar coordinates we have $X = R \cos \Theta$ and $Y = R \sin \Theta$, where Θ has a $\mathsf{U}(0, 2\pi)$ distribution. The cdf of R is given by

$$F(r) = \mathbb{P}(R \leqslant r) = \frac{\pi r^2}{\pi} = r^2, \quad 0 < r < 1 \,.$$

Its inverse is $F^{-1}(u) = \sqrt{u}$, $0 < u < 1$. We can thus generate R via the inverse-transform method as $R = \sqrt{U_1}$, where $U_1 \sim \mathsf{U}(0, 1)$. In addition, we can simulate Θ as $\Theta = 2\pi U_2$, where $U_2 \sim \mathsf{U}(0, 1)$. Note that U_1 and U_2 should be independent draws from $\mathsf{U}(0, 1)$.

Fig. 2.16 Draw a point (X, Y) uniformly on the unit disk
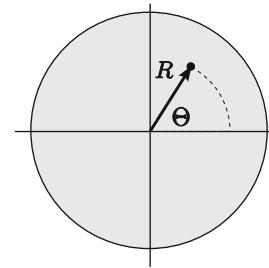
The inverse-transform method holds for general cdfs F. Note that F for discrete random variables is a step function, as illustrated in Fig. 2.17. The algorithm for generating a random variable X from a discrete distribution that takes values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ is thus as follows.

Algorithm 2.3. (Discrete Inverse-Transform Method).

1. Generate $U \sim \mathsf{U}(0, 1)$.
2. Find the smallest positive integer k such that $F(x_k) \geqslant U$ and return $X = x_k$.
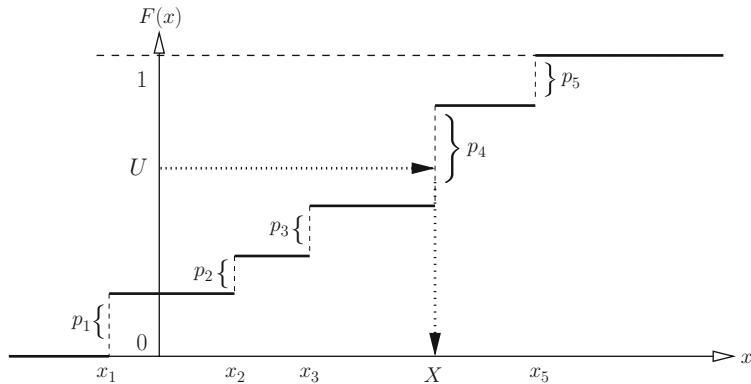
Fig. 2.17 The inverse-transform method for a discrete random variable

Drawing one of the numbers $1, \ldots, n$ according to a probability vector $(p_1, \ldots, p_n)$ can be done in one line of MATLAB code:

```
min(find(cumsum(p) > rand));
```

Here p is the vector of probabilities, such as $(0.3, 0.2, 0.5)$, `cumsum` gives the cumulative vector, e.g., $(0.3, 0.5, 1)$, `find(⋯)` finds the indices i such that the cumulative probability is greater than some random number `rand`, and `min` takes the smallest of these indices.

2.7.3 Acceptance–Rejection Method

The inverse-transform method may not always be easy to implement, in particular when the inverse cdf is difficult to compute. In that case the **acceptance–rejection** method may prove to be useful. The idea of this method is depicted in Fig. 2.18. Suppose we wish to sample from a pdf f. Let g be another pdf such that for some constant $C \geqslant 1$ we have that $C g(x) \geqslant f(x)$ for all x. It is assumed that it is easy to sample from g, for example, via the inverse-transform method.
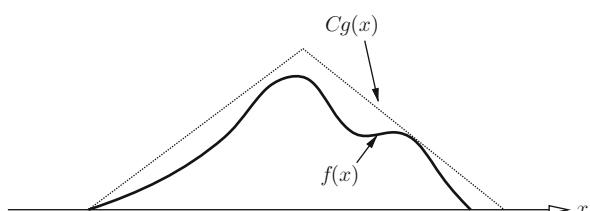
Fig. 2.18 Illustration of the acceptance–rejection method

It is intuitively clear that if a random point (X, Y) is *uniformly* distributed under the graph of f—that is, on the set $\{(x, y) : 0 \leqslant y \leqslant f(x)\}$—then X must have pdf f. To construct such a point, let us first draw a random point (Z, V) by drawing Z from g and then drawing V uniformly on $[0, C g(Z)]$. The point (Z, V) is uniformly distributed under the graph of Cg. If we keep drawing such a point (Z, V) *until it lies under the graph of f*, then the resulting point (X, Y) must be uniformly distributed under the graph of f and hence the X coordinate must have pdf f. This leads to the following algorithm.

Algorithm 2.4. (Acceptance–Rejection Method).

1. Generate $Z \sim g$.
2. Generate $Y \sim \mathsf{U}(0, C\, g(Z))$.
3. If $Y \leqslant f(Z)$, return $X = Z$; otherwise, repeat from Step 1.

Example 2.8 (Generating from the Standard Normal Distribution). To sample from the standard normal pdf via the inverse-transform method requires knowledge of the inverse of the corresponding cdf, which involves numerical integration. Instead, we can use acceptance–rejection. First, observe that the standard normal pdf is symmetric around 0. Hence, if we can generate a random variable X from the **positive normal** pdf (see Fig. 2.19),

$$f(x) = \sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-x^2/2}, \qquad x \geqslant 0 , \tag{2.25}$$

then we can generate a standard normal random variable by multiplying X with 1 or -1, each with probability $1/2$. We can bound $f(x)$ by $C\, g(x)$, where $g(x) = \mathrm{e}^{-x}$ is the pdf of the $\mathsf{Exp}(1)$ distribution. The smallest constant C such that $f(x) \leqslant Cg(x)$ is $\sqrt{2\mathrm{e}/\pi}$.
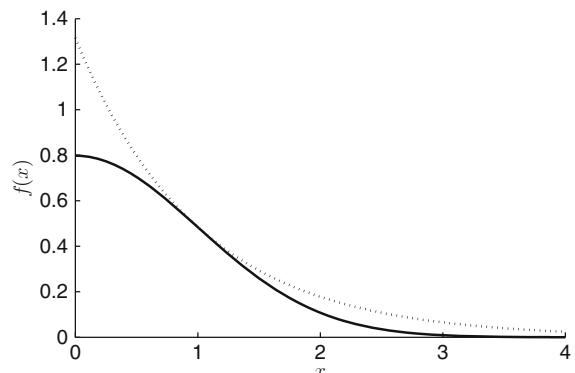
Fig. 2.19 Bounding the positive normal density (*solid curve*) via an Exp(1) pdf (times $C \approx 1.3155$)

Drawing from the Exp(1) distribution can be easily done via the inverse-transform method, noting that the corresponding cdf is the function $1 - e^{-x}, x \geqslant 0$, whose inverse is the function $-\ln(1 - u), u \in (0, 1)$. This gives the following specification of Algorithm 2.4, where f and C are defined above.

Algorithm 2.5. (N(0, 1) Generator).

1. Draw $U_1 \sim \mathsf{U}(0, 1)$, and let $Z = -\ln U_1$.
2. Draw $U_2 \sim \mathsf{U}(0, 1)$, and let $Y = U_2 \, C \, e^{-Z}$.
3. If $Y \leqslant f(Z)$, let $X = Z$ and continue with Step 4. Otherwise, repeat from Step 1.
4. Draw $U_3 \sim \mathsf{U}(0, 1)$ and return $\widetilde{X} = X \, (2 \, \mathrm{I}_{\{U_3 < 1/2\}} - 1)$ as a standard normal random variable.

In Step 1, we have used the fact that if $U \sim \mathsf{U}(0, 1)$, then also $1 - U \sim \mathsf{U}(0, 1)$. In Step 4, $\mathrm{I}_{\{U_3 < 1/2\}}$ denotes the **indicator** of the event $\{U_3 < 1/2\}$, which is 1 if $U_3 < 1/2$ and 0 otherwise. An alternative generation method is given in Algorithm 3.2. In MATLAB normal random variable generation is implemented via the `randn` function.

2.8 Problems

2.1. Two fair dice are thrown and the smallest of the face values, M say, is noted.

(a) Give the discrete pdf of M in table form, as in Table 2.1.
(b) What is the probability that M is at least 3?
(c) Calculate the expectation and variance of M.

2.2. A continuous random variable X has cdf

$$F(x) = \begin{cases} 0, & x < 0 \\ x^2/5, & 0 \leqslant x \leqslant 1 \\ \frac{1}{5}\left(-x^2 + 6x - 4\right), & 1 < x \leqslant 3 \\ 1, & x > 3 \,. \end{cases}$$

(a) Find the corresponding pdf and plot its graph.
(b) Calculate the following probabilities:

    (i) $\mathbb{P}(X \leqslant 2)$.
    (ii) $\mathbb{P}(1 < X \leqslant 2)$.
    (iii) $\mathbb{P}(1 \leqslant X \leqslant 2)$.
    (iv) $\mathbb{P}(X > 1/2)$.

(c) Show that $\mathbb{E}X = 22/15$.

2.3. In this book most random variables are either discrete or continuous; that is, they have either a discrete or a continuous pdf. It is also possible to define random variables that have a mix of discrete and continuous characteristics. A simple example is a random variable X with cdf

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - c\,e^{-x}, & x \geqslant 0 \end{cases}$$

for some fixed $0 < c < 1$.

(a) Sketch the cdf F.
(b) Find the following probabilities:

   (i) $\mathbb{P}(0 \leqslant X \leqslant x), x \geqslant 0$.
   (ii) $\mathbb{P}(0 < X \leqslant x), x \geqslant 0$.
   (iii) $\mathbb{P}(X = x), x \geqslant 0$.

(c) Describe how the inverse-transform method can be used to draw samples from this distribution.

2.4. Let X be a positive random variable with cdf F. Prove that

$$\mathbb{E}X = \int_0^\infty (1 - F(x))\,dx \ . \tag{2.26}$$

2.5. Let X be a random variable that can possibly take values $-\infty$ and ∞ with probabilities $\mathbb{P}(X = -\infty) = a$ and $\mathbb{P}(X = \infty) = b$, respectively. Show that the corresponding cdf F satisfies $\lim_{x\to-\infty} F(x) = a$ and $\lim_{x\to\infty} F(x) = 1 - b$.

2.6. Suppose that in a large population the fraction of left-handers is 12 %. We select at random 100 people from this population. Let X be the number of left-handers among the selected people. What is the distribution of X? What is the probability that at most 7 of the selected people are left-handed?

2.7. Let $X \sim \mathsf{Geom}(p)$. Show that

$$\mathbb{P}(X > k) = (1 - p)^k.$$

2.8. Find the MGF of $X \sim \mathsf{U}[a, b]$.

2.9. Let $X = a + (b - a)U$, where $U \sim \mathsf{U}[0, 1]$. Prove that $X \sim \mathsf{U}[a, b]$. Use this to provide a more elegant proof of Theorem 2.13.

2.10. Show that the exponential distribution is the *only* continuous (positive) distribution that possesses the memoryless property. Hint: show that the memoryless property implies that the tail probability $g(x) = \mathbb{P}(X > x)$ satisfies $g(x + y) = g(x)g(y)$.

2.11. Let $X \sim \text{Exp}(2)$. Calculate the following quantities:

(a) $\mathbb{P}(-1 \leqslant X \leqslant 1)$.
(b) $\mathbb{P}(X > 4)$.
(c) $\mathbb{P}(X > 4 \mid X > 2)$.
(d) $\mathbb{E}X^2$.

2.12. What is the expectation of a random variable X with the following discrete pdf on the set of integer numbers, excluding 0:

$$f(x) = \frac{3}{\pi^2} \frac{1}{x^2}, \quad x \in \mathbb{Z} \setminus \{0\}.$$

What is the pdf of the absolute value $|X|$ and what is its expectation?

2.13. A random variable X is said to have a **discrete uniform distribution** on the set $\{a, a+1, \ldots, b\}$ if

$$\mathbb{P}(X = x) = \frac{1}{b - a + 1}, \quad x = a, a+1, \ldots, b.$$

(a) What is the expectation of X?
(b) Show that $\text{Var}(X) = (b-a)(b-a+2)/12$.
(c) Find the PGF of X.
(d) Describe a simple way to generate X using a uniform number generator.

2.14. Let X and Y be random variables. Prove that if $X \leqslant Y$, then $\mathbb{E}X \leqslant \mathbb{E}Y$.

2.15. A continuous random variable is said to have a **logistic** distribution if its pdf is given by

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad x \in \mathbb{R}. \tag{2.27}$$

(a) Plot the graph of the pdf.
(b) Show that $\mathbb{P}(X > x) = 1/(1 + e^x)$ for all x.
(c) Write an algorithm based on the inverse-transform method to generate random variables from this distribution.

2.16. An electrical component has a lifetime (in years) that is distributed according to an exponential distribution with expectation 3. What is the probability that the component is still functioning after 4.5 years, given that it still works after 4 years? Answer the same question for the case where the component's lifetime is normally distributed with the same expected value and variance as before.

2.17. Consider the pdf given by

$$f(x) = \begin{cases} 4 e^{-4(x-1)}, & x \geqslant 1, \\ 0, & x < 1. \end{cases}$$

(a) If X is distributed according to this pdf f, what is its expectation?
(b) Specify how one can generate a random variable $X \sim f$ using a uniform random number generator.

2.18. Let $X \sim \mathsf{N}(4, 9)$.

(a) Plot the graph of the pdf.
(b) Express the following probabilities in terms of the cdf Φ of the standard normal distribution:

   (i) $\mathbb{P}(X \leqslant 3)$.
   (ii) $\mathbb{P}(X > 4)$.
   (iii) $\mathbb{P}(-1 \leqslant X \leqslant 5)$.

(c) Find $\mathbb{E}[2X + 1]$.
(d) Calculate $\mathbb{E}X^2$.

2.19. Let Φ be the cdf of $X \sim \mathsf{N}(0, 1)$. The integral

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du$$

needs to be evaluated numerically. In MATLAB there are several ways to do this:

(1) If the *Statistics Toolbox* is available, the cdf can be evaluated via the functions `normcdf` or `cdf`. The inverse cdf can be evaluated using `norminv` or `icdf`.
    See also their replacements `cumdf` and `icumdf` in Appendix A.9.
(2) Or one could use the built-in **error function** `erf`, defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-u^2} \, du \, , \quad x \in \mathbb{R} \, .$$

    The inverse of the error function, erf^{-1}, is implemented in MATLAB as `erfinv`.
(3) A third alternative is to use numerical integration (quadrature) via the `quad` function. For example, `quad(@f,0,1)` integrates a MATLAB function `f.m` on the interval $[0, 1]$.

(a) Show that $\Phi(x) = (\mathrm{erf}(x/\sqrt{2}) + 1)/2$.
(b) Evaluate $\Phi(x)$ for $x = 1, 2$, and 3 via (a) the error function and (b) numerical integration of the pdf, using the fact that $\Phi(0) = 1/2$.
(c) Show that the inverse of Φ is given by

$$\Phi^{-1}(y) = \sqrt{2} \, \mathrm{erf}^{-1}(2y - 1) \, , \quad 0 < y < 1 \, .$$

2.20. Based on MATLAB's `rand` and `randn` functions *only*, implement algorithms that generate random variables from the following distributions:

(a) $\mathsf{U}[2, 3]$.
(b) $\mathsf{N}(3, 9)$.
(c) $\mathsf{Exp}(4)$.
(d) $\mathsf{Bin}(10, 1/2)$.
(e) $\mathsf{Geom}(1/6)$.

2.21. The **Weibull** distribution $\mathsf{Weib}(\alpha, \lambda)$ has cdf

$$F(x) = 1 - e^{-(\lambda x)^\alpha}, \quad x \geqslant 0 . \tag{2.28}$$

It can be viewed as a generalization of the exponential distribution. Write a MATLAB program that draws 1000 samples from the $\mathsf{Weib}(2, 1)$ distribution using the inverse-transform method. Give a histogram of the sample.

2.22. Consider the pdf

$$f(x) = c\, e^{-x} x(1 - x), \quad 0 \leqslant x \leqslant 1 .$$

(a) Show that $c = e/(3 - e)$.
(b) Devise an acceptance–rejection algorithm to generate random variables that are distributed according to f.
(c) Implement the algorithm in MATLAB.

2.23. Implement two different algorithms to draw 100 uniformly generated points on the unit disk: one based on Example 2.7 and the other using (two-dimensional) acceptance–rejection.

Chapter 3
Joint Distributions

Often a random experiment is described via more than one random variable. Here are some examples:

1. We randomly select $n = 10$ people and observe their heights. Let $X_1, \ldots, X_n$ be the individual heights.
2. We toss a coin repeatedly. Let $X_i = 1$ if the ith toss is Heads and $X_i = 0$ otherwise. The experiment is thus described by the sequence $X_1, X_2, \ldots$ of Bernoulli random variables.
3. We randomly select a person from a large population and measure his/her weight X and height Y.

How can we specify the behavior of the random variables above? We should not just specify the pdf of the individual random variables, but also say something about the interaction (or lack thereof) between the random variables. For example, in the third experiment above, if the height Y is large, then most likely X is large as well. In contrast, in the first two experiments, it is reasonable to assume that the random variables are "independent" in some way; that is, information about one of the random variables does not give extra information about the others. What we need to specify is the **joint distribution** of the random variables. The theory below for multiple random variables follows a similar path to that of a single random variable described in Sects. 2.1–2.3.

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate the $\{X_i\}$ into a **random vector** $\mathbf{X} = (X_1, \ldots, X_n)$ (row vector) or $\mathbf{X} = (X_1, \ldots, X_n)^{\top}$ (column vector). Recall that the distribution of a *single* random variable X is completely specified by its cumulative distribution function. For *multiple* random variables we have the following generalization.

> **Definition 3.1. (Joint Cdf).** The **joint cdf** of $X_1, \ldots, X_n$ is the function $F : \mathbb{R}^n \to [0, 1]$ defined by
>
> $$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) .$$

Notice that we have used the abbreviation $\mathbb{P}(\{X_1 \leqslant x_1\} \cap \cdots \cap \{X_n \leqslant x_n\}) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n)$ to denote the probability of the intersection of events. We will use this abbreviation throughout the book.

As in the univariate (i.e., single variable) case we distinguish between *discrete* and *continuous* distributions.

3.1   Discrete Joint Distributions

Example 3.1 (Dice Experiment). In a box there are three dice. Die 1 is an ordinary die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces. The experiment consists of selecting a die at random followed by a toss with that die. Let X be the die number that is selected and let Y be the face value of that die. The probabilities $\mathbb{P}(X = x, Y = y)$ in Table 3.1 specify the joint distribution of X and Y. Note that it is more convenient to specify the joint probabilities $\mathbb{P}(X = x, Y = y)$ than the joint cumulative probabilities $\mathbb{P}(X \leqslant x, Y \leqslant y)$. The latter can be found, however, from the former by applying the sum rule. For example, $\mathbb{P}(X \leqslant 2, Y \leqslant 3) = \mathbb{P}(X = 1, Y = 1) + \cdots + \mathbb{P}(X = 2, Y = 3) = 6/18 = 1/3$. Moreover, by that same sum rule, the distribution of X is found by summing the $\mathbb{P}(X = x, Y = y)$ over all values of y—giving the last column of Table 3.1. Similarly, the distribution of Y is given by the column totals in the last row of the table.

Table 3.1 The joint distribution of X (die number) and Y (face value)

		y	
		1	2	3	4	5	6	Σ
	1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
x	2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0	$\frac{1}{3}$
	3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$	$\frac{1}{3}$
	Σ	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

In general, for discrete random variables $X_1, \ldots, X_n$, the joint distribution is easiest to specify via the joint pdf.

> **Definition 3.2. (Discrete Joint Pdf).** The **joint pdf** f of discrete random variables $X_1, \ldots, X_n$ is given by the function
>
> $$f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

We sometimes write $f_{X_1,\dots,X_n}$ instead of f to show that this is the pdf of the random variables $X_1,\dots,X_n$. Or, if $\mathbf{X} = (X_1,\dots,X_n)$ is the corresponding random vector, we can write $f_\mathbf{X}$ instead.

If the joint pdf f is known, we can calculate the probability of any event $\{\mathbf{X} \in B\}$, $B \subset \mathbb{R}^n$, via the sum rule as

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{\mathbf{x} \in B} f(\mathbf{x}) .$$

Compare this with (2.2). In particular, as explained in Example 3.1, we can find the pdf of X_i—often referred to as a **marginal** pdf, to distinguish it from the joint pdf—by summing the joint pdf over all possible values of the other variables:

$$\mathbb{P}(X_i = x) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} f(x_1,\dots,x_{i-1},x,x_{i+1},x_n) . \qquad (3.1)$$

The converse is not true: from the marginal distributions one cannot in general reconstruct the joint distribution. For example, in Example 3.1, we cannot reconstruct the inside of the two-dimensional table if only given the column and row totals.

However, there is an important exception, namely, when the random variables are *independent*. We have so far only defined what independence is for *events*. We can define random variables $X_1,\dots,X_n$ to be independent if events $\{X_1 \in B_1\},\dots,\{X_n \in B_n\}$ are independent for any choice of sets $\{B_i\}$. Intuitively, this means that any information about one of the random variables does not affect our knowledge about the others.

Definition 3.3. (Independence). Random variables $X_1,\dots,X_n$ are called **independent** if for all events $\{X_i \in B_i\}$ with $B_i \subset \mathbb{R}, i = 1,\dots,n$

$$\mathbb{P}(X_1 \in B_1,\dots,X_n \in B_n) = \mathbb{P}(X_1 \in B_1)\cdots\mathbb{P}(X_n \in B_n) . \qquad (3.2)$$

A direct consequence of the above definition is the following important theorem.

Theorem 3.1. (Independence and Product Rule). Random variables $X_1,\dots,X_n$ with joint pdf f are independent if and only if

$$f(x_1,\dots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n) \qquad (3.3)$$

for all $x_1,\dots,x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.

Proof. The theorem is true in both the discrete and continuous case. We only show the discrete case, where (3.3) is a special case of (3.2). It follows that (3.3) is a *necessary* condition for independence. To see that it is also a *sufficient* condition, let $\mathbf{X} = (X_1, \ldots, X_n)$ and observe that

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(\mathbf{X} \in \underbrace{B_1 \times \cdots \times B_n}_{A}) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$

$$= \sum_{\mathbf{x} \in A} f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \sum_{x_1 \in B_1} f_{X_1}(x_1) \cdots \sum_{x_n \in B_n} f_{X_n}(x_n)$$

$$= \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n) .$$

Here $A = B_1 \times \cdots \times B_n$ denotes the Cartesian product of $B_1, \ldots, B_n$. $\quad\square$

Example 3.2 (Dice Experiment Continued). We repeat the experiment in Example 3.1 with three ordinary fair dice. Since the events $\{X = x\}$ and $\{Y = y\}$ are now independent, each entry in the pdf table is $\frac{1}{3} \times \frac{1}{6}$. Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent.

Remark 3.1. An *infinite* sequence $X_1, X_2, \ldots$ of random variables is said to be *independent* if for any finite choice of positive integers $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent. Many statistical models involve random variables $X_1, X_2, \ldots$ that are **independent and identically distributed**, abbreviated as **iid**. We will use this abbreviation throughout this book and write the corresponding model as

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \text{Dist (or } f \text{ or } F) ,$$

where Dist is the common distribution with pdf f and cdf F.

Example 3.3 (Bernoulli Process). Consider the experiment where we toss a biased coin n times, with probability p of Heads. We can model this experiment in the following way. For $i = 1, \ldots, n$ let X_i be the result of the ith toss: $\{X_i = 1\}$ means Heads (or success), and $\{X_i = 0\}$ means Tails (or failure). Also, let

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, n .$$

Finally, assume that $X_1, \ldots, X_n$ are *independent*. The sequence

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \text{Ber}(p)$$

is called a **Bernoulli process** with success probability p. Let $X = X_1 + \cdots + X_n$ be the total number of successes in n trials (tosses of the coin). Denote by B_k the set of all binary vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^{n} x_i = k$. Note that B_k has $\binom{n}{k}$ elements. We have for every $k = 0, \ldots, n$,

$$\mathbb{P}(X = k) = \sum_{\mathbf{x} \in B_k} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

$$= \sum_{\mathbf{x} \in B_k} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = \sum_{\mathbf{x} \in B_k} p^k (1 - p)^{n-k}$$

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

In other words, $X \sim \mathsf{Bin}(n, p)$. Compare this with Example 2.2.   ☞ 24

For the joint pdf of *dependent* discrete random variables we can write, as a
consequence of the product rule (1.5),   ☞ 14

$$f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

$$= \mathbb{P}(X_1 = x_1) \, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \times \cdots$$

$$\cdots \times \mathbb{P}(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) ,$$

assuming that all probabilities $\mathbb{P}(X = x_1), \ldots, \mathbb{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1})$ are
nonzero. The function which maps, *for a fixed x_1*, each variable x_2 to the conditional
probability

$$\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_1 = x_1)} \tag{3.4}$$

is called the **conditional pdf** of X_2 given $X_1 = x_1$. We write it as $f_{X_2 \mid X_1}(x_2 \mid x_1)$.
Similarly, the function $x_n \mapsto \mathbb{P}(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1})$ is the
conditional pdf of X_n given $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$, which is written as
$f_{X_n \mid X_1, \ldots, X_{n-1}}(x_n \mid x_1, \ldots, x_{n-1})$.

Example 3.4 (Generating Uniformly on a Triangle). We uniformly select a point
(X, Y) from the triangle $T = \{(x, y) : x, y \in \{1, \ldots, 6\}, y \leqslant x\}$ in Fig. 3.1.
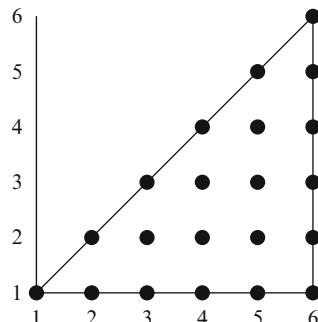
Fig. 3.1 Uniformly select a
point from the triangle

Because each of the 21 points is equally likely to be selected, the joint pdf is constant on T:

$$f(x, y) = \frac{1}{21}, \quad (x, y) \in T .$$

The pdf of X is found by summing $f(x, y)$ over all y. Hence,

$$f_X(x) = \frac{x}{21}, \quad x \in \{1, \ldots, 6\} .$$

Similarly,

$$f_Y(y) = \frac{7 - y}{21}, \quad y \in \{1, \ldots, 6\} .$$

For a fixed $x \in \{1, \ldots, 6\}$ the conditional pdf of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)} = \frac{1/21}{x/21} = \frac{1}{x}, \quad y \in \{1, \ldots, x\} ,$$

which simply means that, given $X = x$, Y has a discrete uniform distribution on $\{1, \ldots, x\}$.

3.1.1   Multinomial Distribution

An important discrete joint distribution is the multinomial distribution. It can be viewed as a generalization of the binomial distribution. We give the definition and then an example of how this distribution arises in applications.

Definition 3.4. (Multinomial Distribution). A random vector $(X_1, X_2, \ldots, X_k)$ is said to have a **multinomial** distribution with parameters n and $p_1, p_2, \ldots, p_k$ (positive and summing up to 1), if

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \qquad (3.5)$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_k = n$. We write $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$.

Example 3.5 (Urn Problem). We independently throw n balls into k urns, such that each ball is thrown in urn i with probability p_i, $i = 1, \ldots, k$; see Fig. 3.2.
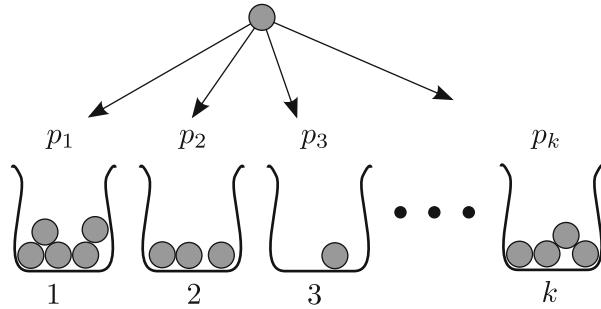
Fig. 3.2 Throwing n balls into k urns with probabilities $p_1, \ldots, p_k$. The random configuration of balls has a multinomial distribution

Let X_i be the total number of balls in urn i, $i = 1, \ldots, k$. We show that $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$. Let $x_1, \ldots, x_k$ be integers between 0 and n that sum up to n. The probability that the *first* x_1 balls fall in the first urn, the *next* x_2 balls fall in the second urn, etc., is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \; .$$

To find the probability that there are x_1 balls in the first urn, x_2 in the second, and so on, we have to multiply the probability above with the number of ways in which we can fill the urns with $x_1, x_2, \ldots, x_k$ balls, i.e., $n!/(x_1! x_2! \cdots x_k!)$. This gives (3.5).

Remark 3.2. Note that for the *binomial* distribution there are only *two* possible urns. Also, note that for each $i = 1, \ldots, k$, $X_i \sim \mathsf{Bin}(n, p_i)$.

3.2 Continuous Joint Distributions

Joint distributions for continuous random variables are usually defined via their joint pdf. The theoretical development below follows very similar lines to both the univariate continuous case in Sect. 2.2.2 and the multivariate discrete case in Sect. 3.1.

Definition 3.5. (Continuous Joint Pdf). Continuous random variables $X_1, \ldots, X_n$ are said to have a **joint pdf** f if

$$\mathbb{P}(a_1 < X_1 \leqslant b_1, \ldots, a_n < X_n \leqslant b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

for all $a_1, \ldots, b_n$.

☞ 28   This implies, similar to the univariate case in (2.3), that the probability of any event pertaining to $\mathbf{X} = (X_1, \ldots, X_n)$—say event $\{\mathbf{X} \in B\}$, where B is some subset of $\mathbb{R}^n$—can be found by *integration*:

$$\mathbb{P}(\mathbf{X} \in B) = \int_B f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n \ . \tag{3.6}$$

☞ 29   As in (2.5) we can interpret $f(x_1, \ldots, x_n)$ as the *density* of the probability distribution at $(x_1, \ldots, x_n)$. For example, in the two-dimensional case, for small $h > 0$,

$$\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + h, \ x_2 \leqslant X_2 \leqslant x_2 + h)$$

$$= \int_{x_1}^{x_1+h} \int_{x_2}^{x_2+h} f(u, v) \, du \, dv \approx h^2 \, f(x_1, x_2) \ .$$

Similar to the discrete multivariate case in (3.1), the marginal pdfs can be recovered from the joint pdf by integrating out the other variables:

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \, dx_1 \ldots dx_{i-1} \, dx_{i+1} \ldots dx_n \ .$$

We illustrate this for the two-dimensional case. We have

$$F_{X_1}(x) = \mathbb{P}(X_1 \leqslant x, X_2 \leqslant \infty) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f(x_1, x_2) \, dx_2 \right) dx_1 \ .$$

By differentiating the last integral with respect to x, we obtain

$$f_{X_1}(x) = \int_{-\infty}^{\infty} f(x, x_2) \, dx_2 \ .$$

It is not possible, in general, to reconstruct the joint pdf from the marginal pdfs. An exception is when the random variables are *independent*; see Definition 3.3. By modifying the arguments in the proof of Theorem 3.3 to the continuous case— basically replacing sums with integrals—it is not difficult to see that the theorem also holds in the continuous case. In particular, continuous random variables $X_1, \ldots, X_n$ are independent if and only if their joint pdf, f say, is the product of the marginal pdfs:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{3.7}$$

for all $x_1, \ldots, x_n$. Independence for an infinite sequence of random variables is
☞ 66   discussed in Remark 3.1.

Example 3.6 (Generating a General iid Sample). Consider the sequence of numbers produced by a uniform random number generator such as MATLAB's `rand` function. A mathematical model for the output stream is $U_1, U_2, \ldots,$ are independent and $U(0, 1)$ distributed; that is,

$$U_1, U_2, \ldots \overset{\text{iid}}{\sim} U(0, 1) \, .$$

Using the inverse-transform method it follows that for any cdf F,

$$F^{-1}(U_1), F^{-1}(U_2), \ldots \overset{\text{iid}}{\sim} F \, .$$

Example 3.7 (Quotient of Two Independent Random Variables). Let X and Y be independent continuous random variables, with $Y > 0$. What is the pdf of the quotient $U = X/Y$ in terms of the pdfs of X and Y? Consider first the cdf of U. We have

$$F_U(u) = \mathbb{P}(U \leqslant u) = \mathbb{P}(X/Y \leqslant u) = \mathbb{P}(X \leqslant Yu)$$

$$= \int_0^\infty \int_{-\infty}^{yu} f_X(x) f_Y(y) \, \mathrm{d}x \, \mathrm{d}y = \int_{-\infty}^u \int_0^\infty y f_X(yz) f_Y(y) \, \mathrm{d}y \, \mathrm{d}z \, ,$$

where we have used the change of variable $z = x/y$ and changed the order of integration in the last equation. It follows that the pdf is given by

$$f_U(u) = \frac{\mathrm{d}}{\mathrm{d}u} F_U(u) = \int_0^\infty y f_X(yu) \, f_Y(y) \, \mathrm{d}y \, . \tag{3.8}$$

As a particular example, suppose that X and V both have a standard normal distribution. Note that X/V has the same distribution as $U = X/Y$, where $Y = |V| > 0$ has a *positive normal* distribution. It follows from (3.8) that

$$f_U(u) = \int_0^\infty y \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2 u^2} \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y$$

$$= \int_0^\infty y \frac{1}{\pi} \mathrm{e}^{-\frac{1}{2}y^2(1+u^2)} \, \mathrm{d}y = \frac{1}{\pi} \frac{1}{1 + u^2}, \quad u \in \mathbb{R} \, .$$

This is the pdf of the *Cauchy* distribution.

Definition 3.6. (Conditional Pdf). Let X and Y have joint pdf f and suppose $f_X(x) > 0$. The **conditional pdf** of Y given $X = x$ is defined as

$$f_{Y|X}(y \,|\, x) = \frac{f(x, y)}{f_X(x)} \quad \text{for all } y \, . \tag{3.9}$$

For the discrete case, this is just a rewrite of (3.4). For the continuous case, the interpretation is that $f_{Y|X}(y \mid x)$ is the density corresponding to the cdf $F_{Y|X}(y \mid x)$ defined by the limit

$$F_{Y|X}(y \mid x) = \lim_{h \downarrow 0} \mathbb{P}(Y \leqslant y \mid x \leqslant X \leqslant x+h) = \lim_{h \downarrow 0} \frac{\mathbb{P}(Y \leqslant y,\ x \leqslant X \leqslant x + h)}{\mathbb{P}(x \leqslant X \leqslant x + h)} \, .$$

In many statistical situations, the conditional and marginal pdfs are known and (3.9) is used to find the joint pdf via

$$f(x, y) = f_X(x)\, f_{Y|X}(y \mid x) \, ,$$

or, more generally for the n-dimensional case,

$$f(x_1, \ldots, x_n) = \\ f_{X_1}(x_1)\, f_{X_2|X_1}(x_2 \mid x_1) \cdots f_{X_n|X_1,\ldots,X_{n-1}}(x_n \mid x_1, \ldots, x_{n-1}) \, , \tag{3.10}$$

☞ 14   which in the discrete case is just a rephrasing of the *product rule* in terms of probability densities. For independent random variables (3.10) reduces to (3.7). Equation (3.10) also shows how one could sequentially generate a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ according to a pdf f, provided that it is possible to generate random variables from the successive conditional distributions, as summarized in the following algorithm.

Algorithm 3.1. (Dependent Random Variable Generation).

1. Generate X_1 from pdf f_{X_1}. Set $t = 1$.
2. While $t < n$, given $X_1 = x_1, \ldots, X_t = x_t$, generate X_{t+1} from the conditional pdf $f_{X_{t+1}|X_1,\ldots,X_t}(x_t \mid x_1, \ldots, x_t)$ and set $t = t + 1$.
3. Return $\mathbf{X} = (X_1, \ldots, X_n)$.

Example 3.8 (Nonuniform Distribution on Triangle). We select a point (X, Y) from the triangle $(0, 0)$-$(1, 0)$-$(1, 1)$ in such a way that X has a uniform distribution on $(0, 1)$ and the conditional distribution of Y given $X = x$ is uniform on $(0, x)$. Figure 3.3 shows the result of 1000 independent draws from the joint pdf $f(x, y) = f_X(x)\, f_{Y|X}(y \mid x)$, generated via Algorithm 3.1. It is clear that the points are not uniformly distributed over the triangle.

Random variable X has a uniform distribution on $(0, 1)$; hence, its pdf is $f_X(x) = 1$ on $x \in (0, 1)$. For any fixed $x \in (0, 1)$, the conditional distribution of Y given $X = x$ is uniform on the interval $(0, x)$, which means that

$$f_{Y|X}(y \mid x) = \frac{1}{x}, \quad 0 < y < x \, .$$
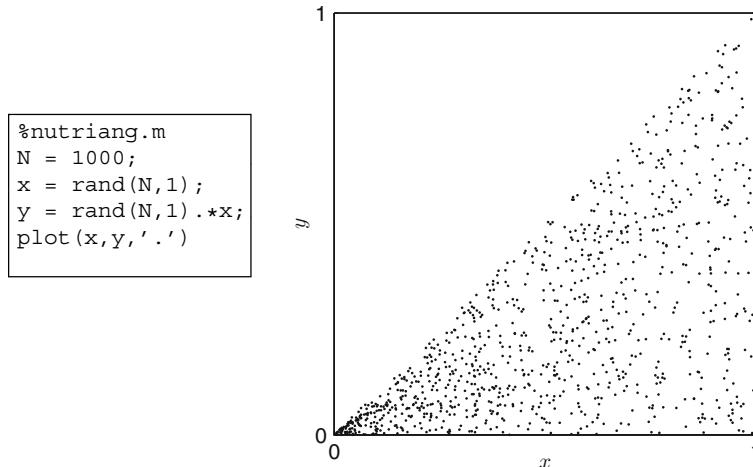
Fig. 3.3 One thousand realizations from the joint density $f(x, y)$, generated using the MATLAB program on the *left*, which implements Algorithm 3.1.

It follows that the joint pdf is given by

$$f(x, y) = f_X(x)\, f_{Y|X}(y \mid x) = \frac{1}{x}, \quad 0 < x < 1, \quad 0 < y < x .$$

From the joint pdf we can obtain the pdf of Y as

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x = \int_y^1 \frac{1}{x}\, \mathrm{d}x = -\ln y, \quad 0 < y < 1 .$$

Finally, for any fixed $y \in (0, 1)$, the conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{-1}{x \ln y}, \quad y < x < 1 .$$

3.3   Mixed Joint Distributions

So far we have only considered joint distributions in which the random variables are all discrete or all continuous. The theory can be extended to mixed cases in a straightforward way. For example, the joint pdf of a discrete variable X and a continuous variable Y is defined as the function $f(x, y)$ such that for all events $\{(X, Y) \in A\}$, where $A \subset \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in A) = \sum_x \int \mathrm{I}_{\{(x,y) \in A\}}\, f(x, y)\, \mathrm{d}y ,$$

where I denotes the indicator. The pdf is often specified via (3.10).

Example 3.9 (Beta Distribution). Let $\Theta \sim \mathsf{U}(0,1)$ and $(X \mid \Theta = \theta) \sim \mathsf{Bin}(n, \theta)$. Using (3.10), the joint pdf of X and Θ is given by

$$f(x, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad \theta \in (0, 1), \ x = 0, 1, \ldots, n \ .$$

By integrating out θ, we find the pdf of X:

$$f_X(x) = \int_0^1 \binom{n}{x} \theta^x (1 - \theta)^{n-x} \mathrm{d}\theta = \binom{n}{x} B(x + 1, n - x + 1) \ ,$$

where B is the **beta function**, defined as

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1 - t)^{\beta-1} \mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \ , \tag{3.11}$$

☞ 48   and Γ is the gamma function in (2.20). The conditional pdf of Θ given $X = x$, where $x \in \{0, \ldots, n\}$, is

$$f_{\Theta|X}(\theta \mid x) = \frac{f(\theta, x)}{f_X(x)} = \frac{\theta^x (1 - \theta)^{n-x}}{B(x + 1, n - x + 1)}, \quad \theta \in (0, 1) \ .$$

The continuous distribution with pdf

$$f(x; \alpha, \beta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)}, \quad x \in (0, 1) \tag{3.12}$$

is called the **beta distribution** with parameters α and β. Both parameters are assumed to be strictly positive. We write $\mathsf{Beta}(\alpha, \beta)$ for this distribution. For this example we have thus $(\Theta \mid X = x) \sim \mathsf{Beta}(x + 1, n - x + 1)$.

3.4   Expectations for Joint Distributions

☞ 31   Similar to the univariate case in Theorem 2.2, the expected value of a real-valued function h of $(X_1, \ldots, X_n) \sim f$ is a weighted average of all values that $h(X_1, \ldots, X_n)$ can take. Specifically, in the continuous case,

$$\mathbb{E}h(X_1, \ldots, X_n) = \int \cdots \int h(x_1, \ldots, x_n) \, f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n \ . \tag{3.13}$$

In the discrete case replace the integrals above with sums.

Two important special cases are the expectation of the *sum* (or more generally affine transformations) of random variables and the *product* of random variables.

Theorem 3.2. (Properties of the Expectation). Let $X_1, \ldots, X_n$ be random variables with expectations $\mu_1, \ldots, \mu_n$. Then,

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1 \mu_1 + \cdots + b_n \mu_n \qquad (3.14)$$

for all constants $a, b_1, \ldots, b_n$. Also, for *independent* random variables,

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1 \mu_2 \cdots \mu_n . \qquad (3.15)$$

Proof. We show it for the continuous case with two variables only. The general case follows by analogy and, for the discrete case, by replacing integrals with sums. Let X_1 and X_2 be continuous random variables with joint pdf f. Then, by (3.13),

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2] = \iint (a + b_1 x_1 + b_2 x_2) \, f(x_1, x_2) \, dx_1 \, dx_2$$

$$= a + b_1 \iint x_1 f(x_1, x_2) \, dx_1 \, dx_2 + b_2 \iint x_2 f(x_1, x_2) \, dx_1 \, dx_2$$

$$= a + b_1 \int x_1 \left(\int f(x_1, x_2) \, dx_2 \right) dx_1 + b_2 \int x_2 \left(\int f(x_1, x_2) \, dx_1 \right) dx_2$$

$$= a + b_1 \int x_1 f_{X_1}(x_1) \, dx_1 + b_2 \int x_2 f_{X_2}(x_2) \, dx_2 = a + b_1 \mu_1 + b_2 \mu_2 .$$

Next, assume that X_1 and X_2 are independent, so that $f(x_1, x_2) = f_{X_1}(x_1) \times f_{X_2}(x_2)$. Then,

$$\mathbb{E}[X_1 X_2] = \iint x_1 x_2 f_{X_1}(x_1) f_{X_2}(x_2) \, dx_1 \, dx_2$$

$$= \int x_1 f_{X_1}(x_1) \, dx_1 \times \int x_2 f_{X_2}(x_2) \, dx_2 = \mu_1 \mu_2 . \qquad \square$$

Definition 3.7. (Covariance). The **covariance** of two random variables X and Y with expectations $\mathbb{E}X = \mu_X$ and $\mathbb{E}Y = \mu_Y$ is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] .$$

The covariance is a measure of the amount of linear dependency between two random variables. A scaled version of the covariance is given by the **correlation coefficient**:

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \; , \tag{3.16}$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. The correlation coefficient always lies between -1 and 1; see Problem 3.16.

☞ 95

For easy reference Theorem 3.3 lists some important properties of the variance and covariance.

Theorem 3.3. (Properties of the Variance and Covariance). For random variables X, Y, and Z, and constants a and b, we have

1. $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
2. $\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X)$.
3. $\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X \, \mathbb{E}Y$.
4. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(aX + bY, Z) = a \, \mathrm{Cov}(X, Z) + b \, \mathrm{Cov}(Y, Z)$.
6. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
7. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \, \mathrm{Cov}(X, Y)$.
8. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Proof. For simplicity of notation we write $\mathbb{E}Z = \mu_Z$ for a generic random variable

☞ 33

Z. Properties 1 and 2 were already shown in Theorem 2.4.

3. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] = \mathbb{E}[XY] - \mu_X\mu_Y$.
4. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[(Y - \mu_Y)(X - \mu_X)] = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(aX + bY, Z) = \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\,\mathbb{E}Z = a\,\mathbb{E}[XZ] - a\,\mathbb{E}X\mathbb{E}Z + b\,\mathbb{E}[YZ] - b\,\mathbb{E}Y\mathbb{E}Z = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.
6. $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)] = \mathbb{E}[(X - \mu_X)^2] = \mathrm{Var}(X)$.
7. By Property 6, $\mathrm{Var}(X + Y) = \mathrm{Cov}(X + Y, X + Y)$. By Property 5, $\mathrm{Cov}(X + Y, X + Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(Y, Y) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, where in the last equation Properties 4 and 6 are used.
8. If X and Y are independent, then $\mathbb{E}[XY] = \mu_X\mu_Y$. Therefore, $\mathrm{Cov}(X, Y) = 0$ follows immediately from Property 3.   □

As a consequence of Properties 2 and 7, we have the following general result for the variance of affine transformations of random variables.

Corollary 3.1. (Variance of an Affine Transformation). Let $X_1, \ldots, X_n$ be random variables with variances $\sigma_1^2, \ldots, \sigma_n^2$. Then,

$$\mathrm{Var}\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \sigma_i^2 + 2 \sum_{i<j} b_i b_j \mathrm{Cov}(X_i, X_j) \qquad (3.17)$$

for any choice of constants a and $b_1, \ldots, b_n$. In particular, for *independent* random variables $X_1, \ldots, X_n$,

$$\mathrm{Var}(a + b_1 X_1 + \cdots + b_n X_n) = b_1^2 \sigma_1^2 + \cdots + b_n^2 \sigma_n^2 . \qquad (3.18)$$

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a random column vector. Sometimes it is convenient to write the expectations and covariances in vector notation.

Definition 3.8. (Expectation Vector and Covariance Matrix). For any random column vector $\mathbf{X}$ we define the **expectation vector** as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^\top .$$

The **covariance matrix** Σ is defined as the matrix whose (i, j)th element is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] .$$

If we define the expectation of a matrix to be the matrix of expectations, then we can write the covariance matrix succinctly as

$$\Sigma = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right] .$$

Definition 3.9. (Conditional Expectation). The **conditional expectation** of Y given $X = x$, denoted $\mathbb{E}[Y \mid X = x]$, is the expectation corresponding to the conditional pdf $f_{Y \mid X}(y \mid x)$. That is, in the continuous case,

$$\mathbb{E}[Y \mid X = x] = \int y \, f_{Y \mid X}(y \mid x) \, \mathrm{d}y .$$

In the discrete case replace the integral with a sum.

Note that $\mathbb{E}[Y \mid X = x]$ is a function of x, say $h(x)$. The corresponding random variable $h(X)$ is written as $\mathbb{E}[Y \mid X]$. The expectation of $\mathbb{E}[Y \mid X]$ is, in the continuous case,

$$\mathbb{E}\mathbb{E}[Y \mid X] = \int \mathbb{E}[Y \mid X = x] f_X(x) \, dx = \int \int y \frac{f(x, y)}{f_X(x)} f_X(x) \, dy \, dx$$

$$= \int y \, f_Y(y) \, dy = \mathbb{E}Y \; .$$

(3.19)

This "stacking" of (conditional) expectations is sometimes referred to as the **tower property**.

Example 3.10 (Nonuniform Distribution on Triangle Continued). In Example 3.8 the conditional expectation of Y given $X = x$, with $0 < x < 1$, is

$$\mathbb{E}[Y \mid X = x] = \frac{1}{2} x \; ,$$

because conditioned on $X = x$, Y is uniformly distributed on the interval $(0, x)$. Using the tower property we find

$$\mathbb{E}Y = \frac{1}{2}\mathbb{E}X = \frac{1}{4} \; .$$

3.5   Functions of Random Variables

Suppose $X_1, \ldots, X_n$ are measurements of a random experiment. What can be said about the distribution of a *function* of the data, say $Z = g(X_1, \ldots, X_n)$, when the joint distribution of $X_1, \ldots, X_n$ is known?

Example 3.11 (Pdf of an Affine Transformation). Let X be a continuous random variable with pdf f_X and let $Z = a + bX$, where $b \neq 0$. We wish to determine the pdf f_Z of Z. Suppose that $b > 0$. We have for any z

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}(X \leqslant (z - a)/b) = F_X((z - a)/b) \; .$$

Differentiating this with respect to z gives $f_Z(z) = f_X((z - a)/b) / b$. For $b < 0$ we similarly obtain $f_Z(z) = f_X((z - a)/b) / (-b)$. Thus, in general,

$$f_Z(z) = \frac{1}{|b|} f_X\left(\frac{z - a}{b}\right) \; .$$

(3.20)

Example 3.12 (Pdf of a Monotone Transformation). Generalizing the previous example, suppose that $Z = g(X)$ for some strictly increasing function g. To find the pdf of Z from that of X we first write

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\left(X \leqslant g^{-1}(z)\right) = F_X\left(g^{-1}(z)\right),$$

where g^{-1} is the inverse of g. Differentiating with respect to z now gives

$$f_Z(z) = f_X(g^{-1}(z)) \frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z) = \frac{f_X(g^{-1}(z))}{g'(g^{-1}(z))}. \qquad (3.21)$$

For strictly decreasing functions, g' needs to be replaced with its negative value.

3.5.1 Linear Transformations

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be a column vector in $\mathbb{R}^n$ and B an $m \times n$ matrix. The mapping $\mathbf{x} \mapsto \mathbf{z}$, with $\mathbf{z} = B\mathbf{x}$, is called a **linear transformation**. Now consider a *random* vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and let

$$\mathbf{Z} = B\mathbf{X}.$$

Then $\mathbf{Z}$ is a random vector in $\mathbb{R}^m$. In principle, if we know the joint distribution of $\mathbf{X}$, then we can derive the joint distribution of $\mathbf{Z}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 3.4. (Expectation and Covariance Under a Linear Transformation). If $\mathbf{X}$ has expectation vector $\boldsymbol{\mu}_\mathbf{X}$ and covariance matrix $\Sigma_\mathbf{X}$, then the expectation vector and covariance matrix of $\mathbf{Z} = B\mathbf{X}$ are given by

$$\boldsymbol{\mu}_\mathbf{Z} = B\boldsymbol{\mu}_\mathbf{X} \qquad (3.22)$$

and

$$\Sigma_\mathbf{Z} = B\Sigma_\mathbf{X}B^\top. \qquad (3.23)$$

Proof. We have $\boldsymbol{\mu_Z} = \mathbb{E}\mathbf{Z} = \mathbb{E}B\mathbf{X} = B\,\mathbb{E}\mathbf{X} = B\boldsymbol{\mu_X}$ and

$$\begin{aligned}
\Sigma_\mathbf{Z} &= \mathbb{E}[(\mathbf{Z} - \boldsymbol{\mu_Z})(\mathbf{Z} - \boldsymbol{\mu_Z})^\top] = \mathbb{E}[B(\mathbf{X} - \boldsymbol{\mu_X})(B(\mathbf{X} - \boldsymbol{\mu_X}))^\top] \\
&= B\,\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu_X})(\mathbf{X} - \boldsymbol{\mu_X})^\top]B^\top \\
&= B\Sigma_\mathbf{X}B^\top \ .
\end{aligned}$$

$\square$

Suppose that B is an *invertible* $n \times n$ matrix. If $\mathbf{X}$ has a joint pdf $f_\mathbf{X}$, what is the joint density $f_\mathbf{Z}$ of $\mathbf{Z}$? Let us consider the continuous case. For any fixed $\mathbf{x}$, let $\mathbf{z} = B\mathbf{x}$. Hence, $\mathbf{x} = B^{-1}\mathbf{z}$. Consider the n-dimensional cube $C = [z_1, z_1 + h] \times \cdots \times [z_n, z_n + h]$. Then, by definition of the joint density for $\mathbf{Z}$, we have

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\,f_\mathbf{Z}(\mathbf{z}) \ .$$

Let D be the image of C under B^{-1}—that is, the parallelepiped of all points $\mathbf{x}$ such that $B\mathbf{x} \in C$; see Fig. 3.4.
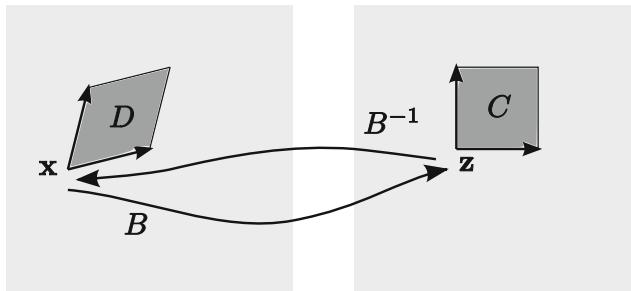
Fig. 3.4 Linear transformation

A basic result from linear algebra is that any matrix B linearly transforms an n-dimensional rectangle with volume V into an n-dimensional parallelepiped with volume $V\,|B|$, where $|B| = |\det(B)|$. Thus, in addition to the above expression for $\mathbb{P}(\mathbf{Z} \in C)$, we also have

$$\mathbb{P}(\mathbf{Z} \in C) = \mathbb{P}(\mathbf{X} \in D) \approx h^n|B^{-1}|\,f_\mathbf{X}(\mathbf{x}) = h^n|B|^{-1}\,f_\mathbf{X}(\mathbf{x}) \ .$$

Equating these two expressions for $\mathbb{P}(\mathbf{Z} \in C)$ and letting h go to 0, we obtain

$$f_\mathbf{Z}(\mathbf{z}) = \frac{f_\mathbf{X}(B^{-1}\mathbf{z})}{|B|}, \quad \mathbf{z} \in \mathbb{R}^n \ . \tag{3.24}$$

3.5.2 General Transformations

We can apply similar reasoning as in the previous subsection to deal with general transformations $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x})$, written out as

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix} .$$

For a fixed $\mathbf{x}$, let $\mathbf{z} = \mathbf{g}(\mathbf{x})$. Suppose $\mathbf{g}$ is invertible; hence, $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{z})$. Any infinitesimal n-dimensional rectangle at $\mathbf{x}$ with volume V is transformed into an n-dimensional parallelepiped at $\mathbf{z}$ with volume $V |J_{\mathbf{g}}(\mathbf{x})|$, where $J_{\mathbf{g}}(\mathbf{x})$ is the *matrix of Jacobi* at $\mathbf{x}$ of the transformation $\mathbf{g}$; that is,

$$J_{\mathbf{g}}(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} .$$

Now consider a random column vector $\mathbf{Z} = \mathbf{g}(\mathbf{X})$. Let C be a small cube around $\mathbf{z}$ with volume h^n. Let D be the image of C under $\mathbf{g}^{-1}$. Then, as in the linear case,

$$h^n \, f_{\mathbf{Z}}(\mathbf{z}) \approx \mathbb{P}(\mathbf{Z} \in C) \approx h^n |J_{\mathbf{g}^{-1}}(\mathbf{z})| \, f_{\mathbf{X}}(\mathbf{x}) .$$

Hence, we have the following result.

> **Theorem 3.5. (Transformation Rule).** Let $\mathbf{X}$ be a continuous n-dimensional random vector with pdf $f_{\mathbf{X}}$ and $\mathbf{g}$ a function from $\mathbb{R}^n$ to $\mathbb{R}^n$ with inverse $\mathbf{g}^{-1}$. Then, $\mathbf{Z} = \mathbf{g}(\mathbf{X})$ has pdf
>
> $$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z})) |J_{\mathbf{g}^{-1}}(\mathbf{z})| , \quad \mathbf{z} \in \mathbb{R}^n . \tag{3.25}$$

Remark 3.3. Note that $|J_{\mathbf{g}^{-1}}(\mathbf{z})| = 1/|J_{\mathbf{g}}(\mathbf{x})|$.

Example 3.13 (Box–Muller Method). The joint distribution of $X, Y \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ is

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} , \quad (x, y) \in \mathbb{R}^2 .$$

In polar coordinates we have

$$X = R \cos \Theta \quad \text{and} \quad Y = R \sin \Theta , \tag{3.26}$$

where $R \geqslant 0$ and $\Theta \in (0, 2\pi)$. What is the joint pdf of R and Θ? Consider the inverse transformation $\mathbf{g}^{-1}$, defined by

$$\begin{pmatrix} r \\ \theta \end{pmatrix} \xmapsto{\mathbf{g}^{-1}} \begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} .$$

The corresponding matrix of Jacobi is

$$J_{\mathbf{g}^{-1}}(r, \theta) = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} ,$$

which has determinant r. Since $x^2 + y^2 = r^2(\cos^2 \theta + \sin^2 \theta) = r^2$, it follows that

$$f_{R,\Theta}(r, \theta) = f_{X,Y}(x, y)\, r = \frac{1}{2\pi} e^{-\frac{1}{2}r^2} r, \quad \theta \in (0, 2\pi), \quad r \geqslant 0 .$$

By integrating out θ and r, respectively, we find $f_R(r) = r\, e^{-r^2/2}$ and $f_\Theta(\theta) = 1/(2\pi)$. Since $f_{R,\Theta}$ is the product of f_R and f_Θ, the random variables R and Θ are independent. This shows how X and Y could be generated: independently generate $R \sim f_R$ and $\Theta \sim \mathsf{U}(0, 2\pi)$ and return X and Y via (3.26). Generation from f_R can ☞ 53   be done via the inverse-transform method. In particular, R has the same distribution as $\sqrt{-2 \ln U}$ with $U \sim \mathsf{U}(0, 1)$. This leads to the following method for generating standard normal random variables.

> **Algorithm 3.2. (Box–Muller Method).**
>
> 1. Generate $U_1, U_2 \overset{\text{iid}}{\sim} \mathsf{U}(0, 1)$.
> 2. Return two independent standard normal variables, X and Y, via
>
> $$\begin{aligned} X &= \sqrt{-2 \ln U_1} \, \cos(2\pi U_2) , \\ Y &= \sqrt{-2 \ln U_1} \, \sin(2\pi U_2) . \end{aligned} \qquad (3.27)$$

3.6   Multivariate Normal Distribution

It is helpful to view a normally distributed random variable as an affine transformation of a standard normal random variable. In particular, if Z has a standard normal ☞ 46   distribution, then $X = \mu + \sigma Z$ has a $\mathsf{N}(\mu, \sigma^2)$ distribution; see Theorem 2.15.

We now generalize this to n dimensions. Let $Z_1, \ldots, Z_n$ be independent and standard normal random variables. The joint pdf of $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}z_i^2} = (2\pi)^{-\frac{n}{2}} \, e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}}, \quad \mathbf{z} \in \mathbb{R}^n \, . \tag{3.28}$$

We write $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I)$, where I is the identity matrix. Consider the affine transformation (i.e., a linear transformation plus a constant vector)

$$\mathbf{X} = \boldsymbol{\mu} + B \, \mathbf{Z} \tag{3.29}$$

for some $m \times n$ matrix B and m-dimensional vector $\boldsymbol{\mu}$. Note that, by Theorem 3.4, $\mathbf{X}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^\top$.

Definition 3.10. (Multivariate Normal Distribution). A random vector $\mathbf{X}$ is said to have a **multivariate normal** or **multivariate Gaussian** distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if it can be written as $\mathbf{X} = \boldsymbol{\mu} + B \, \mathbf{Z}$, where $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I)$ and $BB^\top = \Sigma$. We write $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$.

Suppose that B is an invertible $n \times n$ matrix. Then, by (3.24), the density of $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|B|\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}(B^{-1}\mathbf{y})^\top B^{-1}\mathbf{y}} = \frac{1}{|B|\sqrt{(2\pi)^n}} \, e^{-\frac{1}{2}\mathbf{y}^\top (B^{-1})^\top B^{-1}\mathbf{y}} \, .$$

We have $|B| = \sqrt{|\Sigma|}$ and $(B^{-1})^\top B^{-1} = (B^\top)^{-1}B^{-1} = (BB^\top)^{-1} = \Sigma^{-1}$, so that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \, |\Sigma|}} \, e^{-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1}\mathbf{y}} \, .$$

Because $\mathbf{X}$ is obtained from $\mathbf{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{x} - \boldsymbol{\mu})$ and therefore

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \, |\Sigma|}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n \, . \tag{3.30}$$

Figure 3.5 shows the pdfs of two bivariate (i.e., two-dimensional) normal distributions. In both cases the mean vector is $\boldsymbol{\mu} = (0, 0)^\top$ and the variances (the diagonal elements of Σ) are 1. The correlation coefficients (or, equivalently here, the covariances) are, respectively, $\varrho = 0$ and $\varrho = 0.8$.
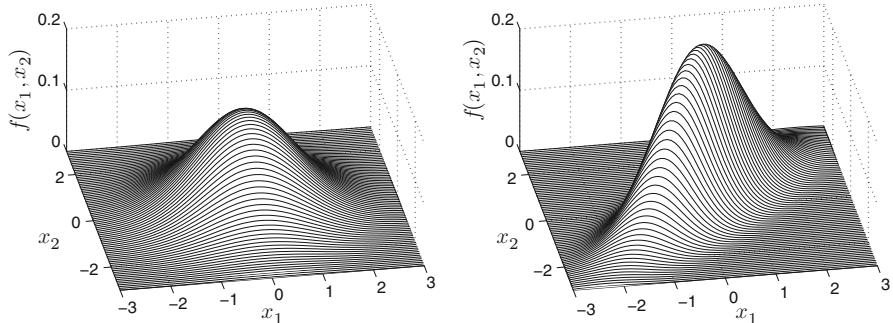
Fig. 3.5 Pdfs of bivariate normal distributions with means zero, variances 1, and correlation coefficients 0 (*left*) and 0.8 (*right*)

Conversely, given a covariance matrix $\Sigma = (\sigma_{ij})$, there exists a unique lower triangular matrix B such that $\Sigma = BB^{\top}$. In MATLAB, the function `chol` accomplishes this so-called **Cholesky factorization**. Note that it is important to use the option `'lower'` when calling this function, as MATLAB produces an upper triangular matrix by default. Once the Cholesky factorization is determined, it is easy to sample from a multivariate normal distribution.

Algorithm 3.3. (Normal Random Vector Generation). To generate N independent draws from a $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ distribution of dimension n carry out the following steps:

1. Determine the lower Cholesky factorization $\Sigma = BB^{\top}$.
2. Generate $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\top}$ by drawing $Z_1, \ldots, Z_n \sim_{\text{iid}} \mathsf{N}(0, 1)$.
3. Output $\mathbf{X} = \boldsymbol{\mu} + B\mathbf{Z}$.
4. Repeat Steps 2 and 3 independently N times.

Example 3.14 (Generating from a Bivariate Normal Distribution). The MATLAB code below draws 1000 samples from the two pdfs in Fig. 3.5. The resulting point clouds are given in Fig. 3.6.

```
%bivnorm.m
N = 1000; rho = 0.8;
Sigma = [1 rho; rho 1];
B=chol(Sigma,'lower');
x=B*randn(2,N);
plot(x(1,:),x(2,:),'.')
```
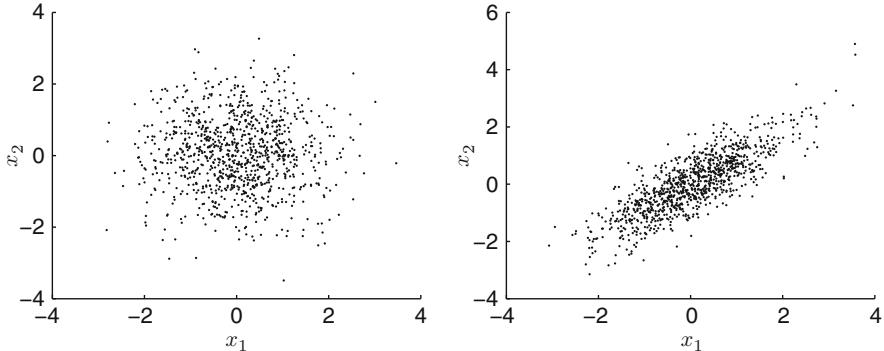
Fig. 3.6 One thousand realizations of bivariate normal distributions with means zero, variances 1, and correlation coefficients 0 (*left*) and 0.8 (*right*)

The following theorem states that any affine combination of independent multivariate normal random variables is again multivariate normal.

> **Theorem 3.6. (Affine Transformation of Normal Random Vectors).** Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_r$ be independent m_i-dimensional normal random vectors, with $\mathbf{X}_i \sim \mathsf{N}(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, \ldots, r$. Then, for any $n \times 1$ vector $\mathbf{a}$ and $n \times m_i$ matrices $B_1, \ldots, B_r$,
>
> $$\mathbf{a} + \sum_{i=1}^{r} B_i \, \mathbf{X}_i \sim \mathsf{N}\left(\mathbf{a} + \sum_{i=1}^{r} B_i \, \boldsymbol{\mu}_i, \ \sum_{i=1}^{r} B_i \, \Sigma_i \, B_i^{\top}\right). \qquad (3.31)$$

Proof. Denote the n-dimensional random vector in the left-hand side of (3.31) by $\mathbf{Y}$. By Definition 3.10, each $\mathbf{X}_i$ can be written as $\boldsymbol{\mu}_i + A_i \mathbf{Z}_i$, where the $\{\mathbf{Z}_i\}$ are independent (because the $\{\mathbf{X}_i\}$ are independent), so that

$$\mathbf{Y} = \mathbf{a} + \sum_{i=1}^{r} B_i \, (\boldsymbol{\mu}_i + A_i \mathbf{Z}_i) = \mathbf{a} + \sum_{i=1}^{r} B_i \, \boldsymbol{\mu}_i + \sum_{i=1}^{r} B_i \, A_i \mathbf{Z}_i \ ,$$

which is an affine combination of independent standard normal random vectors. Hence, $\mathbf{Y}$ is multivariate normal. Its expectation vector and covariance matrix can be found easily from Theorem 3.4.   □

The next theorem shows that the distribution of a subvector of a multivariate normal random vector is again normal.

Theorem 3.7. (Marginal Distributions of Normal Random Vectors). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ be an n-dimensional normal random vector. Decompose $\mathbf{X}$, $\boldsymbol{\mu}$, and Σ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_q \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_p & \Sigma_r \\ \Sigma_r^\top & \Sigma_q \end{pmatrix}, \tag{3.32}$$

where Σ_p is the upper left $p \times p$ corner of Σ and Σ_q is the lower right $q \times q$ corner of Σ. Then, $\mathbf{X}_p \sim \mathsf{N}(\boldsymbol{\mu}_p, \Sigma_p)$.

Proof. Let $BB^\top$ be the lower Cholesky factorization of Σ. We can write

$$\begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_q \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{pmatrix} + \underbrace{\begin{pmatrix} B_p & O \\ C_r & C_q \end{pmatrix}}_{B} \begin{pmatrix} \mathbf{Z}_p \\ \mathbf{Z}_q \end{pmatrix}, \tag{3.33}$$

where $\mathbf{Z}_p$ and $\mathbf{Z}_q$ are independent p- and q-dimensional standard normal random vectors. In particular, $\mathbf{X}_p = \boldsymbol{\mu}_p + B_p \mathbf{Z}_p$, which means that $\mathbf{X}_p \sim \mathsf{N}(\boldsymbol{\mu}_p, \Sigma_p)$, since $B_p B_p^\top = \Sigma_p$.   □

By relabeling the elements of $\mathbf{X}$ we see that Theorem 3.7 implies that *any* subvector of $\mathbf{X}$ has a multivariate normal distribution. For example, $\mathbf{X}_q \sim \mathsf{N}(\boldsymbol{\mu}_q, \Sigma_q)$.

Not only the marginal distributions of a normal random vector are normal but also its *conditional distributions*.

Theorem 3.8. (Conditional Distributions of Normal Random Vectors). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ be an n-dimensional normal random vector with $\det(\Sigma) > 0$. If $\mathbf{X}$ is decomposed as in (3.32), then

$$(\mathbf{X}_q \mid \mathbf{X}_p = \mathbf{x}_p) \sim \mathsf{N}(\boldsymbol{\mu}_q + \Sigma_r^\top \Sigma_p^{-1} (\mathbf{x}_p - \boldsymbol{\mu}_p), \ \Sigma_q - \Sigma_r^\top \Sigma_p^{-1} \Sigma_r). \tag{3.34}$$

As a consequence, $\mathbf{X}_p$ and $\mathbf{X}_q$ are *independent* if and only if they are *uncorrelated*; that is, if $\Sigma_r = O$ (zero matrix).

Proof. From (3.33) we see that

$$(\mathbf{X}_q \mid \mathbf{X}_p = \mathbf{x}_p) = \boldsymbol{\mu}_q + C_r B_p^{-1} (\mathbf{x}_p - \boldsymbol{\mu}_p) + C_q \mathbf{Z}_q,$$

where $\mathbf{Z}_q$ is a q-dimensional multivariate standard normal random vector. It follows that $\mathbf{X}_q$ conditional on $\mathbf{X}_p = \mathbf{x}_p$ has a $\mathsf{N}(\boldsymbol{\mu}_q + C_r B_p^{-1} (\mathbf{x}_p - \boldsymbol{\mu}_p), C_q C_q^\top)$ distribution.

The proof of (3.34) is completed by observing that $\Sigma_r^\top \Sigma_p^{-1} = C_r B_p^\top (B_p^\top)^{-1} B_p^{-1} = C_r B_p^{-1}$, and

$$\Sigma_q - \Sigma_r^\top \Sigma_p^{-1} \Sigma_r = C_r C_r^\top + C_q C_q^\top - C_r B_p^{-1} \underbrace{\Sigma_r}_{B_p C_r^\top} = C_q C_q^\top .$$

If $\mathbf{X}_p$ and $\mathbf{X}_q$ are independent, then they are obviously uncorrelated, as $\Sigma_r = \mathbb{E}[(\mathbf{X}_p - \boldsymbol{\mu}_p)(\mathbf{X}_q - \boldsymbol{\mu}_q)^\top] = \mathbb{E}(\mathbf{X}_p - \boldsymbol{\mu}_p)\,\mathbb{E}(\mathbf{X}_q - \boldsymbol{\mu}_q)^\top = O$. Conversely, if $\Sigma_r = O$, then by (3.34), the conditional distribution of $\mathbf{X}_q$ given $\mathbf{X}_p$ is the same as the unconditional distribution of $\mathbf{X}_q$, that is, $\mathsf{N}(\boldsymbol{\mu}_q, \Sigma_q)$. In other words, $\mathbf{X}_q$ is independent of $\mathbf{X}_p$. $\qquad\square$

Theorem 3.9. (Relationship Between Normal and χ^2 Distributions). If $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ is an n-dimensional normal random with vector with $\det(\Sigma) > 0$, then

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_n^2 . \tag{3.35}$$

Proof. Let $BB^\top$ be the Cholesky factorization of Σ, where B is invertible. Since $\mathbf{X}$ can be written as $\boldsymbol{\mu} + B\mathbf{Z}$, where $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ is a vector of independent standard normal random variables, we have

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})^\top (BB^\top)^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^n Z_i^2 .$$

The moment generating function of $Y = \sum_{i=1}^n Z_i^2$ is given by

$$\mathbb{E}\,\mathrm{e}^{tY} = \mathbb{E}\,\mathrm{e}^{t(Z_1^2 + \cdots + Z_n^2)} = \mathbb{E}\,[\mathrm{e}^{tZ_1^2} \cdots \mathrm{e}^{tZ_n^2}] = \left(\mathbb{E}\,\mathrm{e}^{tZ^2}\right)^n ,$$

where $Z \sim \mathsf{N}(0, 1)$. The moment generating function of Z^2 is

$$\mathbb{E}\,\mathrm{e}^{tZ^2} = \int_{-\infty}^\infty \mathrm{e}^{tz^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \mathrm{e}^{-\frac{1}{2}(1-2t)z^2} \mathrm{d}z = \frac{1}{\sqrt{1 - 2t}} ,$$

so that

$$\mathbb{E}\mathrm{e}^{tY} = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{n}{2}} , \quad t < \frac{1}{2} ,$$

which is the moment generating function of the $\mathsf{Gamma}(n/2, 1/2)$ distribution, that is, the χ_n^2 distribution—see Theorem 2.18. $\qquad\square$

A consequence of Theorem 3.9 is that if $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is n-dimensional standard normal, then the squared length $\|\mathbf{X}\|^2 = X_1^2 + \cdots + X_n^2$ has a χ_n^2 distribution. If instead $X_i \sim \mathsf{N}(\mu_i, 1)$, $i = 1, \ldots$, then $\|\mathbf{X}\|^2$ is said to have a **noncentral χ_n^2 distribution**. This distribution depends on the $\{\mu_i\}$ only through the norm $\|\boldsymbol{\mu}\|$; see Problem 3.22. We write $\|\mathbf{X}\|^2 \sim \chi_n^2(\theta)$, where $\theta = \|\boldsymbol{\mu}\|$ is the **noncentrality parameter**.

Such distributions frequently occur in statistics when considering *projections* of multivariate normal random variables. The proof of the following theorem can be found in Appendix B.4.

Theorem 3.10. (Relationship Between Normal and Noncentral χ^2 Distributions). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, I)$ be an n-dimensional normal random vector and let $\mathscr{V}_k$ and $\mathscr{V}_m$ be linear subspaces of dimensions k and m, respectively, with $k < m \leqslant n$. Let $\mathbf{X}_k$ and $\mathbf{X}_m$ be orthogonal projections of $\mathbf{X}$ onto $\mathscr{V}_k$ and $\mathscr{V}_m$, and let $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_m$ be the corresponding projections of $\boldsymbol{\mu}$. Then, the following holds:

1. The random vectors $\mathbf{X}_k$, $\mathbf{X}_m - \mathbf{X}_k$, and $\mathbf{X} - \mathbf{X}_m$ are independent.
2. $\|\mathbf{X}_k\|^2 \sim \chi_k^2(\|\boldsymbol{\mu}_k\|)$, $\|\mathbf{X}_m - \mathbf{X}_k\|^2 \sim \chi_{m-k}^2(\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_k\|)$, and $\|\mathbf{X} - \mathbf{X}_m\|^2 \sim \chi_{n-m}^2(\|\boldsymbol{\mu} - \boldsymbol{\mu}_m\|)$.

Theorem 3.10 is frequently used in the statistical analysis of *normal linear models*; see Sect. 5.3.1. In typical situations $\boldsymbol{\mu}$ lies in the subspace $\mathscr{V}_m$ or even $\mathscr{V}_k$—in which case $\|\mathbf{X}_m - \mathbf{X}_k\|^2 \sim \chi_{m-k}^2$ and $\|\mathbf{X} - \mathbf{X}_m\|^2 \sim \chi_{n-m}^2$, independently. The (scaled) quotient then turns out to have an F distribution—a consequence of the following theorem.

Theorem 3.11. (Relationship Between χ^2 and F Distributions). Let $U \sim \chi_m^2$ and $V \sim \chi_n^2$ be independent. Then,

$$\frac{U/m}{V/n} \sim \mathsf{F}(m, n) \ .$$

Proof. For notational simplicity, let $c = m/2$ and $d = n/2$. It follows from Example 3.7 that the pdf of $W = U/V$ is given by

$$f_W(w) = \int_0^\infty f_U(wv) \, v \, f_V(v) \, \mathrm{d}v$$

$$= \int_0^\infty \frac{(wv)^{c-1} \, e^{-wv/2}}{\Gamma(c) \, 2^c} \, v \, \frac{v^{d-1} e^{-v/2}}{\Gamma(d) \, 2^d} \, dv$$

$$= \frac{w^{c-1}}{\Gamma(c) \, \Gamma(d) \, 2^{c+d}} \int_0^\infty v^{c+d-1} \, e^{-(1+w)v/2} \, dv$$

$$= \frac{\Gamma(c+d)}{\Gamma(c) \, \Gamma(d)} \, \frac{w^{c-1}}{(1+w)^{c+d}} \, ,$$

where the last equality follows from the fact that the integrand is equal to $\Gamma(\alpha)\lambda^\alpha$ times the density of the $\mathsf{Gamma}(\alpha, \lambda)$ distribution with $\alpha = c + d$ and $\lambda = (1 + w)/2$. The proof is completed by observing that the density of $Z = \frac{n}{m} \frac{U}{V}$ is given by

$$f_Z(z) = f_W(z \, m/n) \, m/n \, .$$

$\square$

Corollary 3.2. (Relationship Between Normal, χ^2, and t Distributions).
Let $Z \sim \mathsf{N}(0, 1)$ and $V \sim \chi_n^2$ be independent. Then,

$$\frac{Z}{\sqrt{V/n}} \sim \mathsf{t}_n \, .$$

Proof. Let $T = Z/\sqrt{V/n}$. Because $Z^2 \sim \chi_1^2$, we have by Theorem 3.11 that $T^2 \sim \mathsf{F}(1, n)$. The result follows now from Theorem 2.19 and the symmetry around 0 of the pdf of T.   ☞ 51 $\square$

3.7   Limit Theorems

Two main results in probability are the *law of large numbers* and *the central limit theorem*. Both are limit theorems involving sums of independent random variables. In particular, consider a sequence $X_1, X_2, \ldots$ of iid random variables with finite expectation μ and finite variance σ^2. For each n define $S_n = X_1 + \cdots + X_n$. What can we say about the (random) sequence of sums $S_1, S_2, \ldots$ or averages $S_1, S_2/2, S_3/3, \ldots$? By (3.14) and (3.18) we have $\mathbb{E}[S_n/n] = \mu$ and $\mathrm{Var}(S_n/n) = \sigma^2/n$. Hence, as n increases, the variance of the (random) average S_n/n goes to 0. Informally, this means that (S_n/n) tends to the constant μ, as $n \to \infty$. This makes intuitive sense, but the important point is that the mathematical theory *confirms* our intuition in this respect. Here is a more precise statement.   ☞ 75

Theorem 3.12. (Weak Law of Large Numbers). If $X_1, \ldots, X_n$ are iid with finite expectation μ and finite variance σ^2, then for all $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(|S_n/n - \mu| > \varepsilon\right) = 0 \,.$$

Proof. Let $Y = (S_n/n - \mu)^2$ and $\delta = \varepsilon^2$. We have

$$\mathrm{Var}(S_n/n) = \mathbb{E}Y = \mathbb{E}[Y \mathrm{I}_{\{Y > \delta\}}] + \mathbb{E}[Y \mathrm{I}_{\{Y \leqslant \delta\}}] \geqslant \mathbb{E}[\delta \, \mathrm{I}_{\{Y > \delta\}}] + 0$$

$$= \delta \, \mathbb{P}(Y > \delta) = \varepsilon^2 \, \mathbb{P}(|S_n/n - \mu| > \varepsilon) \,.$$

Rearranging gives

$$\mathbb{P}(|S_n/n - \mu| > \varepsilon) \leqslant \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\sigma^2}{n \, \varepsilon^2} \,.$$

The proof is concluded by observing that $\sigma^2/(n\varepsilon^2)$ goes to 0 as $n \to \infty$. □

Remark 3.4. In Theorem 3.12 the qualifier "weak" is used to distinguish the result from the *strong* law of large numbers, which states that

$$\mathbb{P}(\lim_{n \to \infty} S_n/n = \mu) = 1 \,.$$

In terms of a computer simulation this means that the probability of drawing a sequence for which the sequence of averages fails to converge to μ is zero. The strong law implies the weak law, but is more difficult to prove in its full generality; see, for example, (Feller 1970).

The central limit theorem describes the approximate distribution of S_n (or S_n/n), and it applies to both continuous and discrete random variables. Loosely, it states that

> *the sum of a large number of iid random variables approx-*
> *imately has a normal distribution.*

Specifically, the random variable S_n has a distribution that is approximately normal, with expectation $n\mu$ and variance $n\sigma^2$. A more precise statement is given next.

Theorem 3.13. (Central Limit Theorem). If $X_1, \ldots, X_n$ are iid with finite expectation μ and finite variance σ^2, then for all $x \in \mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma \sqrt{n}} \leqslant x\right) = \Phi(x) \,,$$

where Φ is the cdf of the standard normal distribution.

Proof. (Sketch) A full proof is out of the scope of this book. However, the main ideas are not difficult. Without loss of generality assume $\mu = 0$ and $\sigma = 1$. This amounts to replacing X_n by $(X_n - \mu)/\sigma$. We also assume, for simplicity, that the moment generating function of X_i is finite in an open interval containing 0, so that we can use Theorem 2.7. We wish to show that the cdf of $S_n/\sqrt{n}$ converges to that ☞ 36 of the standard normal distribution. It can be proved (and makes intuitive sense) that this is equivalent (up to some technical conditions) to demonstrating that the corresponding moment generating functions converge. That is, we wish to show that

$$\lim_{n\to\infty} \mathbb{E}\exp\left(t\frac{S_n}{\sqrt{n}}\right) = e^{\frac{1}{2}t^2}, \quad t \in \mathbb{R},$$

where the right-hand side is the moment generating function of the standard normal distribution. Because $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = \mathrm{Var}(X_1) = 1$, we have by Theorem 2.7 that the moment generation function of X_1 has the following Taylor expansion: ☞ 369

$$M(t) \stackrel{\text{def}}{=} \mathbb{E}\,e^{tX_1} = 1 + t\,\mathbb{E}X_1 + \frac{1}{2}t^2\,\mathbb{E}X_1^2 + o(t^2) = 1 + \frac{1}{2}t^2 + o(t^2),$$

where $o(t^2)$ is a function for which $\lim_{t\downarrow 0} o(t^2)/t^2 = 0$. Because the $\{X_i\}$ are iid, it follows that the moment generating function of $S_n/\sqrt{n}$ satisfies

$$\mathbb{E}\exp\left(t\frac{S_n}{\sqrt{n}}\right) = \mathbb{E}\exp\left(\frac{t}{\sqrt{n}}(X_1 + \cdots + X_n)\right) = \prod_{i=1}^{n} \mathbb{E}\exp\left(\frac{t}{\sqrt{n}}X_i\right)$$

$$= M^n\left(\frac{t}{\sqrt{n}}\right) = \left[1 + \frac{t^2}{2n} + o(t^2/n)\right]^n \longrightarrow e^{\frac{1}{2}t^2}$$

as $n \to \infty$. □

Figure 3.7 shows central limit theorem in action. The left part shows the pdfs of $S_1, \ldots, S_4$ for the case where the $\{X_i\}$ have a $\mathsf{U}[0, 1]$ distribution. The right part shows the same for the $\mathsf{Exp}(1)$ distribution. We clearly see convergence to a bell-shaped curve, characteristic of the normal distribution.
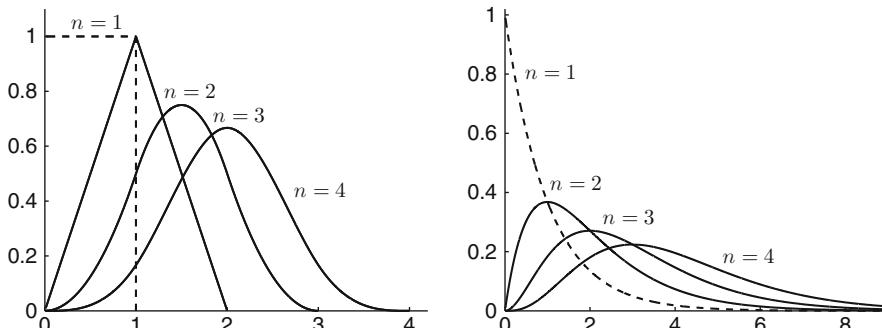
Fig. 3.7 Illustration of the central limit theorem for (*left*) the uniform distribution and (*right*) the exponential distribution

Recall that a binomial random variable $X \sim \mathsf{Bin}(n, p)$ can be viewed as the sum
☞ 66   of n iid $\mathsf{Ber}(p)$ random variables: $X = X_1 + \cdots + X_n$. As a direct consequence
of the central limit theorem it follows that, for large n, $\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k)$,
where $Y \sim \mathsf{N}(np, np(1 - p))$. As a rule of thumb, this normal approximation to the
binomial distribution is accurate if both np and $n(1 - p)$ are larger than 5.

There is also a central limit theorem for random vectors. The multidimensional
version is as follows.

> **Theorem 3.14. (Multivariate Central Limit Theorem).** Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be
> iid random vectors with expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ. For
> large n the random vector $\mathbf{X}_1 + \cdots + \mathbf{X}_n$ approximately has a $\mathsf{N}(n\boldsymbol{\mu}, n\Sigma)$
> distribution.

A more precise formulation of the above theorem is that the average random
vector $\mathbf{Z}_n = (\mathbf{X}_1 + \cdots + \mathbf{X}_n)/n$, when rescaled via $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$, converges in
distribution to a random vector $\mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ as $n \to \infty$. A useful consequence of
this is given next.

> **Theorem 3.15. (Delta Method).** Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ be a sequence of random
> vectors such that $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu}) \to \mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ as $n \to \infty$. Then, for any
> continuously differentiable function $\mathbf{g}$ of $\mathbf{Z}_n$,
>
> $$\sqrt{n}(\mathbf{g}(\mathbf{Z}_n) - \mathbf{g}(\boldsymbol{\mu})) \to \mathbf{R} \sim \mathsf{N}(\mathbf{0}, J \Sigma J^\top) , \qquad (3.36)$$
>
> where $J = J(\boldsymbol{\mu}) = (\partial g_i(\boldsymbol{\mu})/\partial x_j)$ is the Jacobian matrix of $\mathbf{g}$ evaluated at $\boldsymbol{\mu}$.

Proof. (Sketch) A formal proof requires some deeper knowledge of statistical
convergence, but the idea of the proof is quite straightforward. The key step is to
☞ 369   construct the first-order Taylor expansion (see Theorem B.1) of $\mathbf{g}$ around $\boldsymbol{\mu}$, which
yields

$$\mathbf{g}(\mathbf{Z}_n) = \mathbf{g}(\boldsymbol{\mu}) + J(\boldsymbol{\mu})(\mathbf{Z}_n - \boldsymbol{\mu}) + \mathcal{O}(\|\mathbf{Z}_n - \boldsymbol{\mu}\|^2) .$$

As $n \to \infty$, the remainder term goes to 0, because $\mathbf{Z}_n \to \boldsymbol{\mu}$. Hence, the left-
hand side of (3.36) is approximately $J \sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$. For large n this converges to
☞ 80   a random vector $\mathbf{R} = J \mathbf{K}$, where $\mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$. Finally, by Theorem 3.4 , we have
$\mathbf{R} \sim \mathsf{N}(\mathbf{0}, J \Sigma J^\top)$.   □

Example 3.15 (Ratio Estimator). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid copies of a random vector (X, Y) with mean vector (μ_X, μ_Y) and covariance matrix Σ. Denoting the average of the $\{X_i\}$ and $\{Y_i\}$ by $\bar{X}$ and $\bar{Y}$, respectively, what can we say about the distribution of $\bar{X}/\bar{Y}$ for large n?

Let $\mathbf{Z}_n = (\bar{X}, \bar{Y})$ and $\boldsymbol{\mu} = (\mu_X, \mu_Y)$. By the multivariate central limit theorem $\mathbf{Z}_n$ has approximately a $\mathsf{N}(\boldsymbol{\mu}, \Sigma/n)$ distribution. More precisely, $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$ converges to a $\mathsf{N}(\mathbf{0}, \Sigma)$-distributed random vector.

We apply the delta method using the function $g(x, y) = x/y$, whose Jacobian matrix is

$$J(x, y) = \left(\frac{\partial g(x, y)}{\partial x}, \quad \frac{\partial g(x, y)}{\partial y} \right) = \left(\frac{1}{y}, \quad \frac{-x}{y^2} \right) .$$

It follows from (3.36) that $g(\bar{X}, \bar{Y}) = \bar{X}/\bar{Y}$ has approximately a normal distribution with expectation $g(\boldsymbol{\mu}) = \mu_X/\mu_Y$ and variance σ^2/n, where

$$\sigma^2 = J(\boldsymbol{\mu}) \Sigma J^\top(\boldsymbol{\mu}) = \left(\frac{1}{\mu_Y}, \quad \frac{-\mu_X}{\mu_Y^2} \right) \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}(Y) \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_Y} \\ \frac{-\mu_X}{\mu_Y^2} \end{pmatrix}$$

$$= \left(\frac{\mu_X}{\mu_Y} \right)^2 \left(\frac{\mathrm{Var}(X)}{\mu_X^2} + \frac{\mathrm{Var}(Y)}{\mu_Y^2} - 2 \frac{\mathrm{Cov}(X, Y)}{\mu_X \mu_Y} \right) .$$

$$\text{(3.37)}$$

3.8 Problems

3.1. Let U and V be independent random variables with $\mathbb{P}(U = 1) = \mathbb{P}(V = 1) = 1/4$ and $\mathbb{P}(U = -1) = \mathbb{P}(V = -1) = 3/4$. Define $X = U/V$ and $Y = U + V$. Give the joint discrete pdf of X and Y in table form, as in Table 3.1. Are X and Y independent?

3.2. Let $X_1, \ldots, X_4 \sim_{\mathrm{iid}} \mathsf{Ber}(p)$.

(a) Give the joint discrete pdf of $X_1, \ldots, X_4$.
(b) Give the joint discrete pdf of $X_1, \ldots, X_4$ given $X_1 + \cdots + X_4 = 2$.

3.3. Three identical-looking urns each have 4 balls. Urn 1 has 1 red and 3 white balls, urn 2 has 2 red and 2 white balls, and urn 3 has 3 red and 1 white ball. We randomly select an urn with equal probability. Let X be the number of the urn. We then draw 2 balls from the selected urn. Let Y be the number of red balls drawn. Find the following discrete pdfs:

(a) The pdf of X.
(b) The conditional pdf of Y given $X = x$ for $x = 1, 2, 3$.
(c) The joint pdf of X and Y.

(d) The pdf of Y.
(e) The conditional pdf of X given $Y = y$ for $y = 0, 1, 2$.

3.4. We randomly select a point (X, Y) from the triangle $\{(x, y) : x, y \in \{1, \ldots, 6\}, y \leqslant x\}$ (see Fig. 3.1) in the following *nonuniform* way. First, select X discrete uniformly from $\{1, \ldots, 6\}$. Then, given $X = x$, select Y discrete uniformly from $\{1, \ldots, x\}$. Find the conditional distribution of X given $Y = 1$ and its corresponding conditional expectation.

3.5. We randomly and uniformly select a continuous random vector (X, Y) in the triangle $(0, 0)$–$(1, 0)$–$(1, 1)$, the same triangle as in Example 3.8.

(a) Give the joint pdf of X and Y.
(b) Calculate the pdf of Y and sketch its graph.
(c) Specify the conditional pdf of Y given $X = x$ for any fixed $x \in (0, 1)$.
(d) Determine $\mathbb{E}[Y \mid X = 1/2]$.

3.6. Let $X \sim \mathsf{U}[0, 1]$ and $Y \sim \mathsf{Exp}(1)$ be independent.

(a) Determine the joint pdf of X and Y and sketch its graph.
(b) Calculate $\mathbb{P}((X, Y) \in [0, 1] \times [0, 1])$.
(c) Calculate $\mathbb{P}(X + Y < 1)$.

3.7. Let $X \sim \mathsf{Exp}(\lambda)$ and $Y \sim \mathsf{Exp}(\mu)$ be independent.

(a) Show that $\min(X, Y)$ also has an exponential distribution, and determine its corresponding parameter.
(b) Show that

$$\mathbb{P}(X < Y) = \frac{\lambda}{\lambda + \mu} \, .$$

3.8. Let $X \sim \mathsf{Exp}(1)$ and $(Y \mid X = x) \sim \mathsf{Exp}(x)$.

(a) What is the joint pdf of X and Y?
(b) What is the marginal pdf of Y?

3.9. Let $X \sim \mathsf{U}(-\pi/2, \pi/2)$. Show that $Y = \tan(X)$ has a Cauchy distribution.

3.10. Let $X \sim \mathsf{Exp}(3)$ and $Y = \ln(X)$. What is the pdf of Y?

3.11. We draw n numbers independently and uniformly from the interval $[0,1]$ and denote their sum S_n.

(a) Determine the pdf of S_2 and sketch its graph.
(b) What is approximately the distribution of S_{20}?
(c) Approximate the probability that the average of the 20 numbers is greater than 0.6.

3.12. A certain type of electrical component has an exponential lifetime distribution with an expected lifetime of $1/2$ year. When the component fails it is immediately

replaced by a second (new) component; when the second component fails, it is replaced by a third, etc. Suppose there are ten such identical components. Let T be the time that the last of the components fails.

(a) What is the expectation and variance of T?
(b) Approximate, using the central limit theorem, the probability that T exceeds 6 years.
(c) What is the exact distribution of T?

3.13. Let A be an invertible $n \times n$ matrix and let $X_1, \ldots, X_n \sim_{\text{iid}} N(0, 1)$. Define $\mathbf{X} = (X_1, \ldots, X_n)^\top$ and let $(Z_1, \ldots, Z_n)^\top = A\mathbf{X}$. Show that $Z_1, \ldots, Z_n$ are iid standard normal only if $AA^\top = I$ (identity matrix), in other words, only if A is an *orthogonal* matrix. Can you find a geometric interpretation of this?

3.14. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Let $\bar{X} = (X_1 + \cdots + X_n)/n$. Calculate the correlation coefficient of X_1 and $\bar{X}$.

3.15. Suppose that $X_1, \ldots, X_6$ are iid with pdf

$$f(x) = \begin{cases} 3x^2, & 0 \leqslant x \leqslant 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) What is the probability that all $\{X_i\}$ are greater than $1/2$?
(b) Find the probability that at least one of the $\{X_i\}$ is less than $1/2$.

3.16. Let X and Y be random variables.

(a) Express $\text{Var}(-aX + Y)$, where a is a constant, in terms of $\text{Var}(X)$, $\text{Var}(Y)$, and $\text{Cov}(X, Y)$.
(b) Take $a = \text{Cov}(X, Y)/\text{Var}(X)$. Using the fact that the variance in (a) is always nonnegative, prove the following **Cauchy–Schwarz inequality**:

$$(\text{Cov}(X, Y))^2 \leqslant \text{Var}(X) \, \text{Var}(Y) .$$

(c) Show that, as a consequence, the correlation coefficient of X and Y must lie between -1 and 1.

3.17. Suppose X and Y are independent uniform random variables on $[0,1]$. Let $U = X/Y$ and $V = XY$, which means $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$.

(a) Sketch the two-dimensional region where the density of (U, V) is nonzero.
(b) Find the matrix of Jacobi for the transformation $(x, y)^\top \mapsto (u, v)^\top$.
(c) Show that its determinant is $2x/y = 2u$.
(d) What is the joint pdf of U and V?
(e) Show that the marginal pdf of U is

$$f_U(u) = \begin{cases} \frac{1}{2}, & 0 < u < 1 \\ \frac{1}{2u^2}, & u \geqslant 1 \end{cases} . \tag{3.38}$$

3.18. Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $Y = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

(a) Show that

$$Y = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 .$$

(b) Calculate $\mathbb{E}Y$.
(c) Show that $\mathbb{E}Y \to \sigma^2$ as $n \to \infty$.

3.19. Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$, with $\{X_i\} \sim_{\text{iid}} \mathsf{N}(\mu, 1)$. Consider the orthogonal projection, denoted $\mathbf{X}_1$, of $\mathbf{X}$ onto the subspace spanned by $\mathbf{1} = (1, \ldots, 1)^\top$.

(a) Show that $\mathbf{X}_1 = \bar{X}\mathbf{1}$.
(b) Show that $\mathbf{X}_1$ and $\mathbf{X} - \mathbf{X}_1$ are independent.
(c) Show that $\|\mathbf{X} - \mathbf{X}_1\|^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$ has a χ_{n-1}^2 distribution.

Hint: apply Theorem 3.10.

3.20. Let $X_1, \ldots, X_6$ be the weights of six randomly chosen people. Assume each weight is $\mathsf{N}(75, 100)$ distributed (in kg). Let $W = X_1 + \cdots + X_6$ be the total weight of the group. Explain why the distribution of W is equal or not equal to $6X_1$.

3.21. Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ be independent. Show that $X + Y \sim \chi_{m+n}^2$. Hint: use moment generating functions.

3.22. Let $X \sim \mathsf{N}(\mu, 1)$. Show that the moment generation function of X^2 is

$$M(t) = \frac{e^{\mu^2 t/(1-2t)}}{\sqrt{1 - 2t}} \quad t < 1/2 .$$

Next, consider independent random variables $X_i \sim \mathsf{N}(\mu_i, 1)$, $i = 1, \ldots, n$. Use the result above to show that the distribution of $\|\mathbf{X}\|^2$ only depends on n and $\|\boldsymbol{\mu}\|$. Can you find a symmetry argument why this must be so?

3.23. A machine produces cylinders with a diameter which is normally distributed with mean 3.97 cm and standard deviation 0.03 cm. Another machine produces (independently of the first machine) shafts with a diameter which is normally distributed with mean 4.05 cm and standard deviation 0.02 cm. What is the probability that a randomly chosen cylinder fits into a randomly chosen shaft?

3.24. A sieve with diameter d is used to separate a large number of blueberries into two classes: small and large. Suppose that the diameters of the blueberries are normally distributed with an expectation $\mu = 1$ cm and a standard deviation $\sigma = 0.1$ cm.

(a) Find the diameter of the sieve such that the proportion of large blueberries is 30 %.

(b) Suppose that the diameter is chosen such as in (a). What is the probability that out of 1000 blueberries, fewer than 280 end up in the "large" class?

3.25. Suppose X, Y, and Z are independent $N(1, 2)$-distributed random variables. Let $U = X - 2Y + 3Z$ and $V = 2X - Y + Z$. Give the joint distribution of U and V.

3.26. For many of the above problems it is instructive to simulate the corresponding model on a computer in order to better understand the theory.

(a) Generate 10^5 points (X, Y) from the model in Problem 3.6.
(b) Compare the fraction of points falling in the unit square $[0, 1] \times [0, 1]$ with the theoretical probability in Problem 3.6(b).
(c) Do the same for the probability $\mathbb{P}(X + Y < 1)$.

3.27. Simulate 10^5 draws from $U(-\pi/2, \pi/2)$ and transform these using the tangent function, as in Problem 3.9. Compare the histogram of the transformed values with the theoretical (Cauchy) pdf.

3.28. Simulate 10^5 independent draws of (U, V) in Problem 3.17. Verify with a histogram of the U-values that the pdf of U is of the form (3.38).

3.29. Consider the MATLAB experiments in Example 3.14.

(a) Carry out the experiments with $\varrho = 0.4, 0.7, 0.9, 0.99$, and -0.8, and observe how the outcomes change.   ☞ 84
(b) Plot the corresponding pdfs, as in Fig. 3.6.
(c) Give also the contour plots of the pdfs, for $\varrho = 0$ and $\varrho = 0.8$. Observe that the contours are *ellipses*.
(d) Show that these ellipses are of the form

$$x_1^2 + 2\varrho\, x_1\, x_2 + x_2^2 = \text{constant} .$$

Part II
Statistical Modeling and Classical and Bayesian Inference

In Part II of the book we consider the modeling and analysis of random data. We describe various common models for data and discuss the mathematical tools of statistical inference. Both the classical (frequentist) and Bayesian viewpoint of statistics are covered. Classical statistics' main focus is the *likelihood* concept, whereas Bayesian statistics deals primarily with the *posterior distribution* of the model parameters. Both classical and Bayesian methods often involve *Monte Carlo sampling* techniques. It is assumed that the reader is familiar with the probability topics discussed in Part I.

Chapter 4
Common Statistical Models

The conceptual framework for statistical modeling and analysis is sketched in Fig. 4.1. The starting point is some real-life problem (*reality*) and a corresponding set of *data*. On the basis of the data we wish to say something about the real-life problem. The second step consists of finding a probabilistic *model* for the data. This model contains what we know about the reality and how the data were obtained. Within the model we carry out our calculations and analysis. This leads to conclusions about the model. Finally, the conclusions about the model are translated into conclusions about the reality.
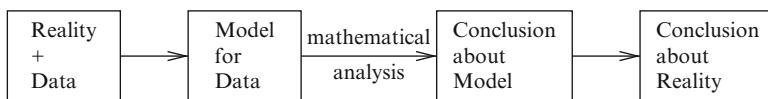
Fig. 4.1 Statistical modeling and analysis

Mathematical statistics uses probability theory and other branches of mathematics to study data. In particular, the data are viewed as realizations of random variables whose joint distribution is specified in advance, possibly up to some unknown parameter(s). The mathematical analysis is then purely about the model and its parameters.

☞ 63

4.1 Independent Sampling from a Fixed Distribution

One of the simplest statistical models is the one where the data $X_1, \ldots, X_n$ are assumed to be independent and identically distributed (iid). We write

☞ 66

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \overset{\circ}{f} \quad \text{or} \quad X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Dist} \, ,$$

to indicate that the random variables form an iid sample from a sampling pdf $\overset{\circ}{f}$
or sampling distribution Dist. Let f denote the joint pdf of $X_1, \ldots, X_n$. Then, by
Theorem 3.1,

$$f(x_1, \ldots, x_n) = \overset{\circ}{f}(x_1) \cdots \overset{\circ}{f}(x_n) .$$

Example 4.1 (Experiments with Iid Samples). Each of the following scenarios
can be modeled via an iid sample.

1. We roll a die 100 times and record at each throw whether a 6 appears or not. Let
   $X_i = 1$ if the ith throw yields a 6 and $X_i = 0$ otherwise, for $i = 1, \ldots, 100$.
   Then,

   $$X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Ber}(p)$$

   for some known or unknown p. For example, if the die is known to be fair, then
   $p = 1/6$.
2. From a large population we select 300 men between 40 and 50 years of age
   and measure their heights. Let X_i be the height of the ith selected person, $i =
   1, \ldots, 300$. Then,

   $$X_1, \ldots, X_{300} \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$$

   for some unknown parameters μ and σ^2.
3. A large marine reserve is divided into 20 similar habitats. In each habitat the
   number octopuses is recorded. Let X_i be the number of octopuses in habitat i,
   $i = 1, \ldots, 20$. Then,

   $$X_1, \ldots, X_{20} \overset{\text{iid}}{\sim} \text{Poi}(\mu)$$

   for some unknown parameter $\mu > 0$.
4. We run a simulation program for a production system for cars and record the total
   production in a day. We repeat this 10 times, each time starting the simulation
   with a different seed. Let X_i be the production per day in the ith simulation,
   $i = 1, \ldots, 10$. Then,

   $$X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \text{Dist}$$

   for some unknown distribution Dist.

Remark 4.1 (About Statistical Modeling). At this point it is good to emphasize a
few points about modeling.

- *Any* model for data is likely to be *wrong*. For example, in Scenario 2 above
  the height would normally be recorded on a discrete scale, say 1000–2200
  (mm). However, samples from a $\text{N}(\mu, \sigma^2)$ can take any real value, including
  negative values. Nevertheless, the normal distribution could be a reasonable

approximation to the real sampling distribution. An important advantage of using a normal distribution is that it has many nice mathematical properties, as described in Sect. 3.6.

- Most statistical models depend on a number of *unknown* parameters. One of the main objectives of *statistical inference*—to be discussed in subsequent chapters—is to gain knowledge of the unknown parameters on the basis of the observed data. Even in Scenario 4 of Example 4.1 the model depends on underlying simulation parameters, although the distribution Dist may not be explicitly known.

- Any model for data needs to be checked for suitability. An important criterion is that data simulated from the model should resemble the observed data—at least for a certain choice of model parameters. This is automatically satisfied for Scenario 4 but should be verified for Scenarios 2 and 3. Model checking and selection is discussed in Sects. 5.3.1, 5.4, 8.6, and 10.1.1.

4.2   Multiple Independent Samples

The single iid sample in Sect. 4.1 is easily generalized to multiple iid samples. The most common models involve Bernoulli and normal random variables.

Example 4.2 (Two-Sample Binomial Model). To assess whether there is a difference between boys and girls in their preference for two brands of cola, say *Sweet* and *Ultra* cola, we select at random 100 boys and 100 girls and ask whether they prefer *Sweet* or *Ultra*. We could model this via two independent Bernoulli samples. That is, for each $i = 1, \ldots, 100$ let $X_i = 1$ if the ith boy prefers *Sweet* over *Ultra* and let $X_i = 0$ otherwise. Similarly, let $Y_i = 1$ if the ith girl prefers *Sweet* over *Ultra*. We thus have the model

$$X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \mathsf{Ber}(p_1) \,,$$

$$Y_1, \ldots, Y_{100} \overset{\text{iid}}{\sim} \mathsf{Ber}(p_2) \,,$$

$$X_1, \ldots, X_{100}, Y_1, \ldots, Y_{100} \ \text{independent, with } p_1 \text{ and } p_2 \text{ unknown} \,.$$

The objective is to assess the difference $p_1 - p_2$ on the basis of the observed values for $X_1, \ldots, X_{100}, Y_1, \ldots, Y_{100}$. Note that it suffices to only record the total number of boys or girls who prefer *Sweet* cola in each group; that is, $X = \sum_{i=1}^{100} X_i$ and $Y = \sum_{i=1}^{100} Y_i$. This gives the **two-sample binomial model**:

$$X \sim \mathsf{Bin}(100, p_1) \,,$$

$$Y \sim \mathsf{Bin}(100, p_2) \,,$$

$$X, Y \ \text{independent, with } p_1 \text{ and } p_2 \text{ unknown} \,.$$

Example 4.3 (Two-Sample Normal Model). From a large population we select 200 men between 25 and 30 years of age and measure their heights. For each person we also record whether the mother smoked during pregnancy or not. Suppose that 60 mothers smoked during pregnancy.

Let $X_1, \ldots, X_{60}$ be the heights (in cm) of the men whose mothers smoked, and let $Y_1, \ldots, Y_{140}$ be the heights of the men whose mothers did not smoke. Then, a possible model is the **two-sample normal model**:

$$X_1, \ldots, X_{60} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_1, \sigma_1^2) \,,$$

$$Y_1, \ldots, Y_{140} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_2, \sigma_2^2) \,,$$

$$X_1, \ldots, X_{60}, Y_1, \ldots, Y_{140} \quad \text{independent} \,,$$

where the model parameters μ_1, μ_2, σ_1^2, and σ_2^2 are unknown. One would typically like to assess the difference $\mu_1 - \mu_2$. That is, does smoking during pregnancy affect the (expected) height of the child? A typical simulation outcome of the model is given in Fig. 4.2, using parameters $\mu_1 = 175, \mu_2 = 170, \sigma_1^2 = 100$, and $\sigma_2^2 = 200$.
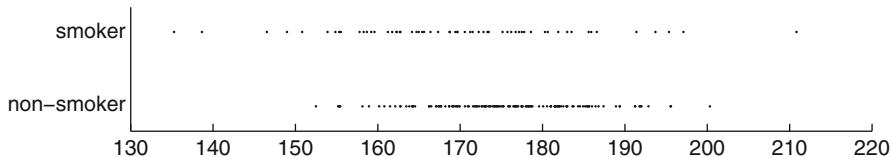
Fig. 4.2 Simulated height data from a two-sample normal model

Instead of dividing the data into two groups, one could further divide the "smoking mother" group into several subgroups according to the intensity of smoking, e.g., rarely, moderately, and heavily, so that in this case the data could be modeled via four independent samples from a normal distribution. This model would, in general, depend on eight unknown parameters—four expectations and four variances.

4.3   Regression Models

Francis Galton observed in an article in 1889 that the heights of adult offspring are, on the whole, more "average" than the heights of their parents. Galton interpreted this as a degenerative phenomenon, using the term *regression*, to indicate this "return to mediocrity." Karl Pearson continued Galton's original work and conducted comprehensive studies comparing various relationships between members of the same family. Figure 4.3 depicts the measurements of the heights of 1078 fathers and their adult sons (one son per father).

The average height of the fathers was 67 inches and of the sons 68 inches. Because sons are on average 1 inch taller than the fathers we could try to "explain" the height of the son by taking the height of his father and adding 1 inch. However,
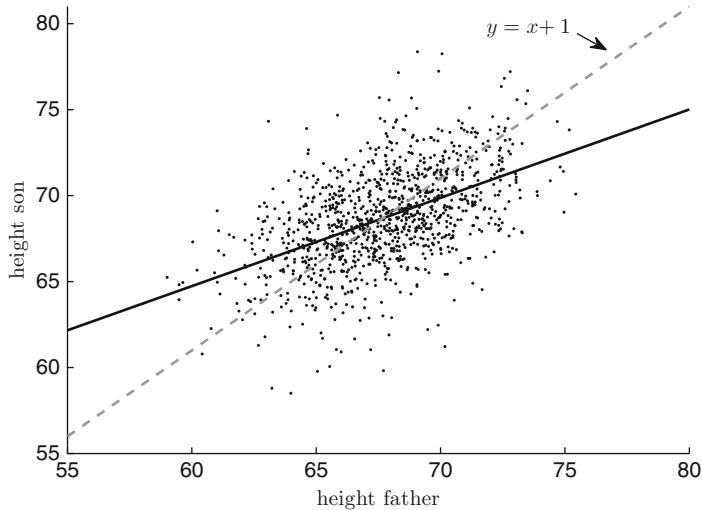
Fig. 4.3 A scatter plot of heights from Pearson's data

the line $y = x + 1$ (dashed) does not seem to predict the height of the sons as accurately as the solid line in Fig. 4.3. This line has a slope less than 1 and demonstrates Galton's "regression" effect. For example, if a father is 5 % taller than average, then his son will be on the whole *less* than 5 % taller than average.

In general, regression analysis is about finding relationships between a number of variables. In particular, there is a **response** variable which we would like to "explain" via one or more **explanatory** variables. Explanatory variables are also called **predictors**, **covariates**, and **independent variables**. In the latter case the response variable is called the **dependent variable**. Regression is usually seen as a functional relationship between *continuous* variables.

4.3.1 Simple Linear Regression

The most basic regression model involves a linear relationship between the response and a single explanatory variable. As in Pearson's height data, we have measurements $(x_1, y_1), \ldots, (x_n, y_n)$ that lie approximately on a straight line. It is assumed that these measurements are outcomes of vectors $(x_1, Y_1), \ldots, (x_n, Y_n)$, where, for each *deterministic* explanatory variable x_i, the response variable Y_i is a *random* variable with

$$\mathbb{E}Y_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n \tag{4.1}$$

for certain *unknown* parameters β_0 and β_1. The (unknown) line

$$y = \beta_0 + \beta_1 x \tag{4.2}$$

is called the **regression line**. To completely specify the model, we need to designate the joint distribution of $Y_1, \ldots, Y_n$. The most common linear regression model is given next. The adjective "simple" refers to the fact that a *single* explanatory variable is used to explain the response.

Definition 4.1. (Simple Linear Regression Model). In a **simple linear regression model** the response data $Y_1, \ldots, Y_n$ depend on explanatory variables $x_1, \ldots, x_n$ via the linear relationship

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \quad i = 1, \ldots, n , \tag{4.3}$$

where $\varepsilon_1, \ldots, \varepsilon_n \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

This formulation makes it even more obvious that we view the responses as random variables which would lie exactly on the regression line, were it not for some "disturbance" or "error" term (represented by the $\{\varepsilon_i\}$).

Note that the simple linear regression model (4.3) is a *Gaussian* model; that is, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ has a multivariate normal distribution. Namely,

☞ 83

$$\mathbf{Y} \sim \mathsf{N}(\beta_0 \mathbf{1} + \beta_1 \mathbf{x}, \ \sigma^2 I) , \tag{4.4}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^\top$, $\mathbf{1}$ is the n-dimensional column vector of 1s, and I is the n-dimensional identity matrix.

4.3.2   Multiple Linear Regression

A linear regression model that contains more than one explanatory variable is called a *multiple linear regression model*.

Definition 4.2.   (Multiple Linear Regression Model). In a **multiple linear regression model** the response data $Y_1, \ldots, Y_n$ depend on d-dimensional explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$, via the linear relationship

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i , \quad i = 1, \ldots, n , \tag{4.5}$$

where $\varepsilon_1, \ldots, \varepsilon_n \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

We can write (4.5) as $Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$. In other words, the data $(\mathbf{x}_i, Y_i)$—where the $\{Y_i\}$ are random—lie approximately on the plane $y = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ for some (typically unknown) constant β_0 and vector $\boldsymbol{\beta}$. Defining $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and A as the matrix

$$
A = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix},
$$

we can reformulate (4.5) as the Gaussian model

$$
\mathbf{Y} \sim \mathsf{N}(\beta_0 \mathbf{1} + A\boldsymbol{\beta}, \ \sigma^2 I) , \tag{4.6}
$$

where $\mathbf{1}$ is the n-dimensional column vector of 1s and I is the n-dimensional identity matrix.

Example 4.4 (Multiple Linear Regression Model). Figure 4.4 gives a realization of the multiple linear regression model

$$
Y_i = x_{i1} + x_{i2} + \varepsilon_i , \quad i = 1, \ldots, 100 ,
$$

where $\varepsilon_1, \ldots, \varepsilon_{100} \sim_{\text{iid}} \mathsf{N}(0, 1/16)$. The fixed vectors $(x_{i1}, x_{i2}), i = 1, \ldots, 100$ of explanatory variables lie in the unit square.
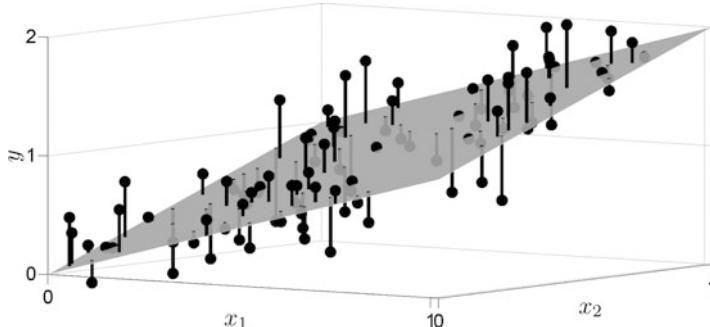
Fig. 4.4 Multiple linear regression

The multiple linear regression model can be viewed as a first-order approximation of the general model

$$
Y_i = b(\mathbf{x}_i) + \varepsilon_i , \quad i = 1, \ldots, n , \tag{4.7}
$$

where $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ and $b(\mathbf{x})$ is some known or unknown function of a d-dimensional vector $\mathbf{x}$ of explanatory variables. To see this, replace $b(\mathbf{x})$ with its

first-order Taylor approximation around some point $\mathbf{x}_0$:

$$b(\mathbf{x}) \approx b(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top \nabla b(\mathbf{x}_0)$$

$$= \underbrace{b(\mathbf{x}_0) - \mathbf{x}_0^\top \nabla b(\mathbf{x}_0)}_{\beta_0} + \mathbf{x}^\top \underbrace{\nabla b(\mathbf{x}_0)}_{\boldsymbol{\beta}} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} \; . \tag{4.8}$$

4.3.3 Regression in General

General regression models not only deal with multiple explanatory variables but also with *nonlinear* relationships between the response and explanatory variables. A broad class of regression models is (similar to (4.7)) of the form

$$Y_i = g(\mathbf{x}_i; \boldsymbol{\beta}) + \varepsilon_i \; , \quad i = 1, \ldots, n \; , \tag{4.9}$$

where $\varepsilon_1, \ldots, \varepsilon_n \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $g(\mathbf{x}; \boldsymbol{\beta})$ is a known function of the explanatory vector $\mathbf{x}$ and the parameter vector $\boldsymbol{\beta}$. Both σ^2 and $\boldsymbol{\beta}$ are assumed to be unknown.

To specify regression models of this form it suffices to report only the functional relationship between the expected response $y = \mathbb{E}Y$ and the explanatory variable (x or $\mathbf{x}$). For the generic model in (4.9) this corresponds to reporting only $y = g(\mathbf{x}; \boldsymbol{\beta})$. We will do this from now on in this section, keeping in mind the general formulation where there are n independent response variables, each with its own explanatory variable and error term.

When $g(\mathbf{x}; \boldsymbol{\beta})$ is *linear* in the parameters (or, equivalently, linear in $\boldsymbol{\beta}$), the model is said to be a **linear regression model**. The obvious examples are the simple linear regression and multiple linear regression models. The following example gives another important class of linear regression models.

Example 4.5 (Polynomial Regression Models). Suppose the expected response y depends on a single explanatory variable x via a polynomial relationship

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d \; . \tag{4.10}$$

This is an example of a **polynomial regression model**. Note that the model is linear in the model parameters $\{\beta_i\}$. In a similar way one can consider polynomial regression models with multiple explanatory variables, as in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \; , \tag{4.11}$$

which defines a second-order polynomial regression model with two explanatory variables. Similar to (4.8), this model can be viewed as a second-order approximation to a general regression model of the form

$$y = b(x_1, x_2)$$

for some known or unknown function b. Polynomial regression models are also called **response surface models**.

Common examples of *nonlinear* regression models are the following:

- **Exponential model** with parameters a and b:

$$y = a\,e^{bx}\,.$$

- **Power law model** with parameters a and b:

$$y = a\,x^{b}\,.$$

- **Logistic model** with parameters a and b and fixed L:

$$y = \frac{L}{1 + e^{a+bx}}\,.$$

- **Weibull model** with parameters a and b:

$$y = 1 - e^{-\frac{x^{b}}{a}}\,.$$

Example 4.6 (Exponential Model). Table 4.1 lists the free chlorine concentration (in mg per liter) in a swimming pool, recorded every 8 h for 4 days.

Table 4.1 Chlorine concentration (in mg/L) as a function of time (hours).

Hours	Concentration	Hours	Concentration
0	1.0056	56	0.3293
8	0.8497	64	0.2617
16	0.6682	72	0.2460
24	0.6056	80	0.1839
32	0.4735	88	0.1867
40	0.4745	96	0.1688
48	0.3563		

A simple chemistry-based model for the chlorine concentration y as a function of time t is

$$y = a\,e^{-b\,t}\,,$$

where a is the initial concentration and $b > 0$ is the *reaction rate*. Figure 4.5 shows that the data closely follow the curve $y = e^{-0.02t}$. A common method for fitting regression curves to data is the *least-squares* method, which will be discussed in Sect. 5.1.2.
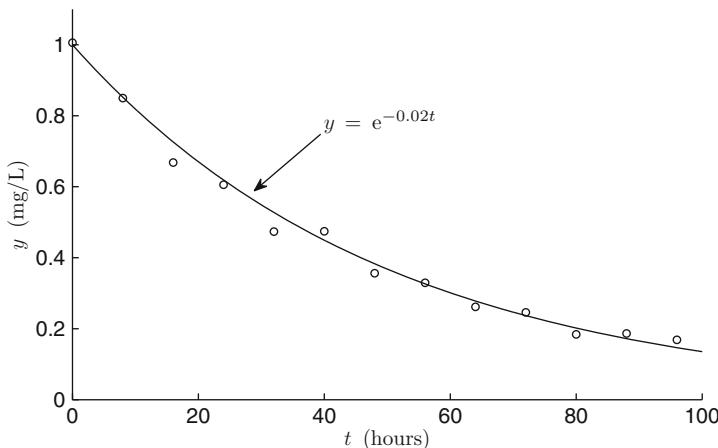
Fig. 4.5 The chlorine concentration seems to have an exponential decay

Another way to deal with nonlinearities in the data is to *transform* the data in order to achieve a linear relationship.

Example 4.7 (Log-Linear Model). Suppose that the expected chlorine concentration in Example 4.6 satisfies $y = a\,e^{-bt}$ for some unknown a and $b > 0$. Then, $\ln y = \ln a - bt$. Hence, there is a *linear* relationship between t and $\ln y$. Thus, if for some given data $(t_1, y_1), \ldots, (t_n, y_n)$ we plot $(t_1, \ln y_1), \ldots, (t_n, \ln y_n)$, these points should approximately lie on a straight line, and hence the simple linear regression model applies. Figure 4.6 illustrates that the transformed data indeed lie approximately on a straight line.
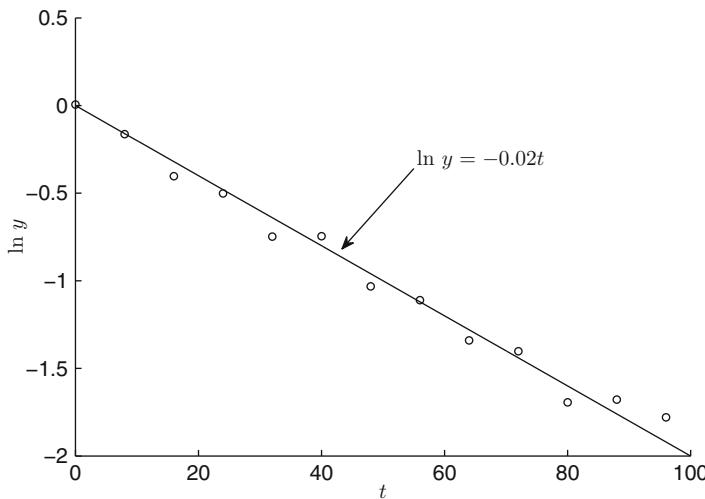
Fig. 4.6 The log-transform of the chlorine concentration can be modeled via a simple linear regression

4.4 Analysis of Variance Models

In this section we discuss models that describe functional relationships between *continuous* response variables and explanatory variables that take values in a *discrete* number of categories, such as yes/no, green/blue/brown, and male/female. Such variables are often called **categorical**. By assigning numerical values to the categories, such as 0/1 and 1/2/3, one can treat them as discrete variables. Models with continuous responses and categorical explanatory variables often arise in **factorial experiments**. These are controlled statistical experiments in which the aim is to assess how a response variable is affected by one or more **factors** tested at several **levels**. A typical example is an agricultural experiment where one wishes to investigate how the yield of a food crop depends on two factors: (1) pesticide, at two levels (yes and no) and (2) fertilizer, at three levels (low, medium, and high). In factorial experiments one usually wishes to collect one or more responses from each of the different combinations of levels. A fictitious data set for the above agricultural experiment with three responses (crop yield) per level pair is given in Table 4.2.

Table 4.2 Crop yield

Pesticide	Fertilizer		
	Low	Medium	High
No	3.23, 3.20, 3.16	2.99, 2.85, 2.77	5.72, 5.77, 5.62
Yes	6.78, 6.73, 6.79	9.07, 9.09, 8.86	8.12, 8.04, 8.31

The main statistical tool to analyze such data is **analysis of variance** (ANOVA), which will be discussed in Sect. 5.3.1. We describe next two common models that are used in such situations.

☞ 142

4.4.1 Single-Factor ANOVA

Consider a response variable which depends on a single factor with d levels. Within each level i there are n_i independent measurements of the response variable. The data thus consist of d independent samples with sizes $n_1, \ldots, n_d$:

$$\underbrace{Y_1, \ldots, Y_{n_1}}_{\text{level 1}}, \ \underbrace{Y_{n_1+1}, \ldots, Y_{n_1+n_2}}_{\text{level 2}}, \ \ldots, \ \underbrace{Y_{n-n_d+1}, \ldots, Y_n}_{\text{level } d}, \qquad (4.12)$$

where $n = n_1 + \cdots + n_d$. An obvious model for the data is that the $\{Y_i\}$ are assumed to be independent and normally distributed with a mean and variance which depend only on the level. Such a model is simply a d-sample generalization of the two-sample normal model in Example 4.3. To be able to analyze the model via ANOVA one needs however the additional model assumption that the *variances are all equal*;

☞ 103

that is, they are the same for each level. Writing Y_{ik} as the response for the kth replication at level i we can define the model as follows.

Definition 4.3.   (**Single-Factor ANOVA Model**). In a **single-factor ANOVA model**, let Y_{ik} be the response for the kth replication at level i. Then,

$$Y_{ik} = \mu_i + \varepsilon_{ik} , \quad k = 1, \ldots, n_i , \ i = 1, \ldots, d , \qquad (4.13)$$

where $\{\varepsilon_{ik}\} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

Instead of (4.13) one often sees the "factor effects" formulation

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} , \quad k = 1, \ldots, n_i , \ i = 1, \ldots, d , \qquad (4.14)$$

where μ is interpreted as the *overall* effect, common to all levels, and $\alpha_i = \mu_i - \mu$ is the *incremental effect* of level i. Although μ can be defined in several ways, it is customary to define it as the expected average response:

$$\mu = \mathbb{E}\left[\frac{Y_1 + \cdots + Y_n}{n}\right] = \frac{\sum_{i=1}^{d} n_i \, \mu_i}{n} ,$$

in which case the $\{\alpha_i\}$ must satisfy the relation

$$\sum_{i=1}^{d} \frac{n_i}{n} \alpha_i = 0 . \qquad (4.15)$$

In particular, for **balanced** designs—where the sample sizes in each group are equal—we have $\sum_{i=1}^{d} \alpha_i = 0$.

Returning to the sequence of response variables $Y_1, \ldots, Y_n$ in (4.12), suppose that for each Y_k we denote the corresponding level by u_k, $k = 1, \ldots, n$. We can then write the model in a form closely resembling a multiple linear regression model, namely,

$$Y_k = \mu_1 \, \mathrm{I}_{\{u_k=1\}} + \cdots + \mu_d \, \mathrm{I}_{\{u_k=d\}} + \varepsilon_k , \quad k = 1, \ldots, n , \qquad (4.16)$$

where $\{\varepsilon_k\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $\mathrm{I}_{\{u=a\}} = 1$ if $u = a$ and 0 otherwise. It follows that the vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ has a multivariate normal distribution with a mean vector whose kth component is $\mu_1 \, \mathrm{I}_{\{u_k=1\}} + \cdots + \mu_d \, \mathrm{I}_{\{u_k=d\}}$ and with covariance matrix $\sigma^2 I$, where I is the n-dimensional identity matrix.

4.4.2 Two-Factor ANOVA

Many designed experiments deal with responses that depend on more than one factor. We consider for simplicity only the two-factor ANOVA model. Models with more than two factors can be formulated analogously.

Suppose Factor 1 has d_1 levels and Factor 2 has d_2 levels. Within each pair of levels (i, j) we assume that there are n_{ij} replications. Let Y_{ijk} be the kth observation at level (i, j). A direct generalization of (4.13) gives the following model.

> **Definition 4.4. (Two-Factor ANOVA Model).** In a **two-factor ANOVA model** let Y_{ijk} be the response for the kth replication at level (i, j). Then,
>
> $$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} , \quad k = 1, \ldots, n_{ij} , i = 1, \ldots, d_1, j = 1, \ldots, d_2 , \quad (4.17)$$
>
> where $\{\varepsilon_{ijk}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$.

Note that the variances of the responses are assumed to be equal to σ^2. To obtain a "factor effects" representation, we can reparameterize model (4.17) as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} ,$$
$$k = 1, \ldots, n_{ij} , i = 1, \ldots, d_1 , j = 1, \ldots, d_2 . \tag{4.18}$$

The parameter μ can be interpreted as the overall mean response, α_i as the incremental effect due to Factor 1 at level i, and β_j as the incremental effect of Factor 2 at level j. The $\{\gamma_{ij}\}$ represents the interaction effects of the two factors. As in the one-factor model, the parameters can be defined in several ways. For the most important *balanced* case (all the n_{ij} are the same), the default choice for the parameters is as follows.

$$\mu = \mathbb{E}\overline{Y}_{\bullet\bullet} = \frac{\sum_i \sum_j \mu_{ij}}{d_1 d_2} . \tag{4.19}$$

$$\alpha_i = \mathbb{E}[\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}] = \frac{\sum_j \mu_{ij}}{d_2} - \mu . \tag{4.20}$$

$$\beta_j = \mathbb{E}[\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}] = \frac{\sum_i \mu_{ij}}{d_1} - \mu . \tag{4.21}$$

$$\gamma_{ij} = \mathbb{E}[Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet}] = \mu_{ij} - \mu - \alpha_i - \beta_j . \tag{4.22}$$

Here, $\overline{Y}_{\bullet\bullet}$ indicates the average of all the $\{Y_{ijk}\}$. Similarly, $\overline{Y}_{i\bullet}$ is the average of all the $\{Y_{ijk}\}$ within level i of Factor 1, and $\overline{Y}_{\bullet j}$ denotes the average of all the $\{Y_{ijk}\}$ within level j of Factor 2. For this case it is easy to see that $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ for all i and j. Note that under these restrictions model (4.18) has the same number of parameters as model (4.17); see Problem 4.5.

One objective of ANOVA is to assess whether the data are best described by a "saturated" model such as (4.18) or if simpler models, with fewer parameters, suffice. For example, a model without interaction terms is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \ .$$

A model where Factor 2 is irrelevant is

$$Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \ .$$

If neither Factor 1 nor Factor 2 has an influence on the response, then the appropriate model would simply be

$$Y_{ijk} = \mu + \varepsilon_{ijk} \ ,$$

that is, $Y_{ijk} \sim_{\mathrm{iid}} \mathsf{N}(\mu, \sigma^2)$.

Remark 4.2 (Blocking). Not all of the factors in an ANOVA model need to be of primary interest to the researcher. Some of the factors are included in the experiment to reduce the variability of the measurements. Such factors are called **nuisance** factors. An example of a nuisance factor in the crop data in Table 4.2 is the *plant location* of the crop. Suppose the data were gathered from three different locations. Different soil conditions in these locations could greatly influence the crop yield and hence the findings of the research. To reduce the effect of plant location, one could take one measurement for each (pesticide, fertilizer, location) triplet. The data in Table 4.2 could represent this situation, where the three measurements for each (pesticide, fertilizer) pair correspond to location 1, 2, and 3. The idea of grouping data into levels of a nuisance factor in order to reduce the experimental error is called **blocking** and is important in the design of controlled experiments.

4.5 Normal Linear Model

The regression model in Sect. 4.3 and the ANOVA models in Sects. 4.4.1 and 4.4.2 are both examples of *normal (or Gaussian) linear models*.

> **Definition 4.5. (Normal Linear Model).** A vector of data $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is said to satisfy a **normal linear model** if
>
> $$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4.23)$$
>
> for some $n \times m$ matrix X (the **design matrix**), an m-dimensional vector of **parameters** $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$, and a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ of iid $N(0, \sigma^2)$ **error** terms.

To see that the simple linear regression model (4.3) is of the form (4.23), take   ☞ 106

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

An equivalent formulation is given in (4.4). Similarly, for the multiple linear regression model (4.5), we have, in view of (4.6),   ☞ 106

$$\mathbf{Y} = \underbrace{\begin{pmatrix} 1 & \mathbf{x}_1^\top \\ 1 & \mathbf{x}_2^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix}}_{\widetilde{\boldsymbol{\beta}}} + \boldsymbol{\varepsilon}.$$

To see that the one-factor ANOVA model is also of the form (4.3), let us define $\mathbf{1}_m$ as the m-dimensional column vector of 1s and $\mathbf{0}_m$ as the vector of 0s. Using the "regression" form (4.16) we can now write the vector $\mathbf{Y}$ as $X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$X = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_d} & \mathbf{0}_{n_d} & \cdots & \mathbf{1}_{n_d} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}.$$

A similar formulation can be given for the multifactor ANOVA case, as illustrated in the following example.

Example 4.8 (ANOVA as a Normal Linear Model). Regression and ANOVA data are often represented in the form of a **spreadsheet**, where each row corresponds to a single measurement and the columns correspond to the response variable and the various factors. Table 4.3 gives such a spreadsheet for the crop yield data in Table 4.2.

Table 4.3 Crop yield data as a spreadsheet

Crop Yield	Pesticide	Fertilizer
3.23	No	Low
3.20	No	Low
3.16	No	Low
2.99	No	Medium
2.85	No	Medium
2.77	No	Medium
5.72	No	High
5.77	No	High
5.62	No	High
6.78	Yes	Low
6.73	Yes	Low
6.79	Yes	Low
9.07	Yes	Medium
9.09	Yes	Medium
8.86	Yes	Medium
8.12	Yes	High
8.04	Yes	High
8.31	Yes	High

The design matrix can be directly constructed from this table. For example, consider the representation (4.17) and define $\boldsymbol{\beta} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})^\top$. With the responses $\{Y_{ijk}\}$ ordered as $(Y_1, \ldots, Y_{18})^\top$ as in Table 4.3, the 18×6 design matrix is given by

$$
X = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix},
$$

where $\mathbf{1} = (1, 1, 1)^\top$ and $\mathbf{0} = (0, 0, 0)^\top$. This may be written in compact notation as $X = I_6 \otimes \mathbf{1}$, where $A \otimes B$ indicates the **Kronecker product** of $A = (a_{ij})$ and B, that is, the block matrix with (i, j)th block $a_{ij} B$. For the "factor effects" representation (4.18), define $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23})^\top$. In this case the design matrix is an 18×12 matrix given by

$$
X = \begin{pmatrix} 1\,1\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,1\,0\,0\,1\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,1\,0\,0\,1\,0\,0\,0 \\ 1\,0\,1\,1\,0\,0\,0\,0\,0\,1\,0\,0 \\ 1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,1 \end{pmatrix}.
$$

Note that in this case the parameters are linearly dependent. For example, $\alpha_2 = -\alpha_1$ and $\gamma_{13} = -(\gamma_{11} + \gamma_{12})$. To retain only 6 linearly independent variables (as in the case (4.17)) one could consider the 6-dimensional parameter vector $\widetilde{\boldsymbol{\beta}} = (\mu, \alpha_1, \beta_1, \beta_2, \gamma_{11}, \gamma_{12})$, which is related to the 12-dimensional parameter vector $\boldsymbol{\beta}$ via the transformation

$$
\underbrace{\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & -1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}}_{M}
\underbrace{\begin{pmatrix}
\mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12}
\end{pmatrix}}_{\widetilde{\boldsymbol{\beta}}}
=
\underbrace{\begin{pmatrix}
\mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23}
\end{pmatrix}}_{\boldsymbol{\beta}}
=
\begin{pmatrix}
\mu \\ \alpha_1 \\ -\alpha_1 \\ \beta_1 \\ \beta_2 \\ -\beta_1 - \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ -\gamma_{11} - \gamma_{12} \\ -\gamma_{11} \\ -\gamma_{12} \\ \gamma_{11} + \gamma_{12}
\end{pmatrix}.
$$

The design matrix corresponding to $\widetilde{\boldsymbol{\beta}}$ is simply $\widetilde{X} = XM$; see also Problem 4.10.

Returning to the general normal linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, observe that $\mathbf{Y}$ has a multivariate normal distribution with mean vector $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 I$, where I is the identity matrix of dimension n. The situation is graphically depicted in Fig. 4.7. Imagine drawing multiple realizations of the random vector $\mathbf{Y}$. These would form a spherically symmetric cloud of points centered around $X\boldsymbol{\beta}$. More precisely, because $\mathbf{Y} \sim \mathsf{N}(X\boldsymbol{\beta}, \sigma^2 I)$, it follows from (3.30) that the joint density of $\mathbf{Y}$ at $\mathbf{y}$ is given by   ☞ 83

$$
f_{\mathbf{Y}}(\mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \, e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^\top (\mathbf{y}-X\boldsymbol{\beta})} = (2\pi\sigma^2)^{-\frac{n}{2}} \, e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-X\boldsymbol{\beta}\|^2} \; .
$$
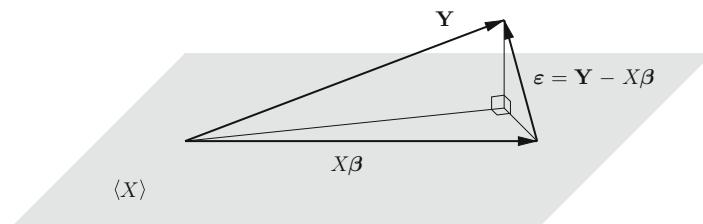
Fig. 4.7 Normal linear model. $\langle X \rangle$ is the subspace of $\mathbb{R}^n$ spanned by the columns of X

4.6 Problems

4.1. Formulate a statistical model for each of the situations below, in terms of one or more iid samples. If a model has more than one parameter, specify which parameter is of primary interest.

a. A shipbuilder buys each week hundreds of tins of paint, labeled as containing 20 L. The builder suspects that the tins contain, on average, less than 20 L and decides to determine the volume of paint in nine randomly chosen tins.
b. An electronics company wishes to examine if the rate of productivity differs significantly between male and female employees involved in assembly work. The time of completion of a certain component is observed for 12 men and 12 women.
c. The head of a mathematics department suspects that lecturers A and B differ significantly in the way they assess student work. To test this, 12 exams are both assessed by lecturer A and B.
d. We wish to investigate if a certain coin is fair. We toss the coin 500 times and examine the results.
e. We investigate the effectiveness of a new teaching method by dividing 20 students into two groups of 10, where the first group is taught by the old method and the second group is taught by the new method. Each student is asked to complete an exam before and after the teaching period.
f. We wish to assess which of two scales is the more sensitive. We measure, for each scale, ten times a standard weight of 1 kg.
g. To investigate if the support for the *Honest* party is the same in two different cities, one hundred voters in each city are asked if they would vote for the *Honest* party or not.
h. In a study on the effectiveness of an advertising campaign, a survey was conducted among 15 retail outlets. For each outlet the sales on a typical Saturday was recorded one month before and one month after the advertising campaign.
i. To focus their marketing of remote-controlled cars an electronics company wishes to investigate who in the end decides to buy: the child or the father. It records who decides in 400 transactions involving a father and a son.

4.2. Formulate appropriate statistical models for the data occurring in the following quality control processes.

a. Consider a packaging line for 500 gm packets of *Yummy* breakfast cereal. The process is monitored by recording each hour the average weight of five randomly selected packets.
b. A mail-order company selects each day at random 50 invoices from the many invoices it receives on a day and has these examined for errors. The number of invoices with errors is recorded.

4.3. An alternative approach to model the height data in Fig. 4.3 is to assume that the observations are outcomes of iid random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$. What would be a suitable two-dimensional distribution?

4.4. Consider a Gaussian model $\mathbf{Y} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$, where $\mathbf{Y}$ is of dimension n. Show that the maximum number of model parameters is $n(n+3)/2$.

4.5. Show that under the restrictions $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$, the factor effects ANOVA model in (4.18) has $d_1 d_2 + 1$ free parameters.

4.6. Verify the relation (4.15).

4.7. For each of the following situations, formulate a regression or ANOVA model.

a. In a study of shipping costs, a company controller has randomly selected 9 air freight invoices from current shippers in order to assess the relationship between shipping costs and distance, for a given volume of goods.
b. We wish to test if three different brands of compact cars have the same average fuel consumption. The fuel consumption for a traveled distance of 100 km is measured for 20 cars of each brand.
c. Heart rates were monitored for 20 laboratory rats during three different stages of sleep.
d. For the last 10 years a peace organization has been keeping record of the yearly military expenditure and gross national product of a country, which appear to be related linearly.
e. We investigate the effectiveness of a new fertilizer by dividing a large patch of land into 20 test plots, each of which is divided into three small sub-plots. In each of the three subplots a different concentration of fertilizer is tested: *weak, moderate*, and *strong*. The product yield for each subplot is recorded.
f. One hundred adults are randomly selected from a large population. The height and weight of each person is recorded, along with their body mass index (i.e., the weight in kilogram divided by the square of the height in meters).

4.8. Let $Y_1, \ldots, Y_n$ be data from the polynomial regression model (4.10), with corresponding explanatory variables $x_1, \ldots, x_n$. Write the model as a Gaussian linear model of the form (4.23).

4.9. Specify the design matrix for the multiple polynomial regression model (4.11), based on n explanatory variable pairs $(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})$.

4.10. Give the 18×6 design matrix corresponding to the parameter vector $\widetilde{\boldsymbol{\beta}}$ for the two-factor ANOVA model in Example 4.8. Verify that the first column, consisting of only 1s, is orthogonal (perpendicular) to all the other columns.

4.11. Table 4.2 was produced using the following parameters: $\mu = 6$, $\sigma = 0.1$, $(\alpha_1, \alpha_2) = (-2, 2)$, $(\beta_1, \beta_2, \beta_3) = (-1, 0, 1)$, and

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} = \begin{pmatrix} 0.2 & -1 & 0.8 \\ -0.2 & 1 & -0.8 \end{pmatrix} .$$

Implement a MATLAB program to draw realizations from the corresponding ANOVA model, producing data similar to that in Table 4.2 .

4.12. The data in Table 4.1 was computer generated from the nonlinear regression model

$$Y_i = e^{-0.02\,t_i} + \varepsilon_i \ ,$$

where $t_i = (i-1)8, i = 1, \ldots, 13$ and $\{\varepsilon_i\} \sim_{\text{iid}} \mathsf{N}(0, (0.03)^2)$. Implement a MATLAB program that generates (new) data from the model. Plot the data and the regression curve as in Fig. 4.5.

Chapter 5
Statistical Inference

Recall the conceptual framework for statistics in Fig. 4.1. *Statistical inference* deals with the middle part of this framework, that is, how to obtain conclusions about the model on the basis of the observed data. The two main approaches to statistical inference are:

- Classical statistics.
- Bayesian statistics.

In *classical statistics* the data vector **x** is viewed as the outcome of a random vector **X** described by a probabilistic model—usually the model is specified up to a (multidimensional) parameter $\boldsymbol{\theta}$; that is, $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$. The statistical inference is then purely concerned with the model and in particular with the parameter $\boldsymbol{\theta}$. For example, on the basis of the data, one may wish to

1. estimate the parameter or
2. perform statistical tests on the parameter.

A main difference between the classical and the *Bayesian* approach is that in the latter case the parameter vector $\boldsymbol{\theta}$ is considered to be *random*. Inference about $\boldsymbol{\theta}$ is carried out by analyzing the conditional pdf $f(\boldsymbol{\theta} \mid \mathbf{x})$—the so-called *posterior pdf*. Bayesian inference is discussed in Chap. 8. For the remainder of this chapter we will explain the main ingredients of the classical approach to statistical inference, starting with a simple motivating example.

Example 5.1 (Biased Coin). We throw a coin 1000 times and observe 570 Heads. Using this information, what can we say about the "fairness" of the coin? The data (or better, *datum*) here is the number $x = 570$. Suppose we view x as the outcome of a random variable X which describes the number of Heads in 1000 tosses. Our statistical model is then

$$X \sim \mathsf{Bin}(1000, p) \ ,$$

where $p \in [0, 1]$ is unknown. Any statement about the fairness of the coin is expressed in terms of p and is assessed via this model. It is important to understand that p will *never be known*. The best we can do is to provide an *estimate* of p. A common-sense estimate of p is simply the proportion of Heads $x/1000 = 0.570$. But how accurate is this estimate? Is it possible that the unknown p could in fact be 0.5? One can make sense of these questions through detailed analysis of the statistical model.

5.1   Estimation

Suppose the distribution of the data $\mathbf{X}$ is completely specified up to an unknown parameter vector $\boldsymbol{\theta}$. The aim is to estimate $\boldsymbol{\theta}$ on the basis of the observed data $\mathbf{x}$ only. Mathematically, the goal is to find function $\mathbf{T} = \mathbf{T}(\mathbf{X})$ of the data $\mathbf{X}$ such that the random vector $\mathbf{T}$ is close to $\boldsymbol{\theta}$. The random variable $\mathbf{T}$ is called an **estimator** of $\boldsymbol{\theta}$. The corresponding outcome $\mathbf{t} = \mathbf{T}(\mathbf{x})$ is the **estimate** of $\boldsymbol{\theta}$. The **bias** of an estimator $\mathbf{T}$ is defined as $\mathbb{E}\mathbf{T} - \boldsymbol{\theta}$. $\mathbf{T}$ is said to be **unbiased** if $\mathbb{E}\mathbf{T} = \boldsymbol{\theta}$. It is important to note that $\mathbf{T}$ is a function of the data only, and not of the parameter. Such a function is called a **statistic**.

Example 5.2 (Iid Sample from a Normal Distribution). Consider the standard
model for data (see Sect. 4.1):

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2) \,,$$

where μ and σ^2 are unknown. The random measurements $\{X_i\}$ could represent the weights of randomly selected teenagers, the heights of the dorsal fin of sharks, the dioxin concentrations in hamburgers, and so on. Suppose, for example, that, with $n = 10$, the observed measurements $x_1, \ldots, x_n$ are

77.01, 71.37, 77.15, 79.89, 76.46, 78.10, 77.18, 74.08, 75.88, 72.63 .

A common-sense *estimate* (a number) for μ is the **sample mean**

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = 75.975 \,. \tag{5.1}$$

Note that the estimate $\bar{x}$ is a function of the data $\mathbf{x} = (x_1, \ldots, x_n)$ only. The corresponding *estimator* (a random variable) is

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \,.$$

To justify why $\bar{x}$ is a good estimate of μ, imagine that we carry out the experiment and the estimation *tomorrow*, obtaining the (random) sample mean $\bar{X}$ as our guess

for μ. From the affine transformation property of the normal distribution (see Theorem 3.6) we see that $\bar{X} \sim \mathsf{N}(\mu, \sigma^2/n)$. Hence, $\bar{X}$ is an unbiased estimator for μ—it is in expectation equal to the unknown μ. Moreover, for large n, the variance of $\bar{X}$ tends to zero, implying that $\bar{X}$ gets closer to μ as the sample size n is increased. To specify exactly how close $\bar{X}$ is to μ one needs to estimate also σ^2, which is discussed in the next section.

Remark 5.1 (Notation). It is customary in statistics to denote the estimate of a parameter $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$; for example, $\widehat{\mu} = \bar{x}$ in the example above. The *same* notation, $\widehat{\boldsymbol{\theta}}$, is often also used for the corresponding (random) *estimator*. It should be clear from the context which meaning is used.

Three systematic approaches to constructing good estimators are the *maximum likelihood method*, the *method of moments*, and *least-squares minimization*. Maximum likelihood estimation is the most powerful of the three and is based on the concept of the *likelihood function*, which plays a central role in statistics. The whole of Chap. 6 is devoted to likelihood methods. In particular, Sect. 6.3 deals with maximum likelihood estimation. The other two estimation procedures are described next.

5.1.1 Method of Moments

Suppose $x_1, \ldots, x_n$ are outcomes from an iid sample $X_1, \ldots, X_n \sim_{\text{iid}} f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is unknown. The *moments* of the sampling distribution can be easily estimated. Namely, if $X \sim f(x; \boldsymbol{\theta})$, then the rth moment of X, that is, $\mu_r(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} X^r$ (assuming it exists), can be estimated through the **sample rth moment**

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r \; .$$

The **method of moments** involves choosing the estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that each of the first k sample moments is matched with the true moments; that is,

$$m_r = \mu_r(\widehat{\boldsymbol{\theta}}), \quad r = 1, 2, \ldots, k \; .$$

In general, this gives a set of nonlinear equations, and so its solution often has to be found numerically. In the following examples, however, the solution can be obtained analytically.

Example 5.3 (Sample Mean and Sample Variance). Suppose that the data are given by $\mathbf{X} = (X_1, \ldots, X_n)^\top$, where the $\{X_i\}$ form an iid sample from a general distribution with mean μ and variance $\sigma^2 < \infty$. Matching the first moment gives the equation

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \widehat{\mu} \ , \tag{5.2}$$

which yields the *sample mean* $\widehat{\mu} = \bar{x}$ (already introduced in Example 5.2) as the method of moments estimate for μ. Matching the second moment, $\mathbb{E}X^2 = (\mathbb{E}X)^2 + \text{Var}(X)$, gives the equation

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 = (\widehat{\mu})^2 + \widehat{\sigma^2} \ . \tag{5.3}$$

The method of moments estimate for σ^2 is therefore

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{5.4}$$

The corresponding estimator turns out to be biased:

$$\mathbb{E}\widehat{\sigma^2} = \mathbb{E}X^2 - \mathbb{E}(\bar{X})^2 = \text{Var}(X) + (\mathbb{E}X)^2 - (\text{Var}(\bar{X}) + (\mathbb{E}\bar{X})^2)$$

$$= \sigma^2 + \mu^2 - \sigma^2/n - \mu^2 = \frac{n-1}{n} \sigma^2 \ .$$

By multiplying $\widehat{\sigma^2}$ with $n/(n-1)$ we obtain an *unbiased* estimator of σ^2, called the **sample variance**, often denoted by S^2:

$$S^2 = \widehat{\sigma^2} \frac{n}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \ . \tag{5.5}$$

The square root of the sample variance $S = \sqrt{S^2}$ is called the **sample standard deviation**.

The method of moments can also be used to estimate parameters of iid random vectors, as illustrated in the following example.

Example 5.4 (Sample Correlation Coefficient). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of a random vector (X, Y) with unknown correlation coefficient $\varrho = \varrho(X, Y)$. Think of iid samples from a bivariate normal distribution. We can estimate ϱ by using the same "moment matching" ideas as in the 1-dimensional case. In particular, write

$$\varrho = \frac{\mathbb{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y} \ , \tag{5.6}$$

where μ_X and μ_Y are the expectations of X and Y, respectively, and σ_X and σ_Y are the standard deviations of X and Y, respectively. We can estimate these parameters

via the corresponding moment estimators, as discussed above. Moreover, we can estimate $\mathbb{E}[XY]$ via the moment estimator

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \ .$$

Hence, we can estimate the numerator of (5.6) as

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \bar{X}\,\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \ .$$

This leads to the following estimator of ϱ:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \ , \tag{5.7}$$

which is called the **sample correlation coefficient**.

5.1.2 Least-Squares Estimation

Least-squares estimation is a simple estimation technique that is particularly useful in regression analysis. In particular, consider the normal linear model (4.23)   ☞ 115

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I) \ ,$$

where the $n \times m$ design matrix $X = (x_{ij})$ is known, but the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ and σ^2 need to be estimated from an outcome $\mathbf{y} = (y_1, \ldots, y_n)^\top$ of $\mathbf{Y}$. We assume that $n > m$; that is, there are at least as many observations as model parameters. The main idea is illustrated in Fig. 5.1: choose the estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ such that the (Euclidean) distance between $X\widehat{\boldsymbol{\beta}}$ and the observed data $\mathbf{y}$ is as small as possible.
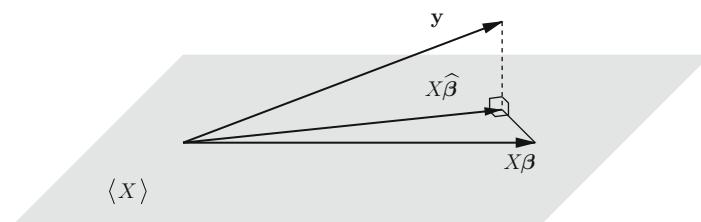
Fig. 5.1 $X\widehat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{y}$ onto the linear space spanned by the columns of the design matrix X

In other words, we seek to minimize $\|\mathbf{y} - X\boldsymbol{\beta}\|$ with respect to $\boldsymbol{\beta}$. This is equivalent to minimizing the squared distance

$$\|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 \,. \tag{5.8}$$

To find the optimal $\beta_1, \ldots, \beta_m$ we take the derivative of (5.8) with respect to each $\beta_k, k = 1, \ldots, m$ and set it equal to 0. This leads to the set of linear equations

$$\frac{\partial \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2}{\partial\beta_k} = -\sum_{i=1}^{n}\left\{2(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)x_{ik}\right\} = 0 \,, k = 1, \ldots, m \,,$$

which can be written in matrix notation as

$$X^\top X\boldsymbol{\beta} = X^\top\mathbf{y} \,. \tag{5.9}$$

These are the so-called **normal equations**. The **rank** of X is the number of linearly independent columns (recall that we assume that the number of columns is less than the number of rows). If X is of *full rank* (i.e., none of the columns can be expressed as a linear combination of the other columns), then $X^\top X$ is *invertible*. In that case,

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top\mathbf{y} \,. \tag{5.10}$$

Note that the matrix $P = X(X^\top X)^{-1}X^\top$ is the *projection matrix* onto the subspace $\langle X \rangle$ spanned by the columns of X—and hence $X\widehat{\boldsymbol{\beta}} = P\mathbf{y}$. Namely, P maps each vector in $\langle X \rangle$ to itself, because $PX = X$; and P maps any vector $\mathbf{v}$ perpendicular to $\langle X \rangle$ to $\mathbf{0}$, because $X^\top\mathbf{v} = \mathbf{0}$. The $m \times n$ matrix

$$X^+ = (X^\top X)^{-1}X^\top$$

is called the (right) **pseudo-inverse** of X, because $X^+ X = I_m$—the m-dimensional identity matrix. We thus have

$$\widehat{\boldsymbol{\beta}} = X^+\mathbf{y} \,. \tag{5.11}$$

Let $\varepsilon_i = Y_i - (X\boldsymbol{\beta})_i$ be the ith component of $\boldsymbol{\varepsilon}$. Note that the $\{\varepsilon_i\}$ form an iid sample from the $\mathsf{N}(0, \sigma^2)$ distribution. To obtain the method of moments estimate of σ^2, we match the second moment of $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ to its sample average

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - (X\widehat{\boldsymbol{\beta}})_i)^2 \,,$$

where we have plugged in the least-squares estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. The estimated errors $u_i = Y_i - (X\widehat{\boldsymbol{\beta}})_i, i = 1, \ldots, n$ are called the **residuals**. Simplifying the above expression using vector notation, we obtain the estimator

$$\widehat{\sigma^2} = \frac{\|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|^2}{n} = \frac{\|\mathbf{u}\|^2}{n}, \qquad (5.12)$$

where $\mathbf{u} = (u_1, \ldots, u_n)^\top$ is the vector of residuals.

Example 5.5 (Simple Linear Regression). For the simple linear regression case we have a design matrix

$$X = (\mathbf{1}\ \mathbf{x}) = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

and a parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. The least-squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}.$$

It is straightforward to write this out to obtain explicit expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (see Problem 5.10), but in practice it is easier to simply solve the normal equations (5.10) numerically. The estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 .$$

By taking the square root of the above expression, one obtains a natural estimator for σ.

   The following MATLAB program draws $N = 100$ samples from a simple linear regression model with parameters $\boldsymbol{\beta} = (6, 13)^\top$ and $\sigma = 2$, where the x-coordinates are evenly spaced on the interval $[0, 1]$. The parameters are estimated in the last two lines of the program. An important thing to keep in mind when solving linear equations is that one should avoid computing costly inverses. In particular, an equation such as $A\mathbf{x} = \mathbf{b}$ should never be solved numerically via $\mathbf{x} = A^{-1}\mathbf{b}$. Instead, use MATLAB's syntax $\mathbf{x} = A \setminus \mathbf{b}$, as in the second-last line of code below. Typical estimates for $\boldsymbol{\beta}$ and σ are $\widehat{\boldsymbol{\theta}} = (6.3, 12.2)^\top$ and $\widehat{\sigma} = 1.86$.

```
%linregest.m
N = 100; x = ( (1:N)/N )';
beta = [6;13]; sigma = 2;  % parameters
X = [ones(N,1),x];         % design matrix
y = X*beta + sigma*randn(N,1);  % draw the data
plot(x,y,'.');                  % plot the data
betahat = X'*X\(X'*y)  % solve the normal equations
sigmahat = norm(y - X*betahat)/sqrt(N)  % estimate for sigma
```

5.2 Confidence Intervals

An essential part of any estimation procedure is to provide an assessment of the *accuracy* of the estimate. Indeed, without information on its accuracy, the estimate itself would be meaningless. Confidence intervals (sometimes called **interval estimates**) provide a precise way of describing the uncertainty in the estimate. In Sect. 6.3.1 we will discuss a systematic approach for constructing (approximate) confidence intervals, based on the *likelihood* concept. The *bootstrap method* (see Sect. 7.3) provides another useful way to construct confidence intervals. The analogue of a confidence interval in Bayesian statistics is the **credible interval**; see Example 8.1.

☞ 174
☞ 203
☞ 228

> **Definition 5.1. (Confidence Interval).** Let $X_1, \ldots, X_n$ be random variables with a joint distribution depending on a parameter $\theta \in \Theta$. Let $T_1 < T_2$ be functions of the data but not of θ. The random interval (T_1, T_2) is called a **stochastic confidence interval** for θ with confidence $1 - \alpha$ if
>
> $$\mathbb{P}_\theta(T_1 < \theta < T_2) \geqslant 1 - \alpha \quad \text{for all } \theta \in \Theta . \tag{5.13}$$
>
> If t_1 and t_2 are the observed values of T_1 and T_2, then the interval (t_1, t_2) is called the **numerical confidence interval** for θ with confidence $1 - \alpha$.

If (5.13) only holds approximately, the interval is called an **approximate confidence interval**. The probability $\mathbb{P}_\theta(T_1 < \theta < T_2)$ is called the **coverage probability**. The subscript θ in $\mathbb{P}_\theta$ indicates that the joint distribution of $X_1, \ldots, X_n$ depends on θ. The coverage probability for an exact $1 - \alpha$ confidence interval is, by definition, at least $1 - \alpha$ for every θ. For approximate $1 - \alpha$ confidence intervals the actual coverage probability could well be less than $1 - \alpha$ for certain choices of θ. An example is given in Problem 5.22.

Remark 5.2. Reducing α widens the confidence interval. A very large confidence interval is not very useful. Common choices for α are $0.01, 0.05$, and 0.1.

We next describe a simple approach to constructing exact or approximate confidence intervals that uses a so-called **pivot** variable $T = T(\mathbf{X}, \theta)$, which is a function of the data $\mathbf{X}$ *and* of the parameter of interest θ, and for which the distribution is known (sometimes only approximately) and does not depend on θ. The construction depends on specific *quantiles* of the pivot distribution. For $\gamma \in (0, 1)$, the γ-**quantile** of a distribution with cdf F is a number z_γ for which $F(z_\gamma) = \gamma$ or, equivalently, $z_\gamma = F^{-1}(\gamma)$. Numerical values for quantiles of various distributions can be obtained in MATLAB via the `icdf` (inverse cdf) function of the Statistics Toolbox or alternatively by using the `icumdf` function in Appendix A.9.

In general, constructing a confidence interval using a pivot variable involves the following steps.

Steps in the Pivot Method

1. Formulate a statistical model for the data $\mathbf{X}$.
2. Choose an appropriate pivot variable $T(\mathbf{X}, \theta)$.
3. Determine the (approximate) distribution of the pivot.
4. Calculate quantiles q_1 and q_2 for the (approximate) pivot distribution such that $\mathbb{P}(q_1 < T(\mathbf{X}, \theta) < q_2) = 1 - \alpha$.
5. Rearrange the event $\{q_1 < T(\mathbf{X}, \theta) < q_2\}$ into $\{T_1 < \theta < T_2\}$ and return (T_1, T_2) as an (approximate) stochastic $1 - \alpha$ confidence interval for θ.

Remark 5.3. For a one-sided confidence interval, such as (T, ∞) or (c, T), where c is fixed, only a single quantile needs to be calculated in Step 4.

Example 5.6 (Confidence Interval for Iid Normal Data). Suppose $X_1, \ldots, X_n \sim_{\text{iid}}$ $\mathsf{N}(\mu, 1)$. We have seen that we can estimate μ with the sample mean $\bar{X}$. Here, $\bar{X} \sim \mathsf{N}(\mu, 1/n)$, so $T = (\bar{X} - \mu)n^{1/2} \sim \mathsf{N}(0, 1)$. Since T depends only on μ and the data, and has a distribution which does not depend on μ, we can use it as a pivot variable. To construct a 95 % confidence interval (hence $\alpha = 0.05$) we consider the $1 - \alpha/2 = 0.975$- and $\alpha/2 = 0.025$-quantiles of the $\mathsf{N}(0, 1)$ distribution, which are 1.96 and -1.96, respectively. Hence, $\mathbb{P}(-1.96 < T < 1.96) = 0.95$. Rearranging $\{-1.96 < (\bar{X} - \mu)n^{1/2} < 1.96\}$ into $\{\bar{X} - 1.96\,n^{-1/2} < \mu < \bar{X} + 1.96\,n^{-1/2}\}$ gives the 0.95 stochastic confidence interval $(\bar{X} - 1.96\,n^{-1/2}, \bar{X} + 1.96\,n^{-1/2})$, sometimes written as $\bar{X} \pm 1.96\,n^{-1/2}$. Thus, if we would repeat the experiment many times, and get many outcomes of the interval $\bar{X} \pm 1.96\,n^{-1/2}$, the true μ would be contained in these intervals in 95 % of the cases.

The remainder of this section is about the construction of (approximate) confidence intervals for a number of standard situations, using appropriate pivots.

5.2.1   Iid Data: Approximate Confidence Interval for μ

Let $X_1, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance
$\sigma^2 < \infty$ (both assumed to be unknown). By the central limit theorem the sample
mean $\bar{X}$ has approximately a $\mathsf{N}(\mu, \sigma^2/n)$ distribution, so $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has
approximately a standard normal distribution. However, this is not yet a pivot
variable for μ, because it still depends on the unknown standard deviation σ. This
can be remedied by substituting σ with the sample standard deviation S_X, which, by
the law of large numbers, will be close to σ for large n. This gives the pivot variable

☞ 90

☞ 89

$$T = \frac{\bar{X} - \mu}{S_X/\sqrt{n}} \overset{\text{approx.}}{\sim} \mathsf{N}(0, 1) . \tag{5.14}$$

For $\gamma \in (0, 1)$, let z_γ denote the γ-quantile of the standard normal distribution.
Rearranging the approximate equality $\mathbb{P}(|T| < z_{1-\alpha/2}) \approx 1 - \alpha$ yields

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{S_X}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{S_X}{\sqrt{n}}\right) \approx 1 - \alpha ,$$

so that

$$\left(\bar{X} - z_{1-\alpha/2}\frac{S_X}{\sqrt{n}}, \ \bar{X} + z_{1-\alpha/2}\frac{S_X}{\sqrt{n}}\right) , \tag{5.15}$$

abbreviated as $\bar{X} \pm z_{1-\alpha/2}S_X/\sqrt{n}$, is an approximate stochastic $1 - \alpha$ confidence
interval for μ.

Since (5.15) is only an asymptotic result, care should be taken when applying it
to cases where the sample size is small or moderate and the sampling distribution is
heavily skewed.

Example 5.7 (Monte Carlo Integration). In **Monte Carlo integration**, random
sampling is used to evaluate complicated integrals. Consider, for example, the
integral

$$\mu = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{|z_1 + z_2 + z_3|} \ e^{-(z_1^2 + z_2^2 + z_3^2)/2} \ dz_1 \, dz_2 \, dz_3 .$$

Defining $X = |Z_1 + Z_2 + Z_3|^{1/2}(2\pi)^{3/2}$, with $Z_1, Z_2, Z_3 \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$, we can write
$\mu = \mathbb{E}X$. In the following MATLAB program we generate an iid sample of $N = 10^6$
copies of X and estimate μ via the corresponding sample mean. A typical outcome
is $\bar{x} = 17.04$ with a 95 % confidence interval $(17.026, 17.054)$.

```
%mcint.m
c = (2*pi)^(3/2); N = 10^6;
H = @(z) c*sqrt(abs(sum(z,2)));
```

```
Z = randn(N,3); X = H(Z);
mX = mean(X); sX = std(X);
R = 1.96*sX/sqrt(N);
fprintf('Estimate = %g, CI = (%g, %g)\n', mX, mX - R, mX + R)
```

5.2.2 Normal Data: Confidence Intervals for μ and σ^2

For the standard model $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$ it is possible to construct *exact* confidence intervals for both μ and σ^2, based on the following result.

Theorem 5.1. (t and χ^2 Statistics for Normal Data). Let $Y_1, \ldots,$ $Y_n \sim_{\text{iid}} \mathsf{N}(0, 1)$ and let $\bar{Y}$ and S_Y^2 be the sample mean and sample variance. Then, $\bar{Y}\sqrt{n} \sim \mathsf{N}(0, 1)$ and $(n-1)S_Y^2 \sim \chi_{n-1}^2$, independently. Moreover,

$$T = \frac{\bar{Y}\sqrt{n}}{S_Y} \sim \mathsf{t}_{n-1} \ . \tag{5.16}$$

Proof. By the linearity property of the normal distribution (see Theorem 3.6), we have $\bar{Y}\sqrt{n} \sim \mathsf{N}(0, 1)$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, and let $\mathbf{Y}_1 = \bar{Y}\mathbf{1}$ be the orthogonal projection of $\mathbf{Y}$ onto $\mathbf{1} = (1, \ldots, 1)^\top$. By Theorem 3.10, $\|\mathbf{Y}_1\|^2 = n\bar{Y}^2$ is independent of $\|\mathbf{Y} - \mathbf{Y}_1\|^2 = (n-1)S_Y^2$, and $\|\mathbf{Y} - \mathbf{Y}_1\|^2 \sim \chi_{n-1}^2$. The result now follows from Corollary 3.2. $\square$

To obtain a stochastic confidence for μ we take the same pivot as in (5.14). Defining $Y_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$, we can write

$$T = \frac{\bar{X} - \mu}{S_X/\sqrt{n}} = \frac{\bar{Y}\sqrt{n}}{S_Y} \ , \tag{5.17}$$

where the $\{Y_i\}$ form an iid sample from the standard normal distribution. By Theorem 5.1, T has a Student's t distribution with $n - 1$ degrees of freedom. We now rearrange, similar to what was done in Sect. 5.2.1, the equality $\mathbb{P}(|T| < t_{n-1;1-\alpha/2}) = 1 - \alpha$, where $t_{n-1;1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the t_{n-1} distribution, to find an exact confidence interval for μ:

$$\bar{X} \pm t_{n-1;1-\alpha/2}\frac{S_X}{\sqrt{n}} \ . \tag{5.18}$$

To obtain an exact confidence interval for σ^2, we can use the pivot

$$\frac{(n-1)S_X^2}{\sigma^2} = (n-1)S_Y^2 \, ,$$

which by Theorem 5.1 has a χ_{n-1}^2 distribution. Note that the corresponding pdf is not symmetric. Let $\chi_{n;\gamma}^2$ be γ-quantile of the χ_n^2 distribution. Then,

$$\mathbb{P}\left(\chi_{n-1;\alpha/2}^2 < \frac{(n-1)S_X^2}{\sigma^2} < \chi_{n-1;1-\alpha/2}^2\right) = 1 - \alpha \, .$$

Rearranging gives:

$$\mathbb{P}\left(\frac{(n-1)S_X^2}{\chi_{n-1;1-\alpha/2}^2} < \sigma^2 < \frac{(n-1)S_X^2}{\chi_{n-1;\alpha/2}^2}\right) = 1 - \alpha \, .$$

Hence, a $(1-\alpha)$ stochastic confidence interval for σ^2 is

$$\left(\frac{(n-1)S_X^2}{\chi_{n-1;1-\alpha/2}^2}, \ \frac{(n-1)S_X^2}{\chi_{n-1;\alpha/2}^2}\right). \tag{5.19}$$

Example 5.8 (Monte Carlo Experiment for Confidence Intervals). The following MATLAB program draws an iid sample of size $n = 10$ from the $N(3, 0.25)$ distribution. It then determines 95 % confidence intervals for μ and σ^2 and checks if the true values are contained in the intervals or not. This is repeated independently 100 times and the total number of times that μ and σ^2 are contained in the confidence intervals is reported. The quantiles for the t and χ^2 distributions are determined here via the inverse cdf function icdf from the MATLAB *Statistics Toolbox*. If the *Statistics Toolbox* is not available, one can replace icdf by icumdf
introduced in Appendix A.9. The values are tq = 2.2622, cq1 = 19.0228, and cq2 = 2.7004. A typical estimate of μ is $\widehat{\mu} = 3.22$, with a 95 % confidence interval $(3.02, 3.41)$. For σ^2 a typical estimate is $\widehat{\sigma^2} = 0.0761$, with a 95 % confidence interval $(0.0360, 0.2535)$. In this case only the second confidence interval contains the true parameter. However, out of the 100 confidence intervals, typically only 95 contain the true parameter.

```
%confintnorm.m
mu = 3; sig = 0.5;    %true parameters
alpha = 0.05; n = 10; mu_count = 0; sig_count = 0;
for k = 1:100
    x = mu + randn(n,1)*sig;    %draw the iid sample
    mu_est = mean(x);        %estimate mu
    sig_est = std(x);        %estimate sigma
    tq = icdf('t',1-alpha/2,n-1);
    mu_lo = mu_est - tq*sig_est/sqrt(n); %low bound CI for mu
```

```
    mu_hi = mu_est +  tq*sig_est/sqrt(n); %upper bound
    cq1 = icdf('chi2',1-alpha/2,n-1);
    cq2 = icdf('chi2',alpha/2,n-1);
    sig_lo = (n-1)*sig_est^2/cq1; %lower bound CI for sigma
    sig_hi = (n-1)*sig_est^2/cq2; %upper bound
    mu_count = mu_count + (mu > mu_lo & mu < mu_hi);
    sig_count = sig_count + (sig^2>sig_lo & sig^2<sig_hi);
end
disp([mu_count, sig_count]) % final counts
```

5.2.3   Two Normal Samples: Confidence Intervals for $\mu_X - \mu_Y$ and σ_X^2/σ_Y^2

Suppose we have two independent samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ from, respectively, a $\mathsf{N}(\mu_X,\sigma_X^2)$ and $\mathsf{N}(\mu_Y,\sigma_Y^2)$ distribution. We wish to make confidence intervals for $\mu_X - \mu_Y$ and σ_X^2/σ_Y^2. The difference $\mu_X - \mu_Y$ tells us how the two *means* relate to each other, and σ_X^2/σ_Y^2 gives an indication how the *variances* relate to each other.

   Constructing a confidence interval for $\mu_X - \mu_Y$ is very similar to the one-sample case *provided* that we make the *extra model assumption* that *the variances of the two samples are the same*. That is, we assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ for some unknown σ^2. The analysis now proceeds as follows. The natural estimator for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$. Next, observe that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{1/m + 1/n}} \sim \mathsf{N}(0, 1) .$$

If σ^2 is unknown, we must replace it with an appropriate estimator in order to obtain a pivot variable for μ. For this we will use the **pooled sample variance**, S_p^2, which is defined as

$$S_p^2 = \frac{(m - 1)S_X^2 + (n - 1)S_Y^2}{m + n - 2} , \tag{5.20}$$

where S_X^2 and S_Y^2 are the sample variances for the $\{X_i\}$ and $\{Y_i\}$, respectively. It is not difficult to show that S_p^2 is an unbiased estimator of σ^2; see Problem 5.9. The following result is the analogue of Theorem 5.1 and is proved in Appendix B.5.   ☞ 371

Theorem 5.2. (*t* Statistic for Two Normal Samples). Let the random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ be defined as above; then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathsf{t}_{m+n-2} \ .$$

Using the pivot T, we find (completely analogously to the one-sample case) the following $1 - \alpha$ stochastic confidence interval for $\mu_X - \mu_Y$:

$$\bar{X} - \bar{Y} \pm t_{m+n-2;1-\alpha/2} \ S_p \sqrt{\frac{1}{m} + \frac{1}{n}} \ . \tag{5.21}$$

If the assumption $\sigma_X^2 = \sigma_Y^2$ is dropped, the pivot method no longer provides the means to obtain an exact confidence interval for $\mu_X - \mu_Y$, although it is easy to construct approximate confidence intervals for large sample sizes; see Problem 5.15.

Next, we turn our attention to a confidence interval for σ_X^2/σ_Y^2. Here, we can employ the pivot

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathsf{F}(m-1, n-1) \ .$$

To see that this pivot has the mentioned F distribution, first observe that, by Theorem 5.1, $(m-1)S_X^2/\sigma_X^2 \sim \chi_{m-1}^2$ and $(n-1)S_Y^2/\sigma_Y^2 \sim \chi_{n-1}^2$, and then apply
☞ 88   Theorem 3.11.

Let $F_{m,n;\gamma}$ denote the γ quantile of the $\mathsf{F}(m, n)$ distribution. Then,

$$\mathbb{P}\left(F_{m-1,n-1;\alpha/2} < \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < F_{m-1,n-1;1-\alpha/2} \right) = 1 - \alpha \ .$$

Rearranging gives the following $(1 - \alpha)$ stochastic confidence interval for σ_X^2/σ_Y^2:

$$\left(\frac{1}{F_{m-1,n-1;1-\alpha/2}} \frac{S_X^2}{S_Y^2}, \ \frac{1}{F_{m-1,n-1;\alpha/2}} \frac{S_X^2}{S_Y^2} \right) \ . \tag{5.22}$$

Example 5.9 (Two-Sample *t* Test). A study of iron deficiency among infants compared breast-fed with formula-fed babies. A sample of 25 breast-fed infants gave a mean blood hemoglobin level of 13.3 and a standard deviation of 1.4, while a sample of 21 formula-fed infants gave a mean and standard deviation of 12.4 and

2.0, respectively. Assuming the hemoglobin levels are normally distributed, is there statistical evidence that the mean hemoglobin levels of the two groups are different?

Let the hemoglobin levels for the breast-fed and formula-fed babies be $X_1, \ldots, X_{25} \sim_{\text{iid}} \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_{21} \sim_{\text{iid}} \mathsf{N}(\mu_Y, \sigma_Y^2)$, respectively. The samples are assumed to be independent of each other. A 95 % numerical confidence interval for σ_X^2/σ_Y^2 is

$$\left(\frac{1}{2.40756} \frac{1.4^2}{2.0^2}, \ \frac{1}{0.42969} \frac{1.4^2}{2.0^2} \right) = (0.2035, \ 1.1404) \,.$$

Because 1 is an element of this interval, there is no reason to believe that σ_X^2 is different from σ_Y^2. We thus assume that the two variances are equal, which allows us to apply (5.21). The pooled sample variance is $s_p^2 = (24(1.4)^2 + 20(2.0)^2)/44 = 2.8873$, and the 0.975 quantile of the t_{44} distribution is 2.0154, so that a 95 % confidence interval for $\mu_X - \mu_Y$ is

$$13.3 - 12.4 \pm 2.0154 \sqrt{2.8873} \sqrt{1/25 + 1/21} = (-0.11, \ 1.91) \,,$$

which contains 0. Hence, on the basis of these data and the assumptions of normality, there is no ground to believe that the expected hemoglobin levels are different for the two groups.

5.2.4 Binomial Data: Approximate Confidence Intervals for Proportions

Suppose we have an outcome x of a random variable X with a $\mathsf{Bin}(n, p)$ distribution. We wish to construct a confidence interval for p. In fact, it is not so easy to find an *exact* confidence interval for p, so we settle for an approximate one. For large n, X has approximately a $\mathsf{N}(np, np(1 - p))$ distribution; see (3.7). The natural estimator for p, that is, $\widehat{p} = X/n$, has therefore approximately a $\mathsf{N}(p, p(1 - p)/n)$ distribution. Thus, using the pivot $(\widehat{p} - p)/\sqrt{p(1 - p)/n}$, we have

$$\mathbb{P}\left(-z_{1-\alpha/2} < \frac{\widehat{p} - p}{\sqrt{p(1 - p)/n}} < z_{1-\alpha/2} \right) \approx 1 - \alpha \,,$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Rearranging gives

$$\mathbb{P}\left(\widehat{p} - z_{1-\alpha/2} \sqrt{\frac{p(1 - p)}{n}} < p < \widehat{p} + z_{1-\alpha/2} \sqrt{\frac{p(1 - p)}{n}} \right) \approx 1 - \alpha \,.$$

This would suggest that we take $\widehat{p} \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ as an approximate $1 - \alpha$ confidence interval for p, were it not for the fact that the bounds still contain the unknown p. However, for large n the estimator $\widehat{p}$ is close to the real p, so that we may replace p with $\widehat{p}$ under the square roots in the expression above. Hence, an approximate $1 - \alpha$ confidence interval for p is

$$\widehat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{p}(1 - \widehat{p})}{n}} . \tag{5.23}$$

Example 5.10 (Approximate Confidence Interval for Proportion). In an opinion poll of 1000 registered voters, 227 voters say they will vote for the *Honest* party. We wish to find a 95 % approximate confidence interval for the proportion p of *Honest* voters of the total population. We hereto view the datum, 227, as the outcome of a random variable X (the number of *Honest* voters out of 1000 registered voters) with a Bin(1000, p) distribution. We have $\widehat{p} = 227/1000 = 0.227$, and $z_{0.975} = 1.96$, so that an approximate 95 % numerical confidence interval for p is

$$0.227 \pm 1.96 \times 0.0132 = (0.20, \ 0.25) .$$

The same methodology can be used to construct approximate confidence intervals for the difference between two proportions. In particular, consider outcomes x and y of two independent random variables $X \sim$ Bin(m, p_X) and $Y \sim$ Bin(n, p_Y). We wish to construct an approximate confidence interval for $p_X - p_Y$. The corresponding estimator is $\widehat{p}_X - \widehat{p}_Y = X/m - Y/n$. As in the one-sample case, for m and n sufficiently large,

$$\mathbb{P}\left(-z_{1-\alpha/2} < \frac{\widehat{p}_X - \widehat{p}_Y - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}} < z_{1-\alpha/2} \right) \approx 1 - \alpha .$$

Rewriting this gives

$$\mathbb{P}\left(\widehat{p}_X - \widehat{p}_Y - z_{1-\alpha/2} \sqrt{\frac{p_X(1 - p_X)}{m} + \frac{p_Y(1 - p_Y)}{n}} < p_X - p_Y \right.$$

$$\left. < \widehat{p}_X - \widehat{p}_Y + z_{1-\alpha/2} \sqrt{\frac{p_X(1 - p_X)}{m} + \frac{p_Y(1 - p_Y)}{n}} \right) \approx 1 - \alpha .$$

By substituting p_X and p_Y with their estimators, we obtain the following approximate $1 - \alpha$ confidence interval for $p_X - p_Y$:

$$\widehat{p}_X - \widehat{p}_Y \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{p}_X(1 - \widehat{p}_X)}{m} + \frac{\widehat{p}_Y(1 - \widehat{p}_Y)}{n}} . \tag{5.24}$$

Example 5.11 (Approximate Confidence Interval for the Difference of Two Proportions). Two groups of men and women are asked whether they would buy *Happy* or *Fun* cola, if they were forced to choose between the two. The results are given in Table 5.1.

Table 5.1 Counts of men and women preferring *Happy* or *Fun* cola

	Men	Women
Happy	55	60
Fun	105	132

The observed proportions of *Happy* cola drinkers among the men and women are $55/160 = 34.4\,\%$ and $60/192 = 31.3\,\%$, respectively. Is this difference statistically significant or due to chance?

We view the data as outcomes of a two-sample binomial model. Specifically, let X be the number of *Happy* cola drinkers among 160 men and Y the number of *Happy* cola drinkers among 192 women. We assume that $X \sim \mathsf{Bin}(160, p_X)$ and $Y \sim \mathsf{Bin}(192, p_Y)$ are independent. To assess the difference between the true proportions p_X and p_Y, we simply evaluate the numerical confidence interval of the form (5.24). We have $\widehat{p}_X = 0.344$, $\widehat{p}_Y = 0.313$, and $z_{0.975} = 1.96$, so that a 95\,% numerical confidence interval for $p_X - p_Y$ is

$$0.031 \pm 0.099 = (-0.07,\ 0.13)\,.$$

This interval contains 0, so there is no evidence that men and women differ in their preference for the two brands of cola.

5.2.5 Confidence Intervals for the Normal Linear Model

Consider the normal linear model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)\,,$$

where X is an $n \times m$ matrix ($m < n$) of full rank m—thus, the columns of X are linearly independent and, as a consequence, the matrix $X^\top X$ has an inverse.

We saw in Sect. 5.1.2 that the parameter vector $\boldsymbol{\beta}$ can be estimated via the estimator

$$\widehat{\boldsymbol{\beta}} = X^+\mathbf{Y} = (X^\top X)^{-1} X^\top \mathbf{Y}\,.$$

Since the random vector $\widehat{\boldsymbol{\beta}}$ is a linear transformation of a normal random vector, it has a multivariate normal distribution. The mean vector and covariance matrix follow from Theorem 3.4:

☞ 80

$$\mathbb{E}\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbb{E}\mathbf{Y} = (X^\top X)^{-1} X^\top X \boldsymbol{\beta} = \boldsymbol{\beta}$$

and

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = (X^\top X)^{-1} X^\top \sigma^2 I ((X^\top X)^{-1} X^\top)^\top = \sigma^2 (X^\top X)^{-1} .$$

Let $\mathbf{a}$ be any m-dimensional vector. A natural estimator for $\theta = \mathbf{a}^\top \boldsymbol{\beta}$ is $\widehat{\theta} = \mathbf{a}^\top \widehat{\boldsymbol{\beta}}$. The following theorem gives an exact confidence interval for θ.

Theorem 5.3. (Confidence Interval for Normal Linear Model). A $1 - \alpha$ stochastic confidence interval for $\theta = \mathbf{a}^\top \boldsymbol{\beta}$ is

$$\widehat{\theta} \pm t_{n-m;1-\alpha/2} \frac{\|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\| \sqrt{\mathbf{a}^\top (X^\top X)^{-1} \mathbf{a}}}{\sqrt{n - m}} , \qquad (5.25)$$

where $t_{n-m;1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the t_{n-m} distribution.

Proof. Being linear in the components of $\boldsymbol{\beta}$, the random variable $\widehat{\theta} = \mathbf{a}^\top \widehat{\boldsymbol{\beta}}$ has a normal distribution, with expectation $\mathbf{a}^\top \boldsymbol{\beta} = \theta$ and variance $\sigma^2 \mathbf{a}^\top (X^\top X)^{-1} \mathbf{a}$. Let

$$\widehat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbf{Y}_m\|^2}{n} ,$$

with $\mathbf{Y}_m = X\widehat{\boldsymbol{\beta}}$, be the least-squares estimator of σ^2. The random variable $\|\mathbf{Y} - \mathbf{Y}_m\|^2/\sigma^2$ has, by Theorem 3.10, a χ^2_{n-m} distribution and is independent of $\mathbf{Y}_m$. Since $\widehat{\boldsymbol{\beta}} = X^+ X\widehat{\boldsymbol{\beta}} = X^+ \mathbf{Y}_m$, we have that $\|\mathbf{Y} - \mathbf{Y}_m\|^2$ is independent of $\widehat{\boldsymbol{\beta}}$. Using Corollary 3.2, we see that the pivot

☞ 88

☞ 89

$$T = \frac{(\widehat{\theta} - \theta)/\sqrt{\mathbf{a}^\top (X^\top X)^{-1} \mathbf{a}}}{\sqrt{\|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|^2/(n - m)}}$$

has a t_{n-m} distribution. By rearranging the identity $\mathbb{P}(|T| < t_{n-m;1-\alpha/2}) = 1 - \alpha$ in the usual way, we arrive at the confidence interval (5.25). $\qquad \square$

Example 5.12 (Confidence Limits in Simple Linear Regression). We continue
Example 5.5 by including confidence intervals, $(l(x), u(x))$ say, of the parameter
$\theta(x) = \beta_0 + \beta_1 x$, for various x. The points $u(x), x \in [0, 1]$ form an upper
confidence curve for the regression line $y = \beta_0 + \beta_1 x$; and $l(x)$ gives the lower
confidence curve. The following MATLAB code, to be appended to the code in
Example 5.5, implements (5.25) and yields a plot of the true regression line and
confidence curves similar to Fig. 5.2.

```
% linregestconf.m
tquant = icdf('t',0.975,N-2) % 0.975 quantile
ucl = zeros(1,N), lcl = zeros(1,N); % upper/lower conf. limits
rl = zeros(1,N);  % (true) regression line
u=0;
for i=1:N
    u = u + 1/N;
    a = [1;u];
    rl(i) = a'*beta;
    ucl(i) = a'*betahat + tquant*norm(y - X*betahat)* ...
            sqrt(a'*inv(X'*X)*a)/sqrt(N-2);
    lcl(i) = a'*betahat - tquant*norm(y - X*betahat)* ...
            sqrt(a'*inv(X'*X)*a)/sqrt(N-2);
end
hold on
plot(x,rl), plot(x,ucl), plot(x,lcl)
hold off
```
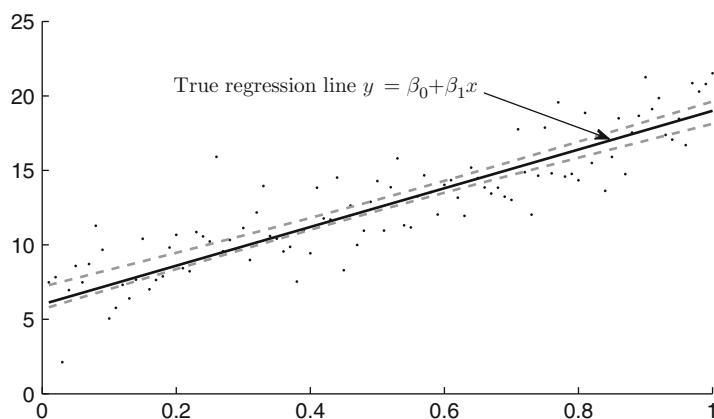

Fig. 5.2 The true regression line (*solid*) and the upper and lower 95 % confidence curves
(*dashed*)

5.3   Hypothesis Testing

Hypothesis testing involves making *decisions* about certain hypotheses on the basis of the observed data. In many cases we have to decide whether the observations are due to "chance" or due to an "effect." Hypothesis testing has traditionally played a prominent role in statistics, and many introductory books still are predominantly about hypothesis testing. Modern statistical analyses, however, especially those based on computer-intensive methods, do not so heavily rely on hypothesis testing anymore, preferring, for example, inference via confidence intervals to inference based on hypothesis tests. In Bayesian statistics hypothesis testing is done in a different way, via *Bayes factors*. We will address the main ideas of classical hypothesis testing in this section.

☞ 252

Suppose the model for the data $\mathbf{X}$ is described by a family of probability distributions that depend on a parameter $\boldsymbol{\theta} \in \Theta$. The aim of **hypothesis testing** is to decide, on the basis of the observed data $\mathbf{x}$, which of two competing hypotheses, $H_0 : \boldsymbol{\theta} \in \Theta_0$ (the **null hypothesis**) and $H_1 : \boldsymbol{\theta} \in \Theta_1$ (the **alternative hypothesis**), holds true, where Θ_0 and Θ_1 are subsets of the parameter space Θ. Traditionally, the null hypothesis and alternative hypothesis do not play equivalent roles. H_0 contains the "status quo" statement and is only rejected if the observed data are very unlikely to have happened under H_0.

The decision whether to reject H_0 or not is dependent on the outcome of a **test statistic $\mathbf{T} = \mathbf{T}(\mathbf{X})$**. For simplicity, we discuss only the one-dimensional case $\mathbf{T} \equiv T$.

The **p-value** is the probability that under H_0 the (random) test statistic takes a value as extreme as or more extreme than the one observed. Let t be the observed outcome of the test statistic T. We consider three types of tests:

- **Left one-sided test**. Here H_0 is rejected for small values of t, and the p-value is defined as $p = \mathbb{P}_{H_0}(T \leqslant t)$.
- **Right one-sided test**: Here H_0 is rejected for large values of t, and the p-value is defined as $p = \mathbb{P}_{H_0}(T \geqslant t)$,
- **Two-sided test**: In this test H_0 is rejected for small or large values of t, and the p-value is defined as $p = \min\{2\mathbb{P}_{H_0}(T \leqslant t), 2\mathbb{P}_{H_0}(T \geqslant t)\}$.

The smaller the p-value, the greater the strength of the evidence against H_0 provided by the data. As a rule of thumb:

$$
\begin{aligned}
&p < 0.10 \quad \text{suggestive evidence,} \\
&p < 0.05 \quad \text{reasonable evidence,} \\
&p < 0.01 \quad \text{strong evidence.}
\end{aligned}
$$

The following decision rule is generally used to decide between H_0 and H_1:

Decision rule: *Reject H_0 if the p-value is smaller than some p_0.*

In general, a statistical test involves the following steps.

Steps for a Statistical Test

1. Formulate a statistical model for the data.
2. Give the null and alternative hypotheses (H_0 and H_1).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject H_0 based on the p-value.

Choosing an appropriate test statistic is akin to selecting a good estimator for the unknown parameter θ. The test statistic should summarize the information about θ and make it possible to distinguish between the alternative hypotheses. The *likelihood ratio test* provides a systematic approach to constructing powerful test statistics; see Sect. 6.4.

Example 5.13 (Blood Pressure). Suppose the systolic blood pressure for white males aged 35–44 is known to be normally distributed with expectation 127 and standard deviation 7. A paper in a public health journal considers a sample of 101 diabetic males and reports a sample mean of 130. Is this good evidence that diabetics have on average a higher blood pressure than the general population?

To assess this, we could ask the question how likely it would be, *if diabetics were similar to the general population*, that a sample of 101 diabetics would have a mean blood pressure this far from 127.

Let us perform the seven steps of a statistical test. A reasonable model for the data is $X_1, \ldots, X_{101} \sim_{\text{iid}} \mathsf{N}(\mu, 49)$. Alternatively, the model could simply be $\bar{X} \sim \mathsf{N}(\mu, 49/101)$, since we only have an outcome of the sample mean of the blood pressures. The null hypothesis (the status quo) is $H_0 : \mu = 127$; the alternative hypothesis is $H_1 : \mu > 127$. We take $\bar{X}$ as the test statistic. Note that we have a right one-sided test here, because we would reject H_0 for high values of $\bar{X}$. Under H_0 we have $\bar{X} \sim \mathsf{N}(12749/101)$. The outcome of $\bar{X}$ is 130, so that the p-value is given by

$$\mathbb{P}(\bar{X} \geqslant 130) = \mathbb{P}\left(\frac{\bar{X} - 127}{\sqrt{49/101}} \geqslant \frac{130 - 127}{\sqrt{49/101}} \right) = \mathbb{P}(Z \geqslant 4.31) = 8.16 \cdot 10^{-6} \,,$$

where $Z \sim \mathsf{N}(0, 1)$. So it is extremely unlikely that the event $\{\bar{X} \geqslant 130\}$ occurs if the two groups are the same with regard to blood pressure. However, the event *has* occurred. Therefore, there is *strong* evidence that the blood pressure of diabetics differs from the general public.

Example 5.14 (Loaded Die). We suspect a certain die to be loaded. Throwing 100 times we observe 25 sixes. Is there enough evidence to justify our suspicion?

We ask ourselves the same type of question as in the previous example: Suppose that the die is fair. What is the probability that out of 100 tosses 25 or more sixes would appear? To calculate this, let X be the number of sixes out of 100. Our model is $X \sim \mathsf{Bin}(100, p)$, with p unknown. We would like to show the hypothesis H_1 : $p > 1/6$; otherwise, we do not reject (accept) the null hypothesis $H_0 : p = 1/6$. Our test statistic is simply X. Under H_0, $X \sim \mathsf{Bin}(1001/6)$, so that the p-value for this right one-sided test is

$$\mathbb{P}(X \geqslant 25) = \sum_{k=25}^{100} \binom{100}{k}(1/6)^k (5/6)^{100-k} \approx 0.0217 .$$

This is quite small. Hence, we have *reasonable* evidence that the die is loaded.

5.3.1   ANOVA for the Normal Linear Model

☞ 115   Hypothesis testing for the normal linear model in Sect. 4.23 is often related to *model selection*. In particular, suppose we have the following model for the data $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$:

$$\mathbf{Y} = \underbrace{X_1 \boldsymbol{\beta}_1 + X_2 \boldsymbol{\beta}_2}_{X\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(0, \sigma^2 I) , \tag{5.26}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown vectors of dimension k and $m - k$, respectively; and X_1 and X_2 are full-rank design matrices of dimensions $n \times k$ and $n \times (m - k)$, respectively. Above we implicitly defined $X = (X_1, X_2)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)$.

Suppose we wish to test the hypothesis $H_0 : \boldsymbol{\beta}_2 = 0$ against $H_1 : \boldsymbol{\beta}_2 \neq 0$.

☞ 125   We saw in Sect. 5.1.2 how to estimate the parameters via least squares. Let $\widehat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ under the full model, and let $\widehat{\boldsymbol{\beta}}_1$ denote the estimate of $\boldsymbol{\beta}_1$ for the reduced model, that is, under H_0. To simplify notation, let $\mathbf{Y}_m = X\widehat{\boldsymbol{\beta}}$ be the projection of $\mathbf{Y}$ onto the space $\langle X \rangle$ spanned by the columns of X; and let $\mathbf{Y}_k = X_1\widehat{\boldsymbol{\beta}}_1$ be the projection of $\mathbf{Y}$ onto the space $\langle X_1 \rangle$ spanned by the columns of X_1 only.

A sensible strategy for deciding upon the reduced or full model is to compare $\|\mathbf{Y} - \mathbf{Y}_k\|$ with $\|\mathbf{Y} - \mathbf{Y}_m\|$ via the quotient of the two. The larger this quotient, the more evidence for the full model. It is more convenient to use instead the equivalent statistic

$$T = \frac{n - m}{m - k} \times \frac{\|\mathbf{Y} - \mathbf{Y}_k\|^2 - \|\mathbf{Y} - \mathbf{Y}_m\|^2}{\|\mathbf{Y} - \mathbf{Y}_m\|^2} = \frac{\|\mathbf{Y}_m - \mathbf{Y}_k\|^2/(m - k)}{\|\mathbf{Y} - \mathbf{Y}_m\|^2/(n - m)} , \tag{5.27}$$

where we have used Pythagoras' theorem in the second equation above, as illustrated in Fig. 5.3.
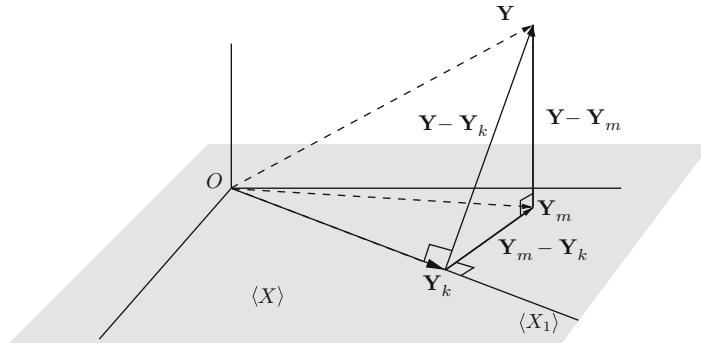
Fig. 5.3 Pythagoras' theorem

Define $\mathbf{X} = \mathbf{Y}/\sigma$ with expectation $\boldsymbol{\mu} = X\boldsymbol{\beta}/\sigma$ and $\mathbf{X}_j = \mathbf{Y}_j/\sigma$ with expectation $\boldsymbol{\mu}_j$, $j = k, m$. Note that $\boldsymbol{\mu} = \boldsymbol{\mu}_m$, and under H_0, $\boldsymbol{\mu}_m = \boldsymbol{\mu}_k$. We can directly apply Theorem 3.10 to find that $\|\mathbf{Y} - \mathbf{Y}_m\|^2/\sigma^2 = \|\mathbf{X} - \mathbf{X}_m\|^2 \sim \chi^2_{n-m}$ and, under H_0, ☞ 88 $\|\mathbf{Y}_m - \mathbf{Y}_k\|^2/\sigma^2 \sim \chi^2_{m-k}$. Moreover, these random variables are independent of each other. It follows from Theorem 3.11 that, under H_0,

$$T \sim \mathsf{F}(m - k, n - m) .$$

We reject H_0 for large values of T. The above methodology is often referred to as **analysis of variance** (ANOVA).

Example 5.15 (Hypothesis Testing for Randomized Block Design). In a randomized block design the data are collected in blocks, in order to reduce variability in the experiment. Consider, for example, the data in Table 5.2, representing the crop yield using four different crop treatments (e.g., strengths of fertilizer) on four different blocks (plots).

Table 5.2 Crop yield

Block	Treatment			
	1	2	3	4
1	9.2988	9.4978	9.7604	10.1025
2	8.2111	8.3387	8.5018	8.1942
3	9.0688	9.1284	9.3484	9.5086
4	8.2552	7.8999	8.4859	8.9485

Let us consider the data first as coming from four different groups, depending only on the level of treatment. A possible model would be the single-factor ANOVA model ☞ 112

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \quad i, k = 1, \ldots, 4 ,$$

with $\{\varepsilon_{ik}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, and $\sum_{i=1}^{4} \alpha_i = 0$. Ordering the $\{Y_{ik}\}$ into a column vector $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots Y_{14}, Y_{21}, \ldots, Y_{44})^\top$, we can write $\mathbf{Y}$ in the form (5.26):

$$\mathbf{Y} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}}_{X_1} \mu + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}}_{X_2} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}}_{\beta_2} + \boldsymbol{\varepsilon} \ ,$$

where $\mathbf{1}$ and $\mathbf{0}$ are vectors of 1s and 0s, respectively. We wish to test whether the treatments make a difference to the crop yield or not. The null hypothesis H_0 : $\alpha_1 = \alpha_2 = \alpha_3 = 0$ is that the treatments have no effect. As a test statistic we use (5.27). For the present model we have $n = 16$, $m = 4$, and $k = 1$. The squared norm $\|\mathbf{Y} - \mathbf{Y}_m\|^2 = \|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|^2$ is often written as SS_{error}, that is, the sum of squares of the error terms. Note that $\|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|^2/(n-m)$ is an unbiased estimator of the variance σ^2 of the model error.

Similarly, $\|\mathbf{Y}_m - \mathbf{Y}_k\|^2$ represents the sum of squares due to the treatment effect and is written as $\text{SS}_{\text{treatment}}$. Our test statistic T in (5.27) can thus be written as

$$T = \frac{\text{SS}_{\text{treatment}}/(m-k)}{\text{SS}_{\text{error}}/(n-m)} \overset{\text{def}}{=} \frac{\text{MS}_{\text{treatment}}}{\text{MS}_{\text{error}}} \ ,$$

where "MS" stands for "mean square." Under H_0 the test statistic T has an $\mathsf{F}(m - k, n - m) = \mathsf{F}(3, 12)$ distribution.

```
%solvehypotcrop1.m
yy = [9.2988, 9.4978, 9.7604, 10.1025;
      8.2111, 8.3387, 8.5018,  8.1942;
      9.0688, 9.1284, 9.3484,  9.5086;
      8.2552, 7.8999, 8.4859,  8.9485];
n = numel(yy); [nrow,ncol] = size(yy); y = yy(:);
X_1 = ones(n,1);
KM = kron(eye(ncol),ones(nrow,1)); X_2 = KM(:,1:ncol-1);
X_2(n-nrow+1:n,:) = -ones(nrow,ncol-1);
X = [X_1 X_2];
m = size(X,2);
betahat = X'*X\(X'*y);
ym = X*betahat;
yk = X_1*mean(y); %omitting treatment effect
k=1;  %number of parameters in reduced model
T = (n-m)/(m-k)*(norm(ym - yk)^2)/norm(y-ym)^2
pval = 1 - cdf('F',T,m-k,n-m)
```

The outcome of T is found to be 0.4724, which gives a p-value of 0.7072. This suggests that the treatment does not have an effect on the crop yield. But what if the

crop yield is not only determined by the treatment levels but also by the blocks? To investigate this, we could describe the data via a two-factor ANOVA model:

$$Y_{ik} = \mu + \alpha_i + \tau_k + \varepsilon_{ik}, \quad i, k = 1, \dots, 4 ,$$

with $\{\varepsilon_{ik}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, and $\sum_{i=1}^{4} \alpha_i = 0$ and $\sum_{i=1}^{4} \tau_i = 0$. Ordering the data in the same way as for the one-factor case, we can write

$$\mathbf{Y} = X_1 \mu + X_2 \boldsymbol{\beta}_2 + \underbrace{\begin{pmatrix} C \\ C \\ C \\ C \end{pmatrix}}_{X_3} \underbrace{\begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}}_{\boldsymbol{\beta}_3} + \boldsymbol{\varepsilon}, \quad \text{with} \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix},$$

and X_1 and X_2 are the same as in the one-factor case. We wish to test first if using such an extended model (as opposed to the previous one-factor model) is justified. In particular, we test if $\tau_1 = 0, \dots, \tau_4 = 0$. We can use again a statistic of the form (5.27). Now the vector $\mathbf{Y}_m$ is the projection of $\mathbf{Y}$ onto the $(m = 7)$-dimensional space spanned by the columns of $X = (X_1, X_2, X_3)$; and $\mathbf{Y}_k$ is the projection of $\mathbf{Y}$ onto the $(k = 4)$-dimensional space spanned by the columns of $X_{12} = (X_1, X_2)$. The test statistic (5.27), which we could write as

$$T_{12} = \frac{\mathrm{MS}_{\text{blocks}}}{\mathrm{MS}_{\text{error}}} ,$$

has under H_0 an $\mathsf{F}(3, 9)$ distribution.

The MATLAB code below, which has to be appended to the first seven lines of code for the one-factor case, calculates the outcome of the test statistic T_{12} and the corresponding p-value. We find $t_{12} = 34.9998$, which gives a p-value 2.73×10^{-5}. This shows that the block effects are extremely important for explaining the data.

Using the extended model—thus with the block effects—we can again test whether the $\{\alpha_i\}$ are all 0 or not. This is done in the last six lines of the code below. The outcome of the test statistic is 4.4878, with a p-value of 0.0346. By including the block effects, we effectively reduce the uncertainty in the model and are able to more accurately assess the effects of the treatments, to conclude that the treatment *does* seem to have an effect on the crop yield. A closer look at the data shows that within each block (row) the crop yield roughly increases with the treatment level.

```
%solvehypotcrop2.m
C = [eye(nrow-1), -ones(nrow-1,1)]'; X_3 = repmat(C,ncol,1);
X = [X_1 X_2 X_3];
m = size(X,2); %number of parameters in full model
betahat = X'*X\(X'*y); %estimate under the full model
```

```
ym = X*betahat;
X_12 = [X_1 X_2]; %omitting the block effect
k = size(X_12, 2); %number of parameters in reduced model
betahat_12 = X_12'*X_12\(X_12'*y);
y_12 = X_12*betahat_12;
T_12=(n-m)/(m-k)*(norm(y-y_12)^2 - norm(y-ym)^2)/norm(y-ym)^2
pval_12 = 1 - cdf('F',T_12,m-k,n-m)
X_13 = [X_1 X_3]; %omitting the treatment effect
k = size(X_13, 2); %number of parameters in reduced model
betahat_13 = X_13'*X_13\(X_13'*y);
y_13 = X_13*betahat_13;
T_13=(n-m)/(m-k)*(norm(y-y_13)^2 - norm(y-ym)^2)/norm(y-ym)^2
pval_13 = 1 - cdf('F',T_13,m-k,n-m)
```

5.4 Cross-Validation

For experimental data it is often the case that several competing models seem equally appropriate. As a concrete example, suppose we observe n independent points in the $x-y$ plane, as depicted in Fig. 5.4. We wish to find a suitable polynomial that fits the data well. To that end, we consider the 5th-order polynomial regression model; see (4.10):

☞ 108

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_5 x_i^5 + \varepsilon_i \ ,$$

where $\{\varepsilon_i\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$. The fitted line is also depicted in Fig. 5.4, which seems to fit the points reasonably well.
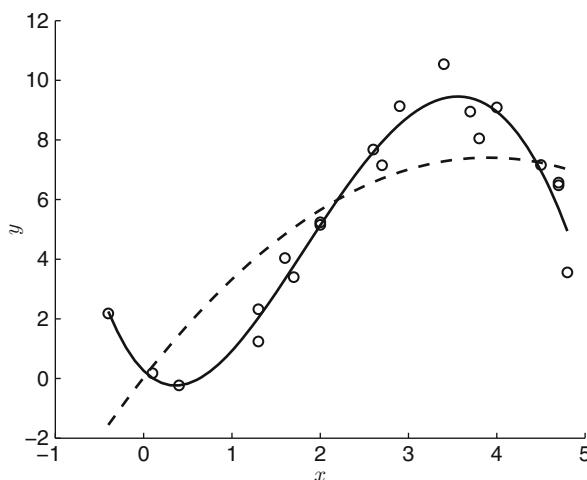
Fig. 5.4 Quadratic (*dotted*) and 5th-order (*solid*) polynomial regression lines

Since the 5th-order polynomial is adequate, we might not need to consider higher-order polynomials. However, it is plausible that a simpler model (e.g., a cubic polynomial) would fit the data almost as well and is therefore more appropriate. One common approach is to test a sequence of hypotheses to determine the exact degree needed. That is, first we estimate the 5th-order polynomial regression model and test the null hypothesis that $\beta_5 = 0$. If the null hypothesis is rejected, we stop and use the 5th-order polynomial. Otherwise, we estimate the 4th-order polynomial regression model and test the null hypothesis that $\beta_4 = 0$. This process is continued until a certain null hypothesis is rejected.

A more thoughtful approach is to select a model based on its predictive performance. After all, one main goal of statistical inference is to predict future observations. One way to assess the predictive ability of a model is to use it to predict a set of observations not used in the estimation. This can be done, for example, by partitioning the data into a "training set" and a "test set." Then, use the "training set" to estimate the model, and its predictive accuracy is assessed by some error measure on the "test set." This is an example of a **cross-validation**.

Fig. 5.5 A graphical representation of a 4-fold cross-validation

More generally, a **K-fold cross-validation** is implemented as follows:

1. Partition the data into K subsamples of equal (or nearly equal) size. Number the subsamples from 1 to K.
2. For $k = 1, \ldots, K$, use all but the kth subsample to estimate the model parameters. Compute the prediction errors for the omitted observations in the kth subsample.
3. Summarize the predictive performance by some error measure, such as the sum of squared errors.

A graphical representation of a 4-fold cross-validation is depicted in Fig. 5.5. For a sample with n observations, we can implement at most an n-fold cross-validation. In fact, this is a popular choice, and it is often called the **leave-one-out cross-validation**.

More specifically, suppose there are n independent observations $y_1, \ldots, y_n$. Let $\widehat{y}_{-k}$ denote the prediction for the kth observation using all the data except y_k. The prediction error $y_k - \widehat{y}_{-k}$ is called a *predicted residual*—in contrast to an ordinary residual, $u_k = y_k - \widehat{y}_k$, which is the difference between an observation and its fitted value obtained using the whole sample. At the end of n iterations, we

obtain the collection of predicted residuals $\{y_k - \widehat{y}_{-k}\}$. One way to summarize them is through the **predicted residual sum of squares** or **PRESS**:

$$\text{PRESS} = \sum_{k=1}^{n} (y_k - \widehat{y}_{-k})^2 .$$

In general, computing the PRESS is computationally intensive as it involves n separate estimations and predictions. For linear models, however, the predicted residuals can be calculated quickly using only the ordinary residuals and the projection matrix.

Theorem 5.4. (PRESS for Linear Models). Consider the normal linear model (4.23)

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I) ,$$

where the $n \times m$ design matrix $X = (x_{ij})$ is known and is of full rank. Given an outcome $\mathbf{y} = (y_1, \ldots, y_n)^\top$ of $\mathbf{Y}$, the fitted values can be obtained as $\widehat{\mathbf{y}} = P\mathbf{y}$, where $P = X(X^\top X)^{-1} X^\top$ is the projection matrix. Then, the predicted residual sum of squares can be written as

$$\text{PRESS} = \sum_{k=1}^{n} \left(\frac{u_k}{1 - p_k} \right)^2 ,$$

where $u_k = y_k - \widehat{y}_k = y_k - (X\widehat{\boldsymbol{\beta}})_k$ is the kth residual and p_k is the kth diagonal element of the projection matrix P.

Proof (Sketch). It suffices to show that the kth predicted residual can be written as $y_k - \widehat{y}_{-k} = u_k/(1 - p_k)$. Let X_{-k} denote the design matrix X with the kth row removed, and define $\mathbf{y}_{-k}$ similarly. Then, the least-squares estimate for $\boldsymbol{\beta}$ using all but the kth observation is $\widehat{\boldsymbol{\beta}}_{-k} = (X_{-k}^\top X_{-k})^{-1} X_{-k}^\top \mathbf{y}_{-k}$. It can be shown (see Problem 5.18) that $\widehat{\boldsymbol{\beta}}_{-k}$ is related to the full-sample least-squares estimate $\widehat{\boldsymbol{\beta}}$ via

$$\widehat{\boldsymbol{\beta}}_{-k} = \widehat{\boldsymbol{\beta}} - \frac{(X^\top X)^{-1} \mathbf{x}_k u_k}{1 - p_k} , \tag{5.28}$$

where $\mathbf{x}_k^\top$ is the kth row of the design matrix X. It follows that the predicted value for the kth observation is given by

$$\widehat{y}_{-k} = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_{-k} = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}} - \frac{\mathbf{x}_k^\top (X^\top X)^{-1} \mathbf{x}_k u_k}{1 - p_k} = \widehat{y}_k - \frac{p_k u_k}{1 - p_k} \,,$$

where we used the fact that $p_k = \mathbf{x}_k^\top (X^\top X)^{-1} \mathbf{x}_k$. The desired result now follows from direct calculation. $\qquad\square$

Example 5.16 (Leave-One-Out Cross-Validation for Polynomial Regressions).
In this example we revisit the polynomial regression example in the beginning of this section. Specifically, given the $n = 20$ points in the x–y plane listed in Table 5.3 (see also Fig. 5.4), we wish to find the simplest polynomial that fits the points well.

Table 5.3 Polynomial regression data

x	y	x	y	x	y	x	y
4.7	6.57	3.7	8.95	4.8	3.56	0.4	−0.23
2.0	5.15	2.0	5.24	1.7	3.40	2.6	7.68
2.7	7.15	3.4	10.54	−0.4	2.18	4.0	9.09
0.1	0.18	1.3	1.24	4.5	7.16	2.9	9.13
4.7	6.48	3.8	8.05	1.3	2.32	1.6	4.04

For this purpose, we consider five different polynomial regression models:

$$Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k + \varepsilon_i$$

for $k = 1, \ldots, 5$, where $\{\varepsilon_i\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$. Since they can all be written as normal linear models, we can use Theorem 5.4 to compute their predicted residual sums of squares. For each of these models, we compute the least-squares estimate and the corresponding PRESS using the MATLAB script below.

```
% polyreg.m
x = [4.7,2,2.7,0.1,4.7,3.7,2,3.4,1.3,3.8,4.8,1.7,...
    -0.4,4.5,1.3,0.4,2.6,4,2.9,1.6]';
y = [6.57,5.15,7.15,0.18,6.48,8.95,5.24,10.54,1.24,...
    8.05,3.56,3.4,2.18,7.16,2.32,-0.23,7.68,9.09,9.13,4.04]';
n = size(x,1);
press = zeros(5,1);
X = ones(n,1);
for k=1:5
    X = [X x.^k];    % construct the design matrix
    P = X*((X'*X)\X');
```

```
    e = y - P*y;
    press(k)  = sum((e./(1-diag(P))).^2);
end
```

The PRESS values for the linear, quadratic, cubic, 4th-, and 5th-order polynomial regression models are, respectively, 117.388, 130.781, 16.0532, 16.3167, and 25.727. Hence, the cubic polynomial regression has the lowest PRESS, indicating that it has the best predictive performance. It illustrates that complex models do not necessarily have better predictive accuracy than simpler models.

5.5   Sufficiency and Exponential Families

A statistic—that is, a function of the data only—is said to be **sufficient** for a parameter (vector) $\boldsymbol{\theta}$ if it captures all the information about $\boldsymbol{\theta}$ contained in the data. Sufficient statistics can be used to *summarize* data, often giving a tremendous reduction in size. To formalize this concept, suppose that $\mathbf{T}(\mathbf{X})$ is a (possibly multidimensional) statistic for $\boldsymbol{\theta}$ such that any inference about $\boldsymbol{\theta}$ depends on the data $\mathbf{X} = (X_1, \ldots, X_n)^\top$ only through the value $\mathbf{T}(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are outcomes such that $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$, then the inference about $\boldsymbol{\theta}$ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed. This observation leads to the following definition.

> **Definition 5.2. (Sufficient Statistic).** A statistic $\mathbf{T}(\mathbf{X})$ is a **sufficient statistic** for $\boldsymbol{\theta}$ if the conditional distribution of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

The workhorse for establishing sufficiency is the following theorem.

> **Theorem 5.5. (Factorization Theorem).** Let $f(\mathbf{x}; \boldsymbol{\theta})$ denote the pdf of the data $\mathbf{X} = (X_1, \ldots, X_n)^\top$. A statistic $\mathbf{T}(\mathbf{X})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions $g(\mathbf{t}, \boldsymbol{\theta})$ and $h(\mathbf{x})$ such that, for all $\mathbf{x}$ and $\boldsymbol{\theta}$,
>
> $$f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})\, h(\mathbf{x}) . \qquad (5.29)$$

Proof. We give the proof only for the case where $\mathbf{X}$ is a discrete random vector. For this case we can write $f(\mathbf{x}; \boldsymbol{\theta})$ as

$$
f(\mathbf{x}; \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x})
$$
$$
= \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}, \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))
$$
$$
= \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \, \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \,.
$$

If $\mathbf{T}(\mathbf{X})$ is a sufficient statistic, then $\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$ does not depend on $\boldsymbol{\theta}$. Consequently, (5.29) holds with $g(\mathbf{t}, \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t})$ and $h(\mathbf{x}) = \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$.

Conversely, suppose that (5.29) holds. We need to show that the conditional probability

$$
\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{t}) = \frac{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}, \mathbf{T}(\mathbf{X}) = \mathbf{t})}{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t})}
$$

does not depend on $\boldsymbol{\theta}$. If $\mathbf{x}$ is a data point such that $\mathbf{T}(\mathbf{x}) \neq \mathbf{t}$, then clearly $\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{t}) = 0$. If $\mathbf{T}(\mathbf{x}) = \mathbf{t}$, then

$$
\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} \mid \mathbf{T}(\mathbf{X}) = \mathbf{t}) = \frac{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x})}{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t})} = \frac{f(\mathbf{x}; \boldsymbol{\theta})}{\sum_{\mathbf{y}:\mathbf{T}(\mathbf{y})=\mathbf{t}} f(\mathbf{y}; \boldsymbol{\theta})}
$$

$$
= \frac{g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) \, h(\mathbf{x})}{\sum_{\mathbf{y}:\mathbf{T}(\mathbf{y})=\mathbf{t}} g(\mathbf{T}(\mathbf{y}), \boldsymbol{\theta}) \, h(\mathbf{y})} = \frac{g(\mathbf{t}, \boldsymbol{\theta}) \, h(\mathbf{x})}{g(\mathbf{t}, \boldsymbol{\theta}) \sum_{\mathbf{y}:\mathbf{T}(\mathbf{y})=\mathbf{t}} h(\mathbf{y})}
$$

$$
= \frac{h(\mathbf{x})}{\sum_{\mathbf{y}:\mathbf{T}(\mathbf{y})=\mathbf{t}} h(\mathbf{y})} \,,
$$

which does not depend on $\boldsymbol{\theta}$. Hence $\mathbf{T}(\mathbf{X})$ is a sufficient statistic. $\square$

Example 5.17 (Sufficient Statistic for Iid Uniform Data). Let $\mathbf{X} = (X_1, \dots, X_n)$ be an iid sample from $\mathsf{U}(0, \theta)$. The pdf of $\mathbf{X}$ is given by

$$
f(\mathbf{x}; \theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } \max\{x_1, \dots, x_n\} \leqslant \theta \text{ and } x_i \geqslant 0, \ i = 1, \dots, n \\ 0 & \text{otherwise} \,. \end{cases}
$$

It follows that $T(\mathbf{X}) = \max(X_1, \dots, X_n)$ is a sufficient statistic for θ.

Example 5.18 (Sufficient Statistic for Iid Normal Data). Let $\mathbf{X} = (X_1, \dots, X_n)$ be an iid sample from $\mathsf{N}(\mu, 1)$. We show that the sample mean $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. Namely, the pdf is

$$f(\mathbf{x}; \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

$$= h(\mathbf{x}) \underbrace{\exp\left(\mu\, n\, \bar{x} - n\, \mu^2/2\right)}_{g(T(\mathbf{x}), \mu)}$$

for some function h, so that the required factorization holds.

The following general class of distributions plays an important role in statistics.

Definition 5.3. (Exponential Family). Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a random vector with pdf $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ is a parameter vector. $\mathbf{X}$ is said to belong to an m-dimensional **exponential family** if there exist real-valued functions $t_i(\mathbf{x})$, $\eta_i(\boldsymbol{\theta})$, $i = 1, \ldots, m \leqslant n$ and $h(\mathbf{x}) > 0$, and a (normalizing) function $c(\boldsymbol{\theta}) > 0$, such that

$$f(\mathbf{x}; \boldsymbol{\theta}) = c(\boldsymbol{\theta})\, \exp\left(\sum_{i=1}^m \eta_i(\boldsymbol{\theta}) t_i(\mathbf{x})\right) h(\mathbf{x}) . \qquad (5.30)$$

The representation of an exponential family is in general not unique. It is often convenient to reparameterize exponential families via the $\{\eta_i\}$, that is, to take $\boldsymbol{\eta} = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta}))^\top$ as the parameter vector rather than $\boldsymbol{\theta}$. The reparameterized pdf is then

$$\widetilde{f}(\mathbf{x}; \boldsymbol{\eta}) = \widetilde{c}(\boldsymbol{\eta})\, e^{\boldsymbol{\eta}^\top \mathbf{t}(\mathbf{x})}\, h(\mathbf{x}) , \qquad (5.31)$$

where $\widetilde{c}(\boldsymbol{\eta})$ is the normalization constant and $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \ldots, t_m(\mathbf{x}))$. Such an exponential family is said to be in **canonical form** or is said to be a **natural exponential family**.

Example 5.19 (Normal Distribution as a Two-Dimensional Exponential Family). The normal distributions $\mathsf{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$ form a two-dimensional exponential family with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)$. To see this, write the logarithm of the pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution as

$$\ln f(x; \boldsymbol{\theta}) = \ln(1/\sqrt{2\pi\sigma^2}) - \frac{1}{2}\frac{(x - \mu)^2}{\sigma^2}$$

$$= \ln(1/\sqrt{2\pi\sigma^2}) - \frac{\mu^2}{2\sigma^2} + x\frac{\mu}{\sigma^2} - x^2\frac{1}{2\sigma^2} ,$$

which shows that we can take $t_1(x) = x$, $t_2(x) = x^2$, $\eta_1(\boldsymbol{\theta}) = \mu/\sigma^2$, and $\eta_2(\boldsymbol{\theta}) = -1/(2\sigma^2)$, with $h(x) = 1$ and $c(\boldsymbol{\theta}) = \exp(-\mu^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$.

Many other families of distributions are of this type, such as the binomial, gamma, beta, geometric, and Poisson distributions, as summarized in Table 5.4.

Table 5.4 Various univariate exponential families

Distr.	$\boldsymbol{\theta}$	$t_1(x),\ t_2(x)$	$c(\boldsymbol{\theta})$	$\eta_1(\boldsymbol{\theta}),\ \eta_2(\boldsymbol{\theta})$	$h(x)$
Beta(α,β)	(α,β)	$\ln x,\ \ln(1-x)$	$1/B(\alpha,\beta)$	$\alpha-1,\ \beta-1$	1
Bin(n,p)	p	$x,\ -$	$(1-p)^n$	$\ln\left(\dfrac{p}{1-p}\right),\ -$	$\dbinom{n}{x}$
Gamma(α,λ)	(α,λ)	$x,\ \ln x$	$\dfrac{\lambda^\alpha}{\Gamma(\alpha)}$	$-\lambda,\ \alpha-1$	1
Geom(p)	p	$x-1,\ -$	p	$\ln(1-p),\ -$	1
N(μ,σ^2)	(μ,σ^2)	$x,\ x^2$	$\dfrac{e^{-\mu^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$	$\dfrac{\mu}{\sigma^2},\ -\dfrac{1}{2\sigma^2}$	1
Poi(λ)	λ	$x,\ -$	$e^{-\lambda}$	$\ln\lambda,\ -$	$\dfrac{1}{x!}$

Sufficiency (and therefore data summarization) is particularly easy to establish for exponential families of distributions. In particular, suppose that $\mathbf{X} = (X_1,\ldots,X_n)^\top$ is an iid sample from the exponential family with pdf

$$\overset{\circ}{f}(x;\boldsymbol{\theta}) = c(\boldsymbol{\theta})\, e^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta})\, t_i(x)}\, \overset{\circ}{h}(x)\,.$$

For simplicity suppose that x is 1-dimensional. By taking the product of the marginal pdfs we obtain the pdf of $\mathbf{X}$:

$$f(\mathbf{x};\boldsymbol{\theta}) = \underbrace{c(\boldsymbol{\theta})^n\, e^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta})\, \sum_{k=1}^n t_i(x_k)}}_{g(\mathbf{T}(\mathbf{x}),\boldsymbol{\theta})}\, \underbrace{\prod_{k=1}^n \overset{\circ}{h}(x_k)}_{h(\mathbf{x})}\,.$$

A direct consequence of the factorization theorem is that

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{k=1}^n t_1(X_k),\ldots,\sum_{k=1}^n t_m(X_k)\right)^\top \tag{5.32}$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Example 5.20 (Sufficient Statistics for Iid Normal Data). As a particular instance of the previous discussion, consider the case where $X_1,\ldots,X_n \sim_{\text{iid}} \mathsf{N}(\mu,\sigma^2)$. It follows from (5.32) and Example 5.19 that $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}),T_2(\mathbf{X}))^\top$, with $T_1(\mathbf{X}) = \sum_{k=1}^n X_k$ and $T_2(\mathbf{X}) = \sum_{k=1}^n X_k^2$, is a sufficient statistic for $\boldsymbol{\theta} = (\mu,\sigma^2)$. This means that for this standard data model, the data can be summarized via only T_1 and T_2.

It is not difficult to see that any 1-to-1 function of a sufficient statistic yields again a sufficient statistic. To see this, suppose that $\mathbf{T}(\mathbf{X})$ is a sufficient statistic and $\widetilde{\mathbf{T}}(\mathbf{X}) = \mathbf{r}(\mathbf{T}(\mathbf{X}))$ is another statistic, with $\mathbf{r}$ being invertible with inverse $\mathbf{r}^{-1}$. By the factorization theorem

$$f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) = g(\mathbf{r}^{-1}(\widetilde{\mathbf{T}}(\mathbf{x})), \boldsymbol{\theta}) h(\mathbf{x}) = \widetilde{g}(\widetilde{\mathbf{T}}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x})$$

for some function $\widetilde{g}$. Thus, the factorization theorem also holds for $\widetilde{\mathbf{T}}$, and therefore the latter is also a sufficient statistic for $\boldsymbol{\theta}$.

Example 5.21 (Sufficient Statistics for Iid Normal Data, Continued). We have seen that $T_1(\mathbf{X}) = \sum_{k=1}^{n} X_k$ and $T_2(\mathbf{X}) = \sum_{k=1}^{n} X_k^2$ are sufficient statistics for $\boldsymbol{\theta} = (\mu, \sigma^2)$ in the standard model for data. The sample mean $\widetilde{T}_1 = \bar{X}$ and the sample variance

$$\widetilde{T}_2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \right)$$

also form a pair of sufficient statistics, because the mapping

$$\widetilde{T}_1 = \frac{T_1}{n}, \qquad \widetilde{T}_2 = \frac{1}{n-1} \left(T_2 - T_1^2/n \right)$$

is invertible.

5.6   Problems

5.1. Find the method of moments estimators for the parameters of the $\mathsf{Geom}(p)$, $\mathsf{Poi}(\lambda)$, and $\mathsf{Gamma}(\alpha, \lambda)$ distributions.

5.2. The **mean square error** (MSE) of a real-valued estimator T is defined as $\mathrm{MSE} = \mathbb{E}_\theta (T - \theta)^2$. It can be used to assess the quality of an estimator: the smaller the MSE, the more efficient the estimator. Show that the MSE can be written as the sum

$$\mathrm{MSE} = (\mathbb{E}_\theta T - \theta)^2 + \mathrm{Var}_\theta(T) .$$

In particular, for an *unbiased* estimator the MSE is simply equal to its variance.

5.3. The normal equations (5.10) can be derived more directly by solving $\nabla_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \mathbf{0}$, where $\nabla_{\boldsymbol{\beta}}$ indicates the gradient with respect to $\boldsymbol{\beta}$. Show, using Sect. B.1, that

$$\nabla_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = 2X^\top (\mathbf{y} - X\boldsymbol{\beta}) .$$

5.4. We wish to estimate the area $a = \mu_1 \mu_2$ of a rectangular plot of land, with length μ_1 and width μ_2. We measure the length and the width twice. There are two natural ways to estimate the unknown constant a. We can either multiply the average width and length, or we can take the average of the two estimated areas. Suppose the measurements are outcomes of independent random variables $X_1, X_2 \sim \mathsf{N}(\mu_1, \sigma^2)$ and $Y_1, Y_2 \sim \mathsf{N}(\mu_2, \sigma^2)$. Here σ describes the accuracy of our measuring instrument. Let

$$T_1 = \frac{X_1 + X_2}{2} \times \frac{Y_1 + Y_2}{2} \quad \text{and} \quad T_2 = \frac{X_1 \times Y_1 + X_2 \times Y_2}{2}.$$

a. Show that T_1 and T_2 are *unbiased* estimators of a.
b. Show that $\mathrm{Var}(X_1 Y_1) = \sigma^2(\sigma^2 + \mu_1^2 + \mu_2^2)$.
c. Derive the variance of T_1 and the variance of T_2 and infer from this which estimator is preferred.

5.5. Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{Exp}(\lambda)$ for some unknown $\lambda > 0$.

a. Show that the method of moments estimator of λ is $1/\bar{X}$.
b. Construct an approximate $1 - \alpha$ stochastic confidence interval for λ, by applying the central limit theorem to $\bar{X}$.

5.6. Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(1, \sigma^2)$ for some unknown $\sigma^2 > 0$.

a. Show that $T = \sum_{i=1}^{n}(X_i - 1)^2/\sigma^2 \sim \chi_n^2$.
b. Construct a $1 - \alpha$ stochastic confidence interval for σ^2 using the pivot T.

5.7. A *buret* is a glass tube with scales that can be used to add a specified volume of a fluid to a receiving vessel. Determine a 95 % confidence interval for the expected volume of *one* drop of water that leaves the buret if the initial volume in the buret is 25.35 (ml), the volume after 50 drops is 22.84, and the volume after 100 drops is 20.36.

5.8. On the label of a certain packet of aspirin it is written that the standard deviation of the amount of aspirin per tablet is 1.0 mg, but we suspect this is not true. To investigate this we take a sample of 25 tablets and find that the sample standard deviation of the amount of aspirin is 1.3 mg. Determine a 95 % numerical confidence interval for σ. Is our suspicion justified?

5.9. Show that S_p in (5.20) is an unbiased estimator of σ^2.

5.10. Show that for the simple linear regression model in Example 5.5 we have   ☞ 127
$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$ and $\widehat{\beta}_1 = S_{xY}/S_{xx}$, where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}).$$

☞ 142 **5.11.** Consider the model selection for the normal linear model in Sect. 5.3.1. We wish to assess how the extended model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$, fits the data, compared to the default model $\mathbf{Y} = \mu\mathbf{1} + \boldsymbol{\varepsilon}$ (i.e., the $\{Y_i\}$ are independent and $\mathsf{N}(\mu, \sigma^2)$ distributed). To do this we can compare the variance of the original data, estimated via $\sum_i (Y_i - \bar{Y})^2/n = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2/n$, with the variance of the fitted data, estimated via $\sum_i (\widehat{Y}_i - \bar{Y})^2/n = \|\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2/n$, where $\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}}$. Note that, in the notation of Fig. 5.3, $\widehat{\mathbf{Y}} = \mathbf{Y}_m$ and $\bar{Y}\mathbf{1} = \mathbf{Y}_k$. The quantity

$$R^2 = \frac{\|\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2} \tag{5.33}$$

is called the **coefficient of determination** of the linear model. Note that R^2 lies between 0 and 1. An R^2 value close to 1 indicates that a large proportion of the variance in the data has been explained by the model.

a. Show that

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{SS}_{\mathrm{total}}} \stackrel{\text{def}}{=} 1 - \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} .$$

Hint: use Pythagoras' theorem, as in Fig. 5.3.

b. For the simple linear regression model in Problem 5.10 show that $R = \sqrt{R^2}$

☞ 125   is equal to the sample correlation coefficient (5.7) —where each X_i is replaced with x_i. Hint: write out $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ using the explicit expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in Problem 5.10.

5.12. A small lead ball is dropped onto a floor from different heights (measured in meters). The times (in seconds) when the ball hits the floor are given in the following table.

height	1	2	3	4
time	0.38	0.67	0.76	0.94

From physics we expect that, ignoring air resistance and the diameter of the ball, the relationship between the time y and the height h is $y = a\sqrt{h}$ for some unknown parameter a. Formulate a plausible statistical model for the data and fit a curve of the form $y = a\sqrt{h}$ to the data using the method of least squares.

5.13. In the past a milk vendor found that 30 % of his milk sales were of a low fat variety. Recently, of his 1500 milk sales, 400 were low fat. Is there any indication of a move toward low fat milk? Give the p-value associated with the test.

5.14. Two lakes are being analyzed with respect to their PCB concentration in fish. The PCB concentration from 10 fish from lake A is given by

     11.5   10.8   11.6   9.4   12.4   11.4   12.2   11.0   10.6   10.8

The concentration from eight fish from lake B is given by

          11.8   12.6   12.2   12.5   11.7   12.1   10.4   12.6

a. Assess whether the true sample variances are the same.
b. Assuming equality of variances, infer whether there is any difference in PCB concentration between the fish from the two lakes.

5.15. Let $X_1, \ldots, X_m \sim_{\text{iid}} \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathsf{N}(\mu_Y, \sigma_Y^2)$ be two independent normal samples with $\sigma_X^2 \neq \sigma_Y^2$. Find a pivot variable of the form

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{V(X_1, \ldots, X_m, Y_1, \ldots, Y_n)}$$

that has approximately (for large m and n) a standard normal distribution, and use this pivot to construct an approximate $1 - \alpha$ confidence interval for $\mu_X - \mu_Y$.

5.16. The Australian Bureau of Statistics reports that during 2003, 48300 babies were born in the state of Queensland. Of these, 24800 were boys and 23500 were girls. Does this suggest that the probability of a male birth is more likely than that of a female birth? Conduct a suitable statistical analysis to find this out.

5.17. Gerrit from Gouda is an exporter of cheese. Gerrit requires that his suppliers produce cheese with an expected percentage fat content (PFC) of 40. From past experience it is known that the PFC has a normal distribution with standard deviation 4. Gerrit selects from each new batch of cheese n cheeses at random and measures their fat content. If the average PFC is less than 39, Gerrit rejects the entire batch.

a. Suppose $n = 5$. Give the distribution of the average PFC of the 5 cheeses.
b. Calculate the probability that Gerrit will reject the batch if the expected PFC is in fact 38.5.
c. Suppose the expected PFC is 38. How large should Gerrit choose n such that the test rejects the batch with a probability of at least 90 %?

5.18. In this problem we prove the identity (5.28).   ☞ 148

a. Suppose A is an $m \times m$ invertible matrix and $\mathbf{b}$ is an $m \times 1$ vector. Show that

$$(A - \mathbf{b}\mathbf{b}^\top)^{-1} = A^{-1} + \frac{A^{-1}\mathbf{b}\mathbf{b}^\top A^{-1}}{1 - \mathbf{b}^\top A^{-1}\mathbf{b}} . \tag{5.34}$$

Hint: by direct computation, show that the right-hand side of (5.34) is indeed the inverse of $A - \mathbf{b}\mathbf{b}^\top$.
b. Using (5.34), show that

$$(X_{-k}^\top X_{-k})^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1}\mathbf{x}_k\mathbf{x}_k^\top (X^\top X)^{-1}}{1 - p_k} , \tag{5.35}$$

where X_{-k} is the design matrix X with the kth row removed, $\mathbf{x}_k^\top$ is the kth row of X, and p_k is the kth diagonal element of the projection matrix $P = X(X^\top X)^{-1}X^\top$.
c. Use (5.35) to show (5.28).

5.19. Let $X_1, \ldots, X_n$ be an iid sample from the pdf

$$f(x; \theta) = \frac{\theta}{1 - \theta} x^{(2\theta-1)/(1-\theta)}, \quad x \in (0, 1), \quad \theta \in (\tfrac{1}{2}, 1) .$$

Show that $\{f(x; \theta)\}$ forms a 1-dimensional exponential family. Show that the joint pdf of $X_1, \ldots, X_n$ forms again a 1-dimensional exponential family. Show that $T = \sum_{i=1}^{n} \ln X_i$ is a sufficient statistic for θ.

5.20. Implement a MATLAB program to estimate

$$\ell = \int_0^1 \int_0^1 \frac{\sin(x) \, e^{-(x+y)}}{\ln(1 + x)} \, dx \, dy$$

via Monte Carlo integration, and give a 95 % confidence interval.

5.21. Implement a MATLAB program to estimate $\ell = \int_{-2}^{2} e^{-x^2/2} \, dx = \int H(x) f(x) \, dx$ via Monte Carlo integration using two different approaches: (1) by taking $H(x) = 4 e^{-x^2/2}$ and f the pdf of the $\mathsf{U}[-2, 2]$ distribution and (2) by taking $H(x) = \sqrt{2\pi} \, \mathrm{I}_{\{-2 \leqslant x \leqslant 2\}}$ and f the pdf of the $\mathsf{N}(0, 1)$ distribution.

a. For both cases estimate ℓ via the estimator $\widehat{\ell} = N^{-1} \sum_{i=1}^{N} H(X_i)$. Use a sample size of $N = 1000$.
b. Give an approximate 95 % confidence interval for ℓ for both cases.
c. Using (b), assess how large N should be such that the width of the confidence interval is less than 0.01, and carry out the simulation with this N. Compare the result with the true (numerical) value of ℓ.

5.22. Consider the approximate confidence interval (5.23) for binomial data.
It is possible to calculate the exact coverage probability via total enumeration. Specifically, define

$$T_1(x) = x/n - z_{1-\alpha/2} \sqrt{(x/n) \times (1 - x/n)/n}$$

and

$$T_2(x) = x/n + z_{1-\alpha/2} \sqrt{(x/n) \times (1 - x/n)/n} .$$

Then, the coverage probability as a function of p is

$$\mathbb{P}_p(T_1(X) < p < T_2(X)) = \sum_{x=0}^{n} \mathrm{I}_{\{T_1(x) < p < T_2(x)\}} \binom{n}{x} p^x (1 - p)^{n-x} .$$

For various n and $\alpha = 0.95$ (so that $z_{1-\alpha/2} = 1.96$) draw the graph of the coverage probability as a function of p and comment on the quality of the coverage (which is aimed to be 95 %).

5.23. In order to investigate the effectiveness of "walking exercises" for babies, 24 babies (of the same age and sex) were randomly divided into four groups. Each group followed a different training program. Table 5.5 shows the age (in months) when the infants first walked alone. Implement a one-factor ANOVA model and compute 95 % confidence intervals for the expected walking age in each group. Test whether the training programs have any effect.

Table 5.5 Walking age of babies (in months)

Group			
A	B	C	D
9	11	11.5	13.25
9.5	10	12	11.5
9.75	10	9	12
10	11.75	11.5	13.5
13	10.5	13.25	11.5
9.5	15	13	11.5

5.24. *Rattus Turpis* is a manufacturer of rat poison. The company wants to investigate if adding artificial flavors to their usual mix of cornmeal with strychnine makes their bait more palatable to the rats. They try three artificial flavors, as well as their usual plain bait. Table 5.6 lists the percentages of bait that is eaten for five different surveys. Does the data suggest that adding artificial flavor makes a difference? Use the two-factor ANOVA program in Example 5.15 to investigate this.

Table 5.6 Percentage of bait eaten

	Flavor			
Survey	Plain	Butter	Beef	Bread
1	13.8	11.7	14.0	12.6
2	12.9	16.7	15.5	13.8
3	25.9	29.8	27.8	25.0
4	18.0	23.1	23.0	16.9
5	15.2	20.2	19.9	13.7

Chapter 6
Likelihood

The concept of *likelihood* is central in statistics. It describes in a precise manner the information about the parameters of the model given the observed data.

> **Definition 6.1. (Likelihood Function).** Let $\mathbf{X}$ be a random vector with pdf $f(\cdot; \boldsymbol{\theta})$ (discrete or continuous) with parameter vector $\boldsymbol{\theta} \in \Theta$. For a given outcome $\mathbf{x}$ of $\mathbf{X}$, the function
>
> $$L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$$
>
> is called the **likelihood function** of $\boldsymbol{\theta}$ based on $\mathbf{x}$.

Note that L is a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$, whereas f is a function of $\mathbf{x}$ for fixed $\boldsymbol{\theta}$.

Example 6.1 (Binomial Likelihood). Let $X \sim \mathsf{Bin}(n, p)$. For a given observation x the likelihood of x under p is given by

$$L(p; x) = f(x; p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad 0 < p < 1 . \tag{6.1}$$

As a particular example, consider the experiment where we flip 100 times a biased coin with success probability p. We know that the total number of successes (say, Heads) in 100 tosses, X, has a $\mathsf{Bin}(100, p)$ distribution. Suppose that $x = 43$ successes were observed. Thus, the likelihood of the observed data as a function of p is

$$L(p; 43) = \binom{100}{43} p^{43} (1 - p)^{57}, \quad 0 < p < 1 ,$$

the graph of which is plotted in Fig. 6.1.

Fig. 6.1 The likelihood function for the Bin$(100, p)$ distribution, with 43 observed successes
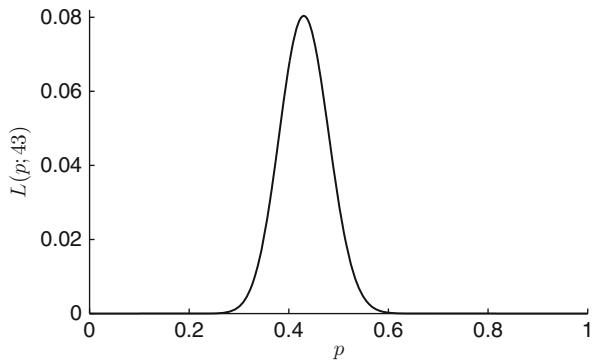
We see that the likelihood is largest for values of p that lie between 0.25 and 0.6. It is very implausible that the current datum was obtained from a p outside this interval. In this sense the likelihood is used to compare the plausibilities of various parameter values.

Example 6.2 (Normal Likelihood). Suppose we are given data $x_1, \ldots, x_n$ from an iid sample $\mathbf{X} = (X_1, \ldots, X_n)$ of the $\mathsf{N}(\mu, \sigma^2)$ distribution, with μ and σ^2 unknown—in this case $\boldsymbol{\theta} = (\mu, \sigma^2)$. The pdf of $\mathbf{X}$ (i.e., the joint pdf of $X_1, \ldots, X_n$) is given by the product of the marginal pdfs; see (3.7). Consequently, the likelihood of the data as a function of the parameters is

$$L(\mu, \sigma^2; \mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2}\right\}$$

for $\mu \in \mathbb{R}$, $\sigma > 0$. As a particular example, suppose $n = 10$ and that the data (computer generated from some $\mathsf{N}(\mu, \sigma^2)$ distribution) are

$$2.39876, \quad -0.149451, \quad -0.770132, \quad 0.87627, \quad -0.0852696,$$
$$1.58494, \quad 1.32772 \quad 1.35611, \quad -0.206479, \quad 0.83773.$$

Figure 6.2 gives the three-dimensional graph and the corresponding contour plot of the likelihood function. Note that the values for $\boldsymbol{\theta}$ for which the likelihood of the data is largest are restricted to an ellipse-like region. The *actual* parameter values for the data were $\mu = 1$ and $\sigma^2 = 1$ in this case.

In general, if $X_1, \ldots, X_n$ is an iid sample from $\mathring{f}(\cdot; \boldsymbol{\theta})$, then the likelihood of the data $\mathbf{x} = (x_1, \ldots, x_n)$ under $\boldsymbol{\theta}$ is the product

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} \mathring{f}(x_i; \boldsymbol{\theta}) . \tag{6.2}$$
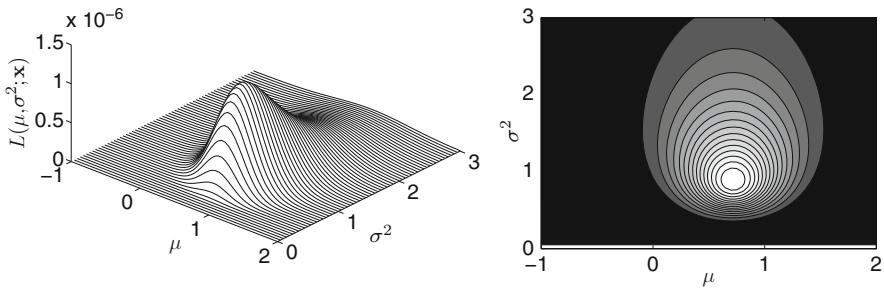
Fig. 6.2 The graph and contour plot of the likelihood function for the $N(\mu, \sigma^2)$ distribution, for the given data

Example 6.3 (Radioactive Source Detection). Suppose a low-intensity radioactive source is emitting particles (in pairs). A screen registers the impact of one particle from each pair. Suppose the position of the source is (a, b) and X is the x-coordinate of the location where a random particle will hit the screen, and let $Y \in (-\pi/2, \pi/2)$ be the angle between the line segments (a, b)-$(a, 0)$ and (a, b)-$(X, 0)$; see Fig. 6.3.
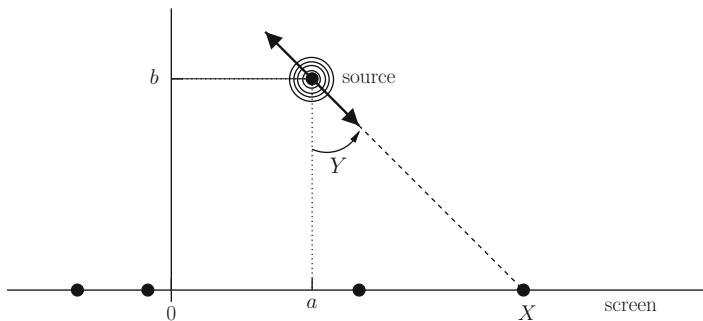
Fig. 6.3 A radioactive source at position (a, b) emits particles in a random direction

Since all angles are equally likely, Y is uniformly distributed in $(-\pi/2, \pi/2)$. Moreover, X and Y are related via $\tan(Y) = \frac{X-a}{b}$. It follows from the transformation formula (3.21) that   ☞ 79

$$f_X(x) = \frac{b}{\pi(b^2 + (x - a)^2)}, \quad x \in \mathbb{R}.$$

In other words, $X = a + bZ$, where Z has a Cauchy distribution; see Problem 6.8.

Suppose that we know that the source is at a distance $b = 1$ from the screen, but we do not know its position a relative to the origin. However, we know the impact positions of 10 particles:

$$1.3615, \quad 3.5616, \quad -14.2411, \quad -4.4950, \quad 2.3014,$$
$$1.1066, \quad -9.3409, \quad 0.3779, \quad 0.9386, \quad -0.1838.$$

Based on these data, what can we say about a? A naive guess is to simply take the mean of the data to estimate the location, which turns out to be -1.8613. This, however, is a fundamentally flawed approach, because the expectation of the distribution of X *does not exist*, namely, $\int_0^\infty x f_X(x) \, dx = \infty$ and $\int_{-\infty}^0 x f_X(x) \, dx = -\infty$, and $\infty - \infty$ is not well defined. Of course the mode of f (the point where f is maximal) is a, but here the mode is not equal to the expectation (which does not exist). A much better approach is to plot the likelihood function for a, which is

$$L(a; \mathbf{x}) = \left(\frac{1}{\pi}\right)^{10} \prod_{i=1}^{10} \frac{1}{1 + (x_i - a)^2} \, .$$

The graph of the likelihood function is given in Fig. 6.4.
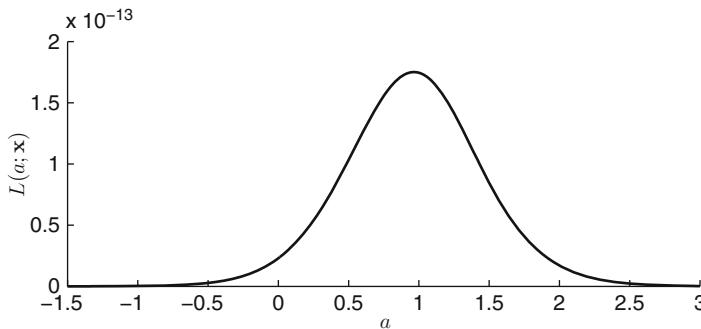
Fig. 6.4 The graph of the likelihood function for the position a of the radioactive source

We see that the most likely position is around 1 and that our initial guess of -1.8613 is extremely unlikely. We also see that the most likely positions fall between roughly -1 and 3. In fact, the actual position was $a = 1$ in this case. So we see that with relatively sparse information we can still make well-founded decisions about a, as long as we use the likelihood.

6.1   Log-Likelihood and Score Functions

Definition 6.2. (Log-Likelihood and Score Functions). Let $\mathbf{X}$ be a random vector with pdf $f(\cdot; \boldsymbol{\theta})$ (discrete or continuous) with parameter vector $\boldsymbol{\theta} \in \Theta$. For a given outcome $\mathbf{x}$ of $\mathbf{X}$, the **log-likelihood function**, denoted l, is the natural logarithm of the likelihood function:

$$l(\boldsymbol{\theta}; \mathbf{x}) = \ln L(\boldsymbol{\theta}; \mathbf{x}) = \ln f(\mathbf{x}; \boldsymbol{\theta}) .$$

Its gradient, denoted $\mathbf{S}$ (column vector), is called the **score function**:

$$\mathbf{S}(\boldsymbol{\theta}; \mathbf{x}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{x}) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}; \boldsymbol{\theta})} . \tag{6.3}$$

Example 6.4 (Binomial Log-Likelihood and Score Functions). For the $\mathrm{Bin}(n, p)$ distribution with observed datum x, the log-likelihood is

$$l(p; x) = \ln\left(\binom{n}{x}\right) + x \ln(p) + (n - x) \ln(1 - p) .$$

Differentiating $l(p; x)$ with respect to p gives the score function

$$S(p; x) = \frac{x}{p} - \frac{n - x}{1 - p} . \tag{6.4}$$

Theorem 6.1. (Log-Likelihood and Score Functions for Iid Data). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from $\overset{\circ}{f}(\cdot; \boldsymbol{\theta})$, and let $\overset{\circ}{l}$ and $\overset{\circ}{\mathbf{S}}$ be, respectively, the log-likelihood and the score function corresponding to $\overset{\circ}{f}$. Then the log-likelihood and score functions of $\boldsymbol{\theta}$ based on an outcome $\mathbf{x}$ of $\mathbf{X}$ are

$$l(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \overset{\circ}{l}(\boldsymbol{\theta}; x_i) \quad \text{and} \quad \mathbf{S}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \overset{\circ}{\mathbf{S}}(\boldsymbol{\theta}; x_i) .$$

Proof. The pdf of $\mathbf{X}$ is $f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \overset{\circ}{f}(x_i; \boldsymbol{\theta})$. Taking the logarithm gives the log-likelihood as the sum of the logarithms of the pdfs. By differentiating this sum we obtain the score function as the sum of the derivatives.   □

Example 6.5 (Log-Likelihood and Score Functions for Normal Iid Data).
Consider the standard model for data: $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$. The log-likelihood function of (μ, σ^2) for a single outcome x is given by the logarithm of the pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution:

$$\mathring{l}(\mu, \sigma^2; x) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2}(x - \mu)^2 / \sigma^2 \ .$$

By differentiating $\mathring{l}$ with respect to μ and σ^2 (note that σ^2 is viewed as a single parameter), we obtain the two components of the score function:

$$\mathring{S}_1(\mu, \sigma^2; x) = \frac{\partial \mathring{l}(\mu, \sigma^2; x)}{\partial \mu} = \frac{x - \mu}{\sigma^2} \ ,$$

and

$$\mathring{S}_2(\mu, \sigma^2; x) = \frac{\partial \mathring{l}(\mu, \sigma^2; x)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2} \frac{(x - \mu)^2}{(\sigma^2)^2} \ .$$

It follows from Theorem 6.1 that the log-likelihood and score functions of (μ, σ^2) based on an outcome $\mathbf{x} = (x_1, \ldots, x_n)$ are given by

$$l(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^{n} \frac{1}{2}(x_i - \mu)^2 / \sigma^2 \ ,$$

$$S_1(\mu, \sigma^2; \mathbf{x}) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \ , \tag{6.5}$$

and

$$S_2(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{(\sigma^2)^2} \ . \tag{6.6}$$

Theorem 6.2. (Score Function for an Exponential Family). The score function for a natural exponential family with pdf $f(\mathbf{x}; \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \, e^{\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{x})} \, h(\mathbf{x})$ is given by

$$\mathbf{S}(\boldsymbol{\theta}; \mathbf{x}) = \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} + \mathbf{t}(\mathbf{x}) \ . \tag{6.7}$$

Proof. The log-likelihood function is $l(\boldsymbol{\theta}; \mathbf{x}) = \ln c(\boldsymbol{\theta}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{x}) + \ln h(\mathbf{x})$. Now take the gradient with respect to $\boldsymbol{\theta}$. □

Example 6.6 (Score Function for Gamma Data). The pdf of the $\mathsf{Gamma}(\alpha, \lambda)$ distribution, where $\alpha, \lambda > 0$, is

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0 .$$

Let us assume that α is known. After the reparameterization $\eta = -\lambda$, we obtain (see Table 5.4) the natural exponential family with pdf

$$\widetilde{f}(x; \eta) = c(\eta) e^{\eta t(x)} h(x) ,$$

where $c(\eta) = (-\eta)^\alpha$ and $t(x) = x$. Here, $h(x)$ does not depend on η (but does depend on the known constant α). Since

$$\frac{c'(\eta)}{c(\eta)} = \frac{\alpha}{\eta} ,$$

we find the score function $\widetilde{S}(\eta; x) = \frac{\alpha}{\eta} + x$. In the original parameter we have (chain rule) $S(\lambda; x) = \widetilde{S}(\eta(\lambda); x) \times \frac{d\eta}{d\lambda} = -\left(\frac{\alpha}{-\lambda} + x\right) = \frac{\alpha}{\lambda} - x$.

6.2 Fisher Information and Cramér–Rao Inequality

Definition 6.3. (Efficient Score). Let $\mathbf{S}(\boldsymbol{\theta}; \mathbf{x})$ be the score function corresponding to an outcome $\mathbf{x}$ of $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$. The *random vector* $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta}; \mathbf{X})$ is called the **efficient score** or simply **score** of $\boldsymbol{\theta}$.

The expected score under $\boldsymbol{\theta}$ is *equal to the zero vector*; namely,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \mathbf{S}(\boldsymbol{\theta}) &= \int \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}; \boldsymbol{\theta})} f(\mathbf{x}; \boldsymbol{\theta}) \, d\mathbf{x} \\ &= \int \nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}) \, d\mathbf{x} = \nabla_{\boldsymbol{\theta}} \int f(\mathbf{x}; \theta) \, d\mathbf{x} = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0} , \end{aligned} \tag{6.8}$$

provided that the interchange of differentiation and integration is justified. This is true for large classes of distributions, including natural exponential families. From now on we simply assume that such an interchange is permitted.

Definition 6.4. (Fisher Information Matrix). For the model $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$, let $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta}; \mathbf{X})$ be the score of $\boldsymbol{\theta}$. The covariance matrix of the random vector $\mathbf{S}(\boldsymbol{\theta})$, denoted by $I(\boldsymbol{\theta})$, is called the **Fisher information matrix**.

Since the expected score is $\mathbf{0}$, we have

$$I(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{S}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})^{\top}] \tag{6.9}$$

and in the one-dimensional case the information number is

$$I(\theta) = \mathbb{E}_{\theta}\left(\frac{\mathrm{d}\ln f(\mathbf{X}; \theta)}{\mathrm{d}\theta}\right)^2 .$$

Example 6.7 (Information Number for Binomial Data). Let $X \sim \mathsf{Bin}(n, p)$. From (6.4) we see that the score is

$$S(p; X) = \frac{X}{p} - \frac{n - X}{1 - p} . \tag{6.10}$$

The information number is therefore

$$\mathrm{Var}_p\left(\frac{X}{p} - \frac{n - X}{1 - p}\right) = \mathrm{Var}_p\left(\frac{X}{p(1 - p)}\right) = \frac{n\, p\, (1 - p)}{p^2\, (1 - p)^2} = \frac{n}{p(1 - p)} . \tag{6.11}$$

For iid samples the score has approximately a multivariate normal distribution that is characterized by the Fisher information of the sampling distribution, as summarized in the following theorem.

Theorem 6.3. (Asymptotic Distribution of the Score). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from $\overset{\circ}{f}(x; \boldsymbol{\theta})$ and let $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta}; \mathbf{X})$ be the score of $\boldsymbol{\theta}$. Then,

1. $\frac{1}{n}\mathbf{S}(\boldsymbol{\theta}) \to \mathbf{0}$ as $n \to \infty$, and
2. $\mathbf{S}(\boldsymbol{\theta}) \overset{\text{approx.}}{\sim} \mathsf{N}(\mathbf{0}, n\,\overset{\circ}{I}(\boldsymbol{\theta}))$ for large n, where $\overset{\circ}{I}(\boldsymbol{\theta})$ is the Fisher information matrix corresponding to $\overset{\circ}{f}$.

Proof. By Theorem 6.1, we can write $\mathbf{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \overset{\circ}{\mathbf{S}}(\boldsymbol{\theta}; X_i)$. Note that the random vectors $\{\overset{\circ}{\mathbf{S}}(\boldsymbol{\theta}; X_i)\}$ are independent and identically distributed with mean $\mathbf{0}$ and

covariance matrix $\overset{\circ}{I}(\boldsymbol{\theta})$. The law of large numbers and the multivariate central limit theorem (see Theorem 3.14) now lead directly to the two properties above.   □

It is sometimes easier to compute the information number in a different way to (6.9), based on the following equality (assuming a one-dimensional parameter θ):

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln f(\mathbf{x}; \theta) = \frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} - \left(\frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} \right)^2 .$$

Multiplying both sides with $f(\mathbf{x}; \theta)$ and integrating with respect to $\mathbf{x}$ gives

$$\mathbb{E}_\theta \frac{\mathrm{d}^2 \ln f(\mathbf{X}; \theta)}{\mathrm{d}\theta^2} = \int \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(\mathbf{x}; \theta) \, \mathrm{d}\mathbf{x} - I(\theta) .$$

Now if we may change the order of differentiation and integration in the integral (allowed for exponential families), then

$$\int \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(\mathbf{x}; \theta) \, \mathrm{d}\mathbf{x} = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} 1 = 0 ,$$

so that the Fisher information number is also given by

$$I(\theta) = -\mathbb{E}_\theta \frac{\mathrm{d}^2 \ln f(\mathbf{X}; \theta)}{\mathrm{d}\theta^2} = -\mathbb{E}_\theta \frac{\mathrm{d}S(\theta; \mathbf{X})}{\mathrm{d}\theta} . \tag{6.12}$$

Example 6.8 (Information Number for Binomial Data Continued). Differentiating the score in (6.10) with respect to p gives

$$\frac{\mathrm{d}S(p; X)}{\mathrm{d}p} = -\frac{X}{p^2} - \frac{n - X}{(1 - p)^2} .$$

The expectation of this random variable (under $X \sim \mathsf{Bin}(n, p)$) is

$$-\frac{np}{p^2} - \frac{n - np}{(1 - p)^2} = -\frac{n}{p(1 - p)} ,$$

which is exactly the negative of the information number found in (6.11).   □

The multidimensional version of (6.12) is

$$I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \, \nabla^2 \ln f(\mathbf{X}; \boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \, \nabla \mathbf{S}(\boldsymbol{\theta}) , \tag{6.13}$$

where $\nabla^2 \ln f(\mathbf{X}; \boldsymbol{\theta})$ is the *Hessian* of $\ln f(\mathbf{X}; \boldsymbol{\theta})$, that is, the (random) matrix

$$\left(\frac{\partial^2 \ln f(\mathbf{X}; \boldsymbol{\theta})}{\partial \theta_i \, \partial \theta_j} \right) = \left(\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_i \, \partial \theta_j} \right) = \left(\frac{\partial S_i(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_j} \right),$$

where S_i denotes the ith component of the score. The following is a direct consequence of Theorem 6.1.

Theorem 6.4. (Information Matrix for Iid Data). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from $\overset{\circ}{f}(x; \boldsymbol{\theta})$, and let $\overset{\circ}{I}(\boldsymbol{\theta})$ be the information matrix corresponding to $X \sim \overset{\circ}{f}(x; \boldsymbol{\theta})$. Then, the information matrix for $\mathbf{X}$ is given by

$$I(\boldsymbol{\theta}) = n \, \overset{\circ}{I}(\boldsymbol{\theta}) .$$

Example 6.9 (Information Matrix for Iid Normal Data). Let $X_1, \ldots, X_n$ be an iid sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Using Example 6.5 and (6.13), we see that the information matrix $\overset{\circ}{I}(\mu, \sigma^2)$ is the expectation of the following matrix of partial derivatives:

$$-\begin{pmatrix} \frac{\partial \overset{\circ}{S}_1(\mu,\sigma^2;X)}{\partial \mu} & \frac{\partial \overset{\circ}{S}_1(\mu,\sigma^2;X)}{\partial \sigma^2} \\ \frac{\partial \overset{\circ}{S}_2(\mu,\sigma^2;X)}{\partial \mu} & \frac{\partial \overset{\circ}{S}_2(\mu,\sigma^2;X)}{\partial \sigma^2} \end{pmatrix} = -\begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{(X-\mu)}{(\sigma^2)^2} \\ -\frac{(X-\mu)}{(\sigma^2)^2} & \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{(\sigma^2)^3} \end{pmatrix}, \tag{6.14}$$

where $X \sim \mathsf{N}(\mu, \sigma^2)$. Taking expectations gives

$$\overset{\circ}{I}(\mu, \sigma^2) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \frac{\sigma^{-4}}{2} \end{pmatrix}. \tag{6.15}$$

By Theorem 6.4 the information matrix corresponding to the whole iid sample is simply a factor n larger: $I(\mu, \sigma^2) = n\overset{\circ}{I}(\mu, \sigma^2)$.

Example 6.10 (Information Matrix for Exponential Families). Consider a natural exponential family with pdf

$$f(\mathbf{x}; \boldsymbol{\theta}) = e^{\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{x}) - \zeta(\boldsymbol{\theta})} h(\mathbf{x}) . \tag{6.16}$$

Then, similar to (6.7),

$$\mathbf{S}(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{t}(\mathbf{x}) - \nabla \zeta(\boldsymbol{\theta}) . \tag{6.17}$$

Since the covariance matrix of a random vector $\mathbf{Z}$ is the same as that of $\mathbf{Z} + \mathbf{a}$ for any constant vector $\mathbf{a}$, we have that the covariance matrix of $\mathbf{S}(\boldsymbol{\theta}; \mathbf{X})$, that is, the information matrix, is simply the covariance matrix of $\mathbf{t}(\mathbf{X})$.

Example 6.11 (Information Number for Location Families). For **location families** $\{f(x; \mu)\}$, that is, when $f(x; \mu) = \widetilde{f}(x - \mu)$ for some fixed pdf $\widetilde{f}$, the Fisher information *does not depend* on μ and is therefore constant. Namely, in this case, the log-likelihood satisfies $l(\mu; x) = \ln \widetilde{f}(x - \mu)$, and the score function is thus a function of $\mu - x$, say $g(x - \mu)$. The variance of the score, that is, the information number, satisfies

$$I(\mu) = \int_{-\infty}^{\infty} S^2(\mu; x) f(x; \mu) \, dx = \int_{-\infty}^{\infty} g^2(x - \mu) \widetilde{f}(x - \mu) \, dx$$
$$= \int_{-\infty}^{\infty} g^2(y) \widetilde{f}(y) \, dy \ ,$$

which does not depend on μ.

The importance of the Fisher information in statistics is corroborated by the famous *Cramér–Rao inequality*.

Theorem 6.5. (Cramér–Rao Information Inequality). Let $\mathbf{X} \sim f(\mathbf{x}; \boldsymbol{\theta})$. The variance of any unbiased estimator $Z = Z(\mathbf{X})$ of $g(\boldsymbol{\theta})$ is bounded from below via

$$\mathrm{Var}(Z) \geqslant (\nabla g(\boldsymbol{\theta}))^{\top} \, I^{-1}(\boldsymbol{\theta}) \, \nabla g(\boldsymbol{\theta}) \ . \qquad (6.18)$$

Proof. We prove only the one-dimensional case. All expectations and variances below are taken with respect to $f(\mathbf{x}; \theta)$. Recall that $S = S(\theta; \mathbf{X}) = \frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta)$ denotes the score and that $\mathrm{Var}(S) = I(\theta)$. The key is to apply the Cauchy–Schwarz inequality:

$$\mathrm{Cov}(Z, S) \leqslant \sqrt{\mathrm{Var}(Z) \mathrm{Var}(S)} \ ,$$

which immediately yields

$$\mathrm{Var}(Z) \geqslant \frac{(\mathrm{Cov}(Z, S))^2}{I(\theta)} \ .$$

Thus, it remains to be shown that $\mathrm{Cov}(Z, S) = g'(\theta)$. This follows from $\mathrm{Cov}(Z, S) = \mathbb{E}[ZS] - \mathbb{E}Z \, \mathbb{E}S = \mathbb{E}ZS$ (because $\mathbb{E}S = 0$) and

$$\mathbb{E}[ZS] = \mathbb{E}\left[Z\, \frac{\frac{\partial}{\partial\theta} f(\mathbf{X};\theta)}{f(\mathbf{X};\theta)} \right] = \int Z(\mathbf{x}) \frac{\partial}{\partial\theta} f(\mathbf{x};\theta)\,\mathrm{d}\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}\theta} \mathbb{E}Z = g'(\theta)\,,$$

assuming that we may change the order of integration and differentiation.   □

6.3   Likelihood Methods for Estimation

Suppose we are given data $\mathbf{x}$ from a model $f(\mathbf{x};\boldsymbol{\theta})$, yielding the likelihood function $L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})$. Although the entire shape of the likelihood function is valuable for our inference about the unknown parameter $\boldsymbol{\theta}$, it is often desirable to summarize the information on the likelihood function into a few key numbers. One of these numbers is the **mode** of the likelihood function, that is, the parameter value $\widehat{\boldsymbol{\theta}}$ for which the function is maximal. This number (or vector of numbers, in the multiparameter case) is in a way our best estimate for $\boldsymbol{\theta}$. It is called the **maximum likelihood estimate** (MLE). Note that $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ is a function of the data $\mathbf{x}$. The corresponding random variable, also denoted $\widehat{\boldsymbol{\theta}}$, is the **maximum likelihood estimator** (also abbreviated as MLE).

Since the natural logarithm is an increasing function, the maximization of $L(\boldsymbol{\theta};\mathbf{x})$ is equivalent (in terms of finding the mode) to the maximization of the log-likelihood $l(\boldsymbol{\theta};\mathbf{x})$. This is often easier, especially when $\mathbf{X}$ is an iid sample from some sampling distribution.

Remark 6.1 (Existence and Uniqueness). MLEs may not always exist (e.g., when estimating a variance with only one data point) or could be non-unique (when the likelihood function attains its maximum at more than one point).

If $l(\boldsymbol{\theta};\mathbf{x})$ is a differentiable function with respect to $\boldsymbol{\theta}$ and the maximum is attained in the *interior* of Θ *and there exists a unique maximum point*, then we can find the MLE of $\boldsymbol{\theta}$ by differentiating $l(\boldsymbol{\theta};\mathbf{x})$ with respect to $\boldsymbol{\theta}$—more precisely, by solving

$$\nabla_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta};\mathbf{x}) = \mathbf{0}\,.$$

In other words, the MLE is obtained by solving the root of the score function, that is, by solving

$$\mathbf{S}(\boldsymbol{\theta};\mathbf{x}) = \mathbf{0}\,. \tag{6.19}$$

In general, solving the above equation only yields a *local* maximum. If the likelihood function is multimodal, there will be more than one point $\boldsymbol{\theta}$ that satisfies (6.19). The evaluation of l at *all* of these points may then identify the *global* maximum.

Example 6.12 (MLE for Binomial Data). Suppose x is an outcome of $X \sim$ Bin(n, p). By (6.4) the MLE is found by solving ☞ 165

$$\frac{x}{p} - \frac{n - x}{1 - p} = 0 \, ,$$

which gives the MLE $\widehat{p} = x/n$ and the corresponding estimator $\widehat{p} = X/n$.

Example 6.13 (MLE for Iid Normal Data). Suppose $x_1, \ldots, x_n$ are the outcomes of $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$. The MLEs follow (see (6.5) and (6.6)) from solving the set of equations

$$\sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0 \, , \tag{6.20}$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{(\sigma^2)^2} = 0 \, , \tag{6.21}$$

giving

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \qquad \text{and} \qquad \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \, . \tag{6.22}$$

We see that the maximum likelihood method and the method of moments yield exactly the same estimates in this case. ☞ 123

Example 6.14 (MLE for the Normal Linear Model). Consider the normal linear model ☞ 115

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \, , \tag{6.23}$$

where X is an $n \times m$ design matrix, $\boldsymbol{\beta}$ an m-dimensional vector of parameters, and $\boldsymbol{\varepsilon}$ a vector of iid $\mathsf{N}(0, \sigma^2)$ error terms. Since $\mathbf{Y} \sim \mathsf{N}(X\boldsymbol{\beta}, \sigma^2 I)$, it follows from (3.30) ☞ 83 that the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 / \sigma^2}$$

for a given outcome $\mathbf{y}$ of $\mathbf{Y}$. Observe that for any *fixed* σ^2 the likelihood $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$, as a function of $\boldsymbol{\beta}$, is maximized by choosing $\boldsymbol{\beta}$ such that $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$ is minimized. But this gives exactly the least-squares estimate of $\boldsymbol{\beta}$; see Sect. 5.1.2. To obtain the ☞ 125 MLE for σ^2 it remains to maximize $L(\widehat{\boldsymbol{\beta}}, \sigma^2; \mathbf{y})$ or, equivalently, solve

$$-\frac{n}{2\sigma^2} + \frac{1}{2} \frac{\|\mathbf{y} - X\widehat{\boldsymbol{\beta}}\|^2}{(\sigma^2)^2} = 0 \, ,$$

where $\widehat{\boldsymbol{\beta}}$ is MLE of $\boldsymbol{\beta}$. This gives the same estimate $\widehat{\sigma^2} = \|\mathbf{y} - X\widehat{\boldsymbol{\beta}}\|^2/n$ as in (5.12). For a generalization to the general regression case (possibly nonlinear), see Problem 6.3.

☞ 127

Example 6.15 (MLE for Exponential Families). For natural exponential families of the form (6.16) the MLE is found by solving

$$\mathbf{t}(\mathbf{x}) - \nabla \zeta(\boldsymbol{\theta}) = \mathbf{t}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, \mathbf{t}(\mathbf{X}) = \mathbf{0} \;, \tag{6.24}$$

where we have used the fact that $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{t}(\mathbf{X}) - \nabla \zeta(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}\mathbf{S}(\boldsymbol{\theta};\mathbf{X}) = \mathbf{0}$; see (6.8). Thus, $\boldsymbol{\theta}$ is chosen such that the observed and expected values of $\mathbf{t}(\mathbf{X})$ are matched.

☞ 167

6.3.1   Score Intervals

The score function is not onlyvaluable for finding point estimates, but it can alsobe used to construct confidence intervals. The key observation here is that for large iid samples the score is approximately normally distributed; see Theorem 6.3. Let us concentrate on the one-dimensional case; that is, θ is real-valued.

☞ 128
☞ 168

Let $\mathbf{X} = (X_1, \ldots, X_n) \sim_{\text{iid}} \overset{\circ}{f}(\cdot; \theta)$ and let $S(\theta; \mathbf{X})$ denote the score. By Theorem 6.3, the pivot variable $S(\theta; \mathbf{X})(n\overset{\circ}{I}(\theta))^{-1/2}$ has approximately a standard normal distribution, and hence

$$\left\{ \theta : -z_{1-\alpha/2} < \frac{S(\theta; \mathbf{X})}{\sqrt{n\overset{\circ}{I}(\theta)}} < z_{1-\alpha/2} \right\}$$

is an approximate $1 - \alpha$ stochastic **confidence set**. We use here "set" instead of "interval" because this set need not be an interval in general.

Example 6.16 (Score Interval for Iid Bernoulli Data). Let $\mathbf{X}$ be an iid sample from $\mathrm{Ber}(p)$. Since the Bernoulli distribution is a special case of the binomial distribution, we can use (6.10) in combination with Theorem 6.1 to find the score $S(p; \mathbf{X}) = \sum_{i=1}^{n}(X_i - p)/(p(1 - p)) = n(\bar{X} - p)/(p(1 - p))$. By a similar reasoning we find the information number $I(p) = n/(p(1 - p))$. So the confidence set becomes

$$\left\{ p : -z_{1-\alpha/2} < \frac{n(\bar{X} - p)}{p(1 - p)} \times \sqrt{\frac{p(1 - p)}{n}} < z_{1-\alpha/2} \right\}$$

$$= \left\{ p : -a < \frac{\bar{X} - p}{\sqrt{p(1 - p)/n}} < a \right\} \;,$$

where we have abbreviated $z_{1-\alpha/2}$ to a. By solving with respect to p the quadratic equation

$$(\bar{X} - p)^2 = a^2 \, p(1 - p)/n \,,$$

this confidence set can be written as the interval $\{T_1 < p < T_2\}$ with

$$T_1 = \frac{a^2 + 2n\bar{X} - a\sqrt{a^2 - 4n(\bar{X} - 1)\bar{X}}}{2\,(a^2 + n)}$$

$$T_2 = \frac{a^2 + 2n\bar{X} + a\sqrt{a^2 - 4n(\bar{X} - 1)\bar{X}}}{2\,(a^2 + n)} \,.$$

This **score interval** has much better coverage behavior than the "standard" confidence interval (5.23) over the complete range of p.   ☞ 136

6.3.2   Properties of the ML Estimator

An important property of the MLE is that it is *invariant* under transformations.

> **Theorem 6.6.   (Invariance of the MLE).** Suppose $\mathbf{X} \sim f(\mathbf{x}; \boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$ and let $\mathbf{g}$ be an invertible function. Then the MLE of $\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\theta})$ is $\widehat{\boldsymbol{\eta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$.

Proof. Let $L(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta})$ be the likelihood function, and let $\widetilde{L}(\boldsymbol{\eta}) = L(\mathbf{g}^{-1}(\boldsymbol{\eta}))$ be the reparameterized likelihood function. The MLE of $\boldsymbol{\eta}$ is, by definition, the number $\widehat{\boldsymbol{\eta}}$ for which $\widetilde{L}(\widehat{\boldsymbol{\eta}})$ is maximal. Since L is maximal for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, the function $L(\mathbf{g}^{-1}(\boldsymbol{\eta}))$ is maximal at $\widehat{\boldsymbol{\eta}}$ for which $\mathbf{g}^{-1}(\widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\theta}}$, which gives $\widehat{\boldsymbol{\eta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$.   □

Remark 6.2. If $\mathbf{g}$ is not invertible, then we can still *define* the MLE of η as $\widehat{\boldsymbol{\eta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$. In effect, this amounts to defining $\widetilde{L}(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}:\mathbf{g}(\boldsymbol{\theta})=\eta} L(\boldsymbol{\theta}; \mathbf{x})$.

Next, we consider the case where $\mathbf{X} = (X_1, \ldots, X_n)$ is an iid sample from some pdf $\overset{\circ}{f}(x; \boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}$ be the ML estimator of $\boldsymbol{\theta}$. The random variable $\widehat{\boldsymbol{\theta}}$ has some nice asymptotic properties.

Theorem 6.7. (Consistency of the MLE). The ML estimator $\widehat{\boldsymbol{\theta}}$ is **consistent**. That is, with probability tending to 1 as $n \to \infty$, the likelihood equation has a root $\widehat{\boldsymbol{\theta}}$ such that for *all* $\varepsilon > 0$

$$\mathbb{P}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > \varepsilon) \to 0 \ .$$

Proof. (Sketch.) Let C_a be a sphere with radius a centered at the true parameter $\boldsymbol{\theta}$. We want to show that for sufficiently small a the probability tends to 1 that

$$l(\boldsymbol{\theta}) > l(\widetilde{\boldsymbol{\theta}})$$

at all points $\widetilde{\boldsymbol{\theta}}$ on the surface of C_a. This can be established as follows. A second-order Taylor expansion of $l(\boldsymbol{\theta})$ around $\boldsymbol{\theta}$, divided by n, yields

☞ 369

$$\frac{1}{n}(l(\widetilde{\boldsymbol{\theta}}; \mathbf{X}) - l(\boldsymbol{\theta}; \mathbf{X})) =$$

$$\frac{1}{n}\mathbf{S}(\boldsymbol{\theta}; \mathbf{X})^{\top}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{n}\frac{1}{2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} H(\boldsymbol{\theta}; \mathbf{X})(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{n} R_n \ , \qquad (6.25)$$

where $\mathbf{S}(\boldsymbol{\theta}; \mathbf{X})$ is the gradient of l (i.e., the score), $H(\boldsymbol{\theta}; \mathbf{X})$ is the Hessian matrix of l (i.e., the matrix of partial derivatives $(\partial^2 l/\partial\theta_i \partial\theta_j)$), and R_n is a random remainder term. By Theorem 6.3, $\mathbf{S}(\boldsymbol{\theta}; \mathbf{X})/n$ converges to the zero vector. Similarly, by Theorem 6.1, $H(\boldsymbol{\theta}; \mathbf{X})$ can be written as the iid sum $\sum_{k=1}^{n} \overset{\circ}{H}(\boldsymbol{\theta}; X_k)$, where $\overset{\circ}{H}(\boldsymbol{\theta}; X_k)$ denotes the matrix of partial derivatives $(\partial^2\overset{\circ}{l}(\boldsymbol{\theta}; X_k)/\partial\theta_i \partial\theta_j)$. Hence, by the law of large numbers and (6.13),

☞ 168
☞ 165

$$\frac{1}{n} H(\boldsymbol{\theta}; \mathbf{X}) \to \mathbb{E}_{\boldsymbol{\theta}} \overset{\circ}{H}(\boldsymbol{\theta}; X) = -\overset{\circ}{I}(\boldsymbol{\theta}) \qquad (6.26)$$

as $n \to \infty$, where $\overset{\circ}{I}$ is the information matrix corresponding to $\overset{\circ}{f}$. Thus, the first and second terms in (6.25) converge to 0 and $-\frac{1}{2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} \overset{\circ}{I}(\boldsymbol{\theta})(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, respectively, as $n \to \infty$. Since the information matrix is positive definite (i.e., $\mathbf{w}^{\top}\overset{\circ}{I}(\boldsymbol{\theta})\mathbf{w} > 0$ for any vector $\mathbf{w}$), the second term is strictly negative. If the remainder term, which depends on the third derivative of l, can be bounded in norm by a constant times a^3/n, then with probability tending to 1 the right-hand side will be less than 0 for a small enough, proving the assertion that $l(\boldsymbol{\theta}) > l(\widetilde{\boldsymbol{\theta}})$ on the surface of the sphere C_a. From this we can conclude that with probability tending to 1 there must be an MLE $\widehat{\boldsymbol{\theta}}$ that lies inside C_a. For a sequence of $a_n \to 0$ we can thus find a sequence of $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$, showing the consistency of the estimator. □

Note that the above theorem only says that there *exists* a sequence of MLEs $\{\widehat{\boldsymbol{\theta}}_n\}$ that converge (in probability) to the true $\boldsymbol{\theta}$. When there are multiple local maxima, a particular sequence $\widehat{\boldsymbol{\theta}}_n$ may in fact converge to a local maximum.

Theorem 6.8. (Asymptotic Distribution of the MLE). Suppose that $\{\widehat{\boldsymbol{\theta}}_n\}$ is a sequence of consistent ML estimators for $\boldsymbol{\theta}$. Then, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges in distribution to a $\mathsf{N}(\boldsymbol{\theta}, \mathring{I}^{-1}(\boldsymbol{\theta}))$-distributed random vector as $n \to \infty$. In other words,

$$\widehat{\boldsymbol{\theta}}_n \overset{\text{approx.}}{\sim} \mathsf{N}(\boldsymbol{\theta}, \mathring{I}^{-1}(\boldsymbol{\theta})/n) .$$

Proof. A sketch of the proof for the one-dimensional case (thus, $\boldsymbol{\theta} = \theta$ is a scalar) is as follows. The key idea is again to take a Taylor expansion, this time a Taylor expansion of $l'(\widehat{\theta}_n)$ around θ:

$$l'(\widehat{\theta}_n) = l'(\theta) + (\widehat{\theta}_n - \theta)l''(\theta) + \frac{1}{2}(\widehat{\theta}_n - \theta)^2 l'''(\theta^*) ,$$

where θ^* lies between θ and $\widehat{\theta}_n$. Since $l'(\widehat{\theta}_n) = 0$ (by definition), it follows that

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{l'(\theta)/\sqrt{n}}{-l''(\theta)/n - (\widehat{\theta}_n - \theta)l'''(\theta^*)/(2n)} . \tag{6.27}$$

The numerator in the right-hand side of (6.27) is $S(\theta; \mathbf{X})/\sqrt{n}$, which by Theorem 6.3 has approximately a $\mathsf{N}(0, \mathring{I}(\theta))$ distribution for large n. The first term of the denominator is $-\frac{1}{n}\sum_{i=1}^{n} \mathring{H}(\theta; X_i)$, which by the law of large numbers converges to $\mathring{I}(\theta)$ (see (6.26)). The second term of the denominator goes to 0 by the consistency property. This shows that either side of (6.27) is approximately $\mathsf{N}(\theta, \mathring{I}^{-1}(\theta))$ distributed. $\qquad\square$

Example 6.17 (Asymptotic Distribution of the Binomial MLE). Let us check if this theorem makes sense for the case where $X_1, \ldots, X_n$ are iid and $\mathsf{Ber}(p)$ distributed. Here the MLE is $\bar{X}$ and the information number for $\mathsf{Ber}(p)$ is $1/(p(1-p))$ (see (6.11) with $n = 1$). Theorem 6.8 states that for large n

$$\bar{X} \overset{\text{approx.}}{\sim} \mathsf{N}\left(p, \frac{p(1-p)}{n}\right),$$

which follows also directly from the normal approximation to the binomial distribution by noting that $n\bar{X} \sim \mathsf{Bin}(n, p)$.

6.4   Likelihood Methods in Statistical Tests

The likelihood does not only provide a systematic way of defining good estimators (via the maximum likelihood principle), but it also yields a systematic way of finding test statistics.

Let $X_1, \ldots, X_n$ be an iid sample from a distribution with unknown parameter $\boldsymbol{\theta}$. Write $\mathbf{X}$ for the corresponding random vector, and let $L(\boldsymbol{\theta}; \mathbf{x})$ be the likelihood function for a given outcome $\mathbf{x}$ of $\mathbf{X}$. Let Θ be the set of possible values for $\boldsymbol{\theta}$. Suppose Θ_0 and Θ_1 are two nonoverlapping subsets of Θ such that $\Theta_0 \cup \Theta_1 = \Theta$. Consider the following hypotheses:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \,,$$

$$H_1 : \boldsymbol{\theta} \in \Theta_1 \,.$$

Definition 6.5. (Generalized Likelihood Ratio). The **generalized likelihood ratio** is defined as the number

$$\lambda = \frac{M_0(\mathbf{x})}{M(\mathbf{x})} \stackrel{\text{def}}{=} \frac{\max_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}; \mathbf{x})}{\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})} \,.$$

Note that $M(\mathbf{x}) = L(\widehat{\boldsymbol{\theta}}; \mathbf{x})$, where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$. Let Λ denote the random generalized likelihood ratio obtained by substituting $\mathbf{X}$ for $\mathbf{x}$. We can use Λ as a test statistic for testing the above hypotheses. The general principle is to *reject* H_0 if Λ is *too small* (left one-sided test). To determine the corresponding p-value $\mathbb{P}(\Lambda \leqslant \lambda)$, we need to know the distribution of Λ under H_0. This is in general a difficult task. However, it is sometimes possible to derive the distribution of a *function* of Λ under H_0, which is then taken as an equivalent test statistic. The new rejection region is no longer necessarily left one-sided.

Example 6.18 (Generalized Likelihood Ratio Test for Iid Normal Data). Suppose $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$, with μ and σ^2 unknown. We wish to test

$$H_0 : \mu = \mu_0 \,,$$

$$H_1 : \mu \neq \mu_0 \,.$$

Hence, $\Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 > 0\}$. The random likelihood function is given by

$$L(\mu, \sigma^2; \mathbf{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}\right) \,.$$

Maximizing L (or $\ln L$) over Θ gives the MLE $(\widehat{\mu}, \widehat{\sigma^2}) = (\bar{X}, \sum_{i=1}^{n}(X_i - \bar{X})^2/n)$. Hence, $M(\mathbf{X}) = L(\widehat{\mu}, \widehat{\sigma^2}; \mathbf{X})$. Optimizing L over Θ_0 gives the estimator $(\mu_0, \widetilde{\sigma^2})$, with

$$\widetilde{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2 .$$

Consequently,

$$\Lambda = \frac{L(\mu_0, \widetilde{\sigma^2}; \mathbf{X})}{L(\widehat{\mu}, \widehat{\sigma^2}; \mathbf{X})} = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \mu_0)^2} \right)^{n/2} = \left(1 + \frac{1}{n-1} T^2 \right)^{-n/2} ,$$

where $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ and S is the sample standard deviation. Rejecting H_0 for small values of Λ is equivalent to rejecting H_0 for large values of $|T|$. By (5.17), T has a t_{n-1} distribution under H_0.

Theorem 6.9. (Asymptotic Distribution of the Generalized Likelihood Ratio). For a k-dimensional parameter space (thus, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$), if the null hypothesis has only one value $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and the alternative hypothesis is $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then for large n (under some mild regularity conditions, which are satisfied for exponential families),

$$-2 \ln \Lambda \overset{\text{approx.}}{\sim} \chi_k^2 .$$

Proof. (Sketch.) This is again an exercise in Taylor expansions. Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$ and let $l(\boldsymbol{\theta})$ be the log-likelihood function. Under H_0

$$-2 \ln \Lambda = -2(l(\boldsymbol{\theta}_0) - l(\widehat{\boldsymbol{\theta}})) .$$

A second-order Taylor expansion at $\boldsymbol{\theta}_0$ around $\widehat{\boldsymbol{\theta}}$ gives

$$l(\boldsymbol{\theta}_0) = l(\widehat{\boldsymbol{\theta}}) + (\nabla l(\widehat{\boldsymbol{\theta}}))^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \nabla^2 l(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathcal{O}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^3) .$$

Because $\nabla l(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\nabla^2 l(\widehat{\boldsymbol{\theta}}) \approx -I(\boldsymbol{\theta}_0)$, where I is the information matrix (of dimension k), we have

$$-2 \ln \Lambda \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) .$$

By Theorem 6.8, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ has approximately a $\mathsf{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$ distribution. Thus, for a large sample size, we have that $-2\ln\Lambda$ is approximately distributed as $\mathbf{X}^\top I(\boldsymbol{\theta}_0)\mathbf{X}$ with $\mathbf{X} \sim \mathsf{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$. From Theorem 3.9 it follows now that $-2\ln\Lambda$ has approximately a χ_k^2 distribution. $\qquad\qquad\qquad\square$

6.5   Newton–Raphson Method

Recall that likelihood maximization often involves solving $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta};\mathbf{x}) = \mathbf{0}$, where $\mathbf{S}(\boldsymbol{\theta})$ is the score function and $\boldsymbol{\theta}$ a k-dimensional parameter vector. The MLE $\widehat{\boldsymbol{\theta}}$ is the solution to this equation; that is, it is the root of $\mathbf{S}(\boldsymbol{\theta})$. It is often not possible to find $\widehat{\boldsymbol{\theta}}$ in an explicit form. In that case one needs to solve the equation $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}$ *numerically*. There exist many standard techniques for root-finding. A well-known method is the **Newton–Raphson** procedure. This is an iterative procedure where, starting from a guess $\boldsymbol{\theta}$, a "better" guess is obtained by approximating the score via a linear function. More precisely, suppose that $\boldsymbol{\theta}$ is our initial guess for $\widehat{\boldsymbol{\theta}}$ (the root of $\mathbf{S}$). If $\widehat{\boldsymbol{\theta}}$ is reasonably close to $\boldsymbol{\theta}$, a first-order Taylor approximation of $S_i = \partial l/\partial\theta_i$ around $\boldsymbol{\theta}$ gives

$$S_i(\widehat{\boldsymbol{\theta}}) \approx S_i(\boldsymbol{\theta}) + [\nabla S_i(\boldsymbol{\theta})]^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad i = 1,\ldots,k ,$$

or, in matrix notation, $\mathbf{S}(\widehat{\boldsymbol{\theta}}) \approx \mathbf{S}(\boldsymbol{\theta}) + H(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where H is the Hessian of the log-likelihood, that is, the matrix of second-order partial derivatives of l. Since $\mathbf{S}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ by definition, we have

$$\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - H^{-1}(\boldsymbol{\theta})\,\mathbf{S}(\boldsymbol{\theta}) .$$

This suggests the following **Newton–Raphson** recursion for finding successively better guesses $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ converging to $\widehat{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - H^{-1}(\boldsymbol{\theta}_t)\,\mathbf{S}(\boldsymbol{\theta}_t) . \tag{6.28}$$

The sequence of successive values is guaranteed to converge to the actual root, provided the function is smooth enough (e.g., has continuous second-order derivatives) and the initial guess is close enough to the root.

Notice that $H(\boldsymbol{\theta}) = H(\boldsymbol{\theta};\mathbf{x})$ depends on the parameter $\boldsymbol{\theta}$ and the data $\mathbf{x}$ and may be quite complicated. On the other hand, the expectation of $H(\boldsymbol{\theta};\mathbf{X})$ under $\boldsymbol{\theta}$ is simply the negative of information matrix $I(\boldsymbol{\theta})$, which does not depend on the data. This suggests the alternative iterative scheme, called **Fisher's scoring method**:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + I^{-1}(\boldsymbol{\theta}_t)\,\mathbf{S}(\boldsymbol{\theta}_t) , \tag{6.29}$$

which may be much easier to implement if the information matrix can be readily evaluated.

Example 6.19 (MLE for Iid Normal Data). Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ is the outcome of an iid sample from the $N(\mu, \sigma^2)$ distribution (both parameters unknown). The score function is given in (6.5)–(6.6). From (6.14) we find that the Hessian matrix is given by

$$H(\mu, \sigma^2; \mathbf{x}) = \sum_{i=1}^{n} \mathring{H}(\mu, \sigma^2; x_i) = \sum_{i=1}^{n} \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{x_i - \mu}{(\sigma^2)^2} \\ -\frac{x_i - \mu}{(\sigma^2)^2} & \frac{1}{2\sigma^4} - \frac{(x_i - \mu)^2}{(\sigma^2)^3} \end{pmatrix},$$

where $\mathring{H}(\mu, \sigma^2; x)$ is the Hessian for the 1-dimensional case. Apart from a starting value, this is all that is required to carry out the Newton–Raphson iteration (6.28). It is easier, however, to apply the recursion (6.29), using the exact expression for the information matrix (see (6.15)):

$$I^{-1}(\mu, \sigma^2) = \left(n \mathring{I}(\mu, \sigma^2) \right)^{-1} = \frac{1}{n} \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

It follows that the Fisher scoring procedure (6.29) involves the following iterative steps:

$$\mu_{t+1} = \mu_t + \frac{1}{n} \sigma_t^2 \sum_{i=1}^{n} \frac{(x_i - \mu_t)}{\sigma_t^2} = \mu_t + \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_t) = \bar{x},$$

$$\sigma_{t+1}^2 = \sigma_t^2 + \frac{2}{n} \sigma_t^4 \left(\frac{-n}{2\sigma_t^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_t)^2}{\sigma_t^4} \right) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_t)^2.$$

Note that, starting from any initial guess, after only two steps we get $\mu_t = \bar{x}$ and $\sigma_t^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, which are the MLEs for μ and σ^2.

Example 6.20 (MLE for the Radioactive Source Detection Example). Let us return to Example 6.3. Suppose we want to find the most likely estimate for the position a of the source. The log-likelihood function is

$$l(a; \mathbf{x}) = -n \ln \pi - \sum_{i=1}^{n} \ln(1 + (x_i - a)^2).$$

Taking the derivative with respect to a gives the score function

$$S(a; \mathbf{x}) = \sum_{i=1}^{n} \frac{2(x_i - a)}{1 + (x_i - a)^2}.$$

The information number is of form $I(a) = n\mathring{I}(a)$, where $\mathring{I}$ is the information number for a single sample. Specifically,

$$\mathring{I}(a) = \mathbb{E}_a S^2(a; X) = \int_{-\infty}^{\infty} \frac{4(x-a)^2}{(1+(x-a)^2)^2} \frac{1}{\pi(1+(x-a)^2)} \, dx$$

$$= \int_{-\infty}^{\infty} \frac{4y^2}{\pi(1+y^2)^3} \, dy \qquad \text{(change of variable } y = x - a\text{)}$$

$$= \frac{1}{2} \, .$$

Thus the information number is constant; this is in agreement with the fact that we are dealing here with a location family of distributions; see Example 6.11. Now (6.29) leads to the scheme

$$a_{t+1} = a_t + \frac{2}{n} \sum_{i=1}^{n} \frac{2(x_i - a_t)}{1 + (x_i - a_t)^2} \, .$$

This is implemented in the following MATLAB code.

```
%lighthousemle.m
x = [1.3615,3.5616,-14.2411,-4.4950,2.3014,1.1066,...
     -9.3409, 0.3779, 0.9386,-0.1838]; %the data
a = 2; %initial guess
n = 10;
for i=1:7
  display(a)
  a = a+4*sum((x-a)./(1 + (x-a).^2))/n; %note vectorization!
end
```

This gives the following succession of values, 2, 1.2626, 0.9536, 0.9654, 0.9648, 0.9648, and 0.9648, so that the MLE is $\hat{a} = 0.9648$, which is remarkably close to the true value $a = 1$.

6.6 Expectation–Maximization (EM) Algorithm

Another useful numerical method for likelihood maximization is the **expectation–maximization** (EM) algorithm.

Suppose that, for a given vector of observations $\mathbf{x} = (x_1, \ldots, x_n)$, we wish to compute the MLE:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, L(\boldsymbol{\theta}; \mathbf{x}) \,, \tag{6.30}$$

where $L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$ is the likelihood function.

One could use a root-finding routine, such as the Newton–Raphson method described in Sect. 6.5, to obtain $\widehat{\boldsymbol{\theta}}$. However, for many problems, computing the score function and the Hessian matrix analytically—required by the Newton–Raphson method—might be difficult. Instead of maximizing the likelihood function directly, the EM algorithm augments the data $\mathbf{x}$ with a suitable vector of *latent* (or hidden) variables $\mathbf{z}$ such that

$$f(\mathbf{x}; \boldsymbol{\theta}) = \int \widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \,.$$

The function of $\boldsymbol{\theta}$

$$\widetilde{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$$

is usually referred to as the **complete-data likelihood** function. The main advantage of the data augmentation step is that it is often possible to introduce latent variables $\mathbf{z}$ in such a way that the maximization of the complete-data likelihood $\widetilde{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ or log-likelihood $\widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \ln \widetilde{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ is much easier than maximizing the original likelihood $L(\boldsymbol{\theta}; \mathbf{x})$ or log-likelihood $l(\boldsymbol{\theta}; \mathbf{x}) = \ln L(\boldsymbol{\theta}; \mathbf{x})$.

Of course, the latent variables $\mathbf{z}$ are not observed, and neither $\widetilde{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ nor $\widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ are available. One feasible approach is to replace it with the expectation $\mathbb{E}_g \widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z})$ with respect to a suitable density $g(\mathbf{z})$. It can be shown (see Problem 6.20) that for all $\boldsymbol{\theta}$ and any density g,

$$\ln f(\mathbf{x}; \boldsymbol{\theta}) \geqslant \mathcal{L}(g, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \int g(\mathbf{z}) \ln \left(\frac{\widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{g(\mathbf{z})} \right) \mathrm{d}\mathbf{z}$$
$$= \mathbb{E}_g \widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z}) - \mathbb{E}_g \ln g(\mathbf{Z}) \,. \tag{6.31}$$

That is, $\mathcal{L}(g, \boldsymbol{\theta})$ is a lower bound for the log-likelihood $l(\boldsymbol{\theta}; \mathbf{x})$. In addition, this lower bound is attained (see Problem 6.20) for

$$g(\mathbf{z}) = \widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{\int \widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \mathrm{d}\mathbf{z}} \,. \tag{6.32}$$

That is, the lower bound is attained when $g(\mathbf{z})$ is taken as the conditional pdf of the latent data $\mathbf{Z}$ given the observed data $\mathbf{X} = \mathbf{x}$.

Algorithm 6.1. (EM Algorithm). Suppose $\boldsymbol{\theta}_0$ is an initial guess for the maximizer. The EM algorithm consists of iterating the following steps for $t = 1, 2, \ldots$:

1. **Expectation Step (E-Step)**: Given the current vector $\boldsymbol{\theta}_{t-1}$, maximize $\mathcal{L}(g, \boldsymbol{\theta}_{t-1})$ as a function of g. It follows from (6.32) that the exact solution is

$$g_t(\mathbf{z}) \stackrel{\text{def}}{=} \widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}_{t-1}) \ .$$

   Compute the expected log-likelihood under g_t:

$$Q_t(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{g_t} \widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z}) \ . \tag{6.33}$$

2. **Maximization Step (M-Step)**: Maximize $\mathcal{L}(g_t, \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. Since $\mathcal{L}(g_t, \boldsymbol{\theta}) = Q_t(\boldsymbol{\theta}) - \mathbb{E}_{g_t} \ln g_t(\mathbf{Z})$, this is equivalent to finding

$$\boldsymbol{\theta}_t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, Q_t(\boldsymbol{\theta}) \ .$$

3. **Stopping Condition**: If, for example, $|l(\boldsymbol{\theta}_t; \mathbf{x}) - l(\boldsymbol{\theta}_{t-1}; \mathbf{x})| \leqslant \varepsilon$ for some small tolerance ε, terminate the algorithm.

A direct consequence of the EM algorithm is that the sequence of log-likelihood values does not decrease with each iteration. In fact, we have

$$l(\boldsymbol{\theta}_{t-1}; \mathbf{x}) = \mathcal{L}(g_t, \boldsymbol{\theta}_{t-1}) \leqslant \mathcal{L}(g_t, \boldsymbol{\theta}_t) \leqslant \mathcal{L}(g_{t+1}, \boldsymbol{\theta}_t) = l(\boldsymbol{\theta}_t; \mathbf{x}) \ , \tag{6.34}$$

where the first and last equalities follow from the definitions of $\mathcal{L}$, g_t, and g_{t+1}, whereas the second and third inequalities follow from the M- and E-steps, respectively. Under certain continuity conditions the sequence $\{\boldsymbol{\theta}_t\}$ is guaranteed to converge to a local maximizer of the log-likelihood ℓ (or the likelihood L). Convergence to a global maximizer (the MLE $\widehat{\boldsymbol{\theta}}$) depends on the appropriate choice for the starting value. Typically, the algorithm is run from different random starting points. Note that (6.34) is useful for debugging computer implementations of the EM algorithm: if likelihood values are observed to decrease at any iteration, then there is an error in the program. For a further discussion of the theoretical and practical aspects of the EM algorithm we refer to McLachlan and Krishnan 2008. We illustrate the EM algorithm via two examples.

Example 6.21 (EM for the Genetic Linkage Experiment). In a genetic linkage experiment, n animals are randomly assigned (by nature) to four categories according to the multinomial distribution with pdf

$$f(x_1, x_2, x_3, x_4; \theta) \propto \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3} \pi_4^{x_4} ,$$

where $n = x_1 + x_2 + x_3 + x_4$ and the cell probabilities are $\pi_1 = 1/2 + \theta/4$, $\pi_2 = (1 - \theta)/4$, $\pi_3 = (1 - \theta)/4$, and $\pi_4 = \theta/4$. Suppose the observed data are given as $\mathbf{x} = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$, and we wish to obtain the MLE for θ.

It is easy to check that the log-likelihood function is given by

$$l(\theta; \mathbf{x}) = x_1 \ln(2 + \theta) + (x_2 + x_3) \ln(1 - \theta) + x_4 \ln \theta + \text{const} .$$

The graph of the log-likelihood function (apart from the constant term) is given in Fig. 6.5. Since this is a univariate problem, the MLE for θ can be obtained by the
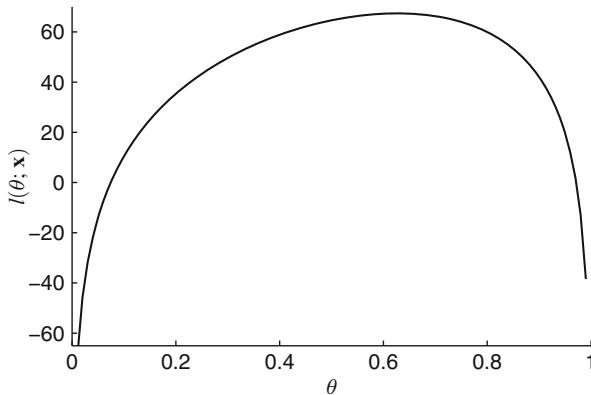
Fig. 6.5 The log-likelihood function for the genetic linkage experiment

grid search or the Newton–Raphson method (see Problem 6.25). In this example we use the EM algorithm to maximize the log-likelihood.

To that end, we augment the observed data as follows: suppose that the first of the original 4 multinomial cells could be split into two subcells having probabilities $1/2$ and $\theta/4$, respectively. Let Z and $X_1 - Z$ be the corresponding split of X_1, and note that Z is not observed. Now the random vector $(Z, X_1 - Z, X_2, X_3, X_4)$ has a multinomial distribution with the following five cell probabilities,

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1 - \theta}{4}, \frac{1 - \theta}{4}, \frac{\theta}{4} \right),$$

and the complete-data log-likelihood can be written as

$$\widetilde{l}(\theta; z, \mathbf{x}) = (x_1 - z + x_4) \ln \theta + (x_2 + x_3) \ln(1 - \theta) + \text{const} .$$

Suppose that θ_{t-1} is the current guess for θ. To implement the E-step, we first derive the conditional density $g_t(\mathbf{z}) = \widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}_{t-1})$. Note that given $X_1 = x_1$, Z has a $\mathsf{Bin}(x_1, p)$ distribution, with success probability

$$p = \frac{1/2}{1/2 + \theta_{t-1}/4} = \frac{2}{2 + \theta_{t-1}} .$$

Recall that for $Y \sim \mathsf{Bin}(n, p)$, we have $\mathbb{E}Y = np$. Hence, we have

$$\mathbb{E}_{g_t} Z = \mathbb{E}[Z \mid X_1 = x_1; \theta_{t-1}] = 2x_1/(2 + \theta_{t-1}) .$$

It follows that

$$Q_t(\theta) = \mathbb{E}_{g_t}\widetilde{l}(\theta; Z, \mathbf{x}) = \left(x_1 + x_4 - \frac{2x_1}{2 + \theta_{t-1}}\right) \ln \theta + (x_2 + x_3) \ln(1 - \theta) + \text{const} .$$

To implement the M-step, we simply solve $\frac{\mathrm{d}}{\mathrm{d}\theta} Q_t(\theta) = 0$ for θ. It is easy to check that the solution is given by

$$\theta_t = \frac{x_1 + x_4 - 2x_1/(2 + \theta_{t-1})}{n - 2x_1/(2 + \theta_{t-1})} .$$

The following MATLAB program implements the EM algorithm to find the MLE for θ, which is estimated to be $\widehat{\theta} = 0.6268$.

```
%genlink.m
x = [ 125 18 20 34 ]'; n = sum(x);
theta = 4*(x(1)/n-1/2);     %% initial guess
err = 1;
while abs(err) > 10^(-5)         %% stopping criterion
    z = 2*x(1)/(2+theta);         %% E-step
    temp = (x(1)+x(4)  - z)/(n-z); %% M-step
    err = theta - temp;
    theta = temp;
end
```

In the next example, we illustrate how one can use the EM algorithm for fitting mixture models. A **mixture pdf** is a pdf of the form

$$f(x) = w_1 f_1(x) + \cdots + w_c f_c(x), \quad w_z \geqslant 0, z = 1, \ldots, c, \quad \sum_{z=1}^{c} w_z = 1 , \quad (6.35)$$

where each f_z is itself a pdf. Such a mixture pdf can be thought of in the following way. Consider two random variables, X and Z, where Z takes values $1, 2, \ldots, c$ with probabilities $w_1, \ldots, w_c$, and conditional on $Z = z$, the random variable X has pdf f_z. By the product rule (3.10), the joint pdf of X and Z is given by $f_{X,Z}(x, z) = w_z f_z(x)$ and the marginal pdf of X is found by summing the joint pdf over the values of z—this gives (6.35).

☞ 72

Example 6.22 (EM for a Gaussian Mixture Model). Let $x_1, \ldots, x_n$ be iid observations drawn from the following *Gaussian* mixture pdf:

$$\overset{\circ}{f}(x; \boldsymbol{\theta}) = \sum_{z=1}^{c} \frac{w_z}{\sigma_z} \varphi\left(\frac{x - \mu_z}{\sigma_z}\right),$$

where φ is the pdf of the $\mathsf{N}(0, 1)$ distribution, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{w})$ with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_c)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_c)$, and $\mathbf{w} = (w_1, \ldots, w_c)$. The likelihood of $\mathbf{x}$ under $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \overset{\circ}{f}(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{z=1}^{c} \frac{w_z}{\sigma_z} \varphi\left(\frac{x_i - \mu_z}{\sigma_z}\right). \qquad (6.36)$$

Such a mixture distribution is often used for modeling unobserved heterogeneity, i.e., the presence of subpopulations that are not identified in the observed data. For example, suppose that x_i is, say, height of the ith student in a class. Further suppose that there are both male and female students in the class, but the genders of the students are not recorded. Then a suitable model for the outcomes is a mixture of two Gaussian distributions.

Direct maximization of the likelihood in (6.36) could be difficult and time-consuming. To simplify the computation, introduce a vector of latent variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$, each Z_i taking values in $\{1, 2, \ldots, c\}$ and such that $(X_i \mid Z_i = z_i) \sim \mathsf{N}(\mu_{z_i}, \sigma_{z_i}^2)$. This gives the complete-data likelihood

$$\widetilde{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{w_{z_i}}{\sigma_{z_i}} \varphi\left(\frac{x_i - \mu_{z_i}}{\sigma_{z_i}}\right). \qquad (6.37)$$

Note that by summing $\widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ over all $\mathbf{z}$ we obtain $f(\mathbf{x}; \boldsymbol{\theta})$. The discussion following (6.35) shows that the latent variable Z_i can be interpreted as the component of the mixture model from which X_i is drawn.

To implement the EM algorithm, suppose that $\boldsymbol{\theta}_{t-1}$ is the current guess for $\boldsymbol{\theta}$. In the E-step we first derive the (discrete) pdf of $\mathbf{Z}$ given the data $\mathbf{X} = \mathbf{x}$:

$$g_t(\mathbf{z}) = \widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}_{t-1}) \propto \widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_{t-1}).$$

Thus, to find g_t we view the right-hand side of (6.37) as a function of $\mathbf{z} = (z_1, \ldots, z_n)$. It follows that under g_t the components of $\mathbf{Z}$ are independent and each Z_i has a (discrete) pdf

$$g_{t,i}(z) \overset{\text{def}}{=} \frac{w_{t-1,z}}{\sigma_{t-1,z}} \varphi\left(\frac{x_i - \mu_{t-1,z}}{\sigma_{t-1,z}}\right) \bigg/ \sum_{k=1}^{c} \frac{w_{t-1,k}}{\sigma_{t-1,k}} \varphi\left(\frac{x_i - \mu_{t-1,k}}{\sigma_{t-1,k}}\right) \qquad (6.38)$$

for $i = 1, \ldots, n$ and $z = 1, \ldots, c$. The expected complete-data likelihood in the E-step is then

$$Q_t(\boldsymbol{\theta}) = \mathbb{E}_{g_t} \widetilde{l}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{z=1}^{c} g_{t,i}(z) \left(\ln w_z - \ln \sigma_z - \frac{(x_i - \mu_z)^2}{2\sigma_z^2}\right) + \text{const}.$$

Next, in the M-step, we maximize $Q_t(\boldsymbol{\theta})$ with respect to $\mathbf{w}$ (under the constraints $\sum_z w_z = 1$, $w_z \geqslant 0$ for all z), $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$. It is easy to check that for $z = 1, \ldots, c$ the solution to $\nabla Q_t(\boldsymbol{\theta}) = \mathbf{0}$ is

$$\begin{aligned}
w_z &= \frac{1}{n} \sum_{i=1}^{n} g_{t,i}(z), \\
\mu_z &= \frac{\sum_{i=1}^{n} g_{t,i}(z) x_i}{\sum_{i=1}^{n} g_{t,i}(z)}, \qquad\qquad (6.39) \\
\sigma_z^2 &= \frac{\sum_{i=1}^{n} g_{t,i}(z)(x_i - \mu_z)^2}{\sum_{i=1}^{n} g_{t,i}(z)}.
\end{aligned}$$

We then set $\boldsymbol{\theta}_t$ according to the values in (6.39) and keep iterating the E-step (6.38) and the M-step (6.39) until convergence is reached.

6.7   Problems

6.1. In a guessing game Albert chooses a number θ between 0 and 10, and the other people have to guess the number; the person whose guess is closest to θ wins. To facilitate the guesswork, Albert draws seven numbers uniformly from the interval $[0, \theta]$ and displays the results to the others. Suppose these seven values (the observed data) are

$$4.3180, 4.8007, 0.6730, 4.8409, 3.3515, 0.5170, 1.4760 .$$

a. Give a model for the data $X_1, \ldots, X_7$. Show that $M = \max\{X_1, \ldots, X_7\}$ is a sufficient statistic for θ.
b. Determine the method of moment estimate of θ. Is the corresponding estimator a function of M?
c. Sketch the graph of the likelihood function, and use it to determine the MLE. Is the corresponding estimator a function of M?
d. Use $T = M/\theta$ as a pivot variable to construct a 95 % numerical confidence interval for θ of the form (m, a) for some $a > m$, where $m = \max\{x_1, \ldots, x_7\}$.

6.2. Let $X_1, \ldots, X_n \sim_{\text{iid}} N(\theta, \theta)$ with an unknown $\theta > 0$. Find the MLE of θ.

6.3. Consider the general regression model

$$Y_i = g(\mathbf{x}_i; \boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, \ldots, n , \tag{6.40}$$

where $\varepsilon_1, \ldots, \varepsilon_n \sim_{\text{iid}} N(0, \sigma^2)$ and $g(\mathbf{x}; \boldsymbol{\beta})$ is a known function of the explanatory vector $\mathbf{x}$ and the parameter vector $\boldsymbol{\beta}$. Both σ^2 and $\boldsymbol{\beta}$ are assumed to be unknown.

a. Show that the MLE of $\boldsymbol{\beta}$ is found by minimizing the sum of the squared deviations between the $\{Y_i\}$ and the $\{g(\mathbf{x}_i; \boldsymbol{\beta})\}$; that is,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} (Y_i - g(\mathbf{x}_i; \boldsymbol{\beta}))^2 .$$

b. Derive the MLE of σ^2.

6.4. For multidimensional parameters $\boldsymbol{\theta}$ it is sometimes useful to draw 1-dimensional graphs for the likelihood function by substituting all parameters *except one* with their MLEs (as a function of the remaining unknown parameter). The function thus obtained is called the **profile likelihood**.

Consider the 10 iid samples from the $N(\mu, \sigma^2)$ distribution given in Example 6.2.

a. Give the formula for the profile likelihood for σ^2.
b. Draw the graph of this profile likelihood. Does its mode correspond to the MLE of σ^2?

6.5. Let $X_1, \ldots, X_n$ be iid random variables with pdf

$$f(x; \theta) = (\theta + 1) x^\theta, \quad 0 \leqslant x \leqslant 1, \quad \theta > -1 .$$

a. Find the method of moment estimator of θ.
b. Find the MLE of θ.

6.6. The weight X (in grams) of an egg is $N(\mu, \sigma^2)$ distributed. Let $\widehat{\mu} = 56.3$ and $\widehat{\sigma} = 7.6$ be the MLEs of μ and σ. Give the MLE of

$$\mathbb{P}(X > 68.5) .$$

6.7. For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, let S^2 be the sample variance and let $\widehat{\sigma^2}$ be the MLE of σ^2.

a. Which of the two is an unbiased estimator of σ^2?
b. Is $\sqrt{S^2}$ an unbiased estimator of σ?
c. Is $\sqrt{\widehat{\sigma^2}}$ the MLE of σ?

6.8. Let $Y \sim \mathsf{U}(-\pi/2, \pi/2)$ and define $Z = \tan(Y)$. Show, using transformation formula (3.21), that Z has a Cauchy distribution.

6.9. The following iid data, 0.685, 2.586, -1.969, -2.673, 1.464, 2.977, -1.120, 1.594, -0.543, 1.505, -1.266, 1.981, have been drawn from a **double exponential distribution**, with pdf

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|}, \quad x \in \mathbb{R} .$$

Find the MLE for λ.

6.10. The **Weibull distribution** with *scale* parameter $\lambda > 0$ and *shape* parameter $\alpha > 0$ has cdf

$$F(x) = 1 - e^{-(\lambda x)^{\alpha}}, \quad x \geqslant 0 .$$

Suppose $x_1, \ldots, x_n$ are the outcomes of an iid sample from a Weibull distribution with shape parameter $\alpha = 2$ and unknown scale parameter λ. Find the MLE of λ.

6.11. Suppose $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{Geom}(p)$. Show that the generalized likelihood ratio method for the hypothesis $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ yields the test statistic

$$\Lambda = \frac{\left(\dfrac{p_0}{1 - p_0} \right)^n (1 - p_0)^{(n\bar{X})}}{\left(\dfrac{1/\bar{X}}{1 - 1/\bar{X}} \right)^n (1 - 1/\bar{X})^{(n\bar{X})}} .$$

What is the approximate distribution of $-2 \ln \Lambda$ for large n?

6.12. Let $X_1, \ldots, X_n$ be an iid sample from the $\mathsf{Bin}(k, p)$ distribution, where k is given but $p \in [0, 1]$ is unknown.

a. Find the MLE $\widehat{p}$ of p.
b. Show that $\widehat{p}$ attains the Cramér–Rao lower bound.
c. Sketch the log-likelihood function for the case where $n = 1$, $k = 10$, and $x_1 = 5$.

6.13. Suppose that 100 observations are taken from the $\mathsf{N}(\mu, 1)$ distribution with an unknown μ. Instead of recording all the observations, one records only whether the observation is less than 0. Suppose that 40 observations are less than 0. What is the MLE for μ based on these observations?

6.14. Let $X_1, \ldots, X_n$ be an iid sample from the $\mathsf{Exp}(1/v)$ distribution, where $v > 0$ is unknown. Let $\mathbf{X} = (X_1, \ldots, X_n)$.

a. Find the score $S(v; \mathbf{X})$.
b. Give the corresponding Fisher information.

c. Find the MLE of v.

d. Give the MLE of $\sin(v)$.

6.15. Let $X_1, \ldots, X_n$ be an iid sample from the distribution with pdf $f(x; \theta)$, where

$$f(x; \theta) = \frac{1}{2\,\theta^3}\, x^2\, e^{-x/\theta}, \quad x > 0, \quad \theta > 0 .$$

a. Show that $\mathbb{E} X_i = 3\,\theta$ and $\mathrm{Var}(X_i) = 3\,\theta^2$.

b. Find a sufficient statistic for the parameter θ using the factorization Theorem 5.5.

c. Find the MLE of θ.

d. Find the Fisher information $I(\theta)$.

e. Give the asymptotic distribution of the MLE of θ.

f. What are the bias and the variance of the MLE of θ?

g. Determine whether or not the MLE of θ attains the Cramér–Rao lower bound.

6.16. An iid sample $X_1, \ldots, X_n$ is taken from the $\mathsf{N}(0, \theta)$ distribution, where $\theta > 0$ is unknown. We wish to test the hypothesis $H_0 : \theta = 3$ against $H_1 : \theta \neq 3$ using an appropriate test statistic.

a. Show that the likelihood ratio test statistic is here a function of

$$T = \sum_{i=1}^n \frac{X_i^2}{3} .$$

b. What is the distribution of T under H_0?

6.17. Verify that the score function corresponding to the observed iid sample $x_1, \ldots, x_n$ from the $\mathsf{Gamma}(\alpha, \lambda)$ distribution is

$$\mathbf{S}(\alpha, \lambda) = \begin{pmatrix} n(\ln \lambda - \psi(\alpha)) + \sum_{i=1}^n \ln x_i \\ \frac{n\alpha}{\lambda} - \sum_{i=1}^n x_i \end{pmatrix} ,$$

and that the corresponding Fisher information matrix is

$$I(\alpha, \lambda) = n \begin{pmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{pmatrix} ,$$

where $\psi'(\alpha)$ is the derivative of the **digamma** function $\psi(x) = \Gamma'(x)/\Gamma(x)$.

6.18. Suppose $x_1, \ldots, x_{10}$ are the outcomes of an iid sample from $\mathsf{Exp}(\theta)$. Construct a score confidence interval for θ with confidence coefficient 0.90 if the sum of the $\{x_i\}$ is 10.

6.19. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random samples from the $\mathsf{Exp}(\lambda)$ and $\mathsf{Exp}(\mu)$ distributions, for unknown λ and μ. Suppose we wish to test the hypothesis $H_0 : \lambda = \mu$ against $H_1 : \lambda \neq \mu$.

a. Find the MLEs for λ and μ.
b. Find the MLEs for $\theta = \lambda = \mu$ under H_0.
c. Show that the following test statistic

$$T = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$$

  can be derived from the generalized likelihood ratio procedure.
d. For large n, T has approximately a normal distribution. Find the parameters of
☞ 92   this distribution via the delta method.

6.20. A useful way to measure how far away a pdf g is from a pdf h is the **Kullback–Leibler (KL) distance** (also called KL divergence or cross-entropy distance), defined as

$$\mathcal{D}(g, h) = \mathbb{E}_g \ln \frac{g(\mathbf{X})}{h(\mathbf{X})} . \qquad (6.41)$$

☞ 183   In the EM algorithm it is used to derive the inequality (6.31) using the following decomposition:

$$\ln f(\mathbf{x}; \boldsymbol{\theta}) = \int g(\mathbf{z}) \ln f(\mathbf{x}; \boldsymbol{\theta}) \, d\mathbf{z}$$

$$= \int g(\mathbf{z}) \ln \left(\frac{\widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})/g(\mathbf{z})}{\widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta})/g(\mathbf{z})} \right) d\mathbf{z}$$

$$= \underbrace{\int g(\mathbf{z}) \ln \left(\frac{\widetilde{f}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{g(\mathbf{z})} \right) d\mathbf{z}}_{\mathcal{L}(g, \boldsymbol{\theta})} + \mathcal{D}(g, \widetilde{f}_{\mathbf{Z}|\mathbf{X}}(\cdot \,|\, \mathbf{x}; \boldsymbol{\theta})) . \qquad (6.42)$$

☞ 33   a. Using Jensen's inequality, show that $\mathcal{D}(g, h) \geqslant 0$ in (6.41).
       b. Verify (6.42) and explain how g should be chosen such that the Kullback–Leibler term in (6.42) is minimized.

⬜   **6.21.** Let $X_1, \ldots, X_n$ be an iid sample from the discrete pdf

$$f(x; \theta) = \frac{\theta^x e^{-\theta}}{x! \, (1 - e^{-\theta})}, \quad x \in \{1, 2, \ldots\}, \quad \theta > 0 .$$

Suppose that an iid sample of size $n = 16$ gives two 5s, four 4s, four 3s, four 2s, and two 1s. Plot the likelihood function and the log-likelihood function of the data and perform a **grid search** to obtain the MLE; that is, of the plotted values, find the θ for which the likelihood (or log-likelihood) is maximal.

6.22. The data 1.1668, 0.0738, 0.7740, 1.0160, 0.4822, 1.4559, 0.1752, 0.5209, 0.1537, 0.2947 are the outcomes of an iid sample $X_1, \ldots, X_{10}$ from the pdf

$$f(x) = c\,(b - x), \quad 0 \leqslant x \leqslant b,$$

where $b > 0$ is unknown and c is a normalization constant.

a. Show that $c = 2/b^2$.
b. Estimate b via the method of moments.
c. Show that the MLE $\widehat{b}$ satisfies

$$\sum_{i=1}^{10} \frac{\widehat{b}}{\widehat{b} - x_i} - 20 = 0, \quad \widehat{b} \geqslant 1.4559,$$

if this equation has a solution. Determine $\widehat{b}$ numerically using MATLAB's root-finding function `fzero`.

6.23. Consider the score interval for the binomial distribution in Example 6.16. As in Problem 5.22 the exact coverage probability can be calculated as a function of p by means of total enumeration. Plot the coverage probability for the score interval and compare it with the "standard" one in Problem 5.22.

6.24. Using Problem 6.17, implement Fisher's scoring method

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + I^{-1}(\boldsymbol{\theta})\,\mathbf{S}(\boldsymbol{\theta})$$

to find the MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\lambda})$ for the following iid data from Gamma(α, λ)):

```
29.7679    12.8406   105.3225    46.6101     75.7135    72.0340
33.9008    35.2510    50.9201    29.8086     32.6963   131.5229
29.1369    61.8774    31.0650    54.4877    103.6889    68.0230
30.1994    48.3140    54.4447    29.2253     27.0242   102.5929
43.0354    96.5552    64.1004    65.3381     89.6879    63.7344
```

Use the method of moment estimates as starting values for the Newton–Raphson scheme. The digamma function ψ is implemented in MATLAB as `psi(x)` and its derivative ψ' as `psi(1,x)` (polygamma function).

6.25. Consider the genetic linkage model in Example 6.21.

a. Show that the score and Hessian functions for θ are given by

$$S(\theta) = \frac{34}{\theta} + \frac{125}{\theta + 2} - \frac{38}{1 - \theta} \quad \text{and} \quad H(\theta) = -\frac{34}{\theta^2} - \frac{125}{(\theta + 2)^2} - \frac{38}{(1 - \theta)^2} \,.$$

b. Implement a Newton–Raphson procedure to find the MLE of θ.
c. Implement a simple grid search to find the MLE.
d. Do the Newton–Raphson, grid search, and EM approaches give the same estimate?

Chapter 7
Monte Carlo Sampling

Monte Carlo sampling—that is, random sampling on a computer—has become an important methodology in modern statistics. By simulating random variables from specified statistical models and probability distributions one can often estimate certain statistical quantities that may otherwise be difficult to obtain. In Sect. 2.7 we already saw how random variables can be generated from common probability distributions via the *inverse-transform* and *acceptance–rejection* methods.

In this chapter we discuss two other important Monte Carlo sampling techniques: the *bootstrap method* and *Markov chain Monte Carlo* (MCMC). The bootstrap method is a sampling procedure in which new samples are generated by resampling the observed data. MCMC is used extensively in Bayesian statistics to sample from complicated multidimensional distributions. Bayesian statistics is introduced in Chap. 8.

The following example illustrates how random sampling can be used to estimate a p-value without having to derive the specific distribution of the test statistic.

Example 7.1 (Estimating a p-Value). Suppose an iid sample of size 4 from a $N(\mu, \sigma^2)$ distribution has a sample mean $\bar{x} = -0.7$ and sample standard deviation $s = 0.4$. We wish to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu < 0$, using the test statistic $T = 2\bar{X}/S$, whereby we reject H_0 if the outcome of T is too small. The observed outcome of T is $t = 2 \times -0.7/0.4 = -3.5$. The corresponding *p-value* is

$$p = \mathbb{P}_{H_0}(T \leqslant -3.5) = \mathbb{E}_{H_0} I_{\{T \leqslant -3.5\}} .$$

This can be estimated by simulating, under H_0, a large iid sample $T_1, \ldots, T_N$ of copies of T and evaluating the sample average

$$\widehat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{\{T_i \leqslant -3.5\}} ,$$

similar to the Monte Carlo integration procedure in Example 5.7. In the MATLAB
program below, each T_i is generated by drawing an iid sample of size 4 from the
standard normal distribution and evaluating T for that sample. The variable `count`
contains the total number of test statistics with a value less than or equal to the
observed value -3.5; the estimate $\widehat{p}$ is simply the value of `count` divided by N.
A typical estimate for $\widehat{p}$ is 0.02. This indicates that there is fairly strong evidence
that H_0 is not true. A huge advantage of this approach is that we do not have to
analyze or evaluate the cdf of the test statistic under H_0; we only have to repeat the
experiment under H_0 many times. See Problem 7.1 for a further discussion.

```
%pvalsim.m
xbar_obs = -0.7; s_obs = 0.4; t_obs = 2*xbar_obs/s_obs;
N = 10^5;
count = 0;
for i=1:N
    x = randn(4,1);
    xbar = mean(x); s = std(x); t = 2*xbar/s;
    count = count + (t <= t_obs);
end
phat = count/N  % estimated p-value
```

Statistical sampling often involves generating an *iid* sample from some specified
discrete or continuous pdf. Two important ways to analyze such data are to use the
empirical cdf and *density estimation*.

7.1   Empirical Cdf

> **Definition 7.1. (Empirical Cdf).** Let $x_1, \ldots, x_N$ be an iid real-valued sam-
> ple from a cdf F. The function
>
> $$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} I_{\{x_i \leqslant x\}}, \quad x \in \mathbb{R}, \tag{7.1}$$
>
> is called the **empirical cdf** of the data.

Here, $I_{\{x_i \leqslant x\}} = 1$ if $x_i \leqslant x$ and 0 otherwise. F_N is a nondecreasing step function
which jumps up by an amount of $1/N$ at each of the points $\{x_i\}$. Moreover, F_N is
right-continuous and bounded between 0 and 1. In other words, F_N is a cdf—see
Sect. 2.1. It is the cdf of a random variable that takes one of the values $x_1, \ldots, x_N$
with equal probability $1/N$, assuming that all the observations are different. In

Fig. 7.1 the empirical cdf is shown of an iid sample of size 10 from the Exp(0.2) distribution. The true cdf is plotted as well.
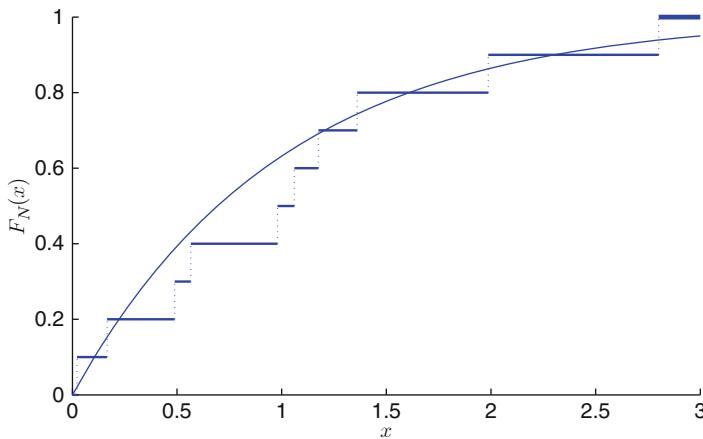
Fig. 7.1 The empirical cdf for a sample of size 10 from the Exp(0.2) distribution and the true cdf

We see that the empirical cdf follows the true cdf quite well. The fit becomes better and better as the sample size increases. In Fig. 7.2 the empirical and true cdfs are shown for the same distribution, but now for a sample size of 200.
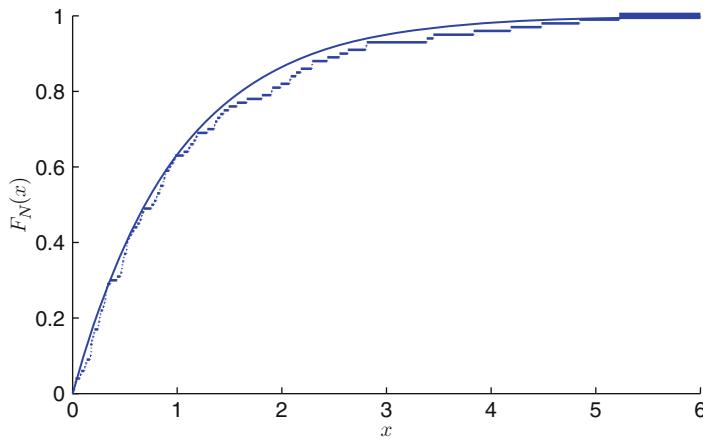
Fig. 7.2 The empirical cdf for a sample of size 200 from the Exp(0.2) distribution and the true cdf

If we order the sample as $x_{(1)} < x_{(2)} < \cdots < x_{(N)}$, then for each $i = 1, \ldots, N$,

$$F_N(x_{(i)}) = \frac{i}{N} \ , \tag{7.2}$$

assuming for simplicity that all $\{x_i\}$ take different values.

If instead of deterministic $\{x_i\}$ we take *random* X_i, in (7.1), then $F_N(x)$ becomes random as well. To distinguish between the deterministic and the random case, let us denote the random empirical cdf by $\widehat{F}_N(x)$. We now have

$$\mathbb{P}\left(\widehat{F}_N(x) = \frac{i}{N} \right) = \mathbb{P}(X_{(i)} \leqslant x, X_{(i+1)} > x) = \binom{N}{i}(F(x))^i\,(1 - F(x))^{N-i} \ . \tag{7.3}$$

To see this, note that the event $\{X_{(i)} \leqslant x\} \cap \{X_{(i+1)} > x\}$ means that exactly i of the N random variables that we draw from F are less than or equal to x. Thus, the event is equivalent to having i successes in N independent Bernoulli experiments with success probability $F(x)$, which leads to (7.3).

Equation (7.3) can be summarized as $N\,\widehat{F}_N(x) \sim \mathsf{Bin}(N, F(x))$. As a consequence,

$$\mathbb{E}\widehat{F}_N(x) = F(x)$$

and

$$\mathrm{Var}(\widehat{F}_N(x)) = F(x)(1 - F(x))/N \ .$$

Moreover, by the law of large numbers and the central limit theorem, we have

$$\mathbb{P}(\lim_{N\to\infty} \widehat{F}_N(x) = F(x)) = 1 \ , \tag{7.4}$$

and

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{\widehat{F}_N(x) - F(x)}{\sqrt{F(x)(1 - F(x))/N}} \leqslant z \right) = \Phi(z) \ , \tag{7.5}$$

where Φ is the cdf of the standard normal distribution.

☞ 136   Exactly as in (5.23) we see that an approximate $1 - \alpha$ confidence interval for $F(x)$ is

$$F_N(x) \pm z_{1-\alpha/2} \sqrt{\frac{F_N(x)(1 - F_N(x))}{N}} \ ,$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Moreover, if we order the observations $x_{(1)} < \cdots < x_{(N)}$, then, by (7.2), an approximate $1 - \alpha$ confidence interval for $F(x_{(i)})$ is

$$\frac{i}{N} \pm z_{1-\alpha/2} \sqrt{\frac{i(1-i/N)}{N^2}} , \quad i = 1, \ldots, N .$$

Example 7.2 (Confidence Interval for a Cdf). In Fig. 7.3 a 90 % confidence interval (hence $z_{1-\alpha/2} = z_{0.95} = 1.645$) is given for the cdf of the $\mathsf{Exp}(1)$ distribution, based on an iid sample of size $N = 60$. The true cdf is given by the smooth line. We see that the true cdf lies convincingly within the confidence curves. However, the actual width of the confidence intervals (as a function of x) is quite sizeable, due to the fact that N is not large.
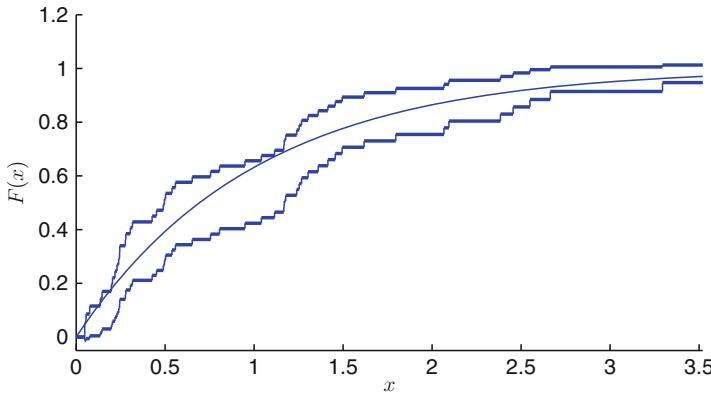
Fig. 7.3 A 90 % confidence interval for the cdf $F(x) = 1 - e^{-x}$, $x \geqslant 0$

Let $X_1, \ldots, X_N \sim_{\text{iid}} F$, where F is continuous and strictly increasing. Define $U_1 = F(X_1), \ldots, U_N = F(X_N)$. From the inverse-transform method (see Sect. 2.7.2) we see that $U_1, \ldots, U_N$ is an iid sample from $\mathsf{U}(0, 1)$. Denote the empirical cdf of the $\{U_i\}$ by $\widehat{G}_N(u)$, and let x and u be related via $x = F^{-1}(u)$ and $u = F(x)$. Then, ☞ 53

$$\widehat{G}_N(u) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{\{U_i \leqslant u\}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{\{F(X_i) \leqslant F(x)\}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{\{X_i \leqslant x\}} = \widehat{F}_N(x) .$$

$\widehat{G}_N$ is called the **reduced empirical cdf**. Note that $N \widehat{G}_N(u) \sim \mathsf{Bin}(N, u)$, irrespective of F. Define the maximum distance between the empirical and the true cdfs as

$$D_N = \sup_{x \in \mathbb{R}} |\widehat{F}_N(x) - F(x)| = \sup_{0 \leqslant u \leqslant 1} |\widehat{G}_N(u) - u| . \tag{7.6}$$

This is called the **Kolmogorov–Smirnov statistic** of the data. Note that this statistic does not depend on F. It can be used to test whether iid samples $X_1, \ldots, X_N$ come from a specified cdf F.

Example 7.3 (Kolmogorov–Smirnov Test). The Weibull distribution $\mathsf{Weib}(\alpha, \lambda)$ has cdf

$$F(x) = 1 - e^{-(\lambda x)^{\alpha}}, \quad x \geqslant 0 .$$

To generate from this distribution, we can use the inverse-transform method: generate $U \sim \mathsf{U}(0, 1)$ and output $X = \frac{1}{\lambda}(-\ln U)^{\frac{1}{\alpha}}$. Note that the $\mathsf{Weib}(1, \lambda)$ distribution is simply the $\mathsf{Exp}(\lambda)$ distribution.

Suppose we have an iid sample from the $\mathsf{Weib}(1.5, 1)$ distribution. Would the Kolmogorov–Smirnov statistic correctly reject the hypothesis H_0 that the sample is from the $\mathsf{Exp}(1)$ distribution? The following MATLAB program carries out this experiment. It generates an iid sample of size $N = 100$ from the $\mathsf{Weib}(1.5, 1)$ distribution. It then evaluates the Kolmogorov–Smirnov statistic. Figure 7.4 shows the reduced empirical cdf $\widehat{G}_N(u)$. The maximum distance between $\widehat{G}_N(u)$ and u is $d_N \approx 0.1458$ in this case. The p-value $\mathbb{P}_{H_0}(D_N > d_N)$ is determined by Monte Carlo simulation, by repeating the experiment $K = 1000$ times under H_0, that is, using $\mathsf{Exp}(1)$ data. The estimated p-value is approximately 0.026. There is thus reasonable to strong evidence to suggest that the true distribution is not $\mathsf{Exp}(1)$.

```
%kolsmirweib.m
rng(1234)
alpha = 1.5;
%generate data
N = 100;
U = rand(1,N);
x = (-log(U)).^(1/alpha);
y = sort(1 - exp(-x));
i=1:N;
%plot empirical cdf and calculate test statistic
stairs([0,y],[0,i/N],'r'), hold on, line([0,1],[0,1]);
dn_up = max(abs(y-i/N));dn_down = max(abs(y-(i-1)/N));
dn = max(dn_up, dn_down);
%use Monte Carlo simulation to find p-value
K = 10000;
for k=1:K
    i=1:N;
    y = sort(rand(1,N));
    DN(k) = max( max(abs(y-i/N)), max(abs(y-(i-1)/N)));
end
p   = sum(DN >= dn)/K
```

Fig. 7.4 The reduced
empirical cdf $\widehat{G}_N(x)$
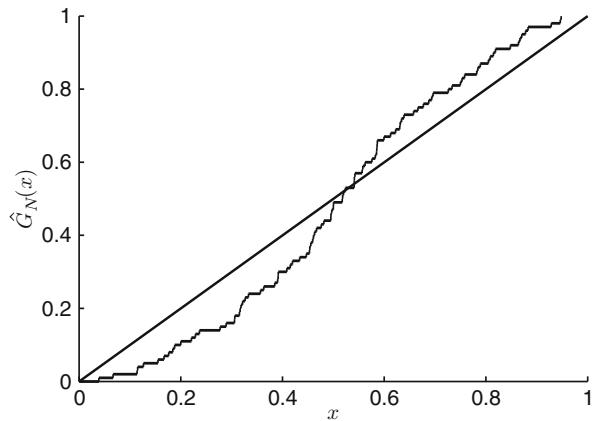
7.2   Density Estimation

Suppose that $x_1, \ldots, x_N$ is an iid sample from some unknown continuous pdf f —
obtained from Monte Carlo sampling, for example. A useful approach to estimate
f from the data is to use a Gaussian kernel density estimator.

Definition 7.2.   (Gaussian Kernel Density Estimator). Let $x_1, \ldots, x_N$ be an
iid sample from a continuous pdf f. The **Gaussian kernel density estimator**
of f with **bandwidth** $h > 0$ is given by

$$\widehat{f}(x; h) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi h^2}} \, \mathrm{e}^{-\frac{(x - X_i)^2}{2h^2}}, \quad x \in \mathbb{R} . \qquad (7.7)$$

The idea is illustrated in Fig. 7.5 for the case of $N = 5$ data points. The Gaussian
kernel density estimate (KDE) is the equally weighed *mixture* (see (6.35)) of N
Gaussian/normal pdfs, where each pdf is centered around a data point and has
variance h^2.

How well the Gaussian KDE $\widehat{f}(\cdot; h)$ fits the true pdf f depends crucially on the
choice of the bandwidth parameter h. If h is too small, the density estimate will be
too spiky; if h is too large, the estimate will be too smooth. An often used *rule of
thumb* is to take

$$h_{\mathrm{Rot}} = \left(\frac{4 \, S^5}{3 \, N} \right)^{4/5} ,$$
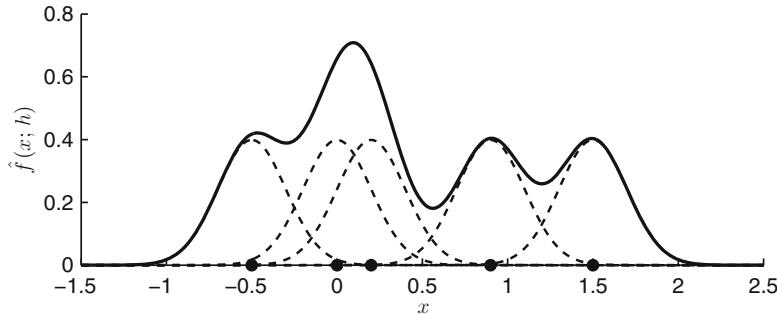
Fig. 7.5 The Gaussian KDE (*solid line*) is the equally weighted mixture of normal pdfs centered around the data and with standard deviation h (*dashed*)

where S is the standard deviation of the data. This choice is based on a mathematical analysis of the discrepancy between $\widehat{f}(\cdot;h)$ and f as $n \to \infty$. There exist many sophisticated modifications of the basic Gaussian KDE in (7.7). In this book we use the fast and reliable **theta KDE** of Botev et al. (2010). The MATLAB function kde.m can be downloaded from http://www.mathworks.com/matlabcentral/fileexchange/14034-kernel-density-estimator.

Example 7.4 (Kernel Density Estimate). The following MATLAB program draws an iid sample from the Exp(1) distribution and constructs a Gaussian KDE. We see in Fig. 7.6 that with an appropriate choice of the bandwidth a good fit to the true pdf can be achieved, except at the boundary $x = 0$. The theta KDE, which can be viewed as a generalization of the Gaussian KDE, does not exhibit this boundary effect. Moreover, it chooses the bandwidth automatically, to achieve a superior fit.

```
%gausthetakde.m
h = 0.1; h2 = h^2; c = 1/sqrt(2*pi)/h; % constants
phi = @(x,x0) exp(-(x-x0).^2/(2*h2)); % unscaled kernel
f = @(x) exp(-x).*(x >= 0);  % True pdf
N = 10^4;   % sample size
x = -log(rand(N,1)); % generate the data

xx = [-0.5:0.01:6];  % plot range
phis = zeros(1,numel(xx));
for i = 1:N
    phis = phis + phi(xx,x(i));
end
phis = c*phis/N;
hold on
plot(xx,phis,'r');        % plot Gaussian KDE
[bandwidth,density,xmesh] = kde(x,2^12,0,max(x));
idx = find(xmesh <= 6);
plot(xmesh(idx),density(idx),'b')  % plot theta KDE
plot(xx,f(xx),'k');       % plot true pdf
```
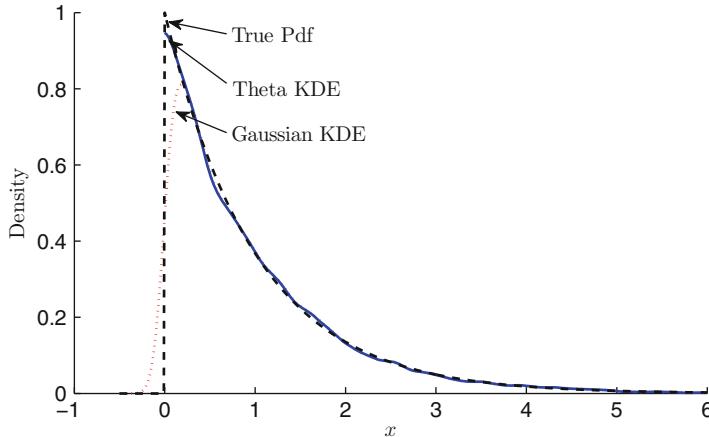
Fig. 7.6 Kernel density estimates for Exp(1)-distributed data

7.3 Resampling and the Bootstrap Method

The idea behind **resampling** is very simple: an iid sample $\mathbf{x} = (x_1, \ldots, x_N)$ from some unknown pdf f represents our best knowledge about f if we make no further *a priori* assumptions about f. So, the best way to "repeat" the experiment is to *resample* the $\{x_i\}$ by drawing from the empirical distribution. The following algorithm is a direct consequence of the inverse-transform method.

Algorithm 7.1. (Sampling from an Empirical Cdf). Let $x_1, \ldots, x_N$ be an iid sample from a continuous cdf F. To generate an iid sample of size M from the empirical cdf F_N, carry out the following steps:

1. Draw $U_1, \ldots, U_M \overset{\text{iid}}{\sim} \mathsf{U}(0, 1)$.
2. Set $I_i = \lceil N U_i \rceil, i = 1, \ldots, M$.
3. Return $x_{I_1}, \ldots, x_{I_M}$.

Here $\lceil x \rceil$ (the *ceiling* of x) is the smallest integer larger than or equal to x. The requirement that F be continuous is to rule out duplicate data points.

By sampling from the empirical cdf we can thus repeat (approximately) the experiment that gave us the original data as many times as we like. This is useful if we want to assess the properties of certain statistics obtained from the data. For example, suppose that the original data $\mathbf{x}$ gave the statistic $t(\mathbf{x})$. By resampling we can gain information about the *distribution* of the corresponding random variable $t(\mathbf{X})$.

Example 7.5 (Resampling Cauchy Data). Suppose we have an iid sample of size $N = 100$ from the Cauchy distribution—that is, with pdf

$$f_X(x) = \frac{1}{\pi(1 + x^2)} , \quad x \in \mathbb{R} ;$$

☞ 181   see also Examples 6.3 and 6.20. We learned from the first example that the sample mean is a poor estimate of the mode (0) of the distribution. By resampling the data we can get a good idea how the *distribution* of the sample mean compares with that of other estimators—for example, the **sample median** of the data. Ordering the data from smallest to largest, $x_{(1)} \leqslant \cdots \leqslant x_{(N)}$, the sample median $\widetilde{x}$ is defined as the "middle" observation; that is, $\widetilde{x} = x_{((N+1)/2)}$ for odd N, and $\widetilde{x} = (x_{(N/2)} + x_{(N/2+1)})/2$ for even N.

Figure 7.7 depicts three graphs. The dashed line is the KDE of $K = 5000$ iid sample means, where each sample mean is obtained from a resampled data set of size $M = 100$ from the original iid Cauchy data of size $N = 100$. Similarly, the solid line represents the KDE of the sample medians.
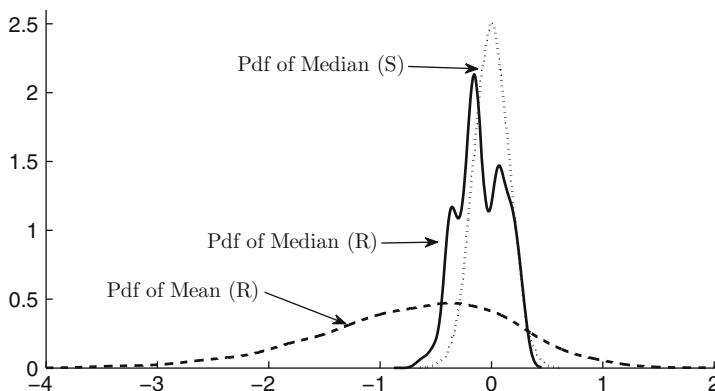
Fig. 7.7 Kernel density estimates for the mean (*dashed line*) and median (*solid line*) of resampled data, as well as for the median of newly sampled data (*dotted line*)

The figure shows that the sample median has much better statistical properties than the sample mean. In particular, the pdf of the sample median (estimated via the KDE) is much less spread out than that of the sample mean and is (here) mostly concentrated in the interval $(-0.5, 0.5)$. For comparison, the figure also shows the KDE of the sample median obtained from $K = 5000$ iid samples from the original distribution (dotted line). Thus, instead of resampling the data, we draw each time the data from scratch. The following MATLAB program can be used to carry out the experiment. See Problem 7.6 for a further discussion of this example.

```
%resampcauchy.m
N = 100; K = 5000;
xorg =  tan(pi*(0.5 - rand(1,N))); % original data
medxorg =  median(xorg);
meanxorg = mean(xorg);
x = zeros(1,N); mx = zeros(1,K);
for i=1:K
    ind = ceil(n*rand(1,N)); % draw random indices
    x = xorg(ind);  % resampling the data (R)
    % x = tan(pi*(0.5 - rand(1,N))); % sampling new
      data (S)
    mx(i) = median(x);
    % mx(i) = mean(x);
end
[bandwidth,density,xmesh]=kde(mx,2^7);
plot(xmesh,density)
```

The **bootstrap method** is a formalization of the resampling idea. Suppose we wish to estimate a number ℓ via some estimator $H = H(\mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_N)$ and the $\{X_i\}$ form an iid sample from some unknown cdf F. It is assumed that H does not depend on the order of the $\{X_i\}$. To assess the quality (e.g., accuracy) of the estimator H, one could draw independent replications $\mathbf{X}_1, \ldots, \mathbf{X}_K$ of $\mathbf{X}$ and find sample estimates for quantities such as the *variance* of the estimator

$$\mathrm{Var}(H) = \mathbb{E}H^2 - (\mathbb{E}H)^2,$$

the *bias*

$$\mathrm{Bias} = \mathbb{E}H - \ell,$$

and the *mean square error* (MSE)

☞ 154

$$\mathrm{MSE} = \mathbb{E}(H - \ell)^2.$$

However, it may be too time-consuming, or simply not feasible, to obtain such replications. An alternative is to resample the original data, as described above. To reiterate, given an outcome $(x_1, \ldots, x_N)$ of $\mathbf{X}$, we draw an iid sample $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ not from F but from the empirical cdf F_N, via Algorithm 7.1 (hence $M = N$ here).

The rationale is that the empirical cdf F_N is close to the actual distribution F and gets closer as N gets larger. Hence, any quantities depending on F, such as $\mathbb{E}_F h(H)$, where h is a function, can be approximated by $\mathbb{E}_{F_N} h(H)$. The latter is usually still difficult to evaluate, but it can be simply estimated via Monte Carlo simulation as

$$\frac{1}{K} \sum_{i=1}^{K} h(H_i^*),$$

where $H_1^*, \ldots, H_K^*$ are independent copies of $H^* = H(\mathbf{X}^*)$. This seemingly self-referent procedure is called **bootstrapping**—alluding to Baron von Münchhausen, who pulled himself out of a swamp by his own bootstraps. As an example, the bootstrap estimate of the expectation of H is

$$\widehat{\mathbb{E}H} = \overline{H}^* = \frac{1}{K}\sum_{i=1}^{K} H_i^* \ ,$$

which is simply the sample mean of $\{H_i^*\}$. Similarly, the bootstrap estimate for $\text{Var}(H)$ is the sample variance

$$\widehat{\text{Var}(H)} = \frac{1}{K-1}\sum_{i=1}^{K}(H_i^* - \overline{H}^*)^2 \ . \tag{7.8}$$

Bootstrap estimators for the bias and MSE are $\overline{H}^* - H$ and $\frac{1}{K}\sum_{i=1}^{K}(H_i^* - H)^2$, respectively. Note that for these estimators the unknown quantity ℓ is replaced with the original estimator H. Confidence intervals can be constructed in the same fashion. We mention two variants: the **normal** method and the **percentile** method. In the normal method, a $1 - \alpha$ confidence interval for ℓ is given by

$$(H \pm z_{1-\alpha/2}S^*) \ ,$$

where S^* is the bootstrap estimate of the standard deviation of H, that is, the square root of (7.8). In the percentile method, the upper and lower bounds of the $1 - \alpha$ confidence interval for ℓ are given by the $1 - \alpha/2$ and $\alpha/2$ quantiles of H, which in turn are estimated via the corresponding sample quantiles of the bootstrap sample $\{H_i^*\}$.

Example 7.6 (Bootstrapping Regression Data). Bootstrapping can be applied to a variety of statistical models, including regression data. Suppose that we have linear regression data $\{(x_i, y_i), i = 1, \ldots, 10\}$ given in Table 7.1.

Table 7.1 Regression data

x	y	x	y
13	5.0768	27	31.4085
16	21.1897	30	26.8648
19	17.1548	33	29.3894
21	22.8325	36	37.4476
24	26.5348	39	44.292

We wish to fit the data with a straight line. The least-squares method gives the following fitted regression line:

$$y = 3.3024\,x + 8.0561 \ .$$

We can assess the quality of this estimate by resampling the pairs $\{(x_i, y_i)\}$ independently and then estimating the regression lines for the resampled data. This is illustrated in Fig. 7.8 for 20 resampled regression lines. We see that there is quite a lot of variability in the estimate.
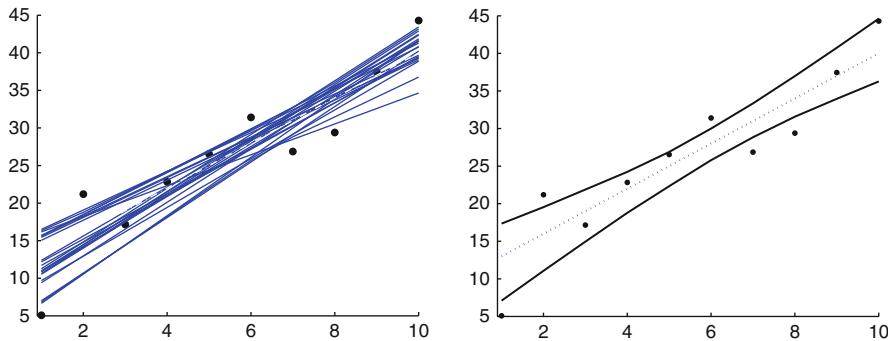

Fig. 7.8 *Left*: the linear regression data (10 points) and 20 resampled regression lines. *Right*: a 90 % bootstrapped confidence interval obtained from 1000 resampled regression lines

Let us determine percentile confidence intervals for the regression line $\beta_1 x + \beta_0$. We carry out the resampling many times, say, 1000 times, and calculate for each x the values $\widehat{\beta}_1 x + \widehat{\beta}_0$. A 90 % bootstrap confidence interval is then obtained by recording the 5 % and the 95 % quantile of these 1000 values for each x. The result is given in the right pane of Fig. 7.8. The straight line through the middle is the estimated regression line. The curved lines form the confidence interval—as a function of x.

Example 7.7 (Bootstrapping the Ratio Estimator). Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ be iid copies of a random vector (X, Y) with mean vector (μ_X, μ_Y) and covariance matrix Σ. Suppose we wish to estimate the ratio μ_X/μ_Y. A straightforward estimator is the so-called **ratio estimator $R = \bar{X}/\bar{Y}$**.

As a particular example, consider the data in Fig. 7.9, where a sample of size $N = 100$ of pairs (x, y) is plotted. The model that was used to generate the data is

$$X \sim \mathsf{N}(11, 25) \quad \text{and} \quad (Y \mid X = x) \sim \mathsf{U}(0, x) .$$

The estimate for μ_X/μ_Y is in this case $\bar{x}/\bar{y} = 2.0359$. But how accurate is this estimate? From Example 3.15 (delta method) we see that the estimator R has approximately a $\mathsf{N}(\mu_X/\mu_Y, \sigma^2/N)$ distribution, where the variance is given in (3.37). By replacing expectations, variances, and covariance with their sample means—that is, by using the method of moments—it is easy to estimate σ^2. The sample means and the covariance matrix of the $\{(X_i, Y_i)\}$ are in this case
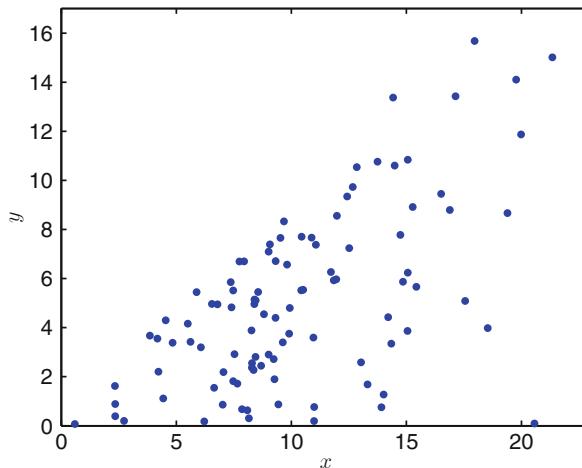
Fig. 7.9 An iid sample from a two-dimensional distribution

$$\bar{x} = 10.3026, \quad \bar{y} = 5.0604, \quad \text{and} \quad \widehat{\Sigma} = \begin{pmatrix} 19.7626 & 9.7052 \\ 9.7052 & 12.9859 \end{pmatrix},$$

which gives $\widehat{\sigma}^2 = 1.3305$. Thus R has approximately a $\mathsf{N}(2.0359, 1.3305)$ distribution. Its pdf is plotted in Fig. 7.10 (dotted graph). The 0.025 and 0.975 quantiles of this distribution give an approximate 95 % confidence interval for μ_X/μ_Y:

$$2.0359 \pm 1.96\sqrt{1.33051/100} = (1.81, 2.26) .$$

The above analysis requires a good deal of mathematical sophistication. In contrast, the application of the bootstrap method for this data is relatively easy: independently resample the data K times and plot a KDE of the ratios, as in the following MATLAB code.

```
%resampratio.m
N = 100; %size of data
K = 50000; %resample size
est = zeros(1,K);
xorg = 11 + 5*randn(1,N); yorg = rand(1,N).*xorg;
  %orig. data
estorg =  mean(xorg)/mean(yorg);
x = zeros(1,N); y = zeros(1,N);
est = zeros(1,K);
for i=1:K
    ind = ceil(N*rand(1,N)); % draw random indices
    x = xorg(ind); y = yorg(ind); % resampled  data
    est(i)  = mean(x)/mean(y);
end
kde(est)
```

Figure 7.10 shows the KDE for the bootstrapped sample of size $K = 50000$. We see that the density estimate is in excellent agreement with that of the delta method.
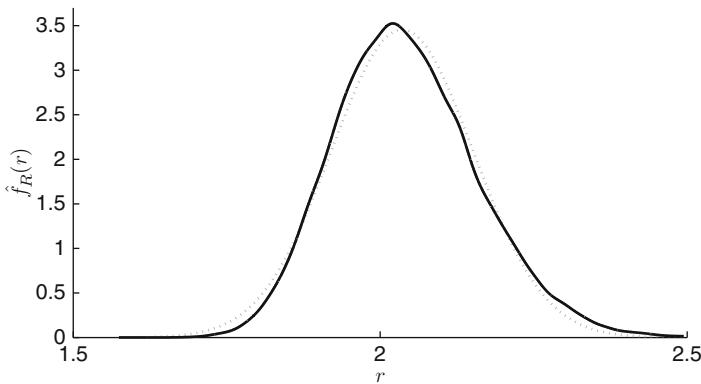
Fig. 7.10 Estimates of the pdf of the ratio estimator $R = \bar{X}/\bar{Y}$ using the delta method (*dotted line*) and the bootstrap method (*solid line*)

7.4 Markov Chain Monte Carlo

MCMC is a Monte Carlo sampling technique for (approximately) generating samples from an arbitrary distribution—often referred to as the **target** distribution. The basic idea is to run a Markov chain long enough such that its limiting distribution is close to the target distribution.

Before we discuss the method in more detail, let us go over some facts about Markov chains.

> **Definition 7.3. (Markov Chain).** A **Markov chain** is a collection $\{X_t, t = 0, 1, 2, \ldots\}$ of random variables (or random vectors) whose futures are conditionally independent of their pasts given their present values. That is,
>
> $$(X_{t+1} \mid X_s, s \leqslant t) \quad \sim \quad (X_{t+1} \mid X_t) \quad \text{for all } t . \tag{7.9}$$

In other words, the conditional distribution of the future variable X_{t+1}, given the entire past $\{X_s, s \leqslant t\}$, is the same as the conditional distribution of X_{t+1} given only the present X_t. Property (7.9) is called the **Markov property**.

The index t in X_t is usually seen as a "time" or "step" parameter. The index set $\{0, 1, 2, \ldots\}$ in the definition above was chosen out of convenience. It can be

replaced by any countable index set. We restrict ourselves to Markov chains for which the conditional pdfs $f_{X_{t+1} \mid X_t}(y \mid x)$ do not depend on t; we abbreviate these as $q(y \mid x)$. The $\{q(y \mid x)\}$ are called the **(one-step) transition densities** of the Markov chain. Note that the random variables or vectors $\{X_t\}$ may be *discrete* (e.g., taking values in some set $\{1, \ldots, r\}$) or *continuous* (e.g., taking values in an interval $[0, 1]$ or $\mathbb{R}^d$). In particular, in the *discrete* case, each $q(y \mid x)$ is a probability: $q(y \mid x) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$.

The distribution of X_0 is called the **initial distribution** of the Markov chain. The one-step transition densities and the initial distribution completely specify the distribution of the random vector $(X_0, X_1, \ldots, X_t)$. Namely, we have by the product rule (3.10) and the Markov property (7.9) that the joint pdf is given by

$$
\begin{aligned}
f_{X_0,\ldots,X_t}&(x_0, \ldots, x_t) \\
&= f_{X_0}(x_0)\, f_{X_1 \mid X_0}(x_1 \mid x_0) \cdots f_{X_t \mid X_{t-1},\ldots,X_0}(x_t \mid x_{t-1}, \ldots, x_0) \\
&= f_{X_0}(x_0)\, f_{X_1 \mid X_0}(x_1 \mid x_0) \cdots f_{X_t \mid X_{t-1}}(x_t \mid x_{t-1}) \\
&= f_{X_0}(x_0)\, q(x_1 \mid x_0)\, q(x_2 \mid x_1) \cdots q(x_t \mid x_{t-1}) \ .
\end{aligned}
$$

This leads to the following generic generation algorithm for Markov chains.

Algorithm 7.2. (Generating a Markov Chain). To generate a Markov chain $X_0, \ldots, X_N$ with transition densities $\{q(y \mid x)\}$ and initial pdf f_{X_0} execute the following steps:

1. Draw $X_0 \sim f_{X_0}$. Set $t = 0$.
2. Given $X_t = x$, draw $X_{t+1} \sim q(\cdot \mid x)$. Set $t = t + 1$.
3. If $t = N$, stop; otherwise, repeat from Step 2.

Example 7.8 (Stepping Stones). Imagine a pond with six stepping stones. From each stone one can step to a neighboring stone with a certain probability, indicated by the graph in Fig. 7.11. Let X_t be the position (stepping stone) after t steps, starting from position 1. Then, $\{X_t, t = 0, 1, 2, \ldots\}$ is a Markov chain. The graph in Fig. 7.11 is called the **transition graph** of the Markov chain. The arc weights indicate the transition probabilities. For example, $q(4 \mid 3) = 0.7$, $q(3 \mid 6) = 0.1$, and $q(3 \mid 4) = 0$.

The following MATLAB program generates the Markov chain for $N = 100$ steps. Note that the transition probabilities have been gathered into a matrix P, with $P(x, y) = q(y \mid x)$. P is called **one-step transition matrix** of the Markov chain. Given that $X_t = x$, state X_{t+1} is generated from the discrete distribution defined by the xth row of P. A typical outcome is depicted in Fig. 7.12. The program also keeps track of the total number of visits to each state—in this case there were 5, 2,

Fig. 7.11 The transition
graph for the Markov chain
$\{X_t, t = 0, 1, 2, \ldots\}$
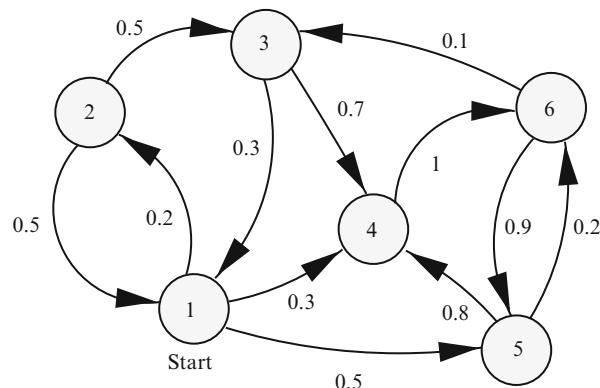
7, 29, 27, and 31 visits to states 1,...,6. We see that the Markov process spends most
of its time in states 4, 5, and 6.

```
%stepstone.m
N = 101; %time index starts at 1, not 0, so add 1.
P = [0,    0.2, 0,    0.3, 0.5, 0;
     0.5, 0,    0.5, 0,    0,    0;
     0.3, 0,    0,    0.7, 0,    0;
     0,    0,    0,    0,    0,    1 ;
     0,    0,    0,    0.8, 0,    0.2;
     0,    0,    0.1, 0,    0.9, 0];
x = zeros(1,N); x(1)= 1;
tot = zeros(1,6); tot(1) = 1;
for t=1:N-1    % generate the Markov chain
    x(t+1)  = min(find(cumsum(P(x(t),:))> rand));
    tot(x(t+1))  = tot(x(t+1)) + 1;
end
plot(0:N-1,x) % plot the path, subtracting 1 from time indices
tot/N  % fractions of visits to the states
```
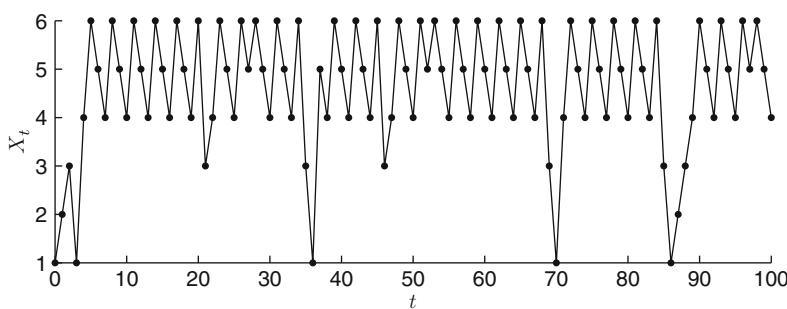

Fig. 7.12 A realization of the stepping stone Markov process $\{X_t, t = 0, 1, 2, \ldots, 100\}$

A Markov chain is said to be **ergodic** if the probability distribution of X_t converges to a fixed distribution as $t \rightarrow \infty$. Ergodicity is a natural property of Markov chains. For example, the Markov chain in Example 7.8 is ergodic. Intuitively, since this Markov chain cannot run off to infinity (which can only happen if the state space is infinite) and since each state can be reached from each other state, the probability $f_{X_t}(x) = \mathbb{P}(X_t = x)$ of encountering the chain in state x at time t far away in the future depends on x but not on t. In general, the pdf $f_{X_t}(x)$ of an ergodic Markov chain converges to a fixed **limiting pdf** $f(x)$ as $t \rightarrow \infty$, irrespective of the starting state. For the discrete case $f(x)$ corresponds to the long-run fraction of times that the Markov process visits x.

The limiting pdf $f(x)$ can be found by solving the **global balance equations**:

$$f(x) = \begin{cases} \sum_y f(y) \, q(x \mid y) & \text{(discrete case)}, \\ \int f(y) \, q(x \mid y) \, \mathrm{d}y & \text{(continuous case)}. \end{cases} \tag{7.10}$$

For the discrete case the rationale behind this is as follows. Since $f(x)$ is the long-run proportion of time that the Markov chain spends in x, the proportion of transitions *out of* x is $f(x)$. This should be balanced with the proportion of transitions *into* state x, which is $\sum_y f(y) \, q(x \mid y)$.

Example 7.9 (Limiting Probabilities for Stepping Stones Example). For the discrete case, the global balance equations can be written in matrix form as $\mathbf{f} = \mathbf{f}P$, where P is the one-step transition matrix and $\mathbf{f}$ the *row* vector of limiting probabilities. This leads to solving the linear equation $\mathbf{f}(I - P) = \mathbf{0}$ or equivalently $(I - P^\top)\mathbf{f}^\top = \mathbf{0}$, where I denotes the identity matrix. In other words, $\mathbf{f}^\top$ lies in the *null space* of $(I - P)^\top$. Also, the components of $\mathbf{f}$ must add to 1. By executing the following lines

```
f = null(eye(6) - P')';
f = f/sum(f);
```

appended to the MATLAB code in Example 7.8, we find the limiting probabilities $\mathbf{f} = (0.0120, 0.0024, 0.0359, 0.2837, 0.3186, 0.3474)$.

In MCMC one is often interested in a stronger type of balance equations. Imagine that we have taken a video of the evolution of the Markov chain, which we may run in forward and reverse time. If we cannot determine whether the video is running forward or backward (we cannot determine any systematic "looping"), the chain is said to be time-reversible or simply **reversible**.

Although not every Markov chain is reversible, each ergodic Markov chain, when run backward, gives another Markov chain—the **reverse Markov chain**— with transition densities $\widetilde{q}(y \mid x) = f(y) \, q(x \mid y)/f(x)$. To see this, first observe that $f(x)$ is the long-run proportion of time spent in x for both the original and reverse Markov chains. Secondly, the "probability flux" from x to y in the reversed chain must be equal to the probability flux from y to x in the original chain, meaning

$f(x)\,\widetilde{q}(y\,|\,x) \;=\; f(y)\,q(x\,|\,y)$, which yields the stated transition probabilities for the reversed chain. In particular, for a *reversible* Markov chain, we have

$$f(x)\,q(y\,|\,x) = f(y)\,q(x\,|\,y) \quad \text{for all } x, y \ . \tag{7.11}$$

These are the **detailed (or local) balance equations**. Note that the detailed balance equations imply the global balance equations. Hence, if a Markov chain is **irreducible** (i.e., every state can be reached from every other state) and there exists a pdf such that (7.11) holds, then $f(x)$ must be the limiting pdf. In the discrete state space case an additional condition is that the chain must be **aperiodic**, meaning that the return times to the same state cannot always be a multiple of some integer $\geqslant 2$; see Problem 7.13.

Example 7.10 (Random Walk on a Graph). Consider a Markov chain that performs a "random walk" on the graph in Fig. 7.13, at each step jumping from the current vertex (node) to one of the adjacent vertices, with equal probability. Clearly this Markov chain is reversible. It is also irreducible and aperiodic. Let $f(x)$ denote the limiting probability that the chain is in vertex x. By symmetry, $f(1) = f(2) = f(7) = f(8)$, $f(4) = f(5)$, and $f(3) = f(6)$. Moreover, by the detailed balance equations, $f(4)/5 = f(1)/3$, and $f(3)/4 = f(1)/3$. It follows that $f(1) + \cdots + f(8) = 4f(1) + 2 \times 5/3\, f(1) + 2 \times 4/3\, f(1) = 10\, f(1) = 1$, so that $f(1) = 1/10$, $f(2) = 2/15$, and $f(4) = 1/6$.

Fig. 7.13 The random walk on this graph is reversible
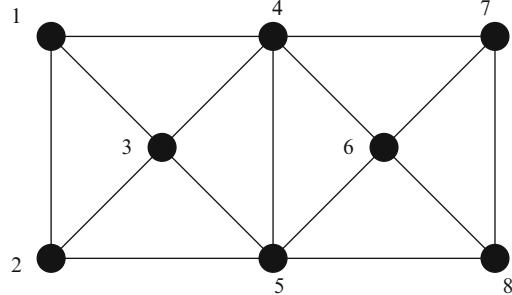
The idea behind MCMC can be summarized as follows. To draw approximately from an arbitrary pdf $f(x)$, run a Markov chain $\{X_t\}$ whose limiting distribution is $f(x)$. Often such a Markov chain is constructed to be reversible, so that the detailed balance equations (7.11) can be used. After a sufficiently long **burn-in period** from 0 to T, say, the random variables $X_{T+1}, X_{T+2}, \ldots$ form an *approximate* and *dependent* sample from $f(x)$.

In the next two sections we discuss two specific MCMC samplers: the Metropolis–Hastings sampler and the Gibbs sampler.

7.5   Metropolis–Hastings Algorithm

Suppose we wish to sample from a discrete pdf $f(x)$, where x takes values in the set $\{1, \ldots, r\}$. Following Metropolis et al. (1953), we construct a Markov chain $\{X_t, t = 0, 1, \ldots\}$ in such a way that its limiting pdf is f. Suppose the Markov chain is in state x at time t. A transition of the Markov chain from state x is carried out in two phases. Similar to the acceptance–rejection method, first, a *trial* or *proposal* state Y is drawn from a transition density $q(\cdot \mid x)$. This state is *accepted* as the new state, with probability $\alpha(x, Y)$, or *rejected* otherwise. In the latter case the chain remains in state x. For any outcome $Y = y$ the one-step transition probabilities of the Markov chain are thus

☞ 55

$$\widetilde{q}(y \mid x) = \begin{cases} q(y \mid x)\, \alpha(x, y), & \text{if } y \neq x \\ 1 - \sum_{z \neq x} q(z \mid x)\, \alpha(x, z), & \text{if } y = x \,. \end{cases} \tag{7.12}$$

By choosing the **acceptance probability** as

$$\alpha(x, y) = \min\left\{ \frac{f(y)\, q(x \mid y)}{f(x)\, q(y \mid x)}, 1 \right\}, \tag{7.13}$$

such a Markov chain can be made (see Problem 7.12) to satisfy the detailed balance equations (7.11):

$$f(x)\, \widetilde{q}(y \mid x) = f(y)\, \widetilde{q}(x \mid y) \quad \text{for all } x, y \,. \tag{7.14}$$

Consequently, if this Markov chain is irreducible and aperiodic, its limiting pdf is $f(x)$.

Note that in order to evaluate the acceptance probability $\alpha(x, y)$ in (7.13), we only need to know the target pdf $f(x)$ *up to a constant*; that is, $f(x) = c\, \bar{f}(x)$ for some known function $\bar{f}(x)$ but unknown constant c.

The extension of the above MCMC approach for generating samples from an arbitrary joint pdf $f(\mathbf{x})$ is straightforward, giving the following algorithm.

Algorithm 7.3. (Metropolis–Hastings Algorithm). Given a transition density $q(\mathbf{y} \mid \mathbf{x})$ and starting from an initial state $\mathbf{X}_0$, repeat the following steps from $t = 1$ to N:

1. Generate $\mathbf{Y} \sim q(\mathbf{y} \mid \mathbf{X}_t)$.
2. Generate $U \sim \mathsf{U}(0, 1)$ and set

(continued)

(continued)

$$\mathbf{X}_{t+1} = \begin{cases} \mathbf{Y}, & \text{if } U \leqslant \alpha(\mathbf{X}_t, \mathbf{Y}) \\ \mathbf{X}_t & \text{otherwise,} \end{cases} \tag{7.15}$$

where

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{ \frac{f(\mathbf{y})\, q(\mathbf{x}\,|\,\mathbf{y})}{f(\mathbf{x})\, q(\mathbf{y}\,|\,\mathbf{x})},\ 1 \right\}. \tag{7.16}$$

The above algorithm produces a sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of *dependent* random vectors, with $\mathbf{X}_t$ approximately distributed according to $f(\mathbf{x})$ for large t.

Since Algorithm 7.3 is of the acceptance–rejection type, its efficiency depends on the acceptance probability $\alpha(\mathbf{x}, \mathbf{y})$. Ideally, one would like the proposal transition density $q(\mathbf{y}\,|\,\mathbf{x})$ to reproduce the desired pdf $f(\mathbf{y})$ as faithfully as possible. Below we consider two particular choices of $q(\mathbf{y}\,|\,\mathbf{x})$.

Example 7.11 (Independence Sampler). The simplest Metropolis-type MCMC algorithm is obtained by choosing the proposal transition density $q(\mathbf{y}\,|\,\mathbf{x})$ to be independent of $\mathbf{x}$; that is, $q(\mathbf{y}\,|\,\mathbf{x}) = g(\mathbf{y})$ for some pdf $g(\mathbf{y})$. Thus, starting from a previous state $\mathbf{X}$, a candidate state $\mathbf{Y}$ is generated from $g(\mathbf{y})$ and accepted with probability

$$\alpha(\mathbf{X}, \mathbf{Y}) = \min\left\{ \frac{f(\mathbf{Y})\, g(\mathbf{X})}{f(\mathbf{X})\, g(\mathbf{Y})},\ 1 \right\}.$$

This procedure is very similar to the acceptance–rejection method of Sect. 2.7.3, and, as in that method, it is important that the proposal distribution g is close to the target f. Note, however, that in contrast to the acceptance–rejection method this **independence sampler** produces *dependent* samples.

As a particular example, consider the pdf

$$f(x) \propto x^2 \exp(-x^2 + \sin(x)), \quad x \in \mathbb{R},$$

where the normalization constant remains unspecified ($\propto$ means "is proportional to"). To sample from this pdf using the independence sampler we choose the symmetric proposal pdf $g(x) = \mathrm{e}^{-|x|}/2$, $x \in \mathbb{R}$. Drawing from this pdf is easy; see Problem 7.16. The program below provides a MATLAB implementation, and Fig. 7.14 shows a KDE of the data (as well as a graph of the true pdf f).
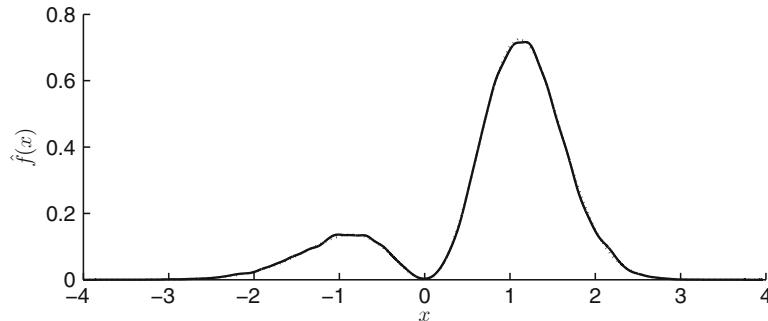
Fig. 7.14 The kernel density estimate $\widehat{f}(x)$ (*smooth curve*), obtained by the independence sampler, is practically indistinguishable from the target pdf $f(x)$ (*dotted curve*)

```
%indepsamp.m
N = 10^5;    % sample size
f = @(x) x.^2.*exp(-x.^2 + sin(x)); % unnormalized target pdf
g = @(x) exp(-abs(x))/2;            % proposal pdf
alpha = @(x,y) min(f(y)*g(x)/(f(x)*g(y)), 1);  % accept. prob.
x = 0; xx = zeros(1,N);
for t = 2:N
    y = -log(rand)*(2*(rand < 1/2) - 1);  % draw a proposal
    if rand < alpha(x,y)
        x = y;
    end
    xx(t) = x;
end
hold on
kde(xx(1:N));      % plot the kde of the data
c = quad(f,-5,5); % determine the normalization constant
tt = [-4:0.1:4]; plot(tt,f(tt)/c,'r') % plot the target pdf
hold off
```

Example 7.12 (Random Walk Sampler). A popular Metropolis–Hastings-type sampler is the **random walk sampler**. Here, the proposal state **Y**, for a given current state **x**, is given by **Y** = **x** + **Z**, where **Z** is typically generated from some spherically symmetric distribution, such as $\mathsf{N}(\mathbf{0}, I)$. In that case the proposal transition density pdf is symmetric; that is, $q(\mathbf{y} \,|\, \mathbf{x}) = q(\mathbf{x} \,|\, \mathbf{y})$. It follows that the acceptance probability is

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{ \frac{f(\mathbf{y})}{f(\mathbf{x})}, \, 1 \right\} . \qquad (7.17)$$

Example 7.13 (Sampling from a Pdf via Random Walk Sampler). Consider the two-dimensional pdf

$$f(x_1, x_2) = c \, \exp(-4(x_2 - x_1^2)^2 + (x_2 - 1)^2), \quad x_1 \in \mathbb{R}, \, x_2 \leqslant 2, \qquad (7.18)$$

where c is an unknown normalization constant. The graph of this pdf (unnormalized) is depicted in Fig. 7.15.

Fig. 7.15 The pdf $f(x_1, x_2)$
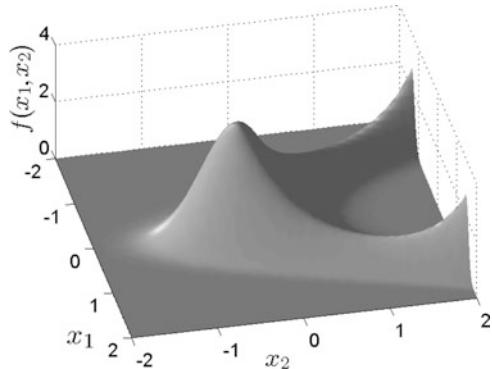
The following MATLAB program implements a random walk sampler to (approximately) draw $N = 10^4$ dependent samples from the pdf f. At each step, given a current state $\mathbf{x}$, a proposal $\mathbf{Y}$ is drawn from the $\mathrm{N}(\mathbf{x}, I)$ distribution. That is, $\mathbf{Y} = \mathbf{x} + \mathbf{Z}$, with $\mathbf{Z}$ bivariate standard normal. We see in Fig. 7.16 that the samples closely follow the contour plot of the pdf, indicating that the sampler works correctly. The starting point for the Markov chain is chosen as $(0, -1)$. Note that the normalization constant c is not used in the program.

```
%rwsamp.m
f=@(X1,X2)exp(-4*(X2-X1.^2).^2+(X2-1).^2).*(X2 < 2);
N = 10000; %sample size
xx = zeros(N,2); x = [0,-1]; xx(1,:) = x;
for i=2:N
    y = x + randn(1,2); %proposal
    alpha = min(f(y(1),y(2))/f(x(1),x(2)),1);
      %acceptance prob
    r = (rand < alpha);
    x = r*y + (1-r)*x; %next value of the Markov chain
    xx(i,:) = x;
end
plot(xx(:,1),xx(:,2),'.')
```

Fig. 7.16 Approximate
samples from pdf f produced
via the random walk sampler
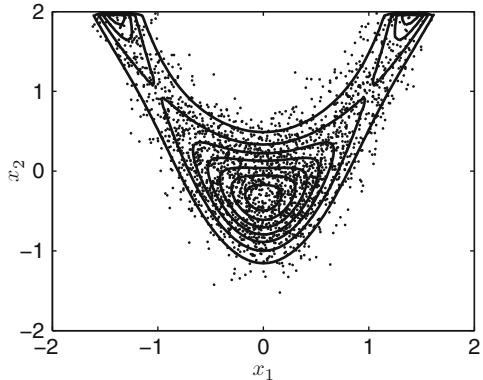
7.6   Gibbs Sampler

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector with joint pdf $f(\mathbf{x})$. Direct sampling from f may be difficult, especially if n is large. However, often sampling from the conditional pdf of X_i given $X_j = x_j$, $j \neq i$ is feasible. Let us denote these one-dimensional pdfs by $f_i(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, $i = 1, \ldots, n$. If drawing from each f_i is easy, then one can use the **Gibbs sampler** to construct a Markov chain $\mathbf{X}_1, \mathbf{X}_2, \ldots$ with limiting pdf f. This Markov chain is generated as follows. As in the Metropolis–Hastings sampler, at each step t, given a current state $\mathbf{X}_t = \mathbf{x}$, a proposal $\mathbf{Y}$ is drawn from a transition density $q_{1 \to n}(\mathbf{y} \mid \mathbf{x})$ given by

$$q_{1 \to n}(\mathbf{y} \mid \mathbf{x}) = f_1(y_1 \mid x_2, \ldots, x_n) f_2(y_2 \mid y_1, x_3, \ldots, x_n) \cdots f_n(y_n \mid y_1, \ldots, y_{n-1}) .$$

Thus, draw Y_1 from the conditional pdf $f_1(y_1 \mid x_2, \ldots, x_n)$, draw Y_2 from $f_2(y_2 \mid y_1, x_3, \ldots, x_n)$, and so on. However, unlike the Metropolis–Hastings sampler, this proposal is *always accepted*; so $\mathbf{X}_{t+1} = \mathbf{Y}$. The algorithm is summarized as follows.

Algorithm 7.4. (Gibbs Sampler). Starting from an initial state $\mathbf{X}_0$, repeat the following steps from $t = 1$ to N:

1. Given the current state $\mathbf{X}_t = \mathbf{x}$, generate $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as follows:

   a. Draw $Y_1 \sim f_1(y_1 \mid x_2, \ldots, x_n)$.
   b. Draw $Y_i \sim f_i(y_i \mid Y_1, \ldots, Y_{i-1}, x_{i+1}, \ldots, x_n)$, $i = 2, \ldots, n-1$.
   c. Draw $Y_n \sim f_n(y_n \mid Y_1, \ldots, Y_{n-1})$.

2. Set $\mathbf{X}_{t+1} = \mathbf{Y}$.

To verify that the Markov chain $\mathbf{X}_0, \mathbf{X}_1, \ldots$ indeed has limiting pdf $f(\mathbf{x})$, we need to check that the global balance equations (7.10) hold. In general the detailed balance equations (7.11) do *not* hold—$f(\mathbf{x})\, q_{1 \to n}(\mathbf{y} \mid \mathbf{x}) \neq f(\mathbf{y})\, q_{1 \to n}(\mathbf{x} \mid \mathbf{y})$. However, a similar result, due to Hammersley and Clifford, does hold: if $q_{n \to 1}(\mathbf{x} \mid \mathbf{y})$ denotes the transition density of the reverse move, in the order $n \to n-1 \to \cdots \to 1$, that is,

$$q_{n \to 1}(\mathbf{x} \mid \mathbf{y})$$
$$= f_n(x_n \mid y_1, \ldots, y_{n-1})\, f_{n-1}(x_{n-1} \mid y_1, \ldots, y_{n-2}, x_n) \cdots f_1(x_1 \mid x_2, \ldots, x_n) ,$$

then

$$f(\mathbf{x})\, q_{1 \to n}(\mathbf{y} \mid \mathbf{x}) = f(\mathbf{y})\, q_{n \to 1}(\mathbf{x} \mid \mathbf{y}) . \tag{7.19}$$

Intuitively, the long-run proportion of transitions $\mathbf{x} \to \mathbf{y}$ for the "forward move" chain is equal to the long-run proportion of transitions $\mathbf{y} \to \mathbf{x}$ for the "reverse move" chain. By integrating (in the continuous case) both sides in (7.19) with respect to $\mathbf{x}$, we see that the global balance equations hold:

$$\int f(\mathbf{x})\, q_{1 \to n}(\mathbf{y} \mid \mathbf{x})\, \mathrm{d}\mathbf{x} = f(\mathbf{y}) .$$

Example 7.14 (Sampling from Pdf via Gibbs Sampler). Consider the two-dimensional pdf

$$f(x_1, x_2) = c\, \mathrm{e}^{-x_1 x_2 - x_1 - x_2}, \quad x_1 \geqslant 0, x_2 \geqslant 0 ,$$

where the normalization constant c remains unspecified. Let (X_1, X_2) be distributed according to f. The conditional pdf of X_1 given $X_2 = x_2$ is

$$f_1(x_1 \mid x_2) \stackrel{\text{def}}{=} f_{X_1 \mid X_2}(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)} \propto f(x_1, x_2) \propto \mathrm{e}^{-x_1(x_2+1)} .$$

It follows that X_1 given $X_2 = x_2$ has an $\mathsf{Exp}(x_2 + 1)$ distribution and, by symmetry, X_2 given $X_1 = x_1$ has an $\mathsf{Exp}(x_1 + 1)$ distribution. Sampling from the joint pdf can thus be established via the Gibbs sampler by alternately generating from $\mathsf{Exp}(x_2+1)$ and $\mathsf{Exp}(x_1 + 1)$, as implemented in the following MATLAB program.

```
%gibbssamp.m
f = @(X1,X2)exp(-(X1.*X2 + X1 +X2)).*(X1 > 0 & X2 > 0);
N = 10^4; x = zeros(N,2); x2 = 1;
for i = 2:N
    x1 = -log(rand)/(x2+1);
    x2 = -log(rand)/(x1+1);
    x(i,:) = [x1,x2];
end
plot(x(:,1),x(:,2),'.')
```

7.7 Problems

7.1. Consider the estimation of the p-value in Example 7.1

a. Under H_0 we have $\mu = 0$, but σ^2 remains unspecified. Why is it allowed to take $\sigma = 1$ to generate the sample $T_1, \ldots, T_N$?

b. Show, using Theorem 5.1, that T under H_0 has a t_3 distribution, and calculate the true p-value.

c. Speed up the given MATLAB code by "vectorizing" the slow for loop.

7.2. Monte Carlo sampling methods are useful for calculating p-values for a **goodness-of-fit test**, where the data $X_1, \ldots, X_k$ is assumed to come from a multinomial distribution $\mathsf{Mnom}(n, p_1, \ldots, p_k)$.

As an example, consider a racetrack with 8 starting boxes. Out of 200 races, the numbers of winning horses that started from boxes $1, 2, \ldots, 8$ are 39, 29, 24, 20, 21, 24, 21, and 22, respectively. Is this an indication that the winning probabilities $p_1, \ldots, p_8$ are not all equal to $1/8$? The test statistic that is typically used for a goodness-of-fit test is

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \ ,$$

where O_i ($=X_i$) is the *observed* number of observations in class i and E_i ($= \mathbb{E}X_i$) is the *expected* number of observations in class i. In this case $E_i = 25$ for all i and the observed counts are given above. The hypothesis $H_0 : p_1 = \ldots = p_8 = 1/8$ is rejected in favor of the negation of H_0 for large values of the test statistic.

a. Write a Monte Carlo sampling program to estimate the p-value for this test. Do you reject the null hypothesis or not? Hint: to draw a vector $\mathbf{X} = (X_1, \ldots, X_8) \sim \mathsf{Mnom}(200, 1/8, \ldots, 1/8)$, you can use winner = ceil(8*rand(200,1)); X = histc(winner,1:8);.

b. It can be shown that under H_0 the test statistic T has approximately a χ_7^2 distribution. Verify this by drawing an iid sample from T and comparing the empirical cdf with that of the χ_7^2 distribution.

7.3. Let F_N be the empirical cdf of $x_1, \ldots, x_N$, and let X be a random variable with cdf F_N. Show that $\mathbb{E}X = \bar{x}$ and $\mathrm{Var}(X) = \sum_{i=1}^{N}(x_i - \bar{x})^2/N$, where $\bar{x} = (x_1 + \cdots + x_N)/N$.

7.4. Consider a mixture pdf

$$f(x) = w_1 f_1(x) + \cdots + w_k f_k(x), \quad w_j \geqslant 0, \; j = 1, \ldots, k, \quad \sum_{j=1}^{k} w_j = 1, \tag{7.20}$$

where each f_j is itself a pdf. Let J be a discrete random variable taking values $1, \ldots, k$ with probabilities $w_1, \ldots, w_k$, respectively. Let X be a random variable such that the conditional pdf of X given $J = j$ is f_j.

a. Show that X has mixture pdf (7.20).
b. Using (a.) describe how one could generate a random variable from the mixture pdf (7.20).
c. Suppose pdf f_j has mean μ_j and variance σ_j^2, $j = 1, \ldots, n$. Express $\mathbb{E}X$ and $\mathrm{Var}(X)$ in terms of these parameters.

7.5. It can be shown that the Kolmogorov–Smirnov statistic D_N in (7.6) satisfies   ☞ 199

$$\lim_{N \to \infty} \mathbb{P}(\sqrt{N}\, D_N \leqslant x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2(kx)^2}, \quad x > 0. \tag{7.21}$$

Compare the estimated p-value in Example 7.3 with an approximated one calculated via (7.21).

7.6. In Example 7.5 we considered the quality of various estimators for the mode of the Cauchy distribution using (re)sampling techniques.   ☞ 203

a. Instead of estimating the pdf of the sample mean using resampled data (dashed line in Fig. 7.7) estimate the pdf of the sample mean by sampling new data from the Cauchy distribution. How do the KDEs compare?
b. Another possible estimator for the mode of the Cauchy distribution is the **trimmed mean** estimator, which is given by the sample mean of all outcomes x_i with $|x_i| \leqslant \beta$, where β is some positive number. Carry out a bootstrap procedure for the trimmed mean with $\beta = 100$ and $\beta = 10$. How do the pdfs compare with those of the sample median and sample mean?

7.7. In Example 7.7 we saw that for a sample size of $N = 100$ the bootstrap and delta method gave identical results for the ratio estimator $\bar{X}/\bar{Y}$. Repeat the analysis and compare the two methods for a sample size $N = 10$, with x-values   ☞ 207

```
16.4321    2.4334    14.3433    7.9650    14.1052
6.7660     0.1430    10.0420    7.1071    13.5305
```

and y-values

```
14.9151    0.4312    11.5407    4.4538    8.7741
0.8462     0.0302     1.7955    1.4568    7.8052
```

7.8. The **median** of a distribution with pdf f is the number m such that $\int_{-\infty}^{m} f(x)\,dx = 1/2$. The data

```
1.4066     1.2917     1.4080     4.2801     1.2136    2.7461
11.1076    0.9247     5.8833    10.2513     3.8285    3.2116
0.5451     0.9896     1.1602     7.7723     0.1702    0.8907
0.2276     3.1197    11.4909     0.6475    11.2279    0.7639
```

form an iid sample from an $\mathsf{Exp}(\lambda)$ distribution.

a. Show that the median of $\mathsf{Exp}(\lambda)$ is $\ln(2)/\lambda$.
b. This suggests that we could estimate λ via the estimator $T = \ln(2)/\widetilde{X}$, where $\widetilde{X}$ is the sample median. Find the corresponding estimate. What is the maximum likelihood estimate of λ?
c. Carry out a bootstrap analysis of both estimators and compare their accuracies.

7.9. The concentration of a certain chemical is measured at times $1, 2, 3, \ldots, 20$. The measurements are

```
18.9506    41.4228    52.0253    63.5451    71.9634
79.0504    80.9685    84.6222    89.6391    93.5085
95.8680    91.3177    97.7423    97.1969    96.7448
96.8155    96.4435    98.2087    98.3126    97.8173
```

(e.g., at time $t = 12$ the concentration is 91.3177). Suppose the data are modeled by the following nonlinear regression model:

$$Y_i = a\,(1 - e^{-b\,t_i}) + \varepsilon_i\ , \quad i = 1, \ldots, n\ , \tag{7.22}$$

where $\{\varepsilon_i\} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ and a, b, and σ^2 are unknown. To fit the model (7.22) to the points $\{(t_i, y_i)\}$ we can apply a *least-squares* approach, where a and b are chosen such that the sum of the squared deviations,

$$r(a, b) = \sum_{i=1}^{n}(y_i - a\,(1 - e^{-b\,t_i}))^2\ , \tag{7.23}$$

is minimized. This requires numerical minimization.

a. Plot the points (t_i, y_i), $i = 1, \ldots, n = 20$.
b. Show that the values $\widehat{a}$ and $\widehat{b}$ that minimize the function r in (7.23) are the maximum likelihood estimates of a and b. Express the maximum likelihood estimate of σ^2 in terms of $\widehat{a}$ and $\widehat{b}$.
c. Implement a MATLAB program to find the optimal values $\widehat{a} = 99.14$ and $\widehat{b} = 0.255$, using the following code snippet (`yorg` and `torg` store the original data):

```
r  =  @(x) sum((yorg - x(1)*(1-exp(-x(2)*torg))).^2);
[xmin,fval]  = fminsearch(r,[100,1])
```

d. To assess how accurate the estimates for a and b are, resample the data 1000 times. For each resampled data set estimate a and b via `fminsearch`, as above. Plot KDEs for the pdfs of $\widehat{a}$ and $\widehat{b}$, and determine 95 % bootstrap intervals for a and b.

7.10. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random samples from the $\text{Exp}(\lambda)$ and $\text{Exp}(\mu)$ distribution, respectively, for unknown λ and μ. Suppose outcomes of $X_1, \ldots, X_n$ are given by the data in Problem 7.8 (so $n = 24$) and outcomes of $Y_1, \ldots, Y_n$ are

23.9618	4.9055	6.0424	0.5870	4.0856	1.6503
10.1976	4.0208	25.9484	15.3954	19.5160	0.5937
11.5481	18.3895	30.4093	7.6527	9.7329	8.6130
6.2353	5.5157	9.9489	21.3850	5.1142	28.2284

The maximum likelihood estimator for $\ell = \lambda/\mu$ is $\sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} X_i$. Find a 95 % bootstrap confidence interval (percentile method) for ℓ.

7.11. Let $X_1, \ldots, X_n$ be an iid sample from a $\text{U}(0, \theta)$ distribution, where $\theta > 0$ is unknown. The maximum likelihood estimator of θ is $M = \max\{X_1, \ldots, X_n\}$. Suppose $M_1^*, \ldots, M_K^*$ is a bootstrap sample of M, based on an outcome $x_1, \ldots, x_n$. Explain why it is a bad idea to construct a confidence interval for θ on the basis of the $\{M_i^*\}$.

7.12. For the Metropolis–Hastings sampler, verify that the local balance equations (7.14) hold if the acceptance probability is chosen as in (7.13). Hint: consider two cases: $f(y)q(x\,|\,y) \leqslant f(x)q(y\,|\,x)$ and $f(y)q(x\,|\,y) \geqslant f(x)q(y\,|\,x)$.

7.13. Let $X_t, t = 0, 1, 2, \ldots$ be a random walk on the graph in Fig. 7.17. From each state the random walk chooses one of the adjacent states with equal probability. The starting state is 1.

a. Is the chain irreducible and aperiodic?
b. Do the local balance equations hold? If so, find the solution $f(1), \ldots, f(6)$.
c. Explain why the probabilities $\mathbb{P}(X_t = x)$, $x = 1, \ldots, 6$, do not converge as $t \to \infty$.

Fig. 7.17 The graph on
which the random walk is
performed
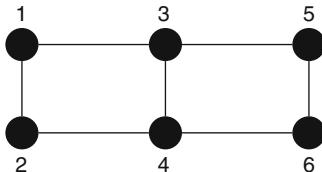
7.14. Run the random walk sampler with a $N(10, 2)$ target distribution and $N(x, 0.01)$ proposal, drawing the initial point from the $N(0, 0.01)$ distribution. Take a sample size of $N = 5000$ and plot $\{X_t\}$ against $t = 1, \ldots, N$. Based on the graph give a rough estimate of the burn-in period.

We can estimate $\ell = \mathbb{E} \ln(X^2)$, where $X \sim N(10, 2)$, by taking the sample average of $\ln(X_{B+1}^2), \ldots, \ln(X_N^2)$, where B is the burn-in size. By independently generating $K = 100$ such estimates, find an approximate 95 % confidence interval for ℓ. Generate 20 such intervals and show that the true value for ℓ (which is $4.58453\ldots$) is contained in these intervals with a probability much smaller than 95 %. Hence, the combination of a burn-in size of $B = 1000$ and a sample size of $N = 5000$ is inadequate to provide an accurate estimate for ℓ.

7.15. Let $\mathscr{X}$ be a finite set on which a *neighborhood* structure is defined; that is, each $\mathbf{x} \in \mathscr{X}$ has a set of neighbors $\mathcal{N}(\mathbf{x})$. Let $n_\mathbf{x}$ be the number of neighbors of $\mathbf{x} \in \mathscr{X}$. Consider a Metropolis–Hastings algorithm with proposal density $q(\mathbf{y} \mid \mathbf{x}) = 1/n_\mathbf{x}$ for all $\mathbf{y} \in \mathcal{N}(\mathbf{x})$. That is, from a current state $\mathbf{x}$, the proposal state is drawn from the set of neighbors with equal probability. Let the acceptance probability be $\alpha(\mathbf{x}, \mathbf{y}) = \min\{n_\mathbf{x}/n_\mathbf{y}, 1\}$.

Assuming the chain is irreducible and aperiodic, what is its limiting distribution?

7.16. Let $U_1, U_2 \sim_{\text{iid}} U(0, 1)$. Explain why $X = -\ln U_1 \times (2I_{\{U_2 \leqslant 1/2\}} - 1)$ has pdf $g(x) = e^{-|x|}/2, x \in \mathbb{R}$.

7.17. A **Langevin Metropolis–Hastings** sampler is a random walk sampler where the proposal state, for a current state $\mathbf{x}$, is given by

$$\mathbf{Y} = \mathbf{x} + \frac{h}{2} \nabla \ln f(\mathbf{x}) + \sqrt{h}\, \mathbf{Z}, \quad \mathbf{Z} \sim N(\mathbf{0}, I) \,,$$

where $h > 0$ is a step size, f is the target pdf, and $\nabla \ln f$ is the gradient of $\ln f$. Note that the proposal distribution is not symmetric around $\mathbf{x}$. Use this sampler to draw $N = 10^5$ dependent samples from the Gamma$(2, 1)$ distribution. Use the kde.m program to assess how well the estimated pdf fits the true pdf. Investigate how the step size h and the length of the burn-in period affect the fit.

Since the `kde.m` program is designed for independent data and MCMC data has typically many duplicate values, the KDEs will typically be very "spiky." To produce smoother estimates, reduce the number of mesh points.

7.18. Let $\mathbf{X} = (X, Y)^\top$ be a random column vector with a bivariate normal distribution with expectation vector $\mathbf{0} = (0, 0)^\top$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix} .$$

a. Show that $(Y \mid X = x) \sim \mathsf{N}(\varrho\, x, 1 - \varrho^2)$ and $(X \mid Y = y) \sim \mathsf{N}(\varrho\, y, 1 - \varrho^2)$.
b. Write a Gibbs sampler to draw 10^4 samples from the bivariate distribution $\mathsf{N}(\mathbf{0}, \Sigma)$ and plot the data for $\varrho = 0, 0.7$, and 0.9.

7.19. Consider the two-dimensional pdf

$$f(\mathbf{x}) = c \, \exp(-(x_1^2 x_2^2 + x_1^2 + x_2^2 - 8x_1 - 8x_2)/2) , \quad \mathbf{x} \in \mathbb{R}^2 . \qquad (7.24)$$

a. Give a 3D plot and a contour plot for this function (ignoring c).
b. Implement a random walk sampler with proposals of the form $\mathbf{Y} = \mathbf{x} + \sigma\mathbf{Z}$, where $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I)$. Start the sampler at the point $(0, 4)$.
c. Plot the progression of the first component of the Markov chain against time, for $\sigma = 0.2$ and $\sigma = 2$. Comment on the difference.
d. Give a KDE of the pdf X_1 if $\mathbf{X} = (X_1, X_2) \sim f$.

7.20. Consider the two-dimensional pdf (7.24) in Problem 7.19.

a. Show that conditional on $X_2 = x_2$, X_1 has a normal distribution with expectation $4/(1 + x_2^2)$ and variance $1/(1 + x_2^2)$.
b. Implement a Gibbs sampler to sample from f.

7.21. In Algorithm 7.4 the vector $\mathbf{X}$ is updated in a *systematic* order: $1, 2, \ldots, n, 1, 2, \ldots$. A variant of the algorithm is to update the coordinates in *random* order. Specifically, Step 1 of the algorithm is replaced by:

Given the current state $\mathbf{X}_t = \mathbf{x}$, generate $\mathbf{Y}$ as follows:

1. Draw J uniformly from $\{1, \ldots, n\}$.
2. Given $J = j$, draw $Y_j \sim f_j(y_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$.
3. For $i \neq j$ set $Y_i = x_i$.

a. Show that, given $\mathbf{X}_t = \mathbf{x}$, $\mathbf{Y}$ has pdf (in the continuous case)

$$q(\mathbf{y} \mid \mathbf{x}) = \frac{1}{n} \frac{f(\mathbf{y})}{\int_{-\infty}^{\infty} f(\mathbf{y}) \, \mathrm{d}y_j} , \tag{7.25}$$

where $\mathbf{y} = (x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n)$.

b. Show that the random-order Gibbs sampler can be viewed as an instance of the Metropolis–Hastings sampler, with transition density $q(\mathbf{y} \mid \mathbf{x})$ given in (7.25) and with acceptance probability $\alpha(\mathbf{x}, \mathbf{y}) = 1$.

Chapter 8
Bayesian Inference

Bayesian statistics is a branch of statistics that is centered around Bayes' formula
(1.8), which is repeated in (8.1) below. To fully appreciate Bayesian inference, it
is important to understand that the type of statistical reasoning here is somewhat
different from that in classical statistics. In particular, model parameters are usually
treated as *random* rather than fixed quantities. Moreover, Bayesian statistics uses a
notation system that deviates from the classical one in two aspects:

1. Pdfs and conditional pdfs always use the *same letter* f (sometimes p is used
   instead of f). For example, instead of writing $f_X(x)$ and $f_Y(y)$ for the pdfs
   of X and Y, one simply writes $f(x)$ and $f(y)$. Similarly, the conditional pdf
   $f_{X\,|\,Y}(x\,|\,y)$ of X given Y is denoted in Bayesian notation as $f(x\,|\,y)$. This
   style of notation can be of great descriptive value, despite its apparent ambiguity,
   and we will use it in this book whenever we work in a Bayesian setting. As an
   example, the Bayesian formula (1.8) in terms of (conditional) pdfs can be written
   in Bayesian notation as

$$f(\mathbf{y}\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\mathbf{y})\,f(\mathbf{y})}{\int f(\mathbf{x}\,|\,\mathbf{y})\,f(\mathbf{y})\,\mathrm{d}\mathbf{y}} \propto f(\mathbf{x}\,|\,\mathbf{y})\,f(\mathbf{y})\,. \tag{8.1}$$

   (Replace the integral with a sum in the discrete case.)
2. In Bayesian statistics the notation does not make a distinction between random
   variables and their outcomes. Both are usually indicated by *lowercase* letters. It
   is assumed that it is clear from the context whether a variable x or θ should be
   interpreted as an outcome (a number) or a random variable.

The general framework for Bayesian statistics is as follows (compare with the
classical framework in Chap. 5): It is assumed that the data vector, $\mathbf{x}$, say, has been
drawn from a conditional pdf $f(\mathbf{x}\,|\,\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a random vector of parameters.
The pdf of $\boldsymbol{\theta}$ conveys the *a priori* (existing beforehand, before any experience)
information about $\boldsymbol{\theta}$. Observing the data $\mathbf{x}$ will affect our knowledge of $\boldsymbol{\theta}$, and the

way to update this information is to use Bayes' formula (8.1). The main concepts are summarized in the following definition.

Definition 8.1. (Prior, Likelihood, and Posterior). Let $\mathbf{x}$ and $\boldsymbol{\theta}$ denote the data and parameters in a Bayesian statistical model.

- The pdf of $\boldsymbol{\theta}$ is called the **prior** pdf.
- The conditional pdf $f(\mathbf{x} \mid \boldsymbol{\theta})$ is called the Bayesian **likelihood** function.
- The central object of interest is the **posterior** pdf $f(\boldsymbol{\theta} \mid \mathbf{x})$ which, by Bayes' theorem, is proportional to the product of the prior and likelihood:

$$f(\boldsymbol{\theta} \mid \mathbf{x}) \propto f(\mathbf{x} \mid \boldsymbol{\theta}) \, f(\boldsymbol{\theta}) \, .$$

The posterior pdf thus conveys the knowledge of $\boldsymbol{\theta}$ after taking into account the information $\mathbf{x}$. Note that the likelihood function in Bayesian statistics differs slightly from that in classical statistics. In Bayesian statistics the likelihood $f(\mathbf{x} \mid \boldsymbol{\theta})$ is a conditional pdf of the data $\mathbf{x}$, whereas in the classical case the likelihood $L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$ is viewed as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$. The posterior pdf can be viewed as a scaled version of the classical likelihood. Indeed, if the prior pdf is constant, then the posterior pdf coincides with the classical likelihood, up to a multiplicative constant.

☞ 161

Example 8.1 (Bayesian Inference for Coin Toss Experiment). Consider the basic random experiment where we toss a biased coin n times. Suppose that the outcomes are $x_1, \ldots, x_n$, with $x_i = 1$ if the ith toss is Heads and $x_i = 0$ otherwise, $i = 1, \ldots, n$. Let θ denote the probability of Heads. We wish to obtain information about θ from the data $\mathbf{x} = (x_1, \ldots, x_n)$. For example, we wish to construct a confidence interval.

The *a priori* information about θ is described by the prior pdf $f(\theta)$. For example, the choice of a *uniform* prior $f(\theta) = 1, 0 \leqslant \theta \leqslant 1$ indicates no prior knowledge about θ. We assume that conditional on θ the $\{x_i\}$ are independent and $\mathsf{Ber}(\theta)$ distributed. Thus, the Bayesian likelihood is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^s (1 - \theta)^{n-s} \, ,$$

where $s = x_1 + \cdots + x_n$ represents the total number of successes. Using a uniform prior gives the posterior pdf

$$f(\theta \mid \mathbf{x}) = c \, \theta^s (1 - \theta)^{n-s} \, , \quad 0 \leqslant \theta \leqslant 1 \, .$$

This is the pdf of the $\mathsf{Beta}(s+1, n-s+1)$ distribution. The normalization constant ☞ 74
is $c = (n+1)\binom{n}{s}$. The graph of the posterior pdf for $n = 100$ and $s = 1$ is given in
Fig. 8.1.

Fig. 8.1 Posterior pdf for θ, with $n = 100$ and $s = 1$
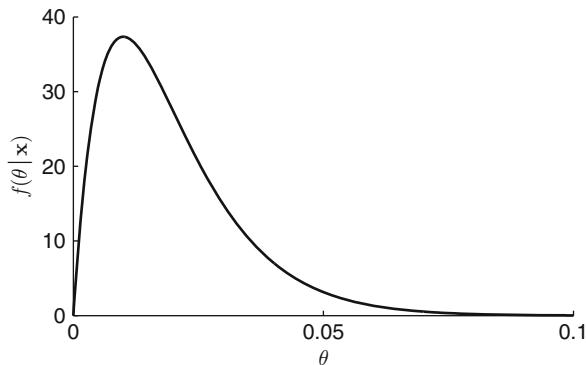
A Bayesian confidence interval, called a **credible interval**, for θ is formed by
taking the appropriate quantiles of the posterior pdf. As an example, suppose that
$n = 100$ and $s = 1$. Then, a left one-sided 95% credible interval for θ is $[0, 0.0461]$,
where 0.0461 is the 0.95 quantile of the $\mathsf{Beta}(2, 100)$ distribution. As an estimate for
θ, one often takes the **posterior mean**, that is, the expectation corresponding to the
posterior pdf. In this case, for general n and s, the posterior mean is $(s + 1)/(s + 1 + n - s + 1) = (s + 1)/(n + 2)$; see also Problem 8.1. An alternative estimate ☞ 256
for θ is the value for which the posterior pdf is maximal—the so-called **posterior
mode**. The posterior mode is here $\widehat{\theta} = s/n$, which coincides with the (classical)
sample mean.

8.1 Hierarchical Bayesian Models

In the coin flipping example both the parameter θ and the data $\mathbf{x}$ are random
variables, and the joint distribution of θ and $\mathbf{x}$ is specified in a "hierarchical" way:

$$\theta \sim f(\theta)$$
$$(\mathbf{x} \mid \theta) \sim f(\mathbf{x} \mid \theta) .$$

By the product rule of probability, the joint pdf is simply the product $f(\theta) f(\mathbf{x} \mid \theta)$,
and the posterior pdf is proportional to this last product (viewed as a function of θ).
For models involving more than one parameter, a similar hierarchical structure is
often used to specify the model. For example, a three-parameter model could be
specified as follows:

$$\alpha \sim f(\alpha)$$
$$(\beta \,|\, \alpha) \sim f(\beta \,|\, \alpha)$$
$$(\gamma \,|\, \alpha, \beta) \sim f(\gamma \,|\, \alpha, \beta)$$
$$(\mathbf{x} \,|\, \alpha, \beta, \gamma) \sim f(\mathbf{x} \,|\, \alpha, \beta, \gamma) \,.$$

That is, first, specify the prior pdf of α, and then given α, specify the pdf of β until finally the likelihood as a function of all the parameters is given. Often in practice the reverse order is used: the likelihood is specified first and the priors are defined last. The hierarchical model approach allows for an easy evaluation of the joint pdf: it is simply the product of the (conditional) pdfs:

☞ 72

$$f(\mathbf{x}, \alpha, \beta, \gamma) = f(\mathbf{x} \,|\, \alpha, \beta, \gamma) \, f(\gamma \,|\, \alpha, \beta) \, f(\beta \,|\, \alpha) \, f(\alpha) \,.$$

To find the posterior

$$f(\alpha, \beta, \gamma \,|\, \mathbf{x}) \,,$$

view $f(\mathbf{x}, \alpha, \beta, \gamma)$ as a function of α, β, and γ for fixed $\mathbf{x}$. To find the marginal posterior pdfs, $f(\alpha \,|\, \mathbf{x})$, $f(\beta \,|\, \mathbf{x})$, $f(\gamma \,|\, \mathbf{x})$, *integrate out* the other parameters. For example,

☞ 70

$$f(\gamma \,|\, \mathbf{x}) = \iint f(\alpha, \beta, \gamma \,|\, \mathbf{x}) \, d\alpha \, d\beta \,.$$

☞ 218   This may not always be easy or feasible. An alternative is to use the *Gibbs sampler* to sample from the posterior pdf. After initializing α, β, γ, iterate the following steps:

1. Draw α from $f(\alpha \,|\, \beta, \gamma, \mathbf{x})$.
2. Draw β from $f(\beta \,|\, \alpha, \gamma, \mathbf{x})$.
3. Draw γ from $f(\gamma \,|\, \alpha, \beta, \mathbf{x})$.

After a (dependent) sample $\{(\alpha_t, \beta_t, \gamma_t)\}$ from $f(\alpha, \beta, \gamma \,|\, \mathbf{x})$ is generated, output only the variables of interest, e.g., only the $\{\alpha_t\}$.

Example 8.2 (Ticket Inspector). A ticket inspector has the option of taking three different routes for inspection of parking violations. Each route is characterized by the time it takes to complete the route and the intensity of ticket violations. Suppose the time t spent on route k is exponentially distributed with mean $k/2$ (hours), $k = 1, 2, 3$. For example, route 2 takes on average 1 h to complete. Suppose further that the number of traffic violations encountered, x, say, has a Poisson distribution with mean $10 \, k \, t$. So if route 3 takes 2 h, an average of 60 tickets will be issued. Suppose that on a particular day the ticket inspector has issued 60 tickets. Which route has she/he taken?

Assuming that our prior information about k is that each of the routes is taken with equal probability, we obtain the following hierarchical model:

$$k \sim \mathsf{DU}\{1, 2, 3\} \quad \text{(discrete uniform)}$$
$$(t \mid k) \sim \mathsf{Exp}(2/k)$$
$$(x \mid k, t) \sim \mathsf{Poi}(10\,k\,t) \;.$$

It follows that the joint pdf is

$$f(k, t, x) = f(k)\,f(t \mid k)\,f(x \mid t, k) = \frac{1}{3}\,\frac{2}{k}\,\mathrm{e}^{-\frac{2}{k}t}\mathrm{e}^{-10kt}\frac{(10kt)^x}{x!} \tag{8.2}$$

for $k = 1, 2, 3$, $t \geqslant 0$, and $x = 0, 1, 2, \ldots$. Note that k and x are discrete random variables and t is continuous. The posterior pdf $f(k, t \mid x = 60)$ is thus of the form

$$f(k, t \mid x = 60) \propto \frac{1}{k}\,\mathrm{e}^{-\frac{2}{k}t}\mathrm{e}^{-10kt}\,(kt)^{60} \;. \tag{8.3}$$

The marginal posterior pdf of k can be found by integrating out t in (8.3). That is, for each $k = 1, 2, 3$, calculate

$$\frac{1}{k}\int_0^\infty \mathrm{e}^{-\frac{2}{k}t}\mathrm{e}^{-10kt}\,(kt)^{60}\,\mathrm{d}t \;,$$

and normalize. Numerical evaluation yields the following posterior probabilities (rounded):

$$0.000353516, \quad 0.30469, \quad \text{and} \quad 0.694957 \;.$$

Hence, we have deduced from Bayes' formula that the most likely route that was followed is route 3. But route 2 is also quite possible. It is very unlikely that route 1 was used.

In a similar manner, to find $f(t \mid x = 60)$, we sum (8.3) with respect to k, giving

$$f(t \mid x = 60) = c^{-1}\,t^{60}\left(3^{59}\mathrm{e}^{-92t/3} + 2^{59}\mathrm{e}^{-21t} + \mathrm{e}^{-12t}\right) \;.$$

The normalization constant (which can be expressed in terms of the gamma function) evaluates to $c \approx 3.481048347 \times 10^{19}$. The graph of this marginal posterior pdf is shown in Fig. 8.2 (solid line).

To (approximately) sample from $f(k, t \mid x = 60)$, we can use the Gibbs sampler (Algorithm 7.4). For this we need to specify:

1. The conditional distribution of k given t and x.
2. The conditional distribution of t given k and x.

Fig. 8.2 Posterior pdf of t given $x = 60$ (*solid line*) and its estimate obtained via Gibbs sampling (*dotted line*)
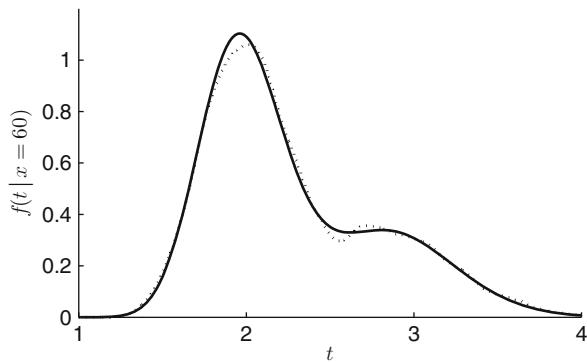
By viewing t and x as constants in (8.2), we see that, given t and x, k has a discrete distribution on $\{1, 2, 3\}$ with probabilities proportional to

$$e^{-12t}, \qquad 2^{x-1}e^{-21t}, \qquad \text{and} \quad 3^{x-1}e^{-92t/3} .$$

Similarly, by viewing k and x as constants in (8.2), we have

$$f(t \mid x, k) \propto t^x \exp\left\{-t\left(\frac{2}{k} + 10k\right)\right\} ,$$

which is the pdf of the Gamma$(x + 1, \frac{2}{k} + 10k)$ distribution. By alternatively sampling from $f(k \mid t, x)$ and $f(t \mid k, x)$ we obtain a dependent sample from $f(k, t \mid x)$. The following MATLAB program implements the Gibbs sampler. It uses

☞ 364   the gamrand function (see Appendix A.8) to sample from the gamma distribution. The burn-in period was ignored. Throughout this chapter we use the function

☞ 202   kde.m to display a kernel density estimate of the simulated data.

```
%ticketinspector.m
n = 10000;
x = 60; %number of tickets
p = [1/3,1/3,1/3] %initial value
kk = zeros(1,n);
tt = zeros(1,n);
k = min(find(cumsum(p) > rand));
for i=1:n
    a = x+ 1;
    b = 2/k + 10*k;
    t = gamrand(a,b);
    tt(i) = t;
    p = [exp(-12*t),2^(x-1)*exp(-21*t), 3^(x-1)*exp(-92*t/3)];
    p = p/sum(p);
    k = min(find(cumsum(p) > rand));
    kk(i) = k;
```

```
end
p1est = sum(kk == 1)/n   %estimate of posterior probability 1
p2est = sum(kk == 2)/n   %estimate of posterior probability 2
p3est = sum(kk == 3)/n   %estimate of posterior probability 3
figure(1)
hist(kk)
figure(2)
kde(tt)
```

Typical outcomes of the posterior probabilities for route k are 10^{-4}, 0.32, and 0.68. These probabilities are in close correspondence with the actual probabilities. The KDE of the posterior pdf of t is given in Fig. 8.2 (dotted line). This is in excellent agreement with the true posterior pdf.

8.2 Common Bayesian Models

The common statistical models in Chap. 4 can also be formulated and analyzed in a Bayesian framework. In this section we give various examples of how this is done. Note that inference in a Bayesian setting depends on the prior information, in contrast to the classical case.

8.2.1 Normal Model with Unknown μ and σ^2

Let $x_1, \ldots, x_n$ be an iid sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$. In classical statistics the model can be written as $\mathbf{x} \sim \mathsf{N}(\mu\mathbf{1}, \sigma^2 I)$, where $\mathbf{1}$ is the n-dimensional vector of 1s and I the n-dimensional identity matrix. To formulate the corresponding Bayesian model, we start with a similar likelihood as in the classical case; that is,

$$(\mathbf{x} \,|\, \mu, \sigma^2) \sim \mathsf{N}(\mu\mathbf{1}, \sigma^2 I) \;.$$

In the Bayesian setting both μ and σ^2 are random, and we need to specify their prior distributions to complete the model. In practice the choice of the prior distribution is governed by two considerations. Firstly, the prior should be simple enough to facilitate the computation or simulation of the posterior pdf. Secondly, the prior distribution should be general enough to model complete ignorance of the parameter of interest. Priors that do not convey any pre-knowledge of the parameter are said to be **uninformative**. The uniform or **flat** prior in Example 8.1 is an example.

For the present model a standard prior for μ is

$$\mu \sim \mathsf{N}(0, \sigma_0^2) , \tag{8.4}$$

where $\sigma_0^2 > 0$ is a constant. The larger the σ_0^2 is, the more uninformative is the prior. A standard prior for σ^2 is

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_0, \lambda_0) , \tag{8.5}$$

where $\alpha_0 > 0$ and $\lambda_0 > 0$ are constants and $\mathsf{InvGamma}(\alpha, \lambda)$ denotes the **inverse-gamma** distribution.

Definition 8.2. (Inverse-Gamma Distribution). A random variable Z is said to have an **inverse-gamma** distribution with **shape** parameter $\alpha > 0$ and **scale** parameter $\lambda > 0$ if its pdf is given by

$$f(z; \alpha, \lambda) = \frac{\lambda^\alpha z^{-\alpha-1} e^{-\lambda z^{-1}}}{\Gamma(\alpha)} , \quad z > 0 . \tag{8.6}$$

This is the pdf of the random variable $Z = 1/X$ with $X \sim \mathsf{Gamma}(\alpha, \lambda)$.

Thus, (8.5) is equivalent to

$$\frac{1}{\sigma^2} \sim \mathsf{Gamma}(\alpha_0, \lambda_0) . \tag{8.7}$$

The smaller the α_0 and λ_0 are, the less informative is the prior. It is further assumed that μ and σ^2 are independent.

The joint pdf of $\mathbf{x}$, μ, and σ^2 is therefore

$$f(\mathbf{x}, \mu, \sigma^2) = f(\mu) \times f(\sigma^2) \times f(\mathbf{x} \mid \mu, \sigma^2)$$

$$= \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma_0^2}\right\}$$

$$\times \frac{\lambda_0^{\alpha_0}(\sigma^2)^{-\alpha_0-1} \exp\left\{-\lambda_0\,(\sigma^2)^{-1}\right\}}{\Gamma(\alpha_0)}$$

$$\times \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right\} .$$

It follows that the posterior pdf is given by

$$f(\mu, \sigma^2 \mid \mathbf{x}) \propto (\sigma^2)^{-n/2-\alpha_0-1} \exp\left\{ -\frac{1}{2} \frac{\sum_i (x_i - \mu)^2}{\sigma^2} - \frac{1}{2} \frac{\mu^2}{\sigma_0^2} - \frac{\lambda_0}{\sigma^2} \right\} \,. \qquad (8.8)$$

To simulate from it using the Gibbs sampler, we need the distributions of both $(\mu \mid \sigma^2, \mathbf{x})$ and $(\sigma^2 \mid \mu, \mathbf{x})$. To find $f(\mu \mid \sigma^2, \mathbf{x})$, view the right-hand side of (8.8) as a function of μ. This gives

$$f(\mu \mid \sigma^2, \mathbf{x}) \propto \exp\left\{ -\frac{n\mu^2 - 2\mu \sum_i x_i}{2\sigma^2} - \frac{1}{2} \frac{\mu^2}{\sigma_0^2} \right\}$$

$$= \exp\left\{ -\frac{1}{2} \left(\frac{(n\mu^2 - 2\mu \sum_i x_i)\sigma_0^2 + \mu^2 \sigma^2}{\sigma^2 \sigma_0^2} \right) \right\}$$

$$= \exp\left\{ -\frac{1}{2} \left(\frac{\mu^2 - 2\mu \sigma_0^2 \sum_i x_i/(n\sigma_0^2 + \sigma^2)}{\sigma^2 \sigma_0^2/(n\sigma_0^2 + \sigma^2)} \right) \right\} \,. \qquad (8.9)$$

This shows that $(\mu \mid \sigma^2, \mathbf{x})$ has a normal distribution with mean $\sigma_0^2 \sum x_i/(n\sigma_0^2 + \sigma^2)$ and variance $\sigma^2 \sigma_0^2/(n\sigma_0^2 + \sigma^2)$. By defining $\kappa_n = \sigma^2/(\sigma_0^2 n)$, we can write this succinctly as

$$(\mu \mid \sigma^2, \mathbf{x}) \sim \mathsf{N}\left(\frac{\bar{x}}{1 + \kappa_n}, \frac{\sigma^2/n}{1 + \kappa_n} \right),$$

where $\bar{x}$ is the sample mean. Similarly, to find $f(\sigma^2 \mid \mu, \mathbf{x})$, view (8.8) as a function of σ^2. This gives

$$f(\sigma^2 \mid \mu, \mathbf{x}) \propto (\sigma^2)^{-n/2-\alpha_0-1} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2/\sigma^2 - \lambda_0/\sigma^2 \right\} \,. \qquad (8.10)$$

In other words,

$$(\sigma^2 \mid \mu, \mathbf{x}) \sim \mathsf{InvGamma}\left(\alpha_0 + n/2, \sum_{i=1}^n (x_i - \mu)^2/2 + \lambda_0 \right) \,.$$

It is interesting to note that in the limit $\sigma_0^2 \to \infty$, $\alpha_0 \to 0$, and $\lambda_0 \to 0$, the right-hand sides of (8.10) and (8.9) define valid probability distributions; namely,

$$(\mu \mid \sigma^2, \mathbf{x}) \sim \mathsf{N}\left(\bar{x}, \sigma^2/n \right)$$

$$(\sigma^2 \mid \mu, \mathbf{x}) \sim \mathsf{InvGamma}\left(n/2, \sum_{i=1}^n (x_i - \mu)^2/2 \right) \,.$$

The two distributions above correspond to the following simplified Bayesian model:

$$f(\mu, \sigma^2) = 1/\sigma^2$$

$$(\mathbf{x} \,|\, \mu, \sigma^2) \sim \mathsf{N}(\mu \mathbf{1}, \sigma^2 I) \,.$$

Here the prior for (μ, σ^2) is **improper**. That is, it is not a pdf in itself, but by obstinately applying Bayes' formula it does yield a proper posterior pdf. In some sense this prior conveys the least amount of information about μ and σ^2.

In the following MATLAB script an iid sample of size $n = 10$ is drawn from the standard normal distribution, and a dependent sample from the posterior distribution for the simplified model is obtained, using the Gibbs sampler with $N = 10^5$ samples.

```
%bayesnorm.m
n = 10;
X = randn(1,n); %generate the data
sample_mean = mean(X);
sample_var = var(X);
sig2 = var(X); mu = sample_mean; %initial state
N = 10^5; %sample size for Gibbs sampler
gibbs_sample = zeros(N,2);
for k=1:N
    mu = sample_mean + sqrt(sig2/n)*randn;   %draw mu
    V = sum((X -mu).^2)/2;
    sig2 = 1/gamrand(n/2,V); %draw sigma^2
    gibbs_sample(k,:) = [mu,sig2];
end
kde(gibbs_sample(:,1)) %kde for posterior of mu
% kde(gibbs_sample(:,2)) %kde for posterior of sigma^2
```

The estimated posterior pdfs of μ and σ^2 are given in Fig. 8.3. In this case the sample mean and sample variance are -0.0562 and 0.7288, respectively. The 0.05 and 0.95 sample quantiles of the simulated posterior values for μ give
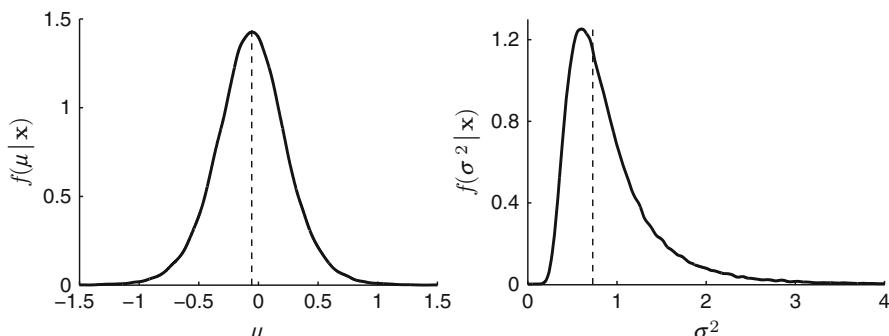
Fig. 8.3 The estimated posterior pdfs of μ and σ^2. The *dashed lines* correspond to the sample mean (*left*) and the sample variance (*right*)

the 90% credible interval $(-0.5522, 0.4389)$. This is in close agreement with the classical confidence interval (5.18), which in this case is $(-0.5510, 0.4387)$. Similarly, an estimated 90% credible interval for σ^2 is $(0.3885, 1.9695)$, which is in close agreement with the classical confidence interval (5.19), which here is $(0.3887, 1.9726)$. See Problem 8.10 for a further discussion of this model.

8.2.2 Bayesian Normal Linear Model

Suppose $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is described via *normal linear model*. That is (see (6.23)) the likelihood is specified by

$$(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \sim \mathsf{N}(X\boldsymbol{\beta}, \sigma^2 I) \,,$$

where $X = (x_{ij})$ is the (known) $n \times m$ design matrix and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ and σ^2 are unknown parameters. Again, both $\boldsymbol{\beta}$ and σ^2 are random in the Bayesian setting, and we need to specify their prior distributions. The prior for σ^2 is the same as in the normal model:

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_0, \lambda_0) \,,$$

with $\alpha_0 > 0$ and $\lambda_0 > 0$ known. A standard prior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \Sigma_0) \,,$$

where Σ_0 is a known covariance matrix and $\boldsymbol{\beta}_0$ a known mean vector. The joint pdf of $\mathbf{y}, \boldsymbol{\beta}$, and σ^2 is thus

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta}) \times f(\sigma^2) \times f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)$$

$$= ((2\pi)^m |\Sigma_0|)^{-1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}$$

$$\times \frac{\lambda_0^{\alpha_0} (\sigma^2)^{-\alpha_0 - 1} \exp\left\{ -\lambda_0 (\sigma^2)^{-1} \right\}}{\Gamma(\alpha_0)}$$

$$\times (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) \right\} \,.$$

It follows that the posterior pdf is given by

$$
f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto \left(\sigma^2\right)^{-n/2 - \alpha_0 - 1} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) \right.
$$
$$
\left. -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \Sigma_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{\lambda_0}{\sigma^2} \right\} . \tag{8.11}
$$

As before, we use the Gibbs sampler to simulate from this posterior pdf. To that end, we need to derive the distributions of both $(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$ and $(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y})$.

Following the same argument in Sect. 8.2.1, we can show that

$$
(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y}) \sim \mathsf{InvGamma}\left(\alpha_0 + n/2, \ (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta})/2 + \lambda_0 \right) .
$$

Next, to find $f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$, view the right-hand side of (8.11) as a function of $\boldsymbol{\beta}$. This gives

$$
f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \Sigma_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}
$$
$$
\propto \exp \left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\beta}^\top X^\top X \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top X^\top \mathbf{y}) + \boldsymbol{\beta}^\top \Sigma_0^{-1} \boldsymbol{\beta}_0 - \frac{1}{2} \boldsymbol{\beta}^\top \Sigma_0^{-1} \boldsymbol{\beta} \right\}
$$
$$
= \exp \left\{ -\frac{1}{2} \left[\boldsymbol{\beta}^\top (X^\top X/\sigma^2 + \Sigma_0^{-1}) \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top (X^\top \mathbf{y}/\sigma^2 + \Sigma_0^{-1} \boldsymbol{\beta}_0) \right] \right\} . \tag{8.12}
$$

Note that the exponent is quadratic in $\boldsymbol{\beta}$, and thus $(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) \sim \mathsf{N}(\boldsymbol{\mu}, D)$ for some mean vector $\boldsymbol{\mu}$ and covariance matrix D. Therefore,

$$
f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^\top D^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}) \right\}
$$
$$
\propto \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\beta}^\top D^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top D^{-1} \boldsymbol{\mu} \right) \right\} .
$$

To determine $\boldsymbol{\mu}$ and D we only need to compare the linear and quadratic terms in (8.12) to those of the $\mathsf{N}(\boldsymbol{\mu}, D)$ density above. This process is sometimes called **completing the squares**. Comparing the quadratic terms gives $D = (X^\top X/\sigma^2 + \Sigma_0^{-1})^{-1}$. Similarly, equating the linear terms in the two expressions gives $D^{-1}\boldsymbol{\mu} = X^\top \mathbf{y}/\sigma^2 + \Sigma_0^{-1} \boldsymbol{\beta}_0$. In summary, we have the following result.

Theorem 8.1. (Conditional Posteriors for the Linear Model). Consider the Bayesian model

$$\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \Sigma_0) \,, \tag{8.13}$$

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_0, \lambda_0) \,, \tag{8.14}$$

$$(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2) \sim \mathsf{N}(X\boldsymbol{\beta}, \sigma^2 I) \,, \tag{8.15}$$

where X is a fixed $n \times m$ design matrix, Σ_0 is a fixed $n \times n$ covariance matrix, $\boldsymbol{\beta}_0$ is a fixed vector, and α_0 and λ_0 are fixed constants. Then,

$$(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}) \sim \mathsf{N}(\boldsymbol{\mu}, D) \,,$$

where $\boldsymbol{\mu} = D(X^\top \mathbf{y}/\sigma^2 + \Sigma_0^{-1}\boldsymbol{\beta}_0)$, with $D = (X^\top X/\sigma^2 + \Sigma_0^{-1})^{-1}$, and

$$(\sigma^2 \,|\, \boldsymbol{\beta}, \mathbf{y}) \sim \mathsf{InvGamma}\left(\alpha_0 + n/2, \ (\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta})/2 + \lambda_0\right) \,.$$

Note that as the prior precision matrix Σ_0^{-1} approaches the zero matrix, the prior for $\boldsymbol{\beta}$ becomes more non-informative. For $\Sigma_0^{-1} = O$ (zero matrix), the prior for $\boldsymbol{\beta}$ is improper. However, the conditional density $f(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y})$ is still a proper pdf. In fact, for $\Sigma_0^{-1} = O$, we have $\boldsymbol{\mu} = DX^\top \mathbf{y}/\sigma^2$, with $D = \sigma^2(X^\top X)^{-1}$, so that

$$(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}) \sim \mathsf{N}(X^+ \mathbf{y}, \sigma^2(X^\top X)^{-1}) \,,$$

where $X^+ = (X^\top X)^{-1} X^\top$ is the (right) pseudo-inverse of X. Using this improper prior for $\boldsymbol{\beta}$, the conditional expectation $\mathbb{E}[\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}]$ therefore coincides with the least-squares estimate in (5.11).

The following corollary presents an important generalization of Theorem 8.1 for the situation where $\mathbf{y} - X\boldsymbol{\beta}$ is an affine transformation of $\mathbf{y}$; that is, $\mathbf{y} - X\boldsymbol{\beta} = \mathbf{a} + A\mathbf{y}$ for some vector $\mathbf{a}$ and matrix A. The result will be heavily relied on in later parts of the book.

Corollary 8.1. (Conditional Posteriors for the Linear Model with General Error Covariance Matrix). Consider the Bayesian model

$$\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \Sigma_0) \,, \tag{8.16}$$

(continued)

(continued)

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_0, \lambda_0) \,, \qquad (8.17)$$

$$(\mathbf{y} - X\boldsymbol{\beta} \mid \boldsymbol{\beta}, \sigma^2) \sim \mathsf{N}(\mathbf{0}, \sigma^2 R) \,, \qquad (8.18)$$

where $\mathbf{y} - X\boldsymbol{\beta}$ is an affine transformation of $\mathbf{y}$, R and Σ_0 are fixed $n \times n$ (covariance) matrices, $\boldsymbol{\beta}_0$ is a fixed vector, and α_0 and λ_0 are fixed constants. Then,

$$(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) \sim \mathsf{N}(\boldsymbol{\mu}, D) \,,$$

where $\boldsymbol{\mu} = D(X^\top R^{-1}\mathbf{y}/\sigma^2 + \Sigma_0^{-1}\boldsymbol{\beta}_0)$, with $D = (X^\top R^{-1}X/\sigma^2 + \Sigma_0^{-1})^{-1}$, and

$$(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y}) \sim \mathsf{InvGamma}\left(\alpha_0 + n/2, \ (\mathbf{y} - X\boldsymbol{\beta})^\top R^{-1}(\mathbf{y} - X\boldsymbol{\beta})/2 + \lambda_0\right).$$

Proof. By assumption we have $\mathbf{y} - X\boldsymbol{\beta} = \mathbf{a} + A\mathbf{y} \stackrel{\text{def}}{=} \mathbf{z}$ for some vector $\mathbf{a}$ and matrix A, where $(\mathbf{z} \mid \sigma^2, \boldsymbol{\beta}) \sim \mathsf{N}(\mathbf{0}, \sigma^2 R)$. It follows that

$$f(\mathbf{y} \mid \sigma^2, \boldsymbol{\beta}) \propto f(\mathbf{z} \mid \sigma^2, \boldsymbol{\beta}) \propto (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^\top R^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right\} \,.$$

The rest of the proof follows exactly the same reasoning as for Theorem 8.1.   □

8.2.3   Bayesian Multinomial Model

In this section we extend the Bayesian analysis of the binomial model in Example 8.1 to the multinomial case. Recall (see Definition 3.4) that a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ has a **multinomial** distribution, with parameters n and $p_1, p_2, \ldots, p_k$ (probabilities summing up to 1), if

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \, x_2! \cdots x_k!} \, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \qquad (8.19)$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_k = n$. We can think of $\mathbf{X} \sim \mathsf{Mnom}(n, \mathbf{p})$ representing the configuration of n balls in k urns when the balls are thrown independently into the urns according to a vector of probabilities $\mathbf{p} = (p_1, \ldots, p_k)$. For the binomial case there are only two urns and $\mathbf{p} = (p, 1-p)$.

Suppose we are given data $\mathbf{x}$ from an $\mathsf{Mnom}(n, \mathbf{p})$ distribution and wish to gain information about $\mathbf{p}$ on the basis of $\mathbf{x}$. Assuming uniform priors, the Bayesian model is

$$f(\mathbf{p}) \propto 1 , \qquad f(\mathbf{x} \mid \mathbf{p}) = \frac{n!}{x_1! \, x_2! \cdots x_k!} \, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} .$$

It follows that the posterior pdf is of the form

$$f(\mathbf{p} \mid \mathbf{x}) \propto p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} , \quad \mathbf{p} \in [0, 1]^k , \quad \sum_{i=1}^{k} p_i = 1 .$$

Since $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} p_i = 1$ we can drop p_k from the analysis and look instead at the posterior pdf of $p_1, \ldots, p_{k-1}$ given $\mathbf{x}$, which is given by

$$f(p_1, \ldots, p_{k-1} \mid \mathbf{x}) \propto p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} \left(1 - \sum_{i=1}^{k-1} p_i \right)^{x_k} ,$$

where $p_i \geqslant 0, i = 1, \ldots, k - 1$, and $\sum_{i=1}^{k-1} p_i \leqslant 1$. This is the pdf of a Dirichlet distribution: $(p_1, \ldots, p_{k-1} \mid \mathbf{x}) \sim \mathsf{Dirichlet}(x_1 + 1, \, x_2 + 1, \, \ldots, x_k + 1)$.

Definition 8.3.   (Dirichlet Distribution). A random vector $\mathbf{Z} = (Z_1, \ldots, Z_m)$ is said to have a **Dirichlet** distribution with **shape** parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m+1})$ if its pdf is given by

$$f(\mathbf{z}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{m+1} \alpha_i\right)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \prod_{i=1}^{m} z_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{m} z_i\right)^{\alpha_{m+1} - 1}, \quad \mathbf{z} \in [0, 1]^m , \quad \sum_{i=1}^{m} z_i \leqslant 1 .$$

We write this distribution as $\mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_{m+1})$ or $\mathsf{Dirichlet}(\boldsymbol{\alpha})$.

The m-dimensional $\mathsf{Dirichlet}(1, \ldots, 1)$ distribution has a constant density on the set $\{\mathbf{z} \in \mathbb{R}^m : z_i \geqslant 0, i = 1, \ldots, m, \sum_{i=1}^{m} z_i \leqslant 1\}$ and thus corresponds to the uniform distribution on that set. The $\mathsf{Dirichlet}(\alpha_1, \alpha_2)$ distribution is the $\mathsf{Beta}(\alpha_1, \alpha_2)$ distribution. Moreover, if $\mathbf{Z} = (Z_1, \ldots, Z_m) \sim \mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_{m+1})$, the marginal distribution of Z_i is $\mathsf{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$; see Problem 8.6. The following theorem shows how one can simulate from the Dirichlet distribution using Gamma random variables.

Theorem 8.2. (Sampling from the Dirichlet Distribution). Let $Y_1, \ldots, Y_{m+1}$ be independent random variables with $Y_i \sim \mathsf{Gamma}(\alpha_i, 1)$, $i = 1, \ldots, m + 1$, and define

$$Z_j = \frac{Y_j}{\sum_{i=1}^{m+1} Y_i}, \quad j = 1, \ldots, m . \tag{8.20}$$

Then, $\mathbf{Z} = (Z_1, \ldots, Z_m) \sim \mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_{m+1})$.

☞ 81   *Proof.* This is a direct consequence of the transformation rule (3.25). In particular, consider the transformation $\mathbf{g} : (y_1, \ldots, y_{m+1})^\top \mapsto (z_1, \ldots, z_{m+1})^\top$ defined by (8.20) and $z_{m+1} = y_1 + \cdots + y_{m+1}$. By rewriting the $\{y_i\}$ in terms of the $\{z_i\}$ we see that the inverse transformation is given by

$$y_i = z_i z_{m+1}, \ i = 1, \ldots, m \qquad \text{and} \qquad y_{m+1} = (1 - (z_1 + \cdots + z_m)) z_{m+1} .$$

The determinant of the corresponding Jacobian matrix is z_{m+1}^m; see Problem 8.5. Using classical notation for clarity and defining $\mathbf{Y} = (Y_1, \ldots, Y_{m+1})^\top$, we have by
☞ 48   the transformation rule and the definition (2.19) of the Gamma pdf:

$$
\begin{aligned}
f_{\mathbf{Z}, Z_{m+1}}(\mathbf{z}, z_{m+1}) &= f_{\mathbf{Y}}(\mathbf{y}) z_{m+1}^m = \frac{\left(\prod_{i=1}^{m+1} y_i^{\alpha_i - 1}\right) e^{-\sum_{i=1}^{m+1} y_i} z_{m+1}^m}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \\
&= \frac{\left(\prod_{i=1}^m y_i^{\alpha_i - 1}\right) y_{m+1}^{\alpha_{m+1} - 1} e^{-z_{m+1}} z_{m+1}^m}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \\
&= \frac{\left(\prod_{i=1}^m z_i^{\alpha_i - 1}\right) \left(1 - \sum_{i=1}^m z_i\right)^{\alpha_{m+1} - 1}}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \underbrace{z_{m+1}^{\left(\sum_{i=1}^{m+1} \alpha_i\right) - 1} e^{-z_{m+1}}}_{(\star)} . \tag{8.21}
\end{aligned}
$$

To obtain the pdf of $\mathbf{Z}$ we need to integrate out z_{m+1} in (8.21). Since $(\star)$ is proportional to the pdf of a $\mathsf{Gamma}(\sum_{i=1}^{m+1} \alpha_i, 1)$ distribution, this integral is $\Gamma(\sum_{i=1}^{m+1} \alpha_i)$, which completes the proof.   □

Example 8.3 (Bayesian Inference for the Multinomial Model). Five hundred people are randomly selected from a large population. They are asked if they like, dislike, or are indifferent to the current anti-smoking campaign. Table 8.1 lists the data.

Let x_{ij} be the count in row i and column j in Table 8.1; for example, $x_{13} = 147$ and $x_{22} = 38$. Denote by $\mathbf{x} = (x_{11}, \ldots, x_{23})$ the vector of counts, and let

Table 8.1 Opinions on smoking campaign, by gender

	Opinion		
Gender	Dislike	Neutral	Like
Male	53	57	147
Female	93	38	113

$\mathbf{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23})$ be the corresponding vector of probabilities. Thus, p_{13} is the probability that a randomly selected person is male and likes the campaign. A natural Bayesian model for the data is that $(\mathbf{x} \mid \mathbf{p}) \sim \mathsf{Mnom}(501, \mathbf{p})$, with a uniform prior for $\mathbf{p}$. It follows that $(p_{11}, \ldots, p_{22} \mid \mathbf{x})$ (i.e., the vector $\mathbf{p}$ with component p_{23} removed) has a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (x_{11} + 1, \ldots, x_{23} + 1)$.

Can we conclude from the data that opinion is independent of gender? For this to be true it must hold that

$$p_{ij} = p_i^{(r)} p_j^{(c)}, \quad i = 1, 2, \quad j = 1, 2, 3,$$

where the row totals $p_i^{(r)} = p_{i1} + p_{i2} + p_{i3}$, $i = 1, 2$ give the probability that a selected person is male ($i = 1$) or female ($i = 2$); similarly, the column totals $p_j^{(c)} = p_{1j} + p_{2j}$, $j = 1, 2, 3$ give the probabilities of the opinions. It thus makes sense to investigate the posterior distribution of

$$a_{ij} = p_{ij} - p_i^{(r)} p_j^{(c)}, \quad i = 1, 2, \quad j = 1, 2, 3 \tag{8.22}$$

and check if 0 lies within a reasonable (say, 95%) credible interval of each a_{ij}. The following MATLAB program generates $N = 10000$ vectors $\mathbf{p}$ drawn from the posterior distribution. For each $\mathbf{p}$ the row and column totals are calculated, and subsequently realizations from the posterior distribution of a_{1j}, $j = 1, 2, 3$ are obtained via (8.22). Since $a_{1j} = -a_{2j}$, it suffices to consider only a_{1j}, $j = 1, 2, 3$. The function dirichrnd, which generates draws from a Dirichlet distribution, is given in Appendix A.8. Kernel density plots of the posterior pdfs are shown in Fig. 8.4. We see that opinion and gender are likely to be *dependent*, as 0 is not contained in, for example, a 0.99 credible interval of the posterior pdf of a_{11}.

```
%multinomex.m
x = [53,57,147,93,38,113];
N = 10000;
p = zeros(N,2,3); a = zeros(N,2,3);
alpha = x + 1;
for i=1:N
    r = dirichrnd(alpha);
    h = [r,1-sum(r)];
    p(i,:,:) = reshape(h',3,2)';
```

```
end
for i=1:2
    p_row(i,:) = sum(p(:,i,:),3);
end
for j=1:3
    p_col(j,:) = sum(p(:,:,j),2);
end
for i=1:2
    for j=1:3
        a(:,i,j) = p(:,i,j) - p_row(i,:)'.*p_col(j,:)';
    end
end
hold on
for j=1:3
    kde(a(:,1,j));
end
```
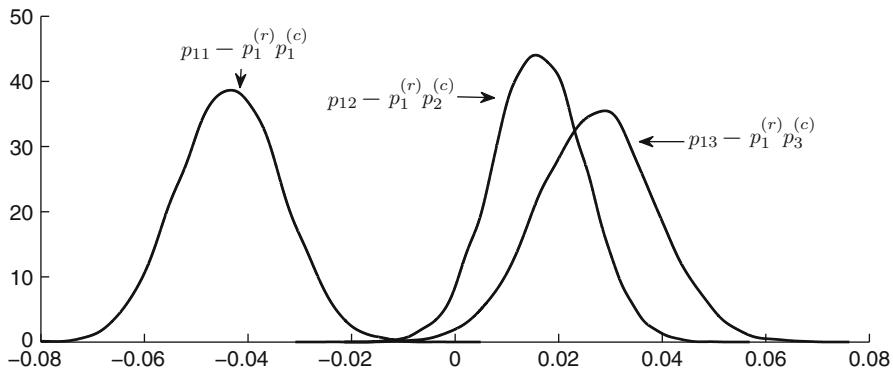

Fig. 8.4 Posterior pdfs of $a_{1j} = p_{1j} - p_1^{(r)} p_j^{(c)}$, $j = 1, 2, 3$, indicating that opinion and gender are not independent

8.3   Bayesian Networks

The formulation and analysis of a Bayesian model can often be facilitated through the use of **Bayesian networks**. Mathematically, a Bayesian network is a **directed acyclic graph**, that is, a collection of **vertices** (nodes) and **arcs** (arrows between nodes) such that arcs, when put head-to-tail, do not create loops. Figure 8.5 shows two directed acyclic graphs ((a) and (b)) and a counterexample (c).
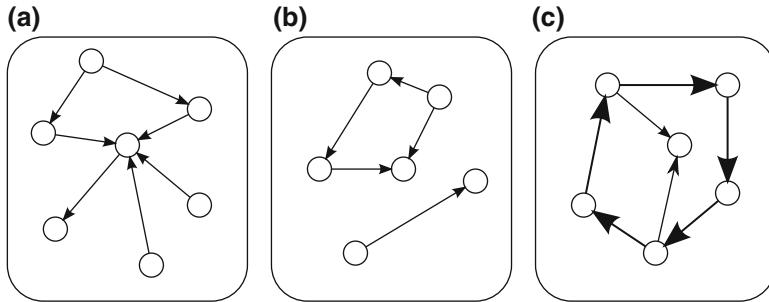
Fig. 8.5 The directed graphs in (**a**) and (**b**) are acyclic. Graph (**c**) has a (directed) cycle and can therefore not represent a Bayesian network

Bayesian networks can be used to graphically represent the joint probability distribution of a collection of random variables. In particular, consider a Bayesian network with vertices labeled $x_1, \ldots, x_n$. Let $\mathcal{P}_j$ denote the set of **parents** of x_j, that is, the vertices x_i for which there exists an arc from x_i to x_j in the graph. We can associate with this network a joint pdf

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j \mid \mathcal{P}_j) .$$

Note that any pdf can be represented by a Bayesian network in this way because, by the product rule (3.10),

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2 \mid x_1) \cdots f(x_n \mid x_1, \ldots, x_{n-1}) .$$

As an example, the left pane of Fig. 8.6 shows a Bayesian network with five variables, representing the following structure for the pdf:

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_2) f(x_4 \mid x_2) f(x_5 \mid x_3, x_4) .$$

In the same figure two small black nodes have been added with labels θ_1 and θ_2. This is a way of representing fixed parameters of the distribution. Thus, in this case, the (classical) pdf is of the form

$$f(x_1, \ldots, x_n) = f(x_1; \theta_1) f(x_2 \mid x_1; \theta_2) f(x_3 \mid x_2) f(x_4 \mid x_2) f(x_5 \mid x_3, x_4) .$$

In the right pane of Fig. 8.6 the corresponding Bayesian model is depicted. It is useful to distinguish between random variables and their observations, by using a dark color or gray scale for the latter one. For example, the right pane of Fig. 8.6 represents the situation where the "data" $x_1, \ldots, x_n$ have been observed. The aim is to find the posterior pdf of θ_1 and θ_2 given the data.
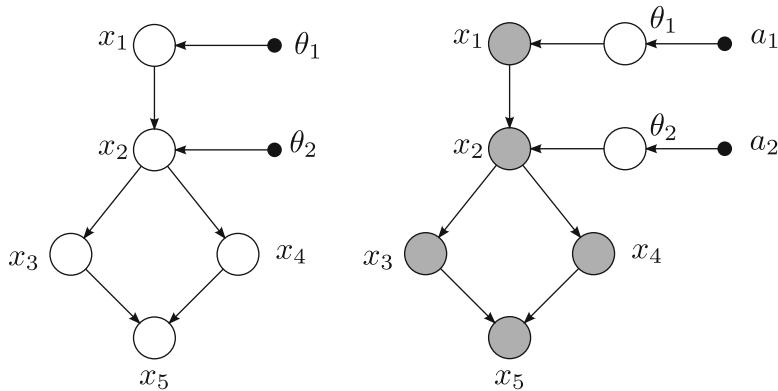
Fig. 8.6 *Left*: a graphical representation of a *classical* statistical model with random variables $x_1, \ldots, x_5$ and fixed parameters θ_1, θ_2. The representation is in the form of a directed acyclic graph (Bayesian network). *Right*: the graphical representation of the corresponding *Bayesian* model with *observed* (i.e., fixed) data $x_1, \ldots, x_n$, indicated by *shaded nodes*. In this case the parameters θ_1 and θ_2 are random and depend on fixed parameters a_1 and a_2 (sometimes called **hyperparameters**)

Figure 8.7 gives two more examples of Bayesian networks. The first corresponds to the ticket inspector model in Example 8.2; the second refers to the normal Bayesian model in Sect. 8.2.1.
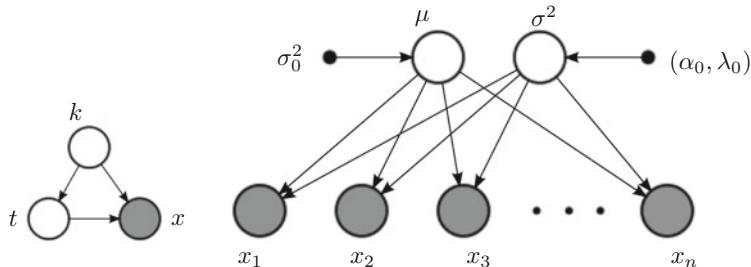
Fig. 8.7 *Left*: the Bayesian network for the ticket inspector model in Example 8.2. *Right*: a representation of the Bayesian model for iid normal data

Example 8.4 (Belief Nets). Bayesian networks are frequently used for medical diagnosis and statistical classification. In this context they are sometimes called *belief nets*. An example belief net is shown in Fig. 8.8. The purpose of this belief net is to determine if a patient is to be diagnosed with heart disease, based on several factors and symptoms. Two important factors in heart disease are smoking and age, and two main symptoms are chest pains and shortness of

breath. The belief net in Fig. 8.8 shows the prior probabilities of smoking and age, the conditional probabilities of heart disease given age and smoking, and the conditional probabilities of chest pains and shortness of breath given heart disease.

Smoking	
Yes	No
0.3	0.7

Age	
≤ 50	> 50
0.6	0.4

	Chest Pains	
H. Disease	Yes	No
Yes	0.2	0.8
No	0.01	0.99

	Shortness O.B.	
H. Disease	Yes	No
Yes	0.3	0.7
No	0.1	0.9

	Heart Disease	
Smoking, Age	Yes	No
Yes, ≤ 50	0.20	0.80
Yes, > 50	0.40	0.60
No, ≤ 50	0.05	0.95
No, > 50	0.15	0.85
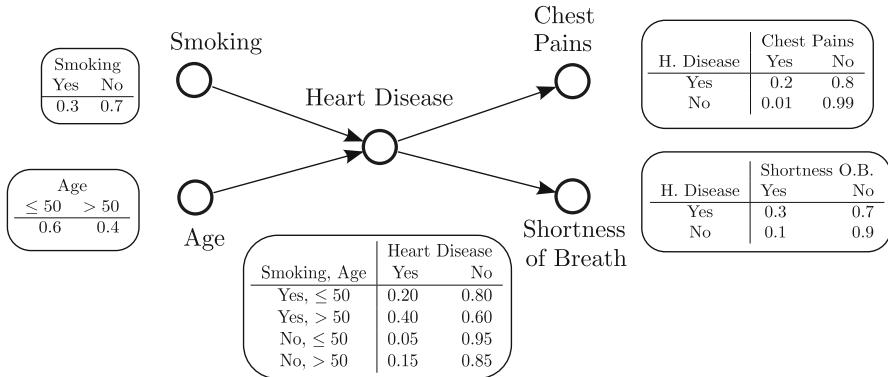
Fig. 8.8 A Bayesian belief net for the diagnosis of heart disease

Suppose a person experiences chest pains and shortness of breath, but we do not know her/his age and if she/he is smoking. How likely is it that she/he has a heart disease?

Define the variables s (smoking), a (age), h (heart disease), c (chest pains), and b (shortness of breath). We assume that s and a are independent. We wish to calculate

$$\mathbb{P}(h = \text{Yes} \,|\, b = \text{Yes}, \, c = \text{Yes}) \,.$$

From the Bayesian network structure, we see that the joint pdf of $s, a, h, c,$ and b can be written as

$$f(s, a, h, c, b) = f(s) f(a) f(h \,|\, s, a) f(c \,|\, h) f(b \,|\, h) \,.$$

It follows that

$$f(h \,|\, b, c) \propto f(c \,|\, h) f(b \,|\, h) \underbrace{\sum_{a,s} f(h \,|\, s, a) f(s) f(a)}_{f(h)} \,.$$

We have

$$f(h = \text{Yes}) = 0.2 \times 0.3 \times 0.6 + 0.4 \times 0.3 \times 0.4$$
$$+ 0.05 \times 0.7 \times 0.6 + 0.15 \times 0.7 \times 0.4 = 0.147 \ .$$

Consequently,

$$f(h = \text{Yes} \mid b = \text{Yes}, c = \text{Yes}) = \beta \times 0.2 \times 0.3 \times 0.147 = \beta \, 0.00882$$

and

$$f(h = \text{No} \mid b = \text{Yes}, c = \text{Yes}) = \beta \times 0.01 \times 0.1 \times (1 - 0.147) = \beta \, 0.000853$$

for some normalization constant β. Thus,

$$f(h = \text{Yes} \mid b = \text{Yes}, c = \text{Yes}) = \frac{0.00882}{0.0882 + 0.000853} = 0.911816 \approx 0.91 \ .$$

8.4 Asymptotic Normality of the Posterior Distribution

We saw in Sect. 6.3.2 various asymptotic properties of the likelihood function. Similar results can be obtained for the posterior pdf. For clarity we identify the (conditional) pdfs by different symbols: f, f_Θ, and $\overset{\circ}{f}$.

> **Theorem 8.3. (Asymptotic Distribution of the Posterior Pdf).** Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an iid sample from $\overset{\circ}{f}(x \mid \theta_0)$, where θ_0 is fixed. The posterior pdf with prior pdf $f_\Theta(\theta)$
>
> $$f(\theta \mid \mathbf{x}) \propto f_\Theta(\theta) \prod_{i=1}^{n} \overset{\circ}{f}(x_i \mid \theta) \qquad (8.23)$$
>
> is approximately normal with mean θ_0 and variance $\overset{\circ}{I}^{-1}(\theta_0)/n$, where $\overset{\circ}{I}(\theta_0)$ is the information number of $\overset{\circ}{f}(x \mid \theta_0)$.

Proof. (Sketch). Let $\widehat{\theta}$ be the mode of the posterior pdf in (8.23). The proof of
Theorem 6.7 can be mimicked to show that $\widehat{\theta}$ is consistent; that is, $\widehat{\theta} \to \theta_0$ as $n \to \infty$. A second-order Taylor expansion of $\ln f(\theta \mid \mathbf{x})$ around $\widehat{\theta}$ gives

$$\ln f(\theta \mid \mathbf{x}) = \ln f(\widehat{\theta} \mid \mathbf{x}) + (\theta - \widehat{\theta})\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f(\widehat{\theta} \mid \mathbf{x}) + \frac{1}{2}(\theta - \widehat{\theta})^2 \frac{\mathrm{d}2}{\mathrm{d}\theta^2} \ln f(\widehat{\theta} \mid \mathbf{x}) + R$$

$$= \ln f(\widehat{\theta} \mid \mathbf{x}) + \frac{n}{2}(\theta - \widehat{\theta})^2 \underbrace{\left(\frac{1}{n}\frac{\mathrm{d}2}{\mathrm{d}\theta^2} \ln\!\big(c(\mathbf{x}) f_\Theta(\widehat{\theta})\big) + \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}2}{\mathrm{d}\theta^2} \ln \mathring{f}(x_i \mid \widehat{\theta}) \right)}_{(\star)} + R \;,$$

where $c(\mathbf{x})$ is the normalization constant of the posterior and R is the remainder term, which includes higher-order polynomials $(\theta - \widehat{\theta})^k$, $k = 3, 4, \ldots$. Note that the linear term in the Taylor expansion can be omitted since the derivative of $\ln(f(\theta \mid \mathbf{x}))$ at $\theta = \widehat{\theta}$ is 0. For large n the first term in $(\star)$ becomes negligible compared to the second one. Moreover, similar to (6.26), the second term converges to $-\mathring{I}(\theta_0)$. Since   ☞ 176
R/n remains bounded as $n \to \infty$ and $\ln f(\widehat{\theta} \mid \mathbf{x})$ is a constant with respect to θ, the posterior pdf $f(\theta \mid \mathbf{x})$ becomes more and more concentrated around θ_0 and tends to the form

$$f(\theta \mid \mathbf{x}) \propto \mathrm{e}^{-\frac{1}{2}(\theta - \theta_0)^2\, n\mathring{I}(\theta_0)} \;,$$

which is the pdf of the $\mathsf{N}(\theta_0, \mathring{I}^{-1}(\theta_0)/n)$ distribution, in accordance with Theorem 6.8.   □   ☞ 177

8.5   Priors and Conjugacy

In Bayesian analysis it is often useful to choose the prior pdf in the same family of distributions as the posterior pdf. Consider for example the binomial model in Example 8.1. Using a uniform prior, the posterior pdf belongs to Beta family of distributions. Suppose we choose the prior in the same family, giving the Bayesian model

$$\theta \sim \mathsf{Beta}(a, b)$$
$$(x \mid \theta) \sim \mathsf{Bin}(n, \theta)$$

for some fixed a and b. By Bayes' formula the posterior pdf satisfies

$$f(\theta \mid x) \propto \theta^{a-1}(1 - \theta)^{b-1}\theta^x(1 - \theta)^{n-x} = \theta^{a+x-1}(1 - \theta)^{b+n-x-1} \;,$$

which corresponds to the $\mathsf{Beta}(a+x, b+n-x)$ distribution. We see that the posterior and prior are in the same family of distributions. This property is called **conjugacy**. The advantage of conjugacy is that only the parameters of the distribution need to be considered. We say that the Beta family is a **conjugate family** for the binomial distribution.

Exponential families provide natural conjugate priors. Recall (see Definition 5.3) that a random variable x is said to belong to an m-dimensional exponential family if its pdf is of the form

☞ 152

$$\mathring{f}(x \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{m} \eta_i(\boldsymbol{\theta}) t_i(x)\right) h(x), \tag{8.24}$$

where we have used the Bayesian notation $\mathring{f}(x \mid \boldsymbol{\theta})$ instead of the classical notation $\mathring{f}(x; \boldsymbol{\theta})$.

Theorem 8.4. (Conjugate Prior for an Exponential Family). Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an iid sample from $\mathring{f}(x \mid \boldsymbol{\theta})$ of the form (8.24). The prior

$$f(\boldsymbol{\theta}) \propto c(\boldsymbol{\theta})^b \exp\left(\sum_{i=1}^{m} \eta_i(\boldsymbol{\theta}) a_i\right), \tag{8.25}$$

where the proportionality constant only depends on $(a_1, \ldots, a_m, b)$, is conjugate to the conditional pdf

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^{m} \eta_i(\boldsymbol{\theta}) \sum_{k=1}^{n} t_i(x_k)\right) \prod_{k=1}^{n} h(x_k). \tag{8.26}$$

Proof. By Bayes' theorem the posterior pdf satisfies

$$f(\boldsymbol{\theta} \mid \mathbf{x}) \propto f(\boldsymbol{\theta}) f(\mathbf{x} \mid \boldsymbol{\theta}) \propto c(\boldsymbol{\theta})^{n+b} \exp\left(\sum_{i=1}^{m} \eta_i(\boldsymbol{\theta})\Big(a_i + \sum_{k=1}^{n} t_i(x_k)\Big)\right),$$

where the proportionality constant does not depend on $\boldsymbol{\theta}$. This shows that the posterior pdf lies in the same $(m + 1)$-dimensional exponential family as the prior (8.25). In particular, if the prior is specified by parameters $(a_1, \ldots, a_m, b)$, then the corresponding parameters for the posterior are $(\widetilde{a}_1, \ldots, \widetilde{a}_m, \widetilde{b})$, with $\widetilde{a}_i = a_i + \sum_{k=1}^{n} t_i(x_k)$, $i = 1, \ldots, m$, and $\widetilde{b} = b + n$.   □

Example 8.5 (Conjugate Prior for Bernoulli Likelihood). In Example 8.1 we are dealing with independent Bernoulli random variables whose joint pdf conditional on θ is

$$f(\mathbf{x} \mid \theta) = \theta^{\sum_{k=1}^{n} x_k} (1 - \theta)^{n - \sum_{k=1}^{n} x_k},$$

which is of the form (8.26), with $m = 1$, $\eta(\theta) = \ln(\theta/(1-\theta))$, $t(x_k) = x_k$, and $c(\theta) = 1 - \theta$. The corresponding conjugate class is therefore of the form

$$c(\theta)^b e^{\eta(\theta)a} = (1-\theta)^b \left(\frac{\theta}{1-\theta} \right)^a \propto \theta^a (1-\theta)^b ,$$

which corresponds to the Beta family of distributions.

Example 8.6 (Conjugate Prior for Poisson Likelihood). Let $x_1, \ldots, x_n$ be an iid sample from the Poisson distribution $\mathsf{Poi}(\lambda)$. This is an exponential family, and the joint pdf can be written as

$$f(\mathbf{x} \,|\, \lambda) = e^{-n\lambda} e^{n\bar{x}\ln\lambda} \prod_{k=1}^{n} \frac{1}{x_k!} .$$

This is of the form (8.26), which suggests a conjugate prior of the form

$$f(\lambda) \propto e^{-b\lambda} e^{a\ln\lambda} = e^{-b\lambda} \lambda^a .$$

This corresponds to the gamma density. In particular, if we take a $\mathsf{Gamma}(a, b)$ prior for λ, that is,

$$f(\lambda) \propto e^{-b\lambda} \lambda^{a-1}$$

(notice λ is the variable here, not the parameter), then the posterior pdf is

$$f(\lambda \,|\, \mathbf{x}) \propto e^{-(n+b)\lambda} \lambda^{a-1+n\bar{x}} ,$$

which corresponds to the $\mathsf{Gamma}(a + n\bar{x}, b + n)$ distribution.

8.6 Bayesian Model Comparison

Under the Bayesian framework, hypothesis testing, or more generally comparing models, is straightforward. Suppose we wish to compare two models M_1 and M_2. Each model $M_i, i = 1, 2$, is formally defined by a likelihood function $f(\mathbf{x} \,|\, \boldsymbol{\theta}_i, M_i)$ and a prior distribution on the model-specific parameter vector $\boldsymbol{\theta}_i$ denoted as $f(\boldsymbol{\theta}_i \,|\, M_i)$. Note that in both the likelihood function and the prior distribution, we make the dependence on the model M_i explicit.

A popular criterion for comparing models M_1 and M_2 is the **Bayes factor** in favor of model M_1 against model M_2:

$$\mathrm{BF}_{12} \overset{\text{def}}{=} \frac{f(\mathbf{x} \,|\, M_1)}{f(\mathbf{x} \,|\, M_2)} ,$$

where

$$f(\mathbf{x} \mid M_i) = \int f(\mathbf{x} \mid \boldsymbol{\theta}_i, M_i) \, f(\boldsymbol{\theta}_i \mid M_i) \, \mathrm{d}\boldsymbol{\theta}_i \qquad (8.27)$$

is the **marginal likelihood** under model M_i, $i = 1, 2$.

The marginal likelihood $f(\mathbf{x} \mid M_i)$ is simply the marginal density of the data $\mathbf{x}$ under model M_i. If the actual data are likely under model M_i, then the associated marginal likelihood will be large, and vice versa. Hence, a Bayes factor BF_{12} greater than 1 indicates that model M_1 better predicts the observed data than M_2. It is therefore taken as evidence in favor of model M_1.

The Bayes factor between the two models is related to their **posterior odds ratio**:

$$\mathrm{PO}_{12} \stackrel{\text{def}}{=} \frac{\mathbb{P}(M_1 \mid \mathbf{x})}{\mathbb{P}(M_2 \mid \mathbf{x})} = \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} \times \frac{f(\mathbf{x} \mid M_1)}{f(\mathbf{x} \mid M_2)} \,,$$

where $\mathbb{P}(M_i)$ and $\mathbb{P}(M_i \mid \mathbf{x})$ are, respectively, the prior and posterior model probabilities of model M_i, $i = 1, 2$. If both models are equally probable *a priori*, i.e., $\mathbb{P}(M_1) = \mathbb{P}(M_2)$, the posterior odds ratio between the two models is then the same as the Bayes factor. In that case, if, for example, $\mathrm{BF}_{12} = 50$, we can say that model M_1 is 50 times more likely than model M_2 given the data.

☞ 242 **Example 8.7 (Comparing Multinomial Models).** In Example 8.3 we investigated if opinions on an anti-smoking campaign are independent of gender. Using the data in Table 8.1, we found evidence that suggests opinions differ by gender. In this example we perform a formal model comparison exercise to quantify the weight of evidence.

Let M_1 denote the multinomial model $(\mathbf{x} \mid \mathbf{p}, M_1) \sim \mathsf{Mnom}(501, \mathbf{p})$, where $\mathbf{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23})$, with a uniform prior for $\mathbf{p}$, or, equivalently, $(p_{11}, \ldots, p_{22} \mid M_1) \sim \mathsf{Dirichlet}(1, \ldots, 1)$. Hence, the prior density is given by

$$f(p_{11}, \ldots, p_{22} \mid M_1) = \Gamma(6) = 5! \,.$$

It follows that the marginal likelihood $f(\mathbf{x} \mid M_1)$ can be directly computed using the definition (8.27):

$$
\begin{aligned}
f(\mathbf{x} \mid M_1) &= \int \frac{501!}{x_{11}! \cdots x_{23}!} \, p_{11}^{x_{11}} \cdots p_{23}^{x_{23}} \times 5! \, \mathrm{d}(p_{11}, \ldots, p_{22}) \\
&= \frac{501! \, 5!}{x_{11}! \cdots x_{23}!} \int p_{11}^{x_{11}} \cdots p_{23}^{x_{23}} \, \mathrm{d}(p_{11}, \ldots, p_{22}) \\
&= \frac{501! \, 5!}{x_{11}! \cdots x_{23}!} \times \frac{\Gamma(x_{11}+1) \cdots \Gamma(x_{23}+1)}{\Gamma(507)} \\
&= \frac{501! \, 5!}{506!} \approx 3.6901 \times 10^{-12} \,.
\end{aligned}
$$

Next, if opinion is independent of gender, we must have

$$p_{ij} = p_i^{(r)} p_j^{(c)}, \quad i = 1, 2, \quad j = 1, 2, 3,$$

where $p_1^{(r)} + p_2^{(r)} = 1$ and $p_1^{(c)} + p_2^{(c)} + p_3^{(c)} = 1$. Let $r_i = x_{i1} + x_{i2} + x_{i3}, i = 1, 2$, and $c_i = x_{1j} + x_{2j}, j = 1, 2, 3$, denote the row and column counts, respectively. Then, the likelihood function under the model M_2 (in which opinion is independent of gender) is given by

$$f(\mathbf{x} \mid \widetilde{\mathbf{p}}, M_2) = \frac{501!}{x_{11}! \cdots x_{23}!} (p_1^{(r)} p_1^{(c)})^{x_{11}} \cdots (p_2^{(r)} p_3^{(c)})^{x_{23}}$$

$$= \frac{501!}{x_{11}! \cdots x_{23}!} (p_1^{(r)})^{r_1} (p_2^{(r)})^{r_2} (p_1^{(c)})^{c_1} (p_2^{(c)})^{c_2} (p_3^{(c)})^{c_3},$$

where $\widetilde{\mathbf{p}} = (p_1^{(r)}, p_2^{(r)}, p_1^{(c)}, p_2^{(c)}, p_3^{(c)})$. Further, we assume independent and uniform priors for $(p_1^{(r)}, p_2^{(r)})$ and $(p_1^{(c)}, p_2^{(c)}, p_3^{(c)})$. Hence, the prior density is

$$f(p_1^{(r)}, p_1^{(c)}, p_2^{(c)} \mid M_2) = \Gamma(2)\Gamma(3) = 2.$$

Following a similar computation as before, the marginal likelihood for model M_2 is given by

$$f(\mathbf{x} \mid M_2) = \frac{2 \times 501!}{x_{11}! \cdots x_{23}!} \frac{\Gamma(r_1 + 1)\Gamma(r_2 + 1)\Gamma(c_1 + 1)\Gamma(c_2 + 1)\Gamma(c_3 + 1)}{\Gamma(r_1 + r_2 + 2)\Gamma(c_1 + c_2 + c_3 + 3)}$$

$$= \frac{2 \times r_1! r_2! c_1! c_2! c_3!}{502 \times x_{11}! \cdots x_{23}! 503!} \approx 9.2122 \times 10^{-15}.$$

Finally, the Bayes factor is $\text{BF}_{12} = f(\mathbf{x} \mid M_1)/f(\mathbf{x} \mid M_2) \approx 400$, showing overwhelming evidence for M_1 against M_2. In other words, given the data, it is highly likely (400 times more so) that opinion varies with gender.

The computation of the marginal likelihood in (8.27) involves "integrating out" all the model parameters, and an analytic expression is often unavailable. In those cases, Monte Carlo methods are required to estimate the marginal likelihood. One popular method to do so using posterior output is the Chib's method (Chib 1995; Chib and Jeliazkov 2001).

However, when comparing *nested* models, i.e., when one model is a restricted version of the other model, the Bayes factor has an alternative expression that can often be easily estimated using posterior output. To set the stage, let M_u denote the *unrestricted* model, where the model parameters are partitioned into two subsets $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\omega})$. Suppose M_r is the *restricted* version of M_u, where $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\omega}_0)$ for some constant vector $\boldsymbol{\omega}_0$. Clearly, comparing M_u and M_r is equivalent to testing the hypothesis $\boldsymbol{\omega} = \boldsymbol{\omega}_0$.

Now, suppose $f(\boldsymbol{\psi}, \boldsymbol{\omega} \mid M_u)$ is the prior distribution under the unrestricted model. Then, the induced prior for $\boldsymbol{\psi}$ under the restricted model M_r is simply the marginal distribution $f(\boldsymbol{\psi} \mid M_r) = \int f(\boldsymbol{\psi}, \boldsymbol{\omega} \mid M_u) \, d\boldsymbol{\omega}$. It turns out that if this induced prior is the same as the conditional prior for $\boldsymbol{\psi}$ given $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, then the Bayes factor is equivalent to the ratio of posterior and prior densities under M_u evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. This is referred to as the **Savage–Dickey density ratio**. The result is summarized in the following theorem. Its proof can be found in, for example, Verdinelli and Wasserman (1995).

Theorem 8.5. (Savage–Dickey Density Ratio). Let M_u denote the unrestricted model with model parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\omega})$, and let M_r be a restricted version of M_u, with $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and free parameter vector $\boldsymbol{\psi}$. Suppose the priors in the two models satisfy

$$f(\boldsymbol{\psi} \mid M_r) = f(\boldsymbol{\psi} \mid \boldsymbol{\omega} = \boldsymbol{\omega}_0, M_u) \,. \tag{8.28}$$

Then, the Bayes factor in favor of model M_r can be written as

$$\mathrm{BF}_{ru} = \frac{f(\boldsymbol{\omega} = \boldsymbol{\omega}_0 \mid \mathbf{x}, M_u)}{f(\boldsymbol{\omega} = \boldsymbol{\omega}_0 \mid M_u)} \,.$$

In particular, (8.28) holds if $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ are *a priori* independent under M_u; that is, $f(\boldsymbol{\psi}, \boldsymbol{\omega} \mid M_u) = f(\boldsymbol{\psi} \mid M_u) f(\boldsymbol{\omega} \mid M_u)$.

Writing the Bayes factor as such a ratio of densities avoids the often difficult task of computing marginal likelihoods. The denominator $f(\boldsymbol{\omega} = \boldsymbol{\omega}_0 \mid M_u)$ can frequently be calculated analytically, when the conditional prior $f(\boldsymbol{\omega} \mid M_u)$ is of a standard form. In addition, the numerator can often be estimated from posterior output of model M_u. In particular, the numerator can be estimated via $\frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\omega} = \boldsymbol{\omega}_0 \mid \mathbf{x}, \boldsymbol{\psi}_i, M_u)$, where $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N$ are posterior draws from model M_u.

☞ 149 **Example 8.8 (Comparing Polynomial Regression Models).** In Example 5.16 we considered five different polynomial regression models for fitting the data in Table 5.3 and compared the models using cross-validation. In this example, we perform a Bayesian model comparison on the same data. Let model M_i denote the ith-order polynomial regression model, $i = 1, \dots, 5$:

$$y_k = \beta_0 + \beta_1 x_k + \cdots + \beta_i x_k^i + \varepsilon_k \,,$$

where $\{\varepsilon_k\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$. Clearly, models $M_1, \dots, M_4$ are all nested within model M_5. To complete the model specification (of model M_5), we take the following independent priors: $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)^\top \sim \mathsf{N}(\mathbf{0}, 100\, I)$ and $\sigma^2 \sim \mathsf{InvGamma}(2, 1)$.

To compare models via the Bayes factor, we can obtain posterior draws from model M_5 and estimate the relevant Savage–Dickey density ratio (since $\beta_0, \ldots, \beta_5$ are independent under the prior, the condition (8.28) is satisfied). For example, model M_3 is obtained by imposing $(\beta_4, \beta_5) = \mathbf{0}$. Hence, the Bayes factor BF_{35} can be written as

$$\mathrm{BF}_{35} = \frac{f((\beta_4, \beta_5) = \mathbf{0} \mid \mathbf{y}, M_5)}{f((\beta_4, \beta_5) = \mathbf{0} \mid M_5)} .$$

Using the properties of the multivariate normal distribution (see Theorem 3.7), the marginal prior $f(\beta_4, \beta_5 \mid M_5)$ is a bivariate normal density and can be evaluated easily. The conditional posterior $f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, M_5)$ is also a normal density, by Theorem 8.1. Hence, the numerator in the ratio can be estimated using posterior draws for σ^2.

☞ 85

☞ 238

The following MATLAB script estimates the log-Bayes factors $\ln \mathrm{BF}_{i5}, i = 1, \ldots, 4$, via the Savage–Dickey density ratio approach. It calls the function `lmvnpdf.m` (see below) that evaluates the log-density of the $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

```matlab
% polyreg_bayes.m
x = [4.7,2,2.7,0.1,4.7,3.7,2,3.4,1.3,3.8,4.8,1.7,...
    -0.4,4.5,1.3,0.4,2.6,4,2.9,1.6]';
y = [6.57,5.15,7.15,0.18,6.48,8.95,5.24,10.54,1.24,...
    8.05,3.56,3.4,2.18,7.16,2.32,-0.23,7.68,9.09,9.13,4.04]';
n = length(x);
X = [ones(n,1) x x.^2 x.^3 x.^4 x.^5];
XX = X'*X;
Xy = X'*y;
m = 6;
N = 10^5;           % Gibbs sample size
V0 = eye(m)*100;    % prior for beta
invV0 = V0\eye(m);
alp0 = 2; lam0 = 1;  % prior for sig2
beta = XX\Xy;
sig2 = sum((y-X*beta).^2)/n;
gibbs_sample = zeros(N,m+1);
lpostden_sample = zeros(N,4);
for k=1:N
    D = (invV0 + XX/sig2)\eye(m);
    betahat = D*(Xy/sig2);
    beta = betahat + chol(D,'lower')*randn(m,1);
    sig2 = 1/gamrand(alp0+n/2, lam0+sum((y-X*beta).^2)/2);
    gibbs_sample(k,:)=[beta' sig2];
    lp1 = lmvnpdf([0 0 0 0]',betahat(3:end),D(3:end,3:end));
    lp2 = lmvnpdf([0 0 0]',betahat(4:end),D(4:end,4:end));
    lp3 = lmvnpdf([0 0]',betahat(5:end),D(5:end,5:end));
    lp4 = lmvnpdf(0,betahat(6),D(6,6));
    lpostden_sample(k,:) = [lp1 lp2 lp3 lp4];
end
lpostden = zeros(4,1);
for i=1:4
```

```
    maxpden = max(lpostden_sample(:,i));
    lpostden(i)=log(mean(exp(lpostden_sample(:,i)-maxpden)))...
        + maxpden;
end
lpriden = zeros(4,1);
lpriden(1) = lmvnpdf([0 0 0 0]',[0 0 0 0]',V0(3:end,3:end));
lpriden(2) = lmvnpdf([0 0 0]',[0 0 0]',V0(4:end,4:end));
lpriden(3) = lmvnpdf([0 0]',[0 0]',V0(5:end,5:end));
lpriden(4) = lmvnpdf(0,0,V0(6,6));
lBF = lpostden - lpriden;
```

```
function lden = lmvnpdf(x,mu,Sig)
m = size(mu,1);
e = x-mu;
c = -m/2*log(2*pi) - sum(log(diag(chol(Sig,'lower'))));
lden = c - .5*e'*(Sig\e);
```

The log-Bayes factors $\ln \mathrm{BF}_{15}, \ldots, \ln \mathrm{BF}_{45}$ are estimated to be, respectively, $-2.08, -2.63, 10.69$, and 5.72. In other words, compared to model M_5, the data favor models M_3 and M_4, but not models M_1 and M_2. Furthermore, note that the Bayes factor BF_{34} can be written as

$$\mathrm{BF}_{34} = \frac{f(\mathbf{y} \mid M_3)}{f(\mathbf{y} \mid M_4)} = \frac{f(\mathbf{y} \mid M_3)}{f(\mathbf{y} \mid M_5)} \times \frac{f(\mathbf{y} \mid M_5)}{f(\mathbf{y} \mid M_4)} = \frac{\mathrm{BF}_{35}}{\mathrm{BF}_{45}}.$$

Hence, an estimate of BF_{34} is $\mathrm{e}^{10.69-5.72} \approx 144$. To conclude, the data decisively prefer the cubic polynomial regression model. If we assume equal prior probabilities for all the models, the cubic polynomial is about 144 times more likely than the next best model ($4th$-order polynomial) given the data.

8.7 Problems

☞ 74   **8.1.** Let $f(x), x \in (0, 1)$ be the pdf of $X \sim \mathsf{Beta}(\alpha, \beta)$.

a. Prove that the derivative of f (or, equivalently, of $\ln f$) has a unique zero at $x^* = (\alpha - 1)/(\alpha + \beta - 2)$ in the interval $(0, 1)$, provided that either $\alpha > 1, \beta > 1$ or $\alpha < 1, \beta < 1$. For which of these two regimes is x^* a maximum point?

☞ 74   b. Show that $\mathbb{E}X = B(\alpha + 1, \beta)/B(\alpha, \beta)$, where B is the *beta* function (3.11).
☞ 48   Using the properties of the *gamma* function (2.20) show that $\mathbb{E}X = \alpha/(\alpha + \beta)$.

8.2. Suppose $x_1 = 1.1065, x_2 = 0.5343, x_3 = 11.1438, x_4 = 0.4893, x_5 = 2.4748$ is an observed iid sample from the $\mathsf{Exp}(\lambda)$ distribution. Consider Bayesian inference for the parameter λ, using an improper prior $f(\lambda) = 1/\lambda$.

a. Show that the posterior pdf of λ has a Gamma$(5,\ 15.7487)$ distribution.
b. Give the expectation of the posterior pdf.

8.3. Let $(x\,|\,\lambda)\ \sim\ \text{Poi}(\lambda)$, and suppose that the prior distribution for λ is Gamma(a, b), where a and b are known. Find the posterior pdf of λ.

8.4. Let $x \sim \text{Gamma}(\alpha, \lambda)$. Show that the pdf of $z = 1/x$ is given by (8.6). ☞ 234

8.5. Consider the transformation $(z_1, \ldots, z_{m+1})^\top \mapsto (y_1, \ldots, y_{m+1})^\top$ defined by $y_i = z_i\, z_{m+1}$, $i = 1, \ldots, m$ and $y_{m+1} = (1 - (z_1 + \cdots + z_m))\, z_{m+1}$. Show that the determinant of the corresponding matrix of Jacobi is z_{m+1}^m. This is used in the proof of Theorem 8.2. ☞ 241

8.6. Let $\mathbf{Z} = (Z_1, \ldots, Z_m) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{m+1})$. Show that the marginal distribution of Z_i is Beta$(\alpha_i, \sum_{j \neq i} \alpha_j)$. Hint: use Theorem 8.2. ☞ 241

8.7. Let $(x\,|\,p) \sim \text{Geom}(p)$. Suppose that the prior distribution of p is U$(0, 1)$.

a. Find the posterior pdf of p.
b. Find the posterior mode.
c. Find the posterior expectation.

8.8. The data 0.4453, 9.2865, 0.4077, 2.0623, 10.4737, 5.7525, 2.7159, 0.1954, 0.1608, and 8.3143 were drawn from an Exp$(1/\theta)$ distribution. Consider a Bayesian model with a constant prior for θ.

a. Show that the posterior distribution of θ is inverse-gamma, and determine the parameters.
b. Determine estimates of the 0.025 and 0.975 quantiles of the posterior distribution, using $N = 10^5$ simulated samples from the posterior distribution.

8.9. Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ is an iid sample from N(μ, σ^2) with *known* variance σ^2. As a prior for μ take the N(μ_0, σ_0^2) distribution for some fixed parameters μ_0 and σ_0^2. The Bayesian model is therefore

$$\mu \sim \text{N}(\mu_0, \sigma_0^2)\ ,$$

$$(x_1, \ldots, x_n\,|\,\mu) \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)\ .$$

Show that the posterior pdf $f(\mu\,|\,\mathbf{x})$ corresponds to the pdf of the N(μ_1, σ_1^2) distribution with

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\ .$$

8.10. Consider the simplified Bayesian model for normal data in Sect. 8.2.1; that is,

$$f(\mu, \sigma^2) = 1/\sigma^2\ ,$$

$$(\mathbf{x}\,|\,\mu, \sigma^2) \sim \text{N}(\mu\mathbf{1}, \sigma^2 I)\ .$$

The joint posterior pdf is

$$f(\mu, \sigma^2 \mid \mathbf{x}) \propto \left(\sigma^2\right)^{-n/2-1} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2}\right\} . \tag{8.29}$$

The marginal posterior pdfs of μ and σ^2 can be obtained by integrating out the other variable.

a. Prove that

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\mu - \bar{x})^2 . \tag{8.30}$$

b. By using (8.30) show that

$$f(\sigma^2 \mid \mathbf{x}) \propto \left(\sigma^2\right)^{-n/2-1/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}\right\} . \tag{8.31}$$

c. Show that (8.31) corresponds to the $\mathsf{InvGamma}\big((n-1)/2,\ s_x^2(n-1)/2\big)$ distribution, where s_x^2 is the classical sample variance of the $\{x_i\}$.

d. Let q_1 and q_2 be the $\gamma/2$ and $1 - \gamma/2$ quantiles of (8.31). Show that the $1 - \gamma$ credible interval (q_1, q_2) is *identical* to the classic confidence interval (5.19) (with α replaced by γ).

e. By using (8.30) and (8.6) show that

$$f(\mu \mid \mathbf{x}) \propto \left(\sum_{i=1}^{n}(x_i - \mu)^2\right)^{-n/2} \propto \left(\frac{(\mu - \bar{x})^2 n}{s_x^2 \nu} + 1\right)^{-(\nu+1)/2} ,$$

where $\nu = n - 1$. Verify that, in view of (2.23), this means that

$$\left(\frac{\mu - \bar{x}}{s_x/\sqrt{n}} \,\Big|\, \mathbf{x}\right) \sim \mathsf{t}_{n-1} .$$

f. Let q_1 and q_2 be the $\gamma/2$ and $1 - \gamma/2$ quantiles of $f(\mu \mid \mathbf{x})$. Show that the $1 - \gamma$
credible interval (q_1, q_2) is *identical* to the classic confidence interval (5.18) (with α replaced by γ).

8.11. In Problem 8.10 compare the simulated densities in Fig. 8.3 with the exact ones. In particular, plot the pdf of $(\sigma^2 \mid \mathbf{x})$, that is, the pdf of the random variable $(n-1)s_x^2 Y$, where $Y \sim \mathsf{InvGamma}((n-1)/2, 1/2)$. Similarly, plot the pdf of $(\mu \mid \mathbf{x})$, that is, of the random variable $\bar{x} + T s_x/\sqrt{n}$, where $T \sim \mathsf{t}_{n-1}$.

8.12. In the *zero-inflated Poisson* model, random data $x_1, \ldots, x_n$ are assumed to be of the form $x_i = r_i y_i$, where the $\{y_i\}$ have a $\mathsf{Poi}(\lambda)$ distribution and the $\{r_i\}$ have a $\mathsf{Ber}(p)$ distribution, all independent of each other. Given an outcome

$\mathbf{x} = (x_1, \ldots, x_n)$, the objective is to estimate both λ and p. Consider the following hierarchical Bayesian model:

$$p \sim \mathsf{U}(0, 1) \, ,$$

$$(\lambda \mid p) \sim \mathsf{Gamma}(a, b) \, ,$$

$$(r_i \mid p, \lambda) \sim \mathsf{Ber}(p) \quad \text{independently} \, ,$$

$$(x_i \mid \mathbf{r}, \lambda, p) \sim \mathsf{Poi}(\lambda \, r_i) \quad \text{independently} \, ,$$

where $\mathbf{r} = (r_1, \ldots, r_n)$ and a and b are known parameters. We wish to sample from the posterior pdf $f(\lambda, p, \mathbf{r} \mid \mathbf{x})$ using the Gibbs sampler.

a. Show that

$$f(\mathbf{r}, \lambda, p \mid \mathbf{x}) \propto \lambda^{a-1} e^{-b\lambda} \prod_{i=1}^{n} e^{-\lambda \, r_i} (\lambda \, r_i)^{x_i} \, p^{r_i} (1 - p)^{1 - r_i} \, .$$

b. Show that

$$(\lambda \mid p, \mathbf{r}, \mathbf{x}) \sim \mathsf{Gamma} \left(a + \sum_{i=1}^{n} x_i, b + \sum_{i=1}^{n} r_i \right),$$

$$(p \mid \lambda, \mathbf{r}, \mathbf{x}) \sim \mathsf{Beta} \left(1 + \sum_{i=1}^{n} r_i, 1 + n - \sum_{i=1}^{n} r_i \right),$$

and, for $k = 1, \ldots, n$,

$$(r_k \mid \lambda, p, \mathbf{x}) \sim \mathsf{Ber} \left(\frac{p \, e^{-\lambda}}{p \, e^{-\lambda} + (1 - p) \, \mathrm{I}_{\{x_k = 0\}}} \right) \, .$$

c. Generate an iid sample of size $n = 100$ for the zero-inflated Poisson model using parameters $p = 0.3$ and $\lambda = 2$.

d. Implement the Gibbs sampler, generate a large (dependent) sample from the posterior distribution, and use this to construct 95% credible intervals for p and λ using the data in (c). Compare these with the true values.

8.13. For a Markov chain $x_1, \ldots, x_n$, the joint pdf is of the form

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_2) \cdots f(x_n \mid x_{n-1}) \, .$$

The corresponding Bayesian network is given in the left pane of Fig. 8.9. An alternative Bayesian network for the same Markov chain is given in the right pane

of the figure, where the arcs have been turned around. Show that both networks represent the same joint pdf. Hint: write $f(x_{t+1} \mid x_t)$ in terms of $f(x_t \mid x_{t+1})$.

Fig. 8.9 Bayesian networks for a Markov chain

8.14. Figure 8.10 shows the Bayesian network for a **hidden Markov model**. Here $x_1, \ldots, x_n$ is a Markov chain on $\{1, \ldots, K\}$, defined by an initial (discrete) pdf $f(x_1)$ and transition probabilities $f(x_t \mid x_{t-1})$, which are here assumed to be known. For each time $t = 1, 2, \ldots, n$ the state of the chain, x_t, remains *hidden*. Instead, a variable y_t is observed, whose (known) distribution depends only on x_t; for example, $(y_t \mid x_t) \sim \mathsf{N}(x_t, 1)$.
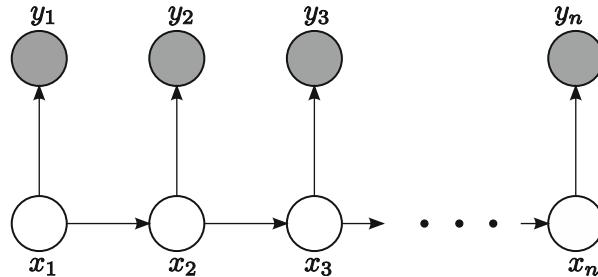
Fig. 8.10 Bayesian network for a hidden Markov model

A typical object of interest for such models is the posterior pdf $f(x_t \mid \mathbf{y}_{1:t})$, where $\mathbf{y}_{1:t} = (y_1, \ldots, y_t)$. That is, we wish to assess the state at time t given all the observations at and before time t.

a. Prove that

$$f(x_t, \mathbf{y}_{1:t}) = \sum_{x_{t-1}} f(x_t, y_t \mid x_{t-1}, \mathbf{y}_{1:t-1}) f(x_{t-1}, \mathbf{y}_{1:t-1}) . \tag{8.32}$$

b. Further, show that

$$f(x_t, y_t \mid x_{t-1}, \mathbf{y}_{1:t-1}) = f(x_t \mid x_{t-1}) f(y_t \mid x_t) . \tag{8.33}$$

c. Express $f(x_1, y_1)$ in terms of $f(x_1)$ and $f(y_1 \mid x_1)$. Explain how, with $f(x_1, y_1)$, (8.32), and (8.33), the posterior distribution of x_t given $\mathbf{y}_{1:t}$ can be determined recursively for $t = 2, 3, \ldots, n$.

8.15. Find an appropriate conjugate family for the $\mathsf{Exp}(\lambda)$ distribution, using
Theorem 8.4.

8.16. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an iid sample from $\mathsf{Exp}(1/\theta)$ for some θ. Show
that $\theta \sim \mathsf{InvGamma}(\alpha_0, \lambda_0)$ is a conjugate prior for this distribution. Determine the
resulting posterior distribution.

8.17. Suppose $f(\boldsymbol{\theta} \mid \mathbf{x})$ is the posterior pdf for some Bayesian estimation problem.
For example, $\boldsymbol{\theta}$ could represent the parameters of a regression model based on the
data $\mathbf{x}$. An important use for the posterior pdf is to make predictions about the
distribution of other random variables. For example, suppose that, conditional on $\mathbf{x}$,
some random vector $\mathbf{y}$ depends on $\boldsymbol{\theta}$ via the conditional pdf $f(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{y} \mid \boldsymbol{\theta})$.
Thus, conditional on $\boldsymbol{\theta}$, the random vector $\mathbf{y}$ is independent of $\mathbf{x}$. The **predictive pdf**
of $\mathbf{y}$ given $\mathbf{x}$ is defined as $f(\mathbf{y} \mid \mathbf{x})$, which can be written as

$$f(\mathbf{y} \mid \mathbf{x}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{x}) \, d\boldsymbol{\theta} \ . \tag{8.34}$$

This can be viewed as the expectation of $f(\mathbf{y} \mid \boldsymbol{\theta})$ under the posterior pdf. Therefore,
we can use Monte Carlo simulation to approximate $f(\mathbf{y} \mid \mathbf{x})$ via

$$f(\mathbf{y} \mid \mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{y} \mid \boldsymbol{\theta}_i) \ ,$$

where the sample $\{\boldsymbol{\theta}_i, i = 1, \ldots, N\}$ is obtained from $f(\boldsymbol{\theta} \mid \mathbf{x})$, for example, via
MCMC.

a. Prove (8.34).
b. As a concrete example, suppose that the iid data $-0.4326, -1.6656$,
   $0.1253, 0.2877$, and -1.1465 come from some $\mathsf{N}(\mu, \sigma^2)$ distribution. Define
   $\boldsymbol{\theta} = (\mu, \sigma^2)$. Let $Y \sim \mathsf{N}(\mu, \sigma^2)$ be a new measurement. Estimate and plot
   the predictive pdf $f(y \mid \mathbf{x})$, using a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ obtained via the Gibbs
   sampler of Example 8.2.1. Take $N = 1000$. Compare this with the "common-
   sense" Gaussian pdf with expectation $\bar{x}$ (sample mean) and variance s^2 (sample
   variance).

8.18. The **bag-of-words** method is a popular procedure for classification. Given are
k *objects* that are each characterized by n *features*. For example, the objects could be
k different people and the features could be various facial measurements, such as the
width of the eyes divided by the distance between the eyes or the ratio of the nose
height and mouth width. The features, $x_1, \ldots, x_n$, say, have a known distribution
and are assumed to be conditionally independent of each other given the object p;
that is, $f(x_1, \ldots, x_n \mid p) = f(x_1 \mid p) \cdots f(x_n \mid p)$. Assuming a uniform prior for p,
the posterior pdf is thus given by

$$f(p \mid x_1, \ldots, x_n) \propto \prod_{i=1}^{n} f(x_i \mid p) .$$

To classify the object on the basis of the features, simply take the p that maximizes the unnormalized posterior pdf.

a. Give the Bayesian network for the joint pdf of $p, x_1, \ldots, x_n$.
b. Suppose the ith feature distribution of object p is $N(\mu_{pi}, \sigma^2)$, $p = 1, \ldots, k$, $i = 1, \ldots, n$. Define $\boldsymbol{\mu}_p = (\mu_{p1}, \ldots, \mu_{pn})$, $p = 1, \ldots, k$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the vector of observed features. Let $p^* = \operatorname{argmin}_p \|\boldsymbol{\mu}_p - \mathbf{x}\|$; that is, among all feature vectors $\{\boldsymbol{\mu}_p\}$, the vector $\boldsymbol{\mu}_{p^*}$ is closest to $\mathbf{x}$. Show that p^* also maximizes the posterior pdf.
c. Next, consider the case where the ith feature of object p is $N(\mu_{pi}, \sigma_{pi}^2)$ distributed. Table 8.2 lists the means μ and standard deviations σ of the normal feature distributions of four objects. The observed features of an object are $(x_1, x_2, x_3) = (1.67, 2.00, 4.23)$. How should this object be classified?

Table 8.2 Feature parameters

	Feature 1		Feature 2		Feature 3	
Object	μ	σ	μ	σ	μ	σ
1	1.6	0.1	2.4	0.5	4.3	0.2
2	1.5	0.2	2.9	0.6	6.1	0.9
3	1.8	0.3	2.5	0.3	4.2	0.3
4	1.1	0.2	3.1	0.7	5.6	0.3

Part III
Advanced Models and Inference

In Part III of the book we consider estimation and inference for a wide variety of advanced models. Topics include generalized linear models with discrete responses, autoregressive–moving average models for time series, Gaussian models for data arising from repeated measurements, and state space models for data exhibiting time-varying persistence and volatility. Both classical and Bayesian estimation of these models are covered. It is assumed that the reader is familiar with the statistical concepts and computational techniques discussed in Part II.

Chapter 9
Generalized Linear Models

The linear models introduced in Chap. 4 deal with *continuous* response variables—
such as height and crop yield—and continuous or discrete explanatory variables.
For example, under a normal linear model, the responses $\{Y_i\}$ are independent
of each other, and each has a normal distribution with mean $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, where
$\mathbf{x}_i^\top$ is the ith row of the design matrix X. However, these continuous models are
obviously not suitable for data that take on *discrete* values. For example, we might
want to analyze women's labor market participation decision (whether to work or
not), voters' opinion of the government (rating on the government performance on
a scale of five), or the choice among a few cereal brands, as a function of one
or more explanatory variables. In this chapter we discuss models that are suitable
for analyzing these discrete response variables. We will first introduce the flexible
framework of *generalized linear models*.

9.1 Generalized Linear Models

Definition 9.1. (Generalized Linear Model). A vector of (response) data
$\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is said to satisfy a **generalized linear model** if the
expectation vector $\boldsymbol{\mu} = \mathbb{E}\mathbf{Y}$ can be written in the form

$$\boldsymbol{\mu} = \mathbf{g}^{-1}(X\boldsymbol{\beta}),$$

where X is an $n \times m$ **design matrix** (i.e., a matrix of explanatory variables),
$\boldsymbol{\beta}$ is an m-dimensional vector of **parameters**, and $\mathbf{g}^{-1}$ is the inverse of a
link function g. The distribution of $\mathbf{Y}$ may depend on additional **dispersion**
parameters that model the randomness in the data that is not explained by the
explanatory variables.

A common assumption for $\mathbf{Y}$ is that its components $Y_1, \ldots, Y_n$ are independent and come from some exponential family. The central focus is the parameter vector $\boldsymbol{\beta}$, which summarizes how the matrix of explanatory variables X affects the response vector $\mathbf{Y}$. By choosing different members of the exponential family and different link functions, the class of generalized linear models can encompass a wide variety of popular models as special cases, some of which are discussed below.

Example 9.1 (Normal Linear Model). The normal linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in Sect. 4.5 is a special case of a generalized linear model. Here, $\boldsymbol{\mu} = X\boldsymbol{\beta}$, so that the link function is simply the identity function: $g(\mathbf{z}) = \mathbf{z}$. The vector $\mathbf{Y}$ has a multivariate normal distribution

$$\mathbf{Y} \sim \mathsf{N}(\boldsymbol{\mu}, \sigma^2 I) \; ,$$

where σ^2 is a dispersion parameter that models the residual randomness in the data.

Example 9.2 (Binary Variable Regression Model). Suppose we are interested in the effectiveness of a certain insecticide. For this purpose an experiment is carried out as follows: the ith insect is exposed to the insecticide with dose level x_i, and we observe Y_i, whether the insect is killed or not. Thus, $Y_i \sim \mathsf{Ber}(\mu_i)$, where $\mu_i = \mathbb{E}Y_i$ is the "success" probability, which has to lie in the interval $(0, 1)$. Let $\mathbf{Y}$ and $\mathbf{x}$ be the response and explanatory vectors. One way to link the expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ to $\mathbf{x}$ is to specify μ_i as

$$\mu_i = F(\beta_0 + \beta_1 x_i)$$

for some cdf F and "regression" parameters β_0 and β_1. Defining the $n \times 2$ design matrix $X = (\mathbf{1} \; \mathbf{x})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, the distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is completely specified by $\boldsymbol{\mu}$, which in turn is determined by $X\boldsymbol{\beta}$. For different choices of F, we have different binary variable models. Common choices for F are (1) the cdf of the standard normal distribution and (2) the cdf of the logistic distribution. These are discussed in detail in the next section. The choice

$$F_{\text{ex}}(x) = 1 - \mathrm{e}^{-\mathrm{e}^x}$$

gives the cdf of the **extreme value distribution**. The corresponding link function for each component is $F_{\text{ex}}^{-1}(z) = \ln(-\ln(1 - z))$. Finally, by taking F as the cdf of the Student's t distribution with parameter ν, we obtain the so-called **t-link model**. One attractive feature of the t-link model is its flexibility; in particular, it includes the popular probit model (see next section) as a limiting case.

9.2   Logit and Probit Models

In this section we discuss two popular specifications for binary data: the **probit** model and the **logit** or **logistic** model. Both models are binary variable regression models of the form discussed in Example 9.2. More precisely, the responses $Y_1, \ldots, Y_n$ are assumed to be independent Bernoulli random variables with success probabilities

$$\mu_i = F(\mathbf{x}_i^\top \boldsymbol{\beta}), \quad i = 1, \ldots, n ,$$

where $\mathbf{x}_i$ is the vector of explanatory variables corresponding to the ith response, $\boldsymbol{\beta}$ is the parameter vector of interest, and F is a cdf.

9.2.1   Logit Model

> **Definition 9.2.   (Logit Model).** Let Y_i denote the ith binary response, and let $\mathbf{x}_i$ represent the vector of explanatory variables and $\boldsymbol{\beta}$ the associated parameter vector. In a **logistic regression** or **logit model**, the $\{Y_i\}$ are independent and $Y_i \sim \mathsf{Ber}(\mu_i)$, with $\mu_i = F(\mathbf{x}_i^\top \boldsymbol{\beta})$, where F is the cdf of **logistic distribution**:
>
> $$F(x) = \frac{1}{1 + e^{-x}} .$$
>
> In other words, the component link function is $g(x) = \ln(x/(1 - x))$.

Example 9.3 (Logit Model). Figure 9.1 shows the outcomes of 500 independent binary response variables for a logistic regression model. The explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$, $i = 1, \ldots, 500$ were chosen uniformly on the unit square, and $\boldsymbol{\beta} = (-8, 8)^\top$. The S-shaped surface depicts the graph of the function

$$p(x_1, x_2) = F(\mathbf{x}^\top \boldsymbol{\beta}) = (1 + \exp(8(x_2 - x_1)))^{-1} .$$

For each given vector of explanatory variables $(x_{i1}, x_{i2})^\top$, the response Y_i is generated from a Bernoulli distribution with success probability $p(x_{i1}, x_{i2})$.

Using the same notation as in Definition 9.2, we now derive the log-likelihood function, the score function, and the information matrix for this model. Since the responses are independent Bernoulli random variables, the log-likelihood function is given by

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \ln \mu_i + (1 - y_i) \ln(1 - \mu_i) \right] ,$$

where $\mu_i = (1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})^{-1}$ and $1 - \mu_i = e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}/(1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})$. It follows that $\ln \mu_i = -\ln(1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})$ and $\ln(1 - \mu_i) = -\mathbf{x}_i^\top \boldsymbol{\beta} - \ln(1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})$. After some algebra, the log-likelihood function can be rewritten as

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[(y_i - 1)\mathbf{x}_i^\top \boldsymbol{\beta} - \ln(1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}) \right] . \tag{9.1}$$

☞ 368 Taking the gradient of the log-likelihood function, we obtain the score function

$$\mathbf{S}(\boldsymbol{\beta}; \mathbf{y}) = \nabla_{\boldsymbol{\beta}}\, l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[(y_i - 1)\mathbf{x}_i + \frac{\mathbf{x}_i\, e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} \right]$$

$$= \sum_{i=1}^{n} \left[y_i - (1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})^{-1} \right] \mathbf{x}_i$$

$$= \sum_{i=1}^{n} (y_i - \mu_i)\, \mathbf{x}_i \; .$$

Differentiating the score function with respect to $\boldsymbol{\beta}$ and multiplying by -1, we obtain the **observed information matrix**:

☞ 168

$$I(\boldsymbol{\beta}; \mathbf{y}) = -\nabla_{\boldsymbol{\beta}}^2\, l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \frac{e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}{(1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}})^2}\, \mathbf{x}_i\, \mathbf{x}_i^\top$$

$$= \sum_{i=1}^{n} \mu_i (1 - \mu_i)\, \mathbf{x}_i\, \mathbf{x}_i^\top \; . \tag{9.2}$$

It is worth noting that the observed information matrix does not depend on the data $\mathbf{y}$, and therefore it coincides with the Fisher information matrix $I(\boldsymbol{\beta})$. Now, ☞ 169 the maximum likelihood estimate can be computed numerically using, say, Fisher's scoring method. Specifically, given an initial value $\boldsymbol{\beta}_0$, for $t = 1, 2, \ldots$, iteratively ☞ 180 compute

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + [I(\boldsymbol{\beta}_{t-1})]^{-1}\mathbf{S}(\boldsymbol{\beta}_{t-1}; \mathbf{y}) \ ,$$

until the sequence $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots$ is found to have converged, using some prefixed convergence criterion. Once the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is obtained, one can readily compute the corresponding asymptotic covariance matrix as $I^{-1}(\widehat{\boldsymbol{\beta}})$; see also Theorem 6.8. ☞ 177

Example 9.4 (MLE for the Logit Model). In the development of drugs, bioassay experiments are often carried out on animals to test the potential toxicity of the drugs. Various dose levels are given to batches of animals, and the animals' responses—typically characterized by a binary outcome, say, alive or dead—are recorded. The aim is to describe the probability of "success," μ, as a function of the dose, x, via a link function $g(\mu) = \beta_0 + \beta_1 x$. In this example we analyze the data with a logit model with

$$\mu = g^{-1}(\beta_0 + \beta_1 x) = (1 + e^{-(\beta_0 + \beta_1 x)})^{-1} \ .$$

The outcomes of such an experiment are given in Table 9.1: a total of 20 animals were tested, five at each of the four dose levels.

Table 9.1 Animal mortality data

Dose (log g/ml)	Number of animals	Number of deaths
−0.863	5	0
−0.296	5	1
−0.053	5	3
0.727	5	5

One obvious quantity of interest is the estimate for β_1. In particular, we are interested to know whether or not it is positive (i.e., if the drug is toxic). In addition, we might also want to learn about the effect of a specific dose level. Since we only have two parameters, we first obtain a contour plot for the likelihood function to get a rough estimate for $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\top}$. From Fig. 9.2 it can be seen that the maximum likelihood estimate for $\boldsymbol{\beta}$ is around $(1, 8)^{\top}$.

We use the following MATLAB code to implement Fisher's scoring method to obtain the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ and the information matrix evaluated at $\widehat{\boldsymbol{\beta}}$.
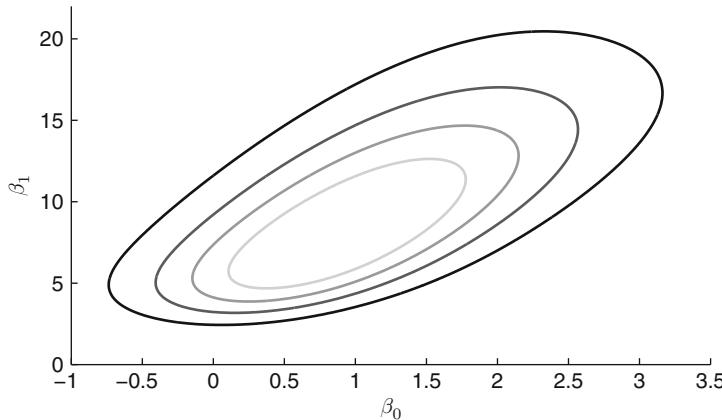
Fig. 9.2 Contour plot for the likelihood function of the parameters in the bioassay example

```
%bioassay.m
y = [0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 1 1 1 1 1]';
x = repmat([-0.863 -0.296 -0.053 0.727],5,1);
X = [ones(20,1) x(:)];       % design matrix
betat = (X'*X)\(X'*y);       % initial guess
S = ones(2,1);               % score
e = 10^(-5);                 % tolerance level
while sum(abs(S)) > e        % stopping criteria
    mu = 1./(1+exp(-X*betat));
    S = sum(repmat((y - mu),1,2).*X)';
    I = X'*diag(mu.*(1-mu))*X; %info matrix
    betat = betat + I\S;
end
V = I\eye(2);                % covariance matrix
```

Note that we have vectorized the computation of the score and information matrix in the code to avoid for loops. For example, the information matrix I in (9.2) can be written as $I = X^\top BX$, with

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{20} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \mu_1(1-\mu_1) & \cdots & & 0 \\ & & \ddots & \\ 0 & & \cdots & \mu_{20}(1-\mu_{20}) \end{pmatrix}.$$

The maximum likelihood estimate for $\boldsymbol{\beta}$ and the associated covariance matrix $V = I^{-1}(\boldsymbol{\beta})$ are

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.873 \\ 7.912 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} 1.081 & 3.833 \\ 3.833 & 25.624 \end{pmatrix}. \tag{9.3}$$

In particular, a 95% (approximate) confidence interval for β_1 is given as $7.912 \pm 1.96\sqrt{25.624}$ or $(-2, 17.83)$. It is interesting to note that we cannot reject the null hypothesis $\beta_1 = 0$ at significance level 0.05, even though the contour plot suggests that most of the mass of the likelihood lies in the region 2–20. One reason might be because the normal distribution is not a good approximation due to the small sample size.

Further, suppose that we are interested in the "success" rate at dose level -0.1 log g/ml. An estimate can be computed as $\widehat{\mu} = (1, -0.1)\widehat{\boldsymbol{\beta}} = 0.082$, or 8.2%.

For *Bayesian* estimation of the logit model, we need to have an efficient way to obtain draws from the posterior distribution $f(\boldsymbol{\beta} \,|\, \mathbf{y})$ for a given prior $f(\boldsymbol{\beta})$. Since the likelihood function for the logit model is highly nonlinear, the posterior distribution is typically nonstandard, and estimation requires more work. One feasible approach to obtaining posterior draws is to use Markov chain Monte Carlo. In particular, we will use an independence sampler (see Example 7.11) with a multivariate Student's t proposal distribution. The reason for sampling from a Student's t proposal is that the samples tend to be less concentrated around the mode of the distribution than is the case for the normal distribution, for example. As a result the samples from the independence sampler tend to be less correlated.

Definition 9.3. (Multivariate Student's t Distribution). An n-dimensional random vector $\mathbf{X}$ is said to have a **multivariate Student's t** distribution with **mean vector** $\boldsymbol{\mu}$ and **scale matrix** Σ if its pdf is given by

$$f(\mathbf{x}; \nu, \boldsymbol{\mu}, \Sigma) = \frac{c}{\sqrt{\det(\Sigma)}}\left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}}, \quad (9.4)$$

where $c = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{n/2}\,\Gamma\left(\frac{\nu}{2}\right)}$ and $\nu > 0$ is the **degrees of freedom** parameter. We write the distribution as $\mathbf{t}_\nu(\boldsymbol{\mu}, \Sigma)$.

Similar to the multivariate $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ distribution, a vector $\mathbf{X} \sim \mathbf{t}_\nu(\boldsymbol{\mu}, \Sigma)$ can be viewed as an affine transformation $\mathbf{X} = \boldsymbol{\mu} + B\mathbf{Z}$ of a random vector $\mathbf{Z} \sim \mathbf{t}_\nu(\mathbf{0}, I)$ from the *standard* multivariate Student's t distribution, where $BB^\top = \Sigma$. To simulate draws from the latter distribution one can use the following theorem. The proof is left as an exercise; see Problem 9.1.

Theorem 9.1. (Generating from the Multivariate Student's t Distribution). Let $\mathbf{R} \sim \mathsf{N}(\mathbf{0}, I)$ and $W \sim \chi_\nu^2$ be independent. Then,

$$\mathbf{Z} = \sqrt{\frac{\nu}{W}}\,\mathbf{R} \sim \mathbf{t}_\nu(\mathbf{0}, I)\,.$$

To sample from the posterior pdf $f(\boldsymbol{\beta} \,|\, \mathbf{y})$ of the logit model, we draw the proposal from a $\mathsf{t}_\nu(\widehat{\boldsymbol{\beta}}, V)$ distribution, where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and V the inverse information matrix evaluated at $\widehat{\boldsymbol{\beta}}$.

Denote the pdf of the $\mathsf{t}_\nu(\widehat{\boldsymbol{\beta}}, V)$ distribution by $f_\mathsf{t}(\boldsymbol{\beta})$. In the independence sampler, given a current draw $\boldsymbol{\beta}$, the candidate $\boldsymbol{\beta}^*$ is accepted with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \min\left\{ \frac{f(\mathbf{y} \,|\, \boldsymbol{\beta}^*) f(\boldsymbol{\beta}^*) f_\mathsf{t}(\boldsymbol{\beta})}{f(\mathbf{y} \,|\, \boldsymbol{\beta}) f(\boldsymbol{\beta}) f_\mathsf{t}(\boldsymbol{\beta}^*)}, 1 \right\} \ ,$$

where $f(\mathbf{y} \,|\, \boldsymbol{\beta})$ is the likelihood function and $f(\boldsymbol{\beta})$ is the prior density.

Example 9.5 (Bayesian Inference for Logit Model). We continue Example 9.4. Taking a uniform prior for $\boldsymbol{\beta}$ (i.e., $f(\boldsymbol{\beta}) \propto 1$), the posterior pdf is proportional to the likelihood function. In other words, Fig. 9.2 is also a contour plot for the posterior distribution. For this example, the posterior pdf is proper even though the prior pdf is not. To compute other useful summary statistics, we use the independence sampler with $\mathsf{t}_\nu(\widehat{\boldsymbol{\beta}}, V)$ proposal, as described above. Note that both the proposal pdf f_t and the likelihood only have to be specified up to a multiplicative normalization constant. In fact, it is easier to specify the natural logarithms of both pdfs (up to an additive constant) and evaluate

$$\varrho(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \ln f(\mathbf{y} \,|\, \boldsymbol{\beta}^*) + \ln f_\mathsf{t}(\boldsymbol{\beta}) - \ln f(\mathbf{y} \,|\, \boldsymbol{\beta}) - \ln f_\mathsf{t}(\boldsymbol{\beta}^*)$$

and accept $\boldsymbol{\beta}^*$ with probability $\min\{\exp(\varrho(\boldsymbol{\beta}, \boldsymbol{\beta}^*)), 1\}$. To obtain a draw from the proposal distribution we first sample $\mathbf{Z} = (Z_1, Z_2)^\top \sim \mathsf{N}(\mathbf{0}, I)$ and $W \sim \chi_\nu^2 = \mathsf{Gamma}(\nu/2, 1/2)$ and return

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} + B \, \mathbf{Z} \, \sqrt{\nu/W} \ , \tag{9.5}$$

where $BB^\top = V$. Then $\boldsymbol{\beta}^*$ follows the desired t distribution, by Theorem 9.1.

The following MATLAB code—to be appended to the code of Example 9.4—implements the independence sampler and is used to obtain 10,000 draws from the posterior distribution after a burn-in period of 500. We use $\nu = 5$, giving samples that are spread out relatively far around the mode $\widehat{\boldsymbol{\beta}}$.

```
%bioassay_bayes.m
B = chol(V,'lower');
burnin = 500;   nloop = 10000+burnin;
store_beta = zeros(nloop,2);
nu = 5;        % df for the proposal
% log posterior density
logf = @(b) sum((y-1).*(X*b)-log((1+exp(-X*b))));
% log density of the t proposal
logprop = @(b) -.5*(nu+2)*log(1+(b-betat)'*(V\(b-betat))/nu);
beta = betat;             % initialize the chain
for i = 1:nloop
```

```
    % candidate draw from the t proposal
    betac = betat+B*randn(2,1)*sqrt(nu/gamrand(nu/2,1/2));
    rho = logf(betac)-logf(beta) + ...
        +logprop(beta)-logprop(betac);
    if exp(rho) > rand
        beta = betac;
    end
    store_beta(i,:) = beta';
end
store_beta = store_beta(burnin+1:end,:);   % discard the burnin
```

The posterior mean and posterior covariance matrix are estimated to be

$$
\mathbb{E}(\boldsymbol{\beta} \mid \mathbf{y}) = \begin{pmatrix} 1.36 \\ 11.98 \end{pmatrix}, \quad \mathrm{Var}(\boldsymbol{\beta} \mid \mathbf{y}) = \begin{pmatrix} 1.25 & 4.47 \\ 4.47 & 35.62 \end{pmatrix}.
$$

It is interesting to note that even though the posterior mode coincides with the maximum likelihood estimate under the flat prior, the posterior mean of β_1 is substantially larger than the corresponding maximum likelihood estimate, reflecting the fact that the marginal distribution of β_1 is positively skewed. Further, a 95% credible interval for β_1 is estimated to be $(3.52, 26.18)$, which excludes the value 0.

9.2.2   Probit Model

Definition 9.4.   (Probit Model). Let Y_i denote the ith binary response, and let $\mathbf{x}_i$ represent the vector of explanatory variables and $\boldsymbol{\beta}$ the associated parameter vector. In a **probit model**, the $\{Y_i\}$ are independent, and $Y_i \sim$ $\mathrm{Ber}(\mu_i)$, with $\mu_i = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$, where Φ is the cdf of the standard normal distribution. That is, the component link function is $g(x) = \Phi^{-1}(x)$.

As in the logit model we first derive the log-likelihood function, score function, and information matrix. Let $\varphi(x)$ denote the pdf of the standard normal distribution. Note that since the standard normal distribution is symmetric around 0, it follows that $\varphi(x) = \varphi(-x)$ and $1 - \Phi(x) = \Phi(-x)$. Now, given the independent Bernoulli

responses and the component link function $g(x) = \Phi^{-1}(x)$, the log-likelihood function for the probit model is

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \ln \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \ln \Phi(-\mathbf{x}_i^\top \boldsymbol{\beta}) \right] . \tag{9.6}$$

The score function is the gradient of the log-likelihood function:

$$\mathbf{S}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\, l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \frac{\varphi(\mathbf{x}_i^\top \boldsymbol{\beta})}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta})} \mathbf{x}_i - (1 - y_i) \frac{\varphi(-\mathbf{x}_i^\top \boldsymbol{\beta})}{\Phi(-\mathbf{x}_i^\top \boldsymbol{\beta})} \mathbf{x}_i \right]$$

$$= \sum_{i=1}^{n} \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}) \left[\frac{y_i}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta})} - \frac{1 - y_i}{\Phi(-\mathbf{x}_i^\top \boldsymbol{\beta})} \right] \mathbf{x}_i .$$

Noting that $\frac{\mathrm{d}}{\mathrm{d}x}\varphi(x) = -x\,\varphi(x)$, we differentiate the score function again with respect to $\boldsymbol{\beta}$ to obtain

$$\nabla_{\boldsymbol{\beta}}^2\, l(\boldsymbol{\beta}; \mathbf{y}) = - \sum_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\beta})\, \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}) \left[\frac{y_i}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta})} - \frac{1 - y_i}{\Phi(-\mathbf{x}_i^\top \boldsymbol{\beta})} \right] \mathbf{x}_i \mathbf{x}_i^\top$$

$$- \sum_{i=1}^{n} \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}) \left[\frac{y_i\, \varphi(\mathbf{x}_i^\top \boldsymbol{\beta})}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta})^2} + \frac{(1 - y_i)\, \varphi(-\mathbf{x}_i^\top \boldsymbol{\beta})}{\Phi(-\mathbf{x}_i^\top \boldsymbol{\beta})^2} \right] \mathbf{x}_i \mathbf{x}_i^\top .$$

Using the fact that $\mathbb{E}(Y_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$, the information matrix is therefore

$$I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\varphi(\mathbf{x}_i^\top \boldsymbol{\beta})^2}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta})\, \Phi(-\mathbf{x}_i^\top \boldsymbol{\beta})} \mathbf{x}_i \mathbf{x}_i^\top .$$

Given the score function and the information matrix, one can then obtain the maximum likelihood estimate via Fisher's scoring method as before.

For a Bayesian analysis, we can sample from the posterior pdf using MCMC, for example, using a similar independence sampler as in the logit model. If we use a normal prior $\boldsymbol{\beta} \sim \mathsf{N}(\mathbf{b}_0, V_0)$, then the logarithm of the posterior pdf $f(\boldsymbol{\beta} \mid \mathbf{y}) \propto f(\boldsymbol{\beta}) f(\mathbf{y} \mid \boldsymbol{\beta})$ is

$$\ln f(\boldsymbol{\beta} \mid \mathbf{y}) = l(\boldsymbol{\beta}; \mathbf{y}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b}_0)^\top V_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0) + \text{const}, \tag{9.7}$$

with $l(\boldsymbol{\beta}; \mathbf{y})$ given in (9.6). From the (dependent) sample of the posterior pdf it is straightforward to estimate the posterior mean, standard deviation, and quantiles. One can also estimate the marginal posterior pdfs $\{f(\beta_i \mid \mathbf{y})\}$, using a kernel density

estimator.

Other quantities of interest include the marginal effects of the covariates, that is, how a change in the covariate affects the response. To make the discussion concrete, let x_j be the jth element of a covariate vector $\mathbf{x}$. If x_j is a continuous explanatory variable, then

$$\frac{\partial}{\partial x_j}\mathbb{E}(Y \mid \boldsymbol{\beta}) = \frac{\partial}{\partial x_j}\Phi(\mathbf{x}^\top\boldsymbol{\beta}) = \varphi(\mathbf{x}^\top\boldsymbol{\beta})\beta_j \; , \tag{9.8}$$

where β_j is the jth element of $\boldsymbol{\beta}$. This depends on both $\boldsymbol{\beta}$ and $\mathbf{x}$. For the "average" marginal effect of x_j one could consider $\varphi(\bar{\mathbf{x}}^\top\boldsymbol{\beta})\beta_j$, where $\bar{\mathbf{x}}$ is the average of the explanatory vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ corresponding to the responses $Y_1, \ldots, Y_n$. Similarly, if x_j is a binary explanatory variable, the average marginal effect of x_j is

$$\Phi(\mathbf{z}_1^\top\boldsymbol{\beta}) - \Phi(\mathbf{z}_0^\top\boldsymbol{\beta}) \; ,$$

where $\mathbf{z}_x = (\bar{x}_1, \ldots, \bar{x}_{j-1}, x, \bar{x}_{j+1}, \ldots, \bar{x}_n)^\top$ for $x \in \{0, 1\}$.

Note that the marginal effect is a (continuous) function of the regression parameter vector $\boldsymbol{\beta}$, and so it is a random variable. Given the posterior draws for $\boldsymbol{\beta}$, the posterior distribution of the marginal effect can be obtained readily.

Example 9.6 (Modeling Extramarital Affairs with Probit Model). Fair (1978) analyzed the decision to have an extramarital affair with a probit model, using surveys conducted by *Psychology Today* and *Redbook*. The data used in this example are obtained from Koop et al. (2007) and contain 601 independent observations. All observations are taken from individuals currently married and married for the first time.

The response is a binary variable that indicates if the respondent has (had) an extramarital affair; the seven explanatory variables are an intercept (CONST), a male indicator (MALE), number of years married (YEAR), a binary variable to indicate if the respondent has children from the marriage (KIDS), a binary variable for classifying one's self as "religious" (RELIGIOUS), years of schooling completed (ED), and a final binary variable denoting whether the person views the marriage as happier than an average marriage (HAPPY).

We first obtain the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ via Fisher's scoring method, as well as the information matrix V evaluated at $\widehat{\boldsymbol{\beta}}$. The following MATLAB code accomplishes this task. If the *Statistics Toolbox* is not available, one can replace normcdf by cumdf introduced in Appendix A.9.

```
%probit_mle.m
load('affair.csv');          % load data
y = affair(:,1);
X = affair(:,2:end);
[n k] = size(X);
    %% find the MLE and the information matrix
S = ones(k,1);               % score
betat = (X'*X)\(X'*y);       % initial guess
```

```
e = 10^(-5);                        % tolerance level
while sum(abs(S)) > e               % stopping criterion
    Xbetat = X*betat;
    phi = normpdf(Xbetat);
    Phi = normcdf(Xbetat);
    S = sum(repmat(y.*phi./Phi-(1-y).*phi./(1-Phi),1,k).*X)';
    d = phi.^2./(Phi.*(1-Phi));
    I = X'*diag(d)*X; % information matrix
    betat = betat + I\S;
end
```

To sample from the posterior distribution we use the same MCMC approach as for the logit model. That is, we use an independence sampler with a $t_\nu(\widehat{\boldsymbol{\beta}}, V)$ proposal distribution. The hyperparameters for the normal prior are chosen as $\mathbf{b}_0 = \mathbf{0}$ and $V_0 = 10I_7$, where I_7 is the identity matrix. This gives a relatively non-informative prior which is centered around zero. The logarithm of the posterior pdf is given in (9.7). The degrees of freedom is set to $\nu = 5$. We use the method described in (9.5) to generate a draw from the proposal distribution, and run the sampler for 5,500 iterations, discarding the first 500 as burn-in.

```
%probit_bayes.m
burnin = 500;
nloop = 5000+burnin;
V = I\eye(k);     % scale matrix for the proposal
B = chol(V,'lower');
nu = 5;           % df for the proposal
b0 = zeros(k,1);  % prior mean
V0 = 10*eye(k);   % prior covariance
% log-posterior density
logf=@(b) y'*log(normcdf(X*b)) + (1-y)'*log(normcdf(-X*b)) ...
    - .5*(b-b0)'*(V0\(b-b0));
% log-proposal density
logprop=@(b) -.5*(k+nu)*log(1+(b-betat)'*(V\(b-betat))/nu);
store_beta = zeros(nloop,k);
beta = betat;
for i = 1:nloop
    % candidate draw from the t proposal
    betac=betat + B*randn(k,1)*sqrt(nu/gamrand(nu/2,1/2));
    rho = logf(betac)-logf(beta) + ...
        + logprop(beta)-logprop(betac);
    if exp(rho) > rand
        beta = betac;
    end
    store_beta(i,:) = beta';
end
store_beta = store_beta(burnin+1:end,:); % discard the burn-in
```

Table 9.2 lists various summary statistics of the posterior distribution, including the means, standard deviations, and 2.5 and 97.5 percentiles, based on the 5000 (dependent) samples from the posterior distribution. Of the six variables (excluding the intercept), only three seem to have a substantial impact on the response. In particular, the 95% credible intervals for the coefficients associated with YEAR, RELIGIOUS, and HAPPY exclude zero, while the other three do not. On average, people reporting themselves as religious or in happy marriages are less likely to have affairs, while the longer someone is in a marriage, the more likely he or she has an affair.

Table 9.2 Coefficient posterior means, standard deviations, and 2.5 and 97.5 percentiles for the probit model

Variable	Mean	Std. dev.	2.5 percentile	97.5 percentile
CONST	−0.728	0.408	−1.528	0.061
MALE	0.150	0.126	−0.098	0.392
YEAR	0.029	0.013	0.004	0.054
KIDS	0.249	0.161	−0.065	0.561
RELIGIOUS	−0.516	0.122	−0.757	−0.277
ED	0.005	0.025	−0.045	0.055
HAPPY	−0.517	0.124	−0.760	−0.269

To assess the quantitative impacts of the covariates we estimate the average marginal effects of the covariates, using the following code:

```
%margeff.m
N = size(store_beta,1);
store_ME = zeros(nloop-burnin,6);
xbar = mean(X)';
for loop=1:N
    beta = store_beta(loop,:)';
                            % ME for continuous variables
    store_ME(loop,[2 5]) = normpdf(xbar'*beta)*beta([3 6]);
    for j = [1 3 4 6];          % ME for discrete variables
        z0 = xbar;  z0(j+1) = 0;
        z1 = xbar;  z1(j+1) = 1;
        store_ME(loop,j) =normcdf(z1'*beta)-normcdf(z0'*beta);
    end
end
```

The summary statistics of the posterior distribution for the marginal effects are reported in Table 9.3. For example, people who report themselves as religious are

15% less likely to have affairs (fixing the other covariates at the sample means), and those who report to be happy in their marriages are 17% less likely.

Table 9.3 Posterior means, standard deviations, and 2.5 and 97.5 percentiles for the marginal effects

Variable	Mean	Std. dev.	2.5 percentile	97.5 percentile
MALE	0.045	0.038	-0.030	0.121
YEAR	0.009	0.004	0.001	0.016
KIDS	0.071	0.046	-0.024	0.157
RELIGIOUS	-0.150	0.035	-0.217	-0.082
ED	0.002	0.008	-0.014	0.017
HAPPY	-0.167	0.043	-0.255	-0.084

9.2.3 Latent Variable Representation

Estimation and inference under the logit and probit models can be simplified by using *data augmentation*. The general idea behind data augmentation is to include "hidden" variables in the model for the data in order to simplify the analysis of the model. A prime example of data augmentation is found in the EM algorithm in Sect. 6.6.

For the logit and probit models data augmentation can be introduced by thinking of an observed binary response in terms of whether or not an underlying continuous latent (i.e., hidden) variable crosses a particular threshold: if it does, then we observe, say, 1; otherwise, we observe 0. The advantage of the latent variable representation is that it is often easier to work with the continuous latent variables than the observed binary variable. To be mathematically precise, consider again the probit model: each binary response Y_i is distributed as $Y_i \sim \mathsf{Ber}(\mu_i)$, where $\mu_i = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$, and $\mathbf{x}_i$ is a vector of covariates.

Now, introduce the independent latent variables $\{Z_i\}$, each of which is distributed according to the normal distribution with mean $\mathbf{x}_i^\top \boldsymbol{\beta}$ and variance 1:

$$Z_i \sim \mathsf{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, 1) . \tag{9.9}$$

These latent variables are then linked to the observed binary variables $\{Y_i\}$ as follows:

$$Y_i = \begin{cases} 1 \text{ if } Z_i > 0 , \\ 0 \text{ if } Z_i \leqslant 0 . \end{cases} \tag{9.10}$$

The values of the binary variables $\{Y_i\}$ are observed and the covariates $\{\mathbf{x}_i\}$ are fixed. However, the latent variables $\{Z_i\}$ are unobserved.

To check that this latent variable representation (9.9) and (9.10) does indeed give the same probit model, we need to show that it implies the same likelihood function. To this end, note that under the latent variable representation, each Y_i is an independent Bernoulli random variable with success probability:

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i > 0) = 1 - \Phi(-\mathbf{x}_i^\top \boldsymbol{\beta}) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}) \,,$$

which is the same success probability under the probit model. Hence, the latent variable representation in (9.9) and (9.10) implies the same probit model.

Introducing more unobserved variables might seem to be odd as they give rise to exactly the same model. However, as it turns out, by augmenting the data with these latent variables, computation becomes more tractable. In fact, we can use the expectation–maximization algorithm discussed in Sect. 6.6 to obtain the maximum likelihood estimate easily.　　　　☞ 182

We first determine the complete-data log-likelihood—using classical rather than Bayesian notation. Since conditional on $\mathbf{Z}$ the vector $\mathbf{Y}$ is deterministic, the joint pdf of $\mathbf{Y}$ and $\mathbf{Z}$ has the same form as the pdf of $\mathbf{Z}$. It follows that

$$l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) = \ln f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\beta}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (z_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \left\{ (\mathbf{x}_i^\top \boldsymbol{\beta})^2 - 2\, z_i\, \mathbf{x}_i^\top \boldsymbol{\beta} \right\} + \text{const} \,. \qquad (9.11)$$

Now, suppose $\boldsymbol{\beta}_{t-1}$ is the current value for $\boldsymbol{\beta}$. To implement the E-step, we derive the conditional density

$$g_t(\mathbf{z}) = f_{\mathbf{Z} \mid \mathbf{Y}}(\mathbf{z} \mid \mathbf{y}; \boldsymbol{\beta}_{t-1}) = \prod_{i=1}^{n} f_{Z_i \mid Y_i}(z_i \mid y_i; \boldsymbol{\beta}_{t-1}) \,,$$

where we use the fact that the latent variables $Z_1, \ldots, Z_n$ are conditionally independent. If $y_i = 1$, the only extra information we have is that $Z_i > 0$. What this means is that given $y_i = 1$, Z_i follows the normal distribution with mean $\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1}$ and variance 1, left-truncated at 0. So, $f_{Z_i \mid Y_i}(z_i \mid y_i = 1; \boldsymbol{\beta}_{t-1}) = 0$ for $z < 0$ and proportional to $\exp(-\frac{1}{2}(z_i - \mathbf{x}_i^\top \boldsymbol{\beta}_{t-1})^2)$ for $z_i \geqslant 0$. We write

$$(Z_i \mid y_i = 1; \boldsymbol{\beta}_{t-1}) \sim \mathsf{TN}_{(0,\infty)}(\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1}, 1) \,. \qquad (9.12)$$

Similarly, if $y_i = 0$, then

$$(Z_i \mid y_i = 0; \boldsymbol{\beta}_{t-1}) \sim \mathsf{TN}_{(-\infty,0)}(\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1}, 1) \,. \qquad (9.13)$$

In particular (see Problem 9.7), we have

$$\mathbb{E}[Z_i \mid y_i = 1; \boldsymbol{\beta}_{t-1}] = \mathbf{x}_i^\top \boldsymbol{\beta}_{t-1} + \frac{\varphi(\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1})}{\Phi(\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1})} \,, \tag{9.14}$$

$$\mathbb{E}[Z_i \mid y_i = 0; \boldsymbol{\beta}_{t-1}] = \mathbf{x}_i^\top \boldsymbol{\beta}_{t-1} - \frac{\varphi(\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1})}{\Phi(-\mathbf{x}_i^\top \boldsymbol{\beta}_{t-1})} \,. \tag{9.15}$$

Writing $v_i = \mathbb{E}_{g_t}[Z_i \mid y_i; \boldsymbol{\beta}_{t-1}]$, it follows from (9.11) that

$$Q_t(\boldsymbol{\beta}) = \mathbb{E}_{g_t} l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{Z}) = -\frac{1}{2} \sum_{i=1}^n \left\{ (\mathbf{x}_i^\top \boldsymbol{\beta})^2 - 2 v_i \, \mathbf{x}_i^\top \boldsymbol{\beta} \right\} + \text{const} \,.$$

Next, to implement the M-step, we simply solve $\nabla Q_t(\boldsymbol{\beta}) = \mathbf{0}$. Since Q_t is
☞ 367 quadratic in $\boldsymbol{\beta}$ we can use the differentiation rules in Sect. B.1 to find (see also
Problem 9.8) the solution:

$$\boldsymbol{\beta}_t = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^n v_i \, \mathbf{x}_i \,. \tag{9.16}$$

Finally, the maximum likelihood estimate for $\boldsymbol{\beta}$ can be obtained by going through
the E- and M-steps iteratively until convergence.

For *Bayesian* estimation, the probit model can be fitted using the Gibbs sampler
with data augmentation. Specifically, if we have draws from the joint posterior pdf
$f(\mathbf{z}, \boldsymbol{\beta} \mid \mathbf{y})$ and retain only the draws for $\boldsymbol{\beta}$, then those draws are from the desired
marginal pdf $f(\boldsymbol{\beta} \mid \mathbf{y})$. Therefore, we can construct a Gibbs sampler by sequentially
drawing from $f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z})$ followed by $f(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\beta})$. As it turns out, both conditional
densities are of standard form and samples from each can be obtained quickly.

For concreteness, assume the prior $\boldsymbol{\beta} \sim \mathsf{N}(\mathbf{0}, \Sigma_0)$. First, to derive $f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z})$ note
that, given the latent vector $\mathbf{z}$, we in fact have a linear regression model; see (9.9).
☞ 238 Hence, using Theorem 8.1 we have

$$(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z}) \sim \mathsf{N}(\widehat{\boldsymbol{\beta}}, D) \,,$$

where

$$D = (X^\top X + \Sigma_0^{-1})^{-1} \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}} = D X^\top \mathbf{z} \,,$$

and X is the design matrix with ith row $\mathbf{x}_i^\top$, $i = 1, \ldots, n$.

Second, the conditional density $f(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^n f(z_i \mid y_i, \boldsymbol{\beta})$ is given in
(9.12) and (9.13). A draw from a truncated normal distribution can be obtained,
☞ 53 say, via the inverse-transform method or (faster) the acceptance–rejection method.

Example 9.7 (Gibbs Sampler for Probit Model). To demonstrate fitting the probit
model using the Gibbs sampler with data augmentation, we revisit Example 9.6. We
use the following function to draw from truncated normal distributions; see also

Problem 9.7. If the *Statistics Toolbox* is not available, one can replace `normcdf` and `norminv` by `cumdf` and `icumdf` introduced in Appendix A.9, respectively.

```
function [draw] = tnormrnd(mu,sigma2,a,b)
N = length(mu);
sigma = sqrt(sigma2);
u = rand(N,1);
p1 = normcdf((a-mu)./sigma);
p2 = normcdf((b-mu)./sigma);
C = norminv(p1+(p2-p1).*u);
draw = mu + sigma.*C;
```

In the main script, we implement a Gibbs sampler by alternatively drawing from $f(\boldsymbol{\beta} \,|\, \mathbf{y}, \mathbf{z})$ and $f(\mathbf{z} \,|\, \mathbf{y}, \boldsymbol{\beta})$. The estimation results are similar to those obtained in Example 9.6, and they are not repeated here.

```
%probit_bayes_gibbs.m
load('affair.csv');
y = affair(:,1);
X = affair(:,2:end);
XX = X'*X;
[n k] = size(X);
V0 = 10*eye(k);        % prior covariance
invV0 = V0\eye(k);
burnin = 500;
nloop = 5000+burnin;
store_beta = zeros(nloop,k);
z = y;                 % initial guess
beta = XX\(X'*z);
                           % compute a few things before the loop
id0 = find(y==0); id1 = find(y==1);
n0 = length(id0); n1=n-n0;
V = (invV0 + XX)\eye(k);       % posterior covariance
for i = 1:nloop
        % sample y^*
    Xb = X*beta;
    z(id0) = tnormrnd(Xb(id0),ones(n0,1),-inf,0);
    z(id1) = tnormrnd(Xb(id1),ones(n1,1),0,inf);
        % sample beta
    dbeta = X'*z;
    beta = V*dbeta + chol(V,'lower')*randn(k,1);
    store_beta(i,:) = beta';
end
store_beta = store_beta(burnin+1:end,:); % discard the burn-in
```

9.3   Poisson Regression

Poisson regression deals with *count* data Y; for example, the number of cars in a household. We are interested in how some observed characteristics $\mathbf{x}$—e.g., household income, number of children in the household, and whether it is a single-parent household—affect the response Y. Since Y takes values on the set of nonnegative integers, one natural specification for Y is the Poisson model $Y \sim$ Poi(μ). In terms of a generalized linear model (see Definition 9.1) it remains to link $\mathbf{x}^\top \boldsymbol{\beta}$ to the mean μ, which has to be positive. One easy way to guarantee this is to specify μ as

$$\mu = e^{\mathbf{x}^\top \boldsymbol{\beta}} \, .$$

This leads to the following definition.

Definition 9.5.   (Poisson Regression Model). Let Y_i denote the ith response (count) and let $\mathbf{x}_i$ represent the vector of explanatory variables and $\boldsymbol{\beta}$ the associated parameter vector. In a **Poisson regression model**, the $\{Y_i\}$ are independent, and $Y_i \sim \mathsf{Poi}(\mu_i)$, with $\mu_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$. In other words, the component link function is $g(x) = \ln x$.

Let $\mathbf{x}_i^\top$ be the ith row of the design matrix X, and let $\mathbf{g} = (g, \ldots, g)^\top$. We see that the distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is completely specified by $\boldsymbol{\mu} = \mathbf{g}^{-1}(X\boldsymbol{\beta})$. In this case no additional dispersion parameters are used.

Example 9.8 (MLE for the Poisson Regression Model).   Suppose we are interested in determining the impact of research and development (R&D) on the number of patents obtained by firms in a certain industry. For this purpose a total of $n = 14$ firms are interviewed. For each firm we record its number of patents obtained over the last 3 years, as well as its R&D budget (in tens of thousands of dollars) over the same period. The data are presented in Table 9.4.

Table 9.4 Number of patents and R&D

Number of patents	R&D budget	Number of patents	R&D budget
6	26	8	29
3	21	2	13
2	19	0	5
1	11	2	3
3	21	6	29
1	16	1	3
1	19	3	21

To investigate the effectiveness of R&D, let Y_i denote the number of patents obtained by the ith firm and let $\mathbf{x}_i = (1, x_i)^\top$ be a 2×1 vector of explanatory variables, where x_i is the ith firm's R&D budget. We consider the Poisson regression $Y_i \sim \text{Poi}(\mu_i)$, where $\mu_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ is a 2×1 vector of regression coefficients.

The log-likelihood function for the Poisson regression model is given by

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^n \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - e^{\mathbf{x}_i^\top \boldsymbol{\beta}} - \ln y_i! \right] \ .$$

Moreover, it can be shown that the score function and the information matrix are, respectively (see Problem 9.2),

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - e^{\mathbf{x}_i^\top \boldsymbol{\beta}}) \mathbf{x}_i \quad \text{and} \quad I(\boldsymbol{\beta}) = \sum_{i=1}^n e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}_i^\top \ .$$

Hence, the maximum likelihood estimate of $\boldsymbol{\beta}$ can be computed using Fisher's scoring method, which is implemented in the following MATLAB script.   ☞ 180

```
% poissonreg.m
y = [6 3 2 1 3 1 1 8 2 0 2 6 1 3]';
RnD = [26 21 19 11 21 16 19 29 13 5 3 29 3 21]';
n = length(y);
X = [ones(n,1) RnD];
betat = (X'*X)\(X'*log(y+.001));    % initial guess
S = ones(2,1);                      % score
e = 10^(-5);                        % tolerance level
while sum(abs(S)) > 10^(-5)         % stopping criteria
    mu = exp(X*betat);
    S = sum(repmat((y - mu),1,2).*X)';
    I =   X'*sparse(1:n,1:n,mu)*X; % information matrix
    betat = betat + I\S;
end
V = I\eye(2);
```

The maximum likelihood estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (-0.795, 0.092)^\top$. In addition, the $(2, 2)$-element of the inverse information matrix evaluated at $\widehat{\boldsymbol{\beta}}$ is 0.0239^2. It follows that an approximate 95% confidence interval for β_2 is $(0.045, 0.139)$, which excludes 0, providing persuasive evidence that R&D has a positive effect on the number of patents.

9.4   Problems

9.1. Prove Theorem 9.1; that is, show that if $\mathbf{R} \sim \mathsf{N}(\mathbf{0}, I)$ and $W \sim$ Gamma$(\nu/2, 1/2)$ are independent, then the random vector $\mathbf{Z} = \sqrt{\nu/W}\,\mathbf{R}$ has pdf

$$f(\mathbf{z}) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\nu/2)(\pi\nu)^{n/2}} \left(1 + \frac{\|\mathbf{z}\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Hint: consider the coordinate transformation

$$\begin{pmatrix} z_1 \\ \vdots \\ w \end{pmatrix} \mapsto \begin{pmatrix} z_1\sqrt{w/\nu} \\ \vdots \\ z_n\sqrt{w/\nu} \\ w \end{pmatrix} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \\ w \end{pmatrix}$$

and determine the determinant of the corresponding Jacobian matrix. Next, apply
☞ 81   the transformation rule (3.25) to find the joint pdf of $(\mathbf{Z}, W)$. Finally integrate out W to obtain the pdf of $\mathbf{Z}$.

9.2. Consider the Poisson regression model in Definition 9.5. Given the data $(y_1, \mathbf{x}_n^\top), \dots (y_n, \mathbf{x}_n^\top)$, show that the log-likelihood function is given by

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^n \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - e^{\mathbf{x}_i^\top \boldsymbol{\beta}} - \ln y_i! \right].$$

Further, show that the score function and the information matrix are, respectively,

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - e^{\mathbf{x}_i^\top \boldsymbol{\beta}})\mathbf{x}_i \quad \text{and} \quad I(\boldsymbol{\beta}) = \sum_{i=1}^n e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \mathbf{x}_i \mathbf{x}_i^\top.$$

9.3. It is generally believed that births by Caesarean section are more frequent in private hospitals than in public ones. To investigate if there is any evidence for this claim, data are collected on the number of Caesarean sections carried out in three private hospitals (type 0) and seven public hospitals (type 1), as well as the total number of births in each of the hospitals. These are presented in Table 9.5.

Use the data to fit a Poisson regression: regress the response variable "number of Caesarean sections" on an intercept, "number of births," and "hospital type". Are births by Caesarean section more frequent in private hospitals? Hint: use the results in Problem 9.2.

Table 9.5 Poisson regression example

Number of Caesarean sections	Number of births	Hospital type	Number of Caesarean sections	Number of births	Hospital type
8	236	0	13	679	1
16	739	1	4	26	0
15	970	1	19	1272	1
23	2371	1	33	3246	1
5	309	1	2	28	0

9.4. Consider again Example 9.6 where we use the probit model to analyze the decision to have an extramarital affair. For a nonreligious, college-educated (16 years of education) male who has married for 10 years with one child from the marriage, and who reports that his marriage is happier than average, what is the probability that he has an extramarital affair? Use the kde.m program to plot a kernel density estimate of the posterior probability.

9.5. In the linear regression model $Y = \mathbf{x}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the parameter vector $\boldsymbol{\beta}$ can be interpreted as the marginal effects of the (continuous) covariates, that is, the rate at which the response changes as the result of an infinitesimal change in the covariate:

$$\boldsymbol{\beta} = \nabla_{\mathbf{x}} \mathbb{E} Y .$$

However, for the probit model the marginal effects depend on *both* the parameter vector $\boldsymbol{\beta}$ and the covariates $\mathbf{x}_i$ in a nonlinear functional form: $\varphi(\mathbf{x}^\top \boldsymbol{\beta})\boldsymbol{\beta}$; see (9.8). What are the marginal effects for the *logit* model?

9.6. In Definition 9.3 the matrix Σ was intentionally called the *scale* matrix rather than *covariance* matrix, because the covariance matrix of $\mathbf{X}$ (i.e., $\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top]$) is not equal to Σ.

(a) Show that the covariance matrix of $\mathbf{X}$ is $\Sigma \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, where $\mathbf{Z}$ has a standard multivariate Student's t distribution.

(b) Use Theorem 9.1 to show that the covariance matrix of $\mathbf{Z}$, that is, $\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, is equal to $c \, \nu \, I$, where I is the identity matrix and

$$c = \int_0^\infty \frac{1}{w} \frac{\left(\frac{1}{2}\right)^{\frac{\nu}{2}} w^{\frac{\nu}{2}-1} e^{-\frac{1}{2}w}}{\Gamma\left(\frac{\nu}{2}\right)} \, dw .$$

(c) Evaluate c.

9.7. Let $Z \sim \mathsf{TN}(\mu, \sigma^2, a, b)$, where $a < b$. Thus, the distribution of Z is that of a random variable $X \sim \mathsf{N}(\mu, \sigma^2)$ conditioned on X lying in $[a, b]$.

(a) Show that the pdf of Z is

$$f_Z(z) = \frac{\varphi((z - \mu)/\sigma)/\sigma}{\Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma)} \, ,$$

where φ and Φ are, respectively, the pdf and cdf of the standard normal distribution.

(b) For $\sigma = 1$, $a = -\infty$, and $b = 0$ show that $\mathbb{E}Z = \mu - \varphi(\mu)/\Phi(-\mu)$.

(c) For $\sigma = 1$, $a = 0$, and $b = \infty$ show that $\mathbb{E}Z = \mu + \varphi(\mu)/\Phi(\mu)$.

(d) Show that the cdf of Z is

$$F_Z(z) = \frac{\Phi((z - \mu)/\sigma) - \Phi((a - \mu)/\sigma)}{\Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma)} \, .$$

(e) Explain why the function `tnormrnd` in Example 9.7 can be used to simulate Z.

9.8. Prove (9.16).

Chapter 10
Dependent Data Models

In the models considered so far the responses $Y_1, \ldots, Y_n$ have been assumed to be independent given the model parameters. Though convenient, this independence assumption is implausible in two common situations. First, in the case of **time series**—observations measured over time—the responses typically exhibit strong serial dependence. For example, high unemployment tends to last for a long period of time; given a high unemployment rate this period, one would expect that the unemployment rates in the next few periods would also be high.

The other situation in which observations are likely to be dependent is when they are measurements on the same or related subjects. For example, learning outcomes of children in the same school tend to be more similar—because they share the same academic environment and come from families of similar socioeconomic backgrounds—than those of other children in a different school.

In this chapter we introduce models that relax the usual independence assumption and are suitable for modeling data that arise in the two aforementioned situations.

10.1 Autoregressive and Moving Average Models

In this section we introduce a widely popular class of simple time series models called **autoregressive-moving average** (ARMA) models. We begin our study of ARMA models with purely autoregressive specifications.

10.1.1 Autoregressive Models

Example 10.1 (Sales and Ads). In this motivating example, we consider a linear regression model in a time series context. Specifically, suppose a company has collected data on its quarterly sales and advertisement expenditures for the last 30 quarters. The data are given in Table 10.1.

Table 10.1 Sales and advertisement expenditures

Time	Sales	Ads	Time	Sales	Ads	Time	Sales	Ads
1	23.45	5.41	11	24.92	4.59	21	21.42	2.61
2	26.03	6.30	12	25.58	5.04	22	30.99	6.59
3	21.73	4.21	13	28.74	5.87	23	20.79	2.65
4	23.66	4.50	14	24.69	3.73	24	22.99	4.43
5	26.54	5.63	15	29.39	6.02	25	23.56	5.18
6	26.62	5.67	16	28.47	5.60	26	23.36	5.27
7	23.59	4.33	17	28.66	5.50	27	19.03	3.52
8	20.05	3.54	18	25.49	4.39	28	24.53	5.76
9	20.67	3.86	19	28.57	5.68	29	19.56	4.30
10	25.97	5.66	20	28.23	6.05	30	20.90	4.45

The marketing manager wonders how the two figures are correlated. To address this question, she considers the following linear regression model:

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t , \quad t = 1, \ldots, 30, \quad \{\varepsilon_t\} \overset{iid}{\sim} N(0, \sigma^2) ,$$

where Y_t is the sales at quarter t and x_t is the corresponding advertisement expenditure. Given the outcomes $y_t, t = 1, \ldots, 30$, the maximum likelihood estimates of the model parameters, $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\sigma}^2$, can be readily computed; see Example 5.5.

As a model diagnostic check, we can compute and plot the *residuals*

$$u_t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t, \quad t = 1, \ldots, 30 .$$

The residuals $\{u_t\}$ are our best guess for the (unknown) error terms $\{\varepsilon_t\}$. If the model is correct the residuals should be approximately iid and normally distributed, because the true error terms behave in this way. A plot of the residuals is given in Fig. 10.1. As the graph shows, the residuals exhibit systemic patterns across time. In particular, they tend to cluster together, indicating that the assumption of serially independent errors might not hold.
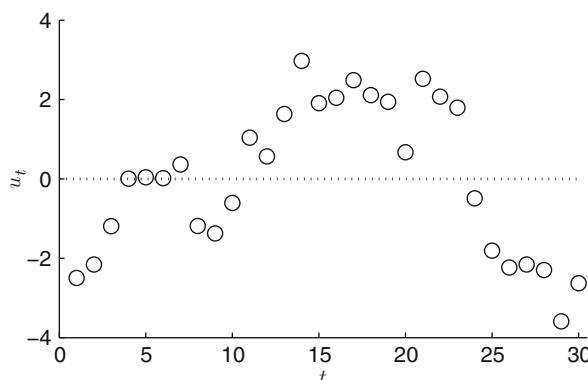
Fig. 10.1 A plot of the residuals of the linear regression model

Instead of the model $\{\varepsilon_t\} \sim_{\text{iid}} N(0, \sigma^2)$ for the errors, one could consider the model where the current error depends on the past errors in a linear way, for example, as in

$$\varepsilon_t = \varrho\, \varepsilon_{t-1} + U_t \, , \quad \{U_t\} \sim_{\text{iid}} N(0, \sigma^2) \, , \tag{10.1}$$

where and ϱ and σ^2 are fixed model parameters.

The model for the errors in (10.1) is an example of an autoregressive model.

Definition 10.1. (Autoregressive Model). In the **pth-order autoregressive (AR(p))** model, the observation at time t depends linearly on the previous p observations $Y_{t-1}, \ldots, Y_{t-p}$:

$$Y_t = \varrho_0 + \varrho_1 Y_{t-1} + \cdots + \varrho_p Y_{t-p} + \varepsilon_t \, , \tag{10.2}$$

$t = 1, \ldots, T$, where $\{\varepsilon_t\} \sim_{\text{iid}} N(0, \sigma^2)$.

To complete the model one needs to specify the **initial conditions**, that is, the probability distribution of the first p observations: $Y_{-p+1}, \ldots, Y_0$. For simplicity it is often assumed that these values are known. An alternative approach is to assume that (10.2) holds for every $t \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and that the time series is **stationary**, meaning that the distribution of $Y_1, Y_2, \ldots$ is the same as that of $Y_{n+1}, Y_{n+2}, \ldots$ for any $n \in \mathbb{Z}$. In particular, the distribution of Y_t does not depend on t (is the same for all t), and the joint distribution of (Y_t, Y_{t+s}) does not depend on t. Stationary AR processes only exist under certain conditions on the $\{\varrho_i\}$.

Example 10.2 (Autocorrelations of AR(1)). Consider an AR(1) time series on $\mathbb{Z}$, governed by

$$Y_t = \varrho Y_{t-1} + \varepsilon_t \, , \quad t \in \mathbb{Z} \, ,$$

where $\{\varepsilon_t\} \sim_{\text{iid}} N(0, \sigma^2)$. By repeated substitution, we have

$$Y_t = \varepsilon_t + \varrho\varepsilon_{t-1} + \varrho^2\varepsilon_{t-2} + \varrho^3\varepsilon_{t-3} + \cdots \, . \tag{10.3}$$

Since the $\{\varepsilon_t\}$ are independent by assumption, the variance of Y_t is finite if $|\varrho| < 1$:

$$\begin{aligned}
\text{Var}(Y_t) &= \text{Var}(\varepsilon_t + \varrho\varepsilon_{t-1} + \varrho^2\varepsilon_{t-2} + \varrho^3\varepsilon_{t-3} + \cdots) \\
&= \sigma^2 + \varrho^2\sigma^2 + \varrho^4\sigma^2 + \varrho^6\sigma^2 + \cdots \\
&= \frac{\sigma^2}{1 - \varrho^2} \, .
\end{aligned}$$

It is worth noting that the variance of Y_t is a constant and does not depend on the time index t. In fact, using the representation (10.3) it is not difficult to see that the time series $\{Y_t\}$ is *stationary* and that each Y_t has a $N(0, \sigma^2/(1 - \varrho^2))$ distribution.

Next, we compute the covariance $\mathrm{Cov}(Y_t, Y_{t-1})$, or the so-called (first-order) **autocovariance**:

$$\mathrm{Cov}(Y_t, Y_{t-1}) = \mathrm{Cov}(\varrho Y_{t-1} + \varepsilon_t, Y_{t-1}) = \mathrm{Cov}(\varrho Y_{t-1}, Y_{t-1})$$

$$= \varrho \mathrm{Var}(Y_{t-1}) = \varrho \frac{\sigma^2}{1 - \varrho^2} ,$$

where the second equality holds because Y_{t-1} is a function of $\varepsilon_s, s \leqslant t - 1$, and is therefore uncorrelated with ε_t. Using a similar argument, one can show that in general the autocovariance of **lag** s is given by

$$R(s) = \mathrm{Cov}(Y_t, Y_{t-s}) = \varrho^s \frac{\sigma^2}{1 - \varrho^2} .$$

Dividing by $\mathrm{Var}(Y_t) = \sigma^2/(1 - \varrho^2)$ gives the **autocorrelation** function:

$$\mathrm{Corr}(Y_t, Y_{t-s}) = \varrho^s , \quad s = 0, 1, 2, \dots .$$

In other words, under the assumption $|\varrho| < 1$, the autocorrelations of the AR(1) decrease geometrically. If ϱ is positive, the autocorrelations monotonously decrease; if ϱ is negative, they alternate in sign. In Fig. 10.2 we plot the autocorrelations of the AR(1) for four different values of ϱ.
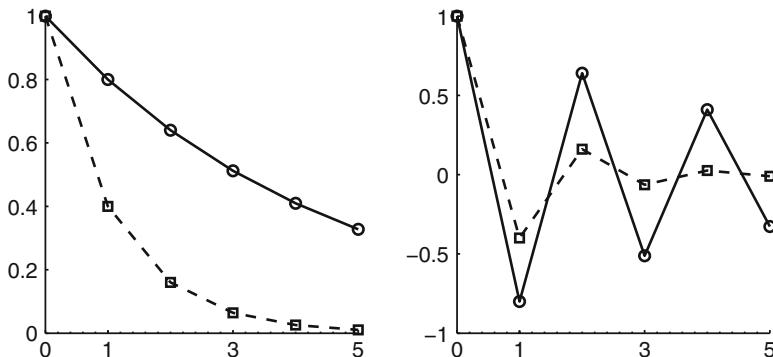
Fig. 10.2 Autocorrelations for the AR(1) model. *Left*: $\varrho = 0.8$ (*solid*) and $\varrho = 0.4$ (*dashed*). *Right*: $\varrho = -0.8$ (*solid*) and $\varrho = -0.4$ (*dashed*)

Remark 10.1 (Estimating Autocovariances). Suppose that $X_1, X_2, \dots, X_T$ is a *stationary* time series with autocovariance function $R(s) = \mathrm{Cov}(X_t, X_{t+s})$. Note that

$R(0) = \text{Var}(X_t)$. The autocovariances can be estimated via their (unbiased) sample averages:

$$\widehat{R}(s) = \frac{1}{T-s-1} \sum_{t=1}^{T-s}(X_t - \bar{X})(X_{t+s} - \bar{X}), \quad s = 0, 1, \ldots, T . \quad (10.4)$$

In order to obtain a meaningful estimate, the lag s should be significantly smaller than T. If the time series is not stationary, it is customary to delete the first K, say, samples, similar to the *burn-in period* for Markov chain Monte Carlo, and view the remaining samples as stationary.

We now turn to the estimation of the model parameters. Under the AR(p) model past observations feed into the current value of the series, where their effects are determined by the vector of AR coefficients $\boldsymbol{\varrho} = (\varrho_0, \varrho_1, \ldots, \varrho_p)^\top$. One main appeal of the AR(p) model is that it is in the form of a normal linear model, and as such, estimation of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\varrho}^\top, \sigma^2)^\top$ is easy. To proceed, let $\mathbf{y} = (y_1, \ldots, y_T)^\top$ be the observed data and denote the initial observations by $\mathbf{Y}_0 = (Y_{-p+1}, \ldots, Y_0)^\top$. Recall that the AR($p$) model determines the conditional pdf of the data given the initial conditions, that is $f_{\mathbf{Y}|\mathbf{Y}_0}(\mathbf{y} \,|\, \mathbf{y}_0 \,;\, \boldsymbol{\theta})$. The likelihood function of $\boldsymbol{\theta}$ for the observed data $\mathbf{y}$ is thus given by

$$L(\boldsymbol{\theta}; \mathbf{y}) = f_{\mathbf{Y}|\mathbf{Y}_0}(\mathbf{y} \,|\, \mathbf{y}_0 \,;\, \boldsymbol{\theta}) f_{\mathbf{Y}_0}(\mathbf{y}_0) .$$

As mentioned before, there are two ways to deal with the initial observations. The first is to assume that the time series is *stationary* and from this assumption we derive the distributions of $\mathbf{Y}_0$ and $\mathbf{Y}$. An easier approach, which we will follow here, is to simply assume that $\mathbf{Y}_0 = \mathbf{y}_0$ is given and specify the likelihood function as

$$L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}_0) = f_{\mathbf{Y}|\mathbf{Y}_0}(\mathbf{y} \,|\, \mathbf{y}_0; \boldsymbol{\theta}) .$$

In typical situations where T is much greater than p it makes little difference for parameter estimation whether or not the initial conditions are explicitly modeled.

Using the results of the normal linear model in Sect. 4.5, we can easily derive the joint density $f_{\mathbf{Y}|\mathbf{Y}_0}(\mathbf{y} \,|\, \mathbf{y}_0 \,;\, \boldsymbol{\theta})$ and hence the likelihood function. To that end, write the AR(p) model in matrix notation:

$$\mathbf{Y} = X\boldsymbol{\varrho} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I) ,$$

where

$$X = \begin{pmatrix} 1 & Y_0 & Y_{-1} & \cdots & Y_{-p+1} \\ 1 & Y_1 & Y_0 & \cdots & Y_{-p+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{T-1} & Y_{T-2} & \cdots & Y_{T-p} \end{pmatrix} .$$

Thus, conditional on $\mathbf{Y}_0 = \mathbf{y}_0$ the random vector $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\varrho}$ (with Y_k in matrix X replaced by y_k for $k = 0, -1, \ldots, -p+1$) has a $\mathsf{N}(\mathbf{0}, \sigma^2 I)$ distribution with density function

$$f_Z(\mathbf{z}) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2}\mathbf{z}^\top \mathbf{z}} .$$

Now, given the outcome $\mathbf{Y} = \mathbf{y}$ and initial conditions $\mathbf{Y}_0 = \mathbf{y}_0$, we have $\mathbf{z} = \mathbf{y} - X\boldsymbol{\varrho}$ (with Y_k in matrix X replaced by y_k for $k = -p+1, \ldots, T$), and the likelihood function is given by

$$L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{y}_0) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\varrho})^\top(\mathbf{y}-X\boldsymbol{\varrho})} .$$

The maximum likelihood estimators of $\boldsymbol{\varrho}$ and σ^2 are given by

$$\widehat{\boldsymbol{\varrho}} = (X^\top X)^{-1} X^\top \mathbf{Y} \qquad \text{and} \qquad \widehat{\sigma^2} = \frac{1}{T}(\mathbf{Y} - X\widehat{\boldsymbol{\varrho}})^\top(\mathbf{Y} - X\widehat{\boldsymbol{\varrho}}) .$$

Finally, Bayesian inference in the AR(p) model can proceed as in Sect. 8.2.2.

Example 10.3 (Modeling Unemployment with AR Models). In this example, we use autoregressive models with different lags to model US quarterly unemployment rates from the first quarter in 2002 to the last quarter in 2011—a total of 40 observations. The data are given in Table 10.2.

Table 10.2 US quarterly unemployment rates from 2002 Q1 to 2011 Q4

Year	Unemployment rate	Year	Unemployment rate	Year	Unemployment rate
2002 Q1	5.7	2006 Q1	4.7	2009 Q1	8.3
Q2	5.8	Q2	4.6	Q2	9.3
Q3	5.7	Q3	4.6	Q3	9.6
Q4	5.9	Q4	4.4	Q4	9.9
2003 Q1	5.9	2007 Q1	4.5	2010 Q1	9.8
Q2	6.1	Q2	4.5	Q2	9.6
Q3	6.1	Q3	4.7	Q3	9.5
Q4	5.8	Q4	4.8	Q4	9.6
2004 Q1	5.7	2008 Q1	5.0	2011 Q1	9.0
Q2	5.6	Q2	5.3	Q2	9.0
Q3	5.4	Q3	6.0	Q3	9.1
Q4	5.4	Q4	6.9	Q4	8.7
2005 Q1	5.3				
Q2	5.1				
Q3	5.5				
Q4	5.5				

We fit two autoregressive models with one and two lags respectively. To that end, we divide the data into two subsets: the first two observations are reserved

as the initial conditions, y_{-1} and y_0, and we explicitly model the remaining 38 observations $y_1, \ldots, y_{38}$. First we fit the following AR(1) model:

$$Y_t = \varrho_0 + \varrho_1 Y_{t-1} + \varepsilon_t \, ,$$

where $\varepsilon_t \sim_{\text{iid}} N(0, \sigma^2)$ for $t = 1, \ldots, 38$. By defining the design matrix X appropriately, the maximum likelihood estimates of ϱ_0, ϱ_1, and σ^2 can be computed easily; see the following MATLAB code.

```
%urate_ar.m
urate = [5.7 5.8 5.7 5.9 5.9 6.1 6.1 5.8 5.7 5.6 5.4 5.4 ...
    5.3 5.1 5.0 5.0 4.7 4.6 4.6 4.4 4.5 4.5 4.7 4.8 5.0 ...
    5.3 6.0 6.9 8.3 9.3 9.6 9.9 9.8 9.6 9.5 9.6 9.0 9.0 ...
    9.1 8.7]';
y = urate(3:end);
T = length(y);
X = [ones(T,1) urate(2:end-1)];
rhohat = (X'*X)\(X'*y);
yhat1 = X*rhohat;                 % fitted values
uhat = y-yhat1;                   % residuals
sig2hat = uhat'*uhat/T;
```

To assess the model fit, we also computed the fitted values under the AR(1) model, as well as the residuals. The fitted values of the AR(1) are plotted in Fig. 10.3 (left panel).
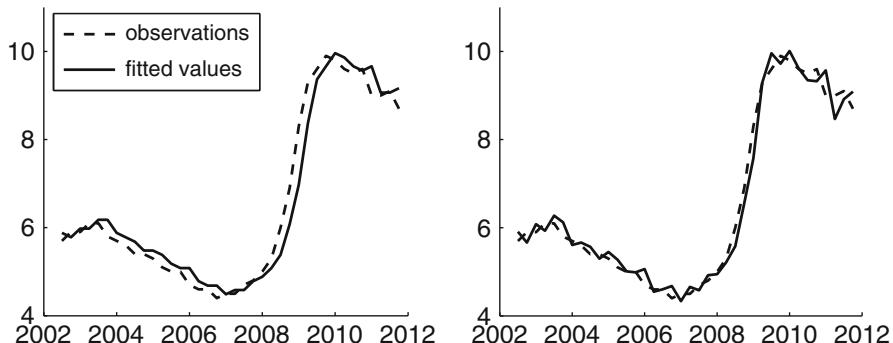
Fig. 10.3 The fitted values of the AR(1) (*left panel*) and the AR(2) (*right panel*)

It can be seen from the graph that the fitted values appear to be quite close to the actual observations. However, the AR(1) model seems to systematically underestimate the unemployment rate before 2007 and then overestimate it after 2007. To check whether the residuals have any serial correlation, we can estimate the autocovariance function via (10.4). In MATLAB we can do this, for example, via the following function.

```
function a = acov(x,s)
a = sum((x(1:end-s)  - mean(x)).*(x(s+1:end)  - mean(x))))/...
        (numel(x) - s - 1);
```

We find the autocorrelation at lag 1 via `acov(uhat,1)/acov(uhat,0)`, which turns out to be 0.7028. Similarly, for lags 2 and 3 we find the autocorrelations 0.5580 and 0.4421, respectively. This indicates that the model assumption of serially independent errors might not be valid, since the autocorrelations of the residuals remain substantial.

Next, we investigate if we can improve the model fit by using the AR(2) model:

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t ,$$

where $\varepsilon_t \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ for $t = 1, \ldots, 38$. By defining the design matrix appropriately, the maximum likelihood estimates for the model parameters and the corresponding fitted values can be obtained easily. The fitted values of the AR(2) are plotted in Fig. 10.3 (right panel), which appear to fit the actual observations better. In this case the lag-1 and lag-2 autocorrelations of the residuals are, respectively, -0.12 and -0.001.

We now return to the modeling situation of Example 10.1, using a more general setting. In particular, we consider a linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} ,$$

where the errors $\{\varepsilon_t\}$ follow an AR(p) model (with $\varrho_0 = 0$):

$$\varepsilon_t = \varrho_1 \varepsilon_{t-1} + \cdots + \varrho_p \varepsilon_{t-p} + U_t , \tag{10.5}$$

where $\{U_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $\varepsilon_{1-p} = \cdots = \varepsilon_0 = 0$. To keep the discussion concrete, we let $p = 1$ and define $\varrho_1 = \varrho$; higher-order AR models can be estimated similarly. Now, rewrite (10.5) in matrix notation:

$$H\boldsymbol{\varepsilon} = \mathbf{U} ,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^{\top}$, $\mathbf{U} = (U_1, \ldots, U_T)^{\top} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$, and

$$H = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\varrho & 1 & 0 & \cdots & 0 \\ 0 & -\varrho & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & -\varrho & 1 \end{pmatrix}$$

is a lower triangular $T \times T$ matrix with ones on the main diagonal. Note that H is **sparse**, i.e., it contains only a small proportion of nonzero elements. Now, since its determinant is 1, H is invertible for any ϱ. By a simple change of variables, we have

$$\mathbf{Y} - X\boldsymbol{\beta} = H^{-1}\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \sigma^2(H^\top H)^{-1}) \ .$$

Noting that the determinant of H is 1, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \varrho, \sigma^2; \mathbf{y}) = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^\top H^\top H(\mathbf{y} - X\boldsymbol{\beta}) \ . \qquad (10.6)$$

If the number of parameters is small, the maximum likelihood estimates can be obtained by numerically maximizing the log-likelihood function in (10.6). But when the dimension of the maximization is large, this approach is time-consuming and sometimes even infeasible.

Here we introduce a method to reduce the dimension of the numerical optimization; see also Example 6.14. First note that if ϱ is known, the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are available analytically (see Problem 10.6):

$$\widehat{\boldsymbol{\beta}} = (X^\top H^\top HX)^{-1}X^\top H^\top H\mathbf{y}, \quad \widehat{\sigma^2} = \frac{1}{T}(\mathbf{y} - X\widehat{\boldsymbol{\beta}})^\top H^\top H(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \ . \quad (10.7)$$

Now, we plug the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\widehat{\sigma^2}$ back into the log-likelihood function to obtain the *profile log-likelihood*, also called the **concentrated log-likelihood**:

$$\widetilde{l}(\varrho; \mathbf{y}) = l(\widehat{\boldsymbol{\beta}}, \varrho, \widehat{\sigma^2}; \mathbf{y}) \ ,$$

which is a function of ϱ only. Thus, we can maximize numerically the profile log-likelihood function to obtain the maximum likelihood estimate $\widehat{\varrho}$. Finally, given $\widehat{\varrho}$, we can use (10.7) to obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ analytically.

Example 10.4 (Sales and Ads Continued). Consider the model for the sales data in Example 10.1:

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \ ,$$
$$\varepsilon_t = \varrho\varepsilon_{t-1} + U_t \ ,$$

where Y_t is the sales in quarter t, x_t is the corresponding ads expenditure, $\varepsilon_0 = 0$, and $\{U_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$. We wish to compute the maximum likelihood estimates for $\boldsymbol{\beta}, \sigma^2$, and ϱ. Note that throughout we will use the matrix notation introduced earlier.

To that end, we first write a MATLAB script to evaluate the profile log-likelihood $\widetilde{l}(\varrho; \mathbf{y}) = l(\widehat{\boldsymbol{\beta}}, \varrho, \widehat{\sigma^2}; \mathbf{y})$.

```
function [l betahat sigma2hat] = AR1_loglike(rho,y,X)
T = length(y);
H = speye(T) - rho*sparse(2:T,1:T-1,ones(1,T-1),T,T);
HH = H'*H;
betahat = (X'*HH*X)\(X'*HH*y);
e = y-X*betahat;
sigma2hat = e'*HH*e/T;
l = -T/2*log(2*pi*sigma2hat) - .5/sigma2hat*e'*HH*e;
```

In the above code, note that the matrix H is constructed by writing

$$
H = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
-\varrho & 1 & 0 & \cdots & 0 \\
0 & -\varrho & 1 & \cdots & 0 \\
\vdots & & & \ddots & \vdots \\
0 & 0 & \cdots & -\varrho & 1
\end{pmatrix}
=
\begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & & & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 1
\end{pmatrix}
- \varrho
\begin{pmatrix}
0 & 0 & 0 & \cdots & 0 \\
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & & & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 0
\end{pmatrix}.
$$

Then, in the main script, we define the function $f(\varrho) = -\widetilde{l}(\varrho; \mathbf{y})$ and use the built-in minimization routine fminsearch to minimize f to obtain the maximum likelihood estimate $\widehat{\varrho}$. Finally, given $\widehat{\varrho}$, we use (10.7) to compute $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ analytically.

```
%sales.m
load 'ads.csv';
y = ads(:,1);
T = length(y);
X = [ones(T,1) ads(:,2)];
f = @(rho) -AR1_loglike(rho,y,X);
rhohat = fminsearch(f,0);
[l betahat sigma2hat] = AR1_loglike(rhohat,y,X);
```

The maximum likelihood estimate of ϱ is 0.95, indicating very strong first-order serial correlation in the error terms $\{\varepsilon_t\}$. The maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are, respectively, $(11.03, 2.32)^\top$ and 0.81. To assess the appropriateness of the model assumption $\{U_t\} \sim_{\text{iid}} N(0, \sigma^2)$, we compute the residuals $\{u_t\}$, which are our best guess for the (unknown) $\{U_t\}$. Recall that under the model we have $H\boldsymbol{\varepsilon} = \mathbf{U}$. Hence, we can obtain the residuals using

$$
\mathbf{u} = \widehat{H}(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) ,
$$

where $\widehat{H}$ is the same as H but with ϱ replaced by its estimate $\widehat{\varrho}$. This is implemented in MATLAB as follows:

```
e = y - X*betahat;
H = speye(T) - rhohat*sparse(2:T,1:T-1,ones(1,T-1),T,T);
u = H*e;
plot(u,'o','MarkerSize',10,'MarkerEdgeColor','k'); box off;
```

A plot of the residuals is given in Fig. 10.4. As the graph shows, the residuals now do not seem to have any systematic patterns across time, indicating no evidence of invalidating the assumption that $\{U_t\}$ are serially independent; see also Problem 10.4.

Fig. 10.4 A plot of the residuals of the linear regression model with AR(1) errors
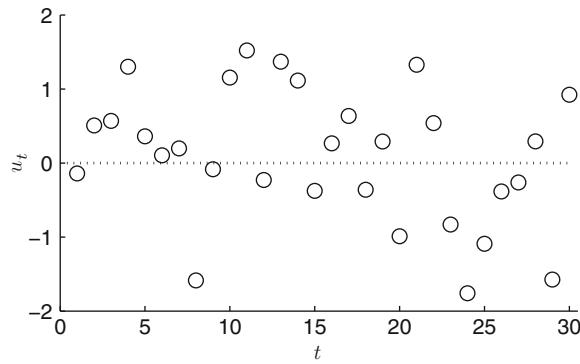
10.1.2 Moving Average Models

In an AR model, dependence of the responses is constructed by defining the current response in terms of a linear combination of previous responses. In contrast, in a moving average (MA) model, the current response Y_t depends on a linear combination of past error terms.

Definition 10.2. (Moving Average Model). In the **qth-order moving average** (MA(q)) model the observation at time t depends linearly on the previous q error terms:

$$Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q} , \tag{10.8}$$

where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$.

A standard way to treat the initial conditions is to assume $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{1-q} = 0$, although relaxing this assumption is straightforward (see Problem 10.3).

Under the MA(q) model, past shocks feed into the current value of the series, where their effects are determined by the signs and magnitudes of the MA coefficients $\psi_1, \ldots, \psi_q$. Unlike the AR(p) model each response always has a finite variance, as the following example shows.

Example 10.5 (Autocorrelations of MA(q)). We investigate the autocorrelation structure implied by the MA(q) model. First, we compute the variance of Y_t for $t > q$:

$$
\begin{aligned}
\mathrm{Var}(Y_t) &= \mathrm{Var}(\varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q}) \\
&= \mathrm{Var}(\varepsilon_t) + \psi_1^2 \mathrm{Var}(\varepsilon_{t-1}) + \cdots + \psi_q^2 \mathrm{Var}(\varepsilon_{t-q}) \\
&= \sigma^2 (1 + \psi_1^2 + \cdots + \psi_q^2) ,
\end{aligned}
$$

which is finite and independent of the time index t for $t > q$. Next, we compute the autocovariance at lag 1 as

$$
\begin{aligned}
\mathrm{Cov}(Y_t, Y_{t-1}) &= \mathrm{Cov}(\varepsilon_t + \cdots + \psi_q \varepsilon_{t-q}, \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q-1}) \\
&= \psi_1 \mathrm{Cov}(\varepsilon_{t-1}, \varepsilon_{t-1}) + \psi_2 \psi_1 \mathrm{Cov}(\varepsilon_{t-2}, \varepsilon_{t-2}) \\
&\quad + \cdots + \psi_q \psi_{q-1} \mathrm{Cov}(\varepsilon_{t-q}, \varepsilon_{t-q}) \\
&= (\psi_1 + \psi_2 \psi_1 + \cdots + \psi_q \psi_{q-1}) \sigma^2 .
\end{aligned}
$$

More generally, it can be shown that (see Problem 10.2)

$$
\mathrm{Cov}(Y_t, Y_{t-j}) = \begin{cases} \sigma^2 \sum_{i=0}^{q-j} \psi_{i+j} \psi_i, & j = 0, \ldots, q , \\ 0, & j > q , \end{cases} \tag{10.9}
$$

where $\psi_0 = 1$. Hence, the autocorrelations are

$$
\mathrm{Corr}(Y_t, Y_{t-j}) = \frac{\sum_{i=0}^{q-j} \psi_{i+j} \psi_i}{1 + \psi_1^2 + \cdots + \psi_q^2}
$$

for $j = 0, \ldots, q$ and $\mathrm{Corr}(Y_t, Y_{t-j}) = 0$ for $j > q$. In contrast to the AR case where the autocorrelation declines geometrically, those of the MA drop to zero after only q lags. As an illustration, in Fig. 10.5 we plot the autocorrelations of two MA(3) models.
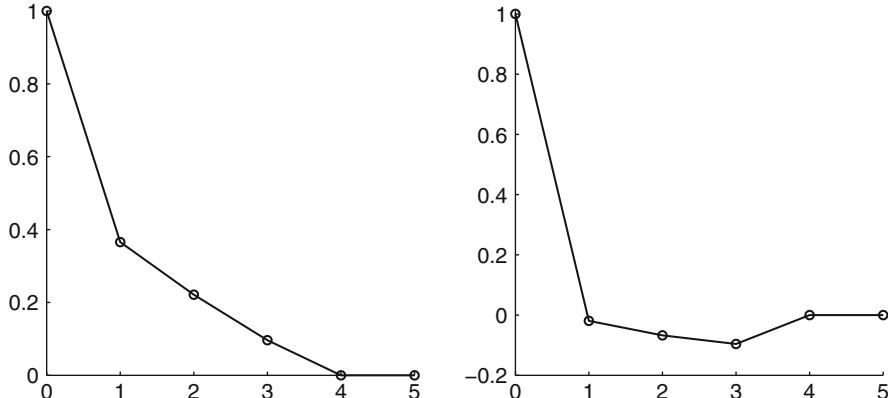
Fig. 10.5 The autocorrelations for the MA(3) with $\psi_1 = 0.8$, $\psi_2 = 0.6$, and $\psi_3 = 0.4$ (*left panel*) and $\psi_1 = -0.8$, $\psi_2 = -0.6$, and $\psi_3 = -0.4$ (*right panel*)

We now turn to estimation of the model parameters. Recall that under the MA(q) model, the responses $Y_1, \ldots, Y_T$ are a linear combination of T normal random variables $\varepsilon_1, \ldots, \varepsilon_T$. Therefore, $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$ has a multivariate normal distribution, and evaluating the log-likelihood should be simple. In practice, however, this approach requires manipulating large matrices, which is often time-consuming. The key to make this approach feasible is to realize that, as in the AR(p) case, the matrices involved in the MA(q) model are sparse, which makes the computation quick.

☞ 83

To keep the discussion concrete, consider evaluating the log-likelihood of the MA(1) model

$$Y_t = \varepsilon_t + \psi\, \varepsilon_{t-1}, \quad t = 1, \ldots T, \quad \varepsilon_0 = 0 .$$

First, we write this model in matrix form:

$$\mathbf{Y} = H\boldsymbol{\varepsilon} , \tag{10.10}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$, and

$$H = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \psi & 1 & 0 & \cdots & 0 \\ 0 & \psi & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & \psi & 1 \end{pmatrix}$$

is a sparse $T \times T$ matrix that contains only $2T - 1$ nonzero elements. It follows that

$$\mathbf{Y} \sim \mathsf{N}(\mathbf{0}, \sigma^2 H H^\top) \ .$$

Noting that the determinant of H is 1, the log-likelihood function is given by

$$l(\psi, \sigma^2; \mathbf{y}) = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathbf{y}^\top (H H^\top)^{-1} \mathbf{y} \ . \qquad (10.11)$$

It is important to note that one need not compute the inverse $(H H^\top)^{-1}$ in order to evaluate the log-likelihood—it is a time-consuming operation. Instead, one only needs to obtain the product $(H H^\top)^{-1} \mathbf{y}$, which can be quickly computed by solving the linear equation

$$H H^\top \mathbf{z} = \mathbf{y}$$

for $\mathbf{z}$. In MATLAB it can be done by using the backslash (\) command. The latter operation is much quicker, especially because $H H^\top$ is a sparse matrix.

Hence, one can evaluate the log-likelihood function $l(\psi, \sigma^2; \mathbf{y})$ quickly without inverting any large matrices. Then, the MLE of ψ and σ^2 can be obtained numerically. To evaluate the log-likelihood function for a general MA(q) model, one only needs to redefine the matrix H appropriately, and everything else follows directly as in the MA(1) case.

Given the method to quickly evaluate the log-likelihood function described above, the maximum likelihood estimate can be computed readily by numerical methods. For Bayesian inference, posterior draws of the parameters can be obtained using the Metropolis–Hastings algorithm.

Example 10.6 (Modeling US Inflation with MA(1)). In this example we model the dynamics of US quarterly inflation rate—computed from the consumer price index (CPI)—using a variant of the MA(1) model. Specifically, given the CPI z_t at time t, we compute the (annualized) inflation rate as $y_t = 400 \ln(z_t / z_{t-1})$. The CPI inflation rate from the second quarter of 1947 to the second quarter of 2011 is plotted in Fig. 10.6. A prominent feature of the CPI inflation is that it exhibits high persistence, in the sense that high (or low) inflation in the past tends to continue into the future. For instance, inflation tends to stay high (and variable) in the late 1970s and early 1980s, but it has become much lower (and less variable) since the mid-1980 until the global financial crisis in 2008.

There are various reasons why one might want to model the past inflation and accurately forecast future inflation. For example, the prices of many financial and real assets—such as bonds, properties, and precious metals—depend on future inflation. Another example is for conducting monetary policy: many central banks have explicit inflation targets, and in order to manage future inflation, it is important to be able to forecast it accurately.

Fig. 10.6 US CPI inflation
rate from 1947Q2 to 2011Q2
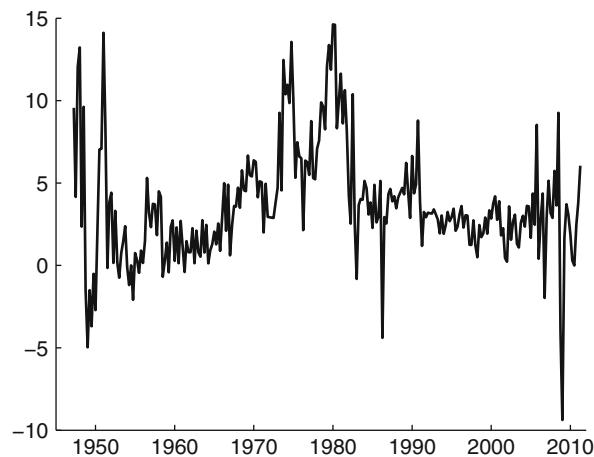
In this example we consider the following model:

$$Y_t = Y_{t-1} + \varepsilon_t + \psi\, \varepsilon_{t-1} \,,$$

where $\varepsilon_1, \ldots, \varepsilon_T \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $\varepsilon_0 = 0$.

This is a variation of the MA(1) model, sometimes called the **first-order integrated moving average** model. Instead of using the inflation rate as our dependent variable to fit the MA(1) model, we use its first difference $\Delta Y_t = Y_t - Y_{t-1}$. Estimation proceeds the same way as in the MA(1) model, with the minor modification of using ΔY_t as the dependent variable.

To obtain the maximum likelihood estimates of σ^2 and ψ, we first need the following function to evaluate the log-likelihood for the MA(1) model as given in (10.11):

```
function l = loglike_MA1(theta,y)
% the log-likelihood function for MA(1)
% input: theta = [psi sigma2]; y = data
psi = theta(1); sigma2 = theta(2);
T = length(y);
H = speye(T)  + psi*sparse(2:T,1:T-1,ones(1,T-1),T,T);
HH = H*H';
l = -T/2*log(2*pi*sigma2)  - .5/sigma2*y'*(HH\y);
```

Then, in the main script, we use the function `loglike_MA1` to compute the maximum likelihood estimates numerically. Specifically, we first construct the function `f` that evaluates the negative log-likelihood for the MA(1) model. (Recall that in MATLAB all optimization routines are framed in terms of minimization.) Then, we minimize `f` using the numerical minimization routine `fminsearch` with starting values 0 for ψ and the sample variance of Δy_t for σ^2.

```
%CPI_MA.m
load USCPI.csv
y0 = USCPI(1);
y = USCPI(2:end);
Dely = y - [y0; y(1:end-1)]; % define the dependent variable
T = length(Dely);
                            % negative of the log-likelihood
f = @(theta) -loglike_MA1(theta,Dely);
theta0 = [0 var(Dely)];       % initial guess
thetahat = fminsearch(f, theta0);
psihat = thetahat(1);
l = loglike_MA1(thetahat,Dely);
```

The maximum likelihood estimates of ψ and σ^2 are, respectively, -0.402 and 5.245, and the corresponding maximized log-likelihood value is -577.74. To assess model fit, we compute the fitted value $\widehat{y}_t$:

$$\widehat{y}_t = y_{t-1} + \widehat{\psi}\, \widehat{u}_{t-1} \;,$$

where $\widehat{\psi}$ is the maximum likelihood estimate of ψ and $\widehat{u}_{t-1}$ is the residual for period $t-1$. Note that the residuals can be computed easily by

```
Hhat = speye(T) + psihat*sparse(2:T,1:T-1,ones(1,T-1),T,T);
uhat = Hhat\Dely;
```

Finally, we plot the fitted versus the observed values of the inflation rates in Fig. 10.7.

Fig. 10.7 The fitted values for the MA(1) model
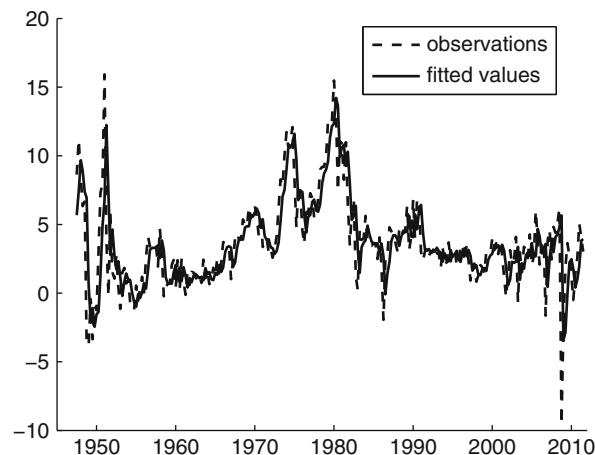
10.1.3   Autoregressive-Moving Average Models

Of course, we can combine autoregressive and moving average models to have more complex autocorrelation patterns.

Definition 10.3.   (Autoregressive-Moving Average Model). In the (p, q)th-order **autoregressive-moving average** (ARMA(p, q)) model the observation at time t depends linearly on the previous p observations as well as the previous q error terms:

$$Y_t = \varrho_0 + \varrho_1 Y_{t-1} + \cdots + \varrho_p Y_{t-p} + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q} \,, \quad (10.12)$$

where $\{\varepsilon_t\} \overset{iid}{\sim} \mathsf{N}(0, \sigma^2)$ for $t = 1, \ldots, T$, and $\varepsilon_0 = \cdots = \varepsilon_{1-q} = 0$.

In matrix notation, we can write the system (10.12) as

$$\mathbf{Y} = X\boldsymbol{\varrho} + H\boldsymbol{\varepsilon} \,,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$, $\boldsymbol{\varrho} = (\varrho_0, \varrho_1, \ldots, \varrho_p)^\top$,

$$X = \begin{pmatrix} 1 & Y_0 & Y_{-1} & \cdots & Y_{-p+1} \\ 1 & Y_1 & Y_0 & \cdots & Y_{-p+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{T-1} & Y_{T-2} & \cdots & Y_{T-p} \end{pmatrix} \,,$$

and H is a lower triangular matrix with ones on the main diagonal, ψ_1 on the first diagonal below the main diagonal, ψ_2 on the second diagonal below the main diagonal, and so on. We define $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)^\top$. Similar to the AR and MA models, we have

$$\mathbf{Y} - X\boldsymbol{\varrho} \sim \mathsf{N}(\mathbf{0}, \sigma^2 H H^\top) \,,$$

and the log-likelihood function $l(\boldsymbol{\varrho}, \boldsymbol{\psi}, \sigma^2; \mathbf{y}, \mathbf{y}_0)$ is given by (with Y_k in matrix X replaced by y_k for $k = -p+1, \ldots, T$):

$$\begin{aligned} l(\boldsymbol{\varrho}, \boldsymbol{\psi}, \sigma^2; \mathbf{y}, \mathbf{y}_0) = & -\frac{T}{2} \ln(2\pi\sigma^2) \\ & -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\varrho})^\top (H H^\top)^{-1} (\mathbf{y} - X\boldsymbol{\varrho}) \,. \end{aligned} \quad (10.13)$$

In principle we can numerically maximize $l(\boldsymbol{\varrho}, \boldsymbol{\psi}, \sigma^2; \mathbf{y}, \mathbf{y}_0)$ to find the maximum likelihood estimates of the parameters. But this approach is time-consuming in this context as the dimension of the parameter vector is typically large. Instead, as discussed earlier, we reduce the dimension of the numerical maximization by first obtaining the *profile log-likelihood*:

$$\widetilde{l}(\boldsymbol{\psi}; \mathbf{y}) = l(\widehat{\boldsymbol{\varrho}}, \boldsymbol{\psi}, \widehat{\sigma^2}; \mathbf{y}, \mathbf{y}_0) \,,$$

where the maximum likelihood estimates of $\boldsymbol{\varrho}$ and σ^2 are available analytically (see Problem 10.6):

$$\widehat{\boldsymbol{\varrho}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \mathbf{y}, \quad \widehat{\sigma^2} = \frac{1}{T}(\mathbf{y} - X\widehat{\boldsymbol{\varrho}})^\top \Sigma^{-1}(\mathbf{y} - X\widehat{\boldsymbol{\varrho}}) \,, \quad (10.14)$$

where $\Sigma = HH^\top$. We then maximize numerically the profile log-likelihood—which is a function of $\boldsymbol{\psi}$ only—to obtain the maximum likelihood estimate $\widehat{\boldsymbol{\psi}}$. Finally, given $\widehat{\boldsymbol{\psi}}$, we use (10.14) to obtain $\widehat{\boldsymbol{\varrho}}$ and $\widehat{\sigma^2}$ analytically.

Example 10.7 (Modeling US Inflation with ARMA(1,1)). In Example 10.6 we fitted the US inflation data with an integrated MA(1) model. In this example we consider a slight generalization by including an intercept and allowing a first-order AR coefficient:

$$Y_t = \varrho_0 + \varrho_1 Y_{t-1} + \varepsilon_t + \psi \varepsilon_{t-1} \,,$$

where $\varepsilon_1, \ldots, \varepsilon_T \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $\varepsilon_0 = 0$. Hence, given the data, the design matrix X is

$$X = \begin{pmatrix} 1 & y_0 \\ 1 & y_1 \\ \vdots & \vdots \\ 1 & y_{T-1} \end{pmatrix} \,.$$

In this example, we compute the maximum likelihood estimates of the model parameters; for a Bayesian treatment of the model, see Problem 10.7.

The MATLAB function `loglike_ARMA11` takes the design matrix X and the outcomes $\mathbf{y}$ and evaluates the profile log-likelihood function at ψ. Note that the function also reports the maximum likelihood estimates of $\boldsymbol{\varrho} = (\varrho_0, \varrho_1)^\top$ and σ^2 given the value of ψ.

```
function [l rhohat sigma2hat] = loglike_ARMA11(psi,X,y)
T = length(y);
H = speye(T) + psi*sparse(2:T,1:T-1,ones(1,T-1),T,T);
HH = H*H';
rhohat = (X'*(HH\X))\(X'*(HH\y));
```

```
uhat = y-X*rhohat;
sigma2hat = uhat'*(HH\uhat)/T;
l = -T/2*log(2*pi*sigma2hat) - .5/sigma2hat*uhat'*(HH\uhat);
```

Then, in the main script, we numerically minimize the negative of the function `loglike_ARMA11` with respect to ψ. Given the maximum likelihood estimate $\widehat{\psi}$, which is calculated as -0.277, we use `loglike_ARMA11` again to compute the maximum likelihood estimates of ϱ and σ^2, which are, respectively, $(0.536, 0.849)^\top$ and 4.91. The corresponding maximized log-likelihood value is -569.15.

```
%CPI_ARMA.m
load USCPI.csv
y0 = USCPI(1);
y = USCPI(2:end);
T = length(y);
X = [ones(T,1) [y0; y(1:end-1)]];
f = @(psi) -loglike_ARMA11(psi,X,y);
psihat = fminsearch(f,0);
[l rhohat sigma2hat] = loglike_ARMA11(psihat,X,y);
```

Compared with the integrated MA(1) model in Example 10.6, it is not obvious that the ARMA(1,1) is a better model. Although it does fit the data better—its maximized log-likelihood value is -569.15 compared to -577.74, the corresponding value for the integrated MA(1) model—it is also more complex and has more parameters.

To compare these two models while taking into account both goodness of fit and model complexity, we make use of two popular information criteria: **Akaike information criterion** (AIC) and the **Bayesian information criterion** (BIC); see Problem 10.4. The AIC and BIC for the integrated MA(1) model are, respectively, -1159.5 and -1166.6, whereas the corresponding values for the ARMA(1,1) model are -1146.3 and -1160.5. Hence, both information criteria suggest that ARMA(1,1) is a better model for the inflation data.

10.2   Gaussian Models

As discussed in the introduction, an important case where observations are likely to be dependent is when there are measurements on related subjects. One convenient class of models for dependent data is *Gaussian models*, where the data, say, $Y_1, \ldots, Y_n$, are distributed according to a multivariate normal (i.e., Gaussian) distribution:

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$$

for some known or unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. In order for the model to be meaningful for statistical analysis one usually needs to impose extra structure on the parameters $\boldsymbol{\mu}$ and Σ.

As a first illustration, consider an extension of the two-sample normal model in Example 4.3, where in addition to measuring the heights of men whose mothers smoked (group 1) and did not smoke (group 2), we also measure the weights. The data can then be described by a vector $\mathbf{Y} = (X_1, V_1, \ldots X_{60}, V_{60}, Y_1, W_1, \ldots, Y_{140}, W_{140})^\top$, where (X_i, V_i) is the (height,weight) of person i in group 1 and (Y_i, W_i) the (height,weight) of person i in group 2. The vector $\mathbf{Y}$ can be modeled with a $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ distribution, where $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \ldots, \mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \ldots, \mu_{21}, \mu_{22})^\top$ (60 pairs μ_{11}, μ_{12} followed by 140 pairs μ_{21}, μ_{22}), and the covariance matrix Σ is block-diagonal with 2×2 blocks. The first 60 blocks on the diagonal (all the same) correspond to the covariance matrix of the height X and weight W of a person from the first group, which are clearly not independent. Similarly the remaining 140 diagonal blocks (all the same) correspond to the covariance matrix of the height Y and weight W of a person from the second group. This Gaussian model has only 10 parameters, as opposed to the possibly 80600 parameters of the general multivariate Gaussian model; see also
Problem 4.4.

Recall from Sect. 3.6 some important properties of the multivariate normal distribution. Let $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$:

1. All the marginal distributions are Gaussian.
2. Conditional distributions are Gaussian.
3. Any affine combination $b_0 + \sum_{i=1}^{n} b_i X_i$ has a normal distribution.
4. To simulate $\mathbf{X}$:

    a. Derive the *Cholesky decomposition* $\Sigma = AA^\top$.
    b. Generate $Z_1, \ldots, Z_n \sim_{\text{iid}} \mathsf{N}(0, 1)$. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$.
    c. Return $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$.

10.2.1   Gaussian Graphical Model

Since a Gaussian distribution is fully characterized by the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ, it suffices to specify these two quantities to construct a Gaussian model. It is often convenient to represent the covariance structure of a Gaussian model via a graph: a **Gaussian graphical model**. This graph is similar to
a Bayesian network in Sect. 8.3, but is *undirected*.

The purpose of a Gaussian graphical model is to summarize the conditional independence properties of the variables. In particular, in a Gaussian graphical model of a random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$, the nodes represent the components $X_1, \ldots, X_n$. Two nodes are connected by an undirected edge if and only if the corresponding variables are *conditionally dependent given all the other values*.

Recall that X_i and X_j are conditionally *independent* if (using Bayesian notation for simplicity)

$$f(x_i, x_j \mid x_k, k \neq i, j) = f(x_i \mid x_k, k \neq i, j) \, f(x_j \mid x_k, k \neq i, j) \, .$$

Because **X** is Gaussian, its pdf is given by

$$f(\mathbf{x}) = (2\pi)^{-n/2} \sqrt{\det(\Lambda)} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Lambda(\mathbf{x}-\boldsymbol{\mu})} \, ,$$

where $\Lambda = (\lambda_{ij})$ is the inverse of the covariance matrix Σ, called the **precision matrix**. Therefore, the conditional joint pdf of X_i and X_j is

$$f(x_i, x_j \mid x_k, k \neq i, j) \propto \exp\left(-\frac{1}{2}(\lambda_{ii} x_i^2 + 2x_i \, a + \lambda_{ij} x_i x_j + \lambda_{jj} x_j^2 + 2x_j \, b)\right) \, ,$$

where a and b may depend on $x_k, k \neq i, j$. This shows that X_i and X_j are conditionally independent given $\{X_k, k \neq i, j\}$, if and only if $\lambda_{ij} = 0$. Consequently, (i, j) is an edge in the graphical model if and only if $\lambda_{ij} \neq 0$. In typical applications (e.g., in image analysis), each vertex in the graphical model only has a small number of adjacent vertices. In such cases the precision matrix is thus sparse, and the Gaussian vector can be generated efficiently using, for example, sparse Cholesky factorization.

For a sparse precision matrix the following algorithm is more efficient than Algorithm 3.3 for generating independent samples.   ☞ 84

Algorithm 10.1. (Multivariate Normal Vector Generation Using the Precision Matrix). To generate N independent draws from a $\mathsf{N}(\boldsymbol{\mu}, \Lambda^{-1})$ distribution of dimension n carry out the following steps:

1. Determine the lower Cholesky factorization $\Lambda = DD^\top$.
2. Generate $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ by drawing $Z_1, \ldots, Z_n \sim_{\text{iid}} \mathsf{N}(0, 1)$.
3. Solve $\mathbf{Y}$ from $\mathbf{Z} = D^\top \mathbf{Y}$.
4. Output $\mathbf{X} = \boldsymbol{\mu} + \mathbf{Y}$.
5. Repeat Steps 2–4 independently N times.

Example 10.8 (Gaussian Graphical Model). In Fig. 10.8 a Gaussian graphical model is depicted for $n = 8$ normal random variables, divided into three groups. Variables in different groups are independent of each other. The first group contains only X_1, which is independent of all the other variables. In the second group, X_2, X_3, and X_4, each variable is conditionally dependent of the other two. In the last group, $X_5, \ldots, X_8$, each variable is conditionally dependent on one or two variables, while conditionally independent of the rest. For example, X_6 is conditionally independent of X_8 given X_7.

Fig. 10.8 An example of a
Gaussian graphical model
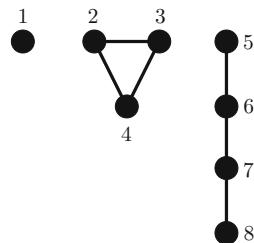
The covariance and precision matrices have the following structure:

$$
\Sigma = \begin{pmatrix}
\boxed{*} & & \\
 & \begin{matrix} * & * & * \\ * & * & * \\ * & * & * \end{matrix} & \\
 & & \begin{matrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{matrix}
\end{pmatrix}, \qquad
\Lambda = \begin{pmatrix}
\boxed{*} & & \\
 & \begin{matrix} * & * & * \\ * & * & * \\ * & * & * \end{matrix} & \\
 & & \begin{matrix} * & * & 0 & 0 \\ * & * & * & 0 \\ 0 & * & * & * \\ 0 & 0 & * & * \end{matrix}
\end{pmatrix}.
$$

Namely, the (i, j)th element of the covariance matrix Σ is the covariance between X_i and X_j. Hence, Σ consists of three diagonal blocks, each corresponding to the covariances among the variables within each group. The precision matrix has nonzero entries λ_{ij} precisely when (i, j) is an edge in the graph. Notice that here the precision matrix is sparser than the covariance matrix.

10.2.2   Random Effects

In the ANOVA models introduced in Chap. 4, it was assumed that the "effects" of the factors—that is, the parameters μ, α_j, and β_k—are *fixed* (deterministic). In a variety of situations it is more appropriate to assume that certain model parameters are *random*. The following example illustrates the idea.

Example 10.9 (One-Factor Random Effects ANOVA Model). To investigate whether geographical location is important in the effectiveness of a new type of herbicide, a researcher selects ten locations from a large number of possible locations within a country. At each location the herbicide is applied to three similar test plots. Each test plot is divided in half. One half (randomly selected) receives the herbicide and the other half is left untreated. The difference in crop yield for each plot is measured, giving 30 measurements (response variables) in total. The experimental design is here *hierarchical* in structure: first the locations are chosen,

and then the measurements are taken. The selection of the location could be modeled via independent random variables

$$\mu_1, \ldots, \mu_{10} \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma_\mu^2) \, ,$$

representing the expected difference in crop yields at the ten locations, where μ and σ_μ^2 are fixed parameters. Given the $\{\mu_i\}$, the actual difference in crop yield for the kth crop at location i could be modeled as

$$(Y_{ik} \mid \mu_i) \sim \mathsf{N}(\mu_i, \sigma^2), \text{ independently for } i = 1, \ldots, 10 \, ,$$

where σ^2 is fixed.

For this designed experiment the researcher is not interested *per se* in statements about the ten selected locations, but rather in conclusions pertaining to *all* possible geographical locations—in particular, regarding the parameters μ and σ_μ^2. For example, is the treatment effective across the country ($\mu > 0$)? Is geographical location much more important than measurement error in explaining the variability in the measurements (σ_μ^2 is much greater than σ^2)?

We summarize the one-factor random effects ANOVA model as follows.

Definition 10.4.   (One-Factor Random Effects ANOVA Model). Let Y_{ik} be the response for the kth replication at level i. Then

$$Y_{ik} = \mu_i + \varepsilon_{ik} \, , \quad k = 1, \ldots, n_i \, , \ i = 1, \ldots, d \, , \tag{10.15}$$

where

$$\mu_1, \ldots, \mu_d \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma_\mu^2) \, ,$$

independent of $\{\varepsilon_{ik}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$.

The model is again Gaussian, but, due to the hierarchical formulation, the $\{Y_{ik}\}$ are no longer independent within the ith level. By the model assumptions in (10.15), the responses $Y_{i1}, \ldots, Y_{in_i}$ are independent conditional on the random effect μ_i. However, marginalized over μ_i, the covariance between, say, Y_{ij} and Y_{ik}, $j \neq k$, is nonzero:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(\mu_i + \varepsilon_{ij}, \mu_i + \varepsilon_{ik}) = \sigma_\mu^2 \, .$$

If we define $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, then each $\mathbf{Y}_i$ is independent but not identically distributed (denoted by "ind" below) as

$$\mathbf{Y}_i \overset{\text{ind}}{\sim} \mathsf{N}(\mu \mathbf{1}_{n_i}, \sigma^2 I_{n_i} + \sigma_\mu^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top), \quad i = 1, \ldots, d \, ,$$

where $\mathbf{1}_{n_i}$ is an $n_i \times 1$ column of ones and I_{n_i} is the n_i-dimensional identity matrix. To see this, first note that, by definition,

$$\mathbf{Y}_i = \mu_i \mathbf{1}_{n_i} + \boldsymbol{\varepsilon}_i \; ,$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_1, \dots, \varepsilon_{n_i})^\top \sim \mathsf{N}(\mathbf{0}, \sigma^2 I_{n_i})$. In other words, $\mathbf{Y}_i$ is an affine transformation of normal random variables and therefore has a normal distribution. In addition, it is easy to check that its expectation is $\mathbb{E}\mathbf{Y}_i = \mu \mathbf{1}_{n_i}$, and its covariance matrix is

$$\mathrm{Cov}(\mathbf{Y}_i) = \mathrm{Var}(\mu_i)\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top + \mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \sigma_\mu^2 \mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top + \sigma^2 I_{n_i} \; .$$

Hence, the claim follows.

From the above discussion, we have also derived the log-likelihood function for the one-factor random effects model in (10.15). More specifically, given the outcomes $\mathbf{Y}_1 = \mathbf{y}_1, \dots, \mathbf{Y}_d = \mathbf{y}_d$, the log-likelihood function is given by

$$\begin{aligned}
l(\mu, \sigma_\mu^2, \sigma^2; \mathbf{y}) = &-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{d}\ln|\Sigma_i| \\
&-\frac{1}{2}\sum_{i=1}^{d}(\mathbf{y}_i - \mu\mathbf{1}_{n_i})^\top \Sigma_i^{-1}(\mathbf{y}_i - \mu\mathbf{1}_{n_i}) \; ,
\end{aligned} \tag{10.16}$$

where $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_d^\top)^\top$ and $\Sigma_i = \sigma_\mu^2 \mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top + \sigma^2 I_{n_i}$.

Since the log-likelihood function is of low dimension, the maximum likelihood estimates of μ, σ_μ^2, and σ^2 can be obtained quickly by numerically maximizing the log-likelihood. For a Bayesian analysis of the one-factor random effects model, see Problem 10.11.

Example 10.10 (One-Factor Random Effects ANOVA Model Continued). Consider again Example 10.9 in which we investigate if geographical location is important in the effectiveness of a new type of herbicide. Suppose the researcher has carried out the experiments and collected data from ten randomly selected locations. Specifically, at each location the differences in crop yield (kg) for three test plots are measured. The results are reported in Table 10.3.

Table 10.3 Differences in crop yield (kg)

Location	Difference in crop yield			Location	Difference in crop yield		
1	22.6	20.5	20.8	6	14.5	10.5	12.3
2	22.6	21.2	20.5	7	20.8	19.1	21.3
3	17.3	16.2	16.6	8	17.4	18.6	18.6
4	21.4	23.7	23.2	9	25.1	24.8	24.9
5	20.9	22.2	22.6	10	14.9	16.3	16.6

To compute the maximum likelihood estimates of μ, σ_μ^2, and σ^2, we first write a MATLAB script to evaluate the log-likelihood function $l(\mu, \sigma_\mu^2, \sigma^2; \mathbf{y})$. Note that in the code below the outcomes are stored as a matrix, where each row contains the experimental results in one of the randomly selected locations.

```
function l = sfran_loglike(mu,sigma2_mu,sigma2,y)
[d ni] = size(y);
Sigmai = sigma2*eye(ni) + sigma2_mu*ones(ni,ni);
l = -(ni*d)/2*log(2*pi) - d/2*log(det(Sigmai));
for i=1:d
    yi = y(i,:)';
    l = l - .5*(yi-mu)'*(Sigmai\(yi-mu));
end
```

Next, in the main script, we load the data and define a trivariate function that evaluates the negative of the log-likelihood function `sfran_loglike`. The new function is then passed to the built-in minimization routine `fminsearch` to compute the minimizer. For the starting values for μ, σ_μ^2, and σ^2, we use

$$\bar{y} = \frac{1}{d}\sum_{i=1}^{d}\bar{y}_i, \quad s^2 = \frac{1}{d-1}\sum_{i=1}^{d}(\bar{y}_i - \bar{y})^2, \quad \bar{s}^2 = \frac{1}{d}\sum_{i=1}^{d}s_i^2,$$

where $\bar{y}_i$ and s_i^2 are, respectively, the sample mean and sample variance of the outcomes in location i.

```
%sfran.m
y = [ 22.6 20.5 20.8; 22.6 21.2 20.5; 17.3 16.2 16.6; ...
      21.4 23.7 23.2; 20.9 22.2 22.6; 14.5 10.5 12.3; ...
      20.8 19.1 21.3; 17.4 18.6 18.6; 25.1 24.8 24.9; ...
      14.9 16.3 16.6];
f = @(theta) -sfran_loglike(theta(1),theta(2),theta(3),y);
ybar = mean(y,2);
theta0 = [mean(ybar) var(ybar) mean(var(y,0,2))];
thetahat = fminsearch(f,theta0);
```

The maximum likelihood estimates of μ, σ_μ^2, and σ^2 are 19.6, 12.19, and 1.167, respectively. For this example the herbicide seems to be quite effective.

The two-factor random effects ANOVA model can be defined similarly. Below we use the "factor effects" representation.

Definition 10.5. (Two-Factor Random Effects ANOVA Model). Let Y_{ijk} be the response for the kth replication at cell (i, j). Then

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} ,$$

$$k = 1, \ldots, n_{ij} , \ i = 1, \ldots, d_1 , \ j = 1, \ldots, d_2 , \tag{10.17}$$

where μ is a fixed constant and the following random variables are independent of each other:

$$\{\alpha_i\} \overset{\text{iid}}{\sim} N(0, \sigma_\alpha^2) , \qquad \{\beta_j\} \overset{\text{iid}}{\sim} N(0, \sigma_\beta^2) ,$$

$$\{\gamma_{ij}\} \overset{\text{iid}}{\sim} N(0, \sigma_\gamma^2), \qquad \{\varepsilon_{ijk}\} \overset{\text{iid}}{\sim} N(0, \sigma^2) .$$

Note that there are much fewer parameters than for the corresponding fixed effects model. Also, there are no restrictions such as $\sum_i \alpha_i = 0$ on the parameters.

To derive the likelihood function, we first rewrite (10.17) in matrix form. To that end, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_1})^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_2})^\top$, and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{d_1 d_2})^\top$. Arrange the responses $\{Y_{ijk}\}$ and errors $\{\varepsilon_{ijk}\}$ as $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, where $n = \sum_{i,j} n_{ij}$. Then,

$$\mathbf{Y} = \mu\mathbf{1} + X_{\boldsymbol{\alpha}}\boldsymbol{\alpha} + X_{\boldsymbol{\beta}}\boldsymbol{\beta} + X_{\boldsymbol{\gamma}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} , \tag{10.18}$$

where $\mathbf{1}$ an n-dimensional vector of 1s, and the design matrices $X_{\boldsymbol{\alpha}}$, $X_{\boldsymbol{\beta}}$, and $X_{\boldsymbol{\gamma}}$ are appropriately defined. Again, $\mathbf{Y}$ is an affine transformation of normal random variables and therefore has a multivariate normal distribution. Its mean is $\mathbb{E}\mathbf{Y} = \mu\mathbf{1}$ and its covariance matrix is given by

$$\Sigma = \sigma_\alpha^2 X_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}^\top + \sigma_\beta^2 X_{\boldsymbol{\beta}} X_{\boldsymbol{\beta}}^\top + \sigma_\gamma^2 X_{\boldsymbol{\gamma}} X_{\boldsymbol{\gamma}}^\top + \sigma^2 I .$$

Hence, given the outcomes $\mathbf{Y} = \mathbf{y}$, the log-likelihood function for the two-factor ANOVA model is

$$l(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma^2; \mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma|$$

$$-\frac{1}{2}(\mathbf{y} - \mu\mathbf{1})^\top \Sigma^{-1} (\mathbf{y} - \mu\mathbf{1}) . \tag{10.19}$$

Calculation of the maximum likelihood estimates involves a 5-dimensional maximization problem, which may be time-consuming, but can still be done numerically. There are two computational issues that are worth mentioning. First, note that evaluation of the log-likelihood function (10.19) involves calculating the log determinant $\ln |\Sigma|$. When the dimension n is large, computing the determinant

$|\Sigma|$ first and then taking the log might lead to substantial rounding error. Instead, consider the following equivalent calculations: obtain the Cholesky factor C of Σ. Since C is a lower triangular matrix, its determinant is equal to the product of the diagonal elements, say, $c_{11}, \ldots, c_{nn}$. It follows then that

$$\ln |\Sigma| = 2 \ln |C| = 2 \sum_{i=1}^{n} \ln c_{ii} \ .$$

The second issue concerns the restrictions on the variance parameters σ_α^2, σ_β^2, σ_γ^2, and σ^2—since they represent variances, they have to be positive. In other words, computing their maximum likelihood estimates is in fact a constrained maximization problem, and ignoring the restrictions might lead to numerical errors. One solution to this problem is to reparameterize in terms of

$$\eta_\alpha = \ln(\sigma_\alpha^2), \quad \eta_\beta = \ln(\sigma_\beta^2), \quad \eta_\gamma = \ln(\sigma_\gamma^2), \quad \eta = \ln(\sigma^2) \ ,$$

and maximize the log-likelihood (10.19)

$$l(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma^2; \mathbf{y}) = l(\mu, e^{\eta_\alpha}, e^{\eta_\beta}, e^{\eta_\gamma}, e^\eta; \mathbf{y})$$

with respect to μ, η_α, η_β, η_γ, and η. Once the maximum likelihood estimates of the new parameters are obtained, those for the original parameterization can be computed easily. The following example illustrates these points.

Example 10.11 (Two-Factor Random Effects ANOVA Model). In this example we investigate the breeding value of a set of five sires in raising pigs. Each sire is mated to a random group of dams, and the mating produces a litter of pigs whose characteristics are measured. In particular, the average daily gain of two piglets in each litter (in pounds) over a given period of time is recorded. The outcomes are reported in Table 10.4.

Table 10.4 Average daily gain of two piglets in each litter (in pounds)

Sire	Dam	Gain	Sire	Dam	Gain
1	1	1.39	3	2	0.95
1	1	1.29	3	2	0.96
1	2	1.12	4	1	0.82
1	2	1.16	4	1	0.92
2	1	1.52	4	2	1.18
2	1	1.62	4	2	1.20
2	2	1.88	5	1	1.47
2	2	1.87	5	1	1.41
3	1	1.24	5	2	1.57
3	1	1.18	5	2	1.65

The model we consider is

$$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \varepsilon_{ijk} ,$$

where α_i is the random effect associated with the ith sire and γ_{ij} is the random effect associated with the ith sire and jth dam. To write the model in matrix form, let $\mathbf{Y} = (Y_{111}, Y_{112}, Y_{121}, Y_{122}, \ldots, Y_{521}, Y_{522})^\top$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_5)^\top$, and $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{51}, \gamma_{52})^\top$. Then,

$$\mathbf{Y} = \mu \mathbf{1}_{20} + X_{\boldsymbol{\alpha}} \boldsymbol{\alpha} + X_{\boldsymbol{\gamma}} \boldsymbol{\gamma} + \boldsymbol{\varepsilon} ,$$

where $X_{\boldsymbol{\alpha}} = I_5 \otimes \mathbf{1}_4$, $X_{\boldsymbol{\gamma}} = I_{10} \otimes \mathbf{1}_2$, $\otimes$ is the Kronecker product, $\mathbf{1}_p$ is a $p \times 1$ vector of ones, and I_q is the q-dimensional identity matrix.

As in the previous example, we first write a MATLAB script to evaluate the log-likelihood function parameterized in terms of μ, $\eta_{\boldsymbol{\alpha}} = \ln(\sigma_{\boldsymbol{\alpha}}^2)$, $\eta_{\boldsymbol{\gamma}} = \ln(\sigma_{\boldsymbol{\gamma}}^2)$, and $\eta = \ln(\sigma^2)$.

```
function l = sfran2_loglike(mu,eta_alpha,eta_gamma, ...
    eta,y,Xalpha,Xgamma)
sigma2_alpha = exp(eta_alpha);
sigma2_gamma = exp(eta_gamma);
sigma2 = exp(eta);
n = length(y);
Sigma = sigma2*speye(n) + sigma2_alpha*(Xalpha*Xalpha')+...
    sigma2_gamma*(Xgamma*Xgamma');
l = -n/2*log(2*pi) - sum(log(diag(chol(Sigma)))) + ...
    - .5*(y-mu)'*(Sigma\(y-mu));
```

Then, in the main script, we maximize the log-likelihood function numerically with respect to μ, $\eta_{\boldsymbol{\alpha}}$, $\eta_{\boldsymbol{\gamma}}$, and η. The maximizer is then transformed to get the maximum likelihood estimates of the original parameters. The estimates of μ, $\sigma_{\boldsymbol{\alpha}}^2$, $\sigma_{\boldsymbol{\gamma}}^2$, and σ^2 are, respectively, 1.32, 0.0537, 0.0318, and 0.023. For this data set, the sires seem to be the most important factor in explaining the variation in the average daily gain of the piglets.

```
%sfran2.m
y = [1.39 1.29 1.12 1.16 1.52 1.62 1.88 1.87 1.24 1.18 ...
     0.95 0.96 0.82 0.92 1.18 1.20 1.47 1.41 1.57 1.65]';
Xalpha = kron(speye(5),ones(4,1));
Xgamma = kron(speye(10),ones(2,1));
f = @(theta) -sfran2_loglike(theta(1),theta(2),theta(3), ...
    theta(4),y,Xalpha,Xgamma);
yhat = mean(reshape(y,4,5));
theta0 = [mean(y) log(var(yhat)) log(var(y)/3) log(var(y)/3)];
thetahat = fminsearch(f,theta0);
```

10.2.3   Gaussian Linear Mixed Models

It is also possible to combine fixed and random factors. This leads to the so-called **mixed models**. A general formulation for such models is given below. We first discuss an example.

Example 10.12 (Two-Factor Mixed ANOVA Model).   Suppose the experiment in Example 10.9 is modified in the following way. Each of the 30 test plots is subjected to three different treatments of herbicide: (1) the new herbicide, (2) a standard herbicide, and (3) no herbicide. Specifically, each test plot is divided into three subplots, and the three treatments are assigned in a random (and uniform) way to the subplots. The crop yield is recorded for each of the subplots. For each of the 10 plots there are thus 9 measurements—3 for each treatment. There are now two factors to consider: herbicide and location. The first is a *fixed* factor, and the second is a *random* factor. Denoting by Y_{ijk} the kth crop yield at location j, with treatment i, we obtain the two-factor mixed ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \,,$$
$$i = 1, 2, 3, \quad j = 1, 2, \ldots, 10, \quad k = 1, 2, 3 \,,$$

where μ is a constant, $\alpha_1, \ldots, \alpha_3$ are the fixed incremental effects of the herbicide, and $\beta_1, \ldots, \beta_{10}$ are the random incremental effects due to location. As in the fixed ANOVA case, we impose the restriction $\sum_i \alpha_i = 0$. The random incremental effects are modeled via $\{\beta_j\} \sim_{\text{iid}} \mathsf{N}(0, \sigma_\beta^2)$. The measurement errors $\{\varepsilon_{ijk}\}$ are assumed to be independent of each other and of the $\{\beta_j\}$ and are all $\mathsf{N}(0, \sigma_\varepsilon^2)$ distributed, for some fixed σ_ε^2. Finally, for the terms $\{\gamma_{ij}\}$ there are two common model choices. The simplest one is to assume that $\{\gamma_{ij}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma_\gamma^2)$. However, this introduces a subtle problem regarding the interpretation of α_i as an "incremental effect" due to treatment i. To circumvent this difficulty one often imposes the restriction

$$\sum_i \gamma_{ij} = 0 \quad \text{for all } j \,.$$

The latter is called a **restricted** mixed ANOVA model, as opposed to the former **unrestricted** model.

Here we give a general formulation for the linear mixed models.

Definition 10.6. (Gaussian Linear Mixed Model).   Let $\mathbf{Y}$ be an $n \times 1$ vector of responses and then

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{U} + \boldsymbol{\varepsilon} \,, \tag{10.20}$$

(continued)

(continued)

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, X is an $n \times p$ design matrix for the fixed effects, and Z is an $n \times q$ design matrix for the random effects. In addition, $\mathbf{U}$ and $\boldsymbol{\varepsilon}$ are independent of each other, and

$$\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \Sigma_{\mathbf{U}}) , \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I) .$$

The covariance matrix of $\mathbf{Y}$ is $\Sigma = \sigma^2 I + Z \Sigma_{\mathbf{U}} Z^\top$. Hence, given the outcome $\mathbf{Y} = \mathbf{y}$, the log-likelihood function is

$$l(\boldsymbol{\beta}, \sigma^2, \Sigma_{\mathbf{U}}; \mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) . \quad (10.21)$$

Unless the dimension of the log-likelihood function is low, direct maximization could be time-consuming. Instead, dimension reduction techniques such as using the profile likelihood can be applied to speed up the estimation. Alternatively, the linear mixed model can be estimated using the Gibbs sampler, as the following example illustrates.

Example 10.13 (Gaussian Linear Mixed Model). In an experiment concerning the growth rate of rats, 30 different rats are weighed at five different points in time—8, 15, 22, 29, and 36 days since birth. Using Bayesian notation, let y_{ik} denote the weight of the ith rat at the kth measurement, and let x_{ik} denote the corresponding age of the rat. Then,

$$x_{i1} = 8 , \quad x_{i2} = 15 , \quad x_{i3} = 22 , \quad x_{i4} = 29 , \quad x_{i5} = 36$$

for $i = 1, \ldots, 30$. The data are taken from Gelfand et al. 1990, and the growth curves are depicted in Fig. 10.9.
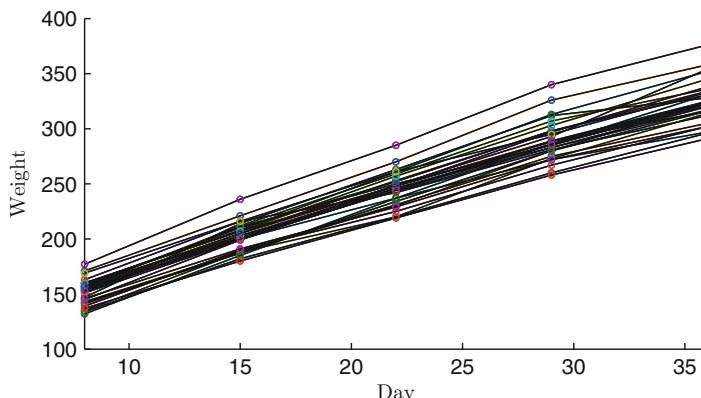
Fig. 10.9 Growth curves for 30 rats

In the model, we allow for individual-specific variation in the initial birth weight but assume the same growth rate:

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \alpha_i + \varepsilon_{ik} \, ,$$

where $\{\alpha_i\} \sim_{\text{iid}} \mathsf{N}(0, \sigma_\alpha^2)$ and $\{\varepsilon_{ik}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ are independent, $i = 1, \ldots, 30$, $k = 1, \ldots, 5$.

To write the model in the form in (10.20), let

$$\mathbf{y} = (y_{11}, \ldots, y_{15}, y_{21}, \ldots, y_{25}, \ldots, y_{30,1}, \ldots, y_{30,5})^\top,$$

and define $\boldsymbol{\varepsilon}$ similarly. Further, let $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{30})^\top$, $Z = I \otimes \mathbf{1}_5$, $X = \mathbf{1}_{30} \otimes (\mathbf{1}_5, \mathbf{x}_i)$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})^\top$, and let $\mathbf{1}_m$ be an $m \times 1$ vector of ones. Then,

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \, ,$$

where $\boldsymbol{\alpha} \sim \mathsf{N}(\mathbf{0}, \sigma_\alpha^2 I)$ and $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$ are independent. The actual outcomes of the experiment are reported in Table 10.5.

Table 10.5 Weight measurements of rats

Rat	Weights					Rat	Weights				
i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

To perform a Bayesian analysis, consider the following independent priors:

$$\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, 100\, I), \quad \sigma_\alpha^2 \sim \mathsf{InvGamma}(3, 100), \quad \sigma^2 \sim \mathsf{InvGamma}(3, 100) \, ,$$

where $\boldsymbol{\beta}_0 = (100, 10)^\top$. The degrees of freedom parameters for the inverse-gamma distributions are chosen to be small so that the prior variances are large. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ enter the likelihood additively, we sample them in one block to improve efficiency. Specifically, we consider the 3-block Gibbs sampler: (1) simulate from $f(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \sigma_\alpha^2)$, (2) simulate from $f(\sigma^2 \mid \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\alpha^2)$, and (3) simulate from $f(\sigma_\alpha^2 \mid \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$. Steps 2 and 3 are straightforward as the two
☞ 238 conditional distributions are both inverse-gamma (see Theorem 8.1):

$$(\sigma^2 \mid \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\alpha^2) \sim \mathsf{InvGamma}(78, \lambda) \ ,$$

$$(\sigma_\alpha^2 \mid \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) \sim \mathsf{InvGamma}(18, \lambda_\alpha) \ ,$$

where $\lambda = 100 + (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\alpha})^\top (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\alpha})/2$ and $\lambda_\alpha = 100 + \boldsymbol{\alpha}^\top \boldsymbol{\alpha}/2$. For Step 1, let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$. Then, the prior for $\boldsymbol{\gamma}$ is $\mathsf{N}(\boldsymbol{\gamma}_0, V_{\boldsymbol{\gamma}})$, where

$$\boldsymbol{\gamma}_0 = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \end{pmatrix}, \quad V_{\boldsymbol{\gamma}} = \begin{pmatrix} 100\,I & 0 \\ 0 & \sigma_\alpha^2 I \end{pmatrix} \ .$$

Note that the covariance matrix $V_{\boldsymbol{\gamma}}$ is in fact diagonal. In addition, the linear mixed model can be written as

$$\mathbf{y} = W\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \ ,$$

☞ 238 where $W = (X, Z)$. Hence, using Theorem 8.1, we have

$$(\boldsymbol{\gamma} \mid \mathbf{y}, \sigma^2, \sigma_\alpha^2) \sim \mathsf{N}(\widehat{\boldsymbol{\gamma}}, K_{\boldsymbol{\gamma}}^{-1}) \ ,$$

where

$$K_{\boldsymbol{\gamma}} = W^\top W/\sigma^2 + V_{\boldsymbol{\gamma}}^{-1} \ , \quad \widehat{\boldsymbol{\gamma}} = K_{\boldsymbol{\gamma}}^{-1}(V_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\gamma}_0 + W^\top \mathbf{y}/\sigma^2) \ .$$

It is important to realize that although $\boldsymbol{\gamma}$ is high-dimensional, sampling from its conditional distribution is quick if we use sparse matrix routines and avoid inverting
☞ 307 the large precision matrix; see Algorithm 10.1.

The following MATLAB code implements the Gibbs sampler discussed above to fit the experimental data.

```
%linmix.m
load 'rats.csv';
[d ni] = size(rats);
n = d*ni;
y = reshape(rats',n,1);
nloop = 11000;
burnin = 1000;
    % storage
store_beta = zeros(nloop-burnin,2);
```

```
store_alpha = zeros(nloop-burnin,d);
store_var = zeros(nloop-burnin,2);
    % priors
beta0 = [100 10]';
invVbeta = [1/100 1/100];
gamma0 = [beta0;sparse(d,1)];
nu_alpha = 3; lam_alpha = 100;
nu = 3; lam = 100;
    % initialize the Markov chain
sigma2 = 100;
sigma2_alpha = 100;
    % compute a few things before the loop
Z = kron(speye(d),ones(ni,1));
xi = [8 15 22 29 36]';
X = kron(ones(d,1),[ones(ni,1) xi]);
W = [X Z];
WW = W'*W;
Wy = W'*y;
newnu_alpha = d/2 + nu_alpha;
newnu = n/2 + nu;
for loop=1:nloop
        % sample alpha and beta
    invVgamma = sparse(1:d+2,1:d+2, ...
        [invVbeta 1/sigma2_alpha*ones(1,d)]);
    invDgamma = invVgamma + WW/sigma2;
    gammahat = invDgamma\(invVgamma*gamma0 + Wy/sigma2);
    gamma = gammahat + chol(invDgamma,'lower')'\randn(d+2,1);
    beta = gamma(1:2);
    alpha = gamma(3:end);
        % sample sigma2_alpha
    newlam_alpha = lam_alpha + sum(alpha.^2)/2;
    sigma2_alpha = 1/gamrand(newnu_alpha, newlam_alpha);
        % sample sigma2
    newlam = lam + sum((y-W*gamma).^2)/2;
    sigma2 = 1/gamrand(newnu, newlam);
        % storage
    if loop>burnin
        i = loop-burnin;
        store_beta(i,:) = beta';
        store_alpha(i,:) = alpha';
        store_var(i,:) = [sigma2 sigma2_alpha];
    end
end
betahat = mean(store_beta)';
alphahat = mean(store_alpha)';
varhat = mean(store_var)';
```

In Table 10.6 we report the posterior means, standard deviations, and quantiles for selected parameters. The results indicate that there is substantial variation in initial birth weight—the posterior mean of σ_α^2 is about three times the estimate corresponding to the measurement error σ^2. Inference on individual random effects

can also be easily carried out. For example, the birth weight of the 14th rat is estimated to be between 19.90 and 34.73 above average with probability of 80 %.

Table 10.6 Posterior means, standard deviations, and quantiles for selected parameters

Parameter	Post. mean	Post. std.	Post. 0.1 quantile	Post. 0.9 quantile
β_0	106.10	2.73	102.82	112.58
β_1	6.19	0.07	6.11	6.35
σ^2	63.97	8.16	54.38	85.90
σ_α^2	170.89	46.62	119.98	313.66
α_{14}	25.01	4.19	19.90	34.73

10.3   Problems

10.1. Calculate the lag-1, lag-2, and lag-3 autocorrelations of the residuals in Example 10.1.

10.2. Prove Equation (10.9); that is, show that under the MA(q) model, the autocovariances of the responses are given by

$$\text{Cov}(Y_t, Y_{t-j}) = \begin{cases} \sigma^2 \sum_{i=0}^{q-j} \psi_{i+j} \psi_i, & j = 0, \ldots, q , \\ 0, & j > q , \end{cases}$$

where $\psi_0 = 1$.

10.3. Consider again the MA(q) model:

$$Y_t = \varepsilon_t + \psi \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q} .$$

In this exercise we relax the standard assumption that $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{1-q} = 0$. Instead, we assume $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ for $t = 1 - q, \ldots, T$. Derive the likelihood function for the outcome $\mathbf{Y} = \mathbf{y}$, where $\mathbf{Y} = (Y_1, \ldots, Y_T)^\top$ and $\mathbf{y} = (y_t, \ldots, y_T)^\top$.

10.4. In many situations the data can be described by several competing models, and the question then is which model is "the best." Complex models tend to fit the data better, but they run the risk of overfitting. Hence, we want a measure that awards goodness of fit while penalizing model complexity. Two popular selection criteria that explicitly take this trade-off into account are the AIC and the BIC; see, for example, Bishop (2006) for a detailed introduction. Given a model defined by the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$, where $\boldsymbol{\theta}$ is a $p \times 1$ vector of model parameters and $\mathbf{y}$ is an $n \times 1$ vector of outcomes, the two information criteria are defined as follows:

$$\text{AIC} = 2l(\boldsymbol{\theta}; \mathbf{y}) - 2p ,$$

$$\text{BIC} = 2l(\boldsymbol{\theta}; \mathbf{y}) - p \ln n .$$

The only difference between the two information criteria is the penalty term: BIC tends to penalize complex model more heavily when $n \geqslant 8$. Given a set of competing models, the preferred model is the one with the maximum AIC/BIC value.

Use the two information criteria to compare the linear regression models in Examples 10.1 and 10.4.

10.5. Show that the log-likelihood function of the ARMA(p, q) model in (10.12) is given by (10.13). That is,

$$l(\boldsymbol{\varrho}, \boldsymbol{\psi}, \sigma^2; \mathbf{y}, \mathbf{y}_0) = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\varrho})^{\top}(HH^{\top})^{-1}(\mathbf{y} - X\boldsymbol{\varrho}) \ .$$

10.6. Consider the linear regression model with general covariance matrix $\sigma^2 \Sigma$:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \qquad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \Sigma) \ .$$

Suppose Σ is a symmetric invertible constant matrix. Show that the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}} = (X^{\top}\Sigma^{-1}X)^{-1}X^{\top}\Sigma^{-1}\mathbf{y}, \quad \widehat{\sigma^2} = \frac{1}{T}(\mathbf{y} - X\widehat{\boldsymbol{\beta}})^{\top}\Sigma^{-1}(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \ .$$

10.7. We revisit Example 10.7 on fitting the inflation data with an ARMA(1,1) model:

$$Y_t = \varrho_0 + \varrho_1 y_{t-1} + \varepsilon_t + \psi\varepsilon_{t-1} \ ,$$

where $\varepsilon_1, \ldots, \varepsilon_T \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$ and $\varepsilon_0 = 0$. Specifically, we consider a Bayesian analysis of the model using the following independent priors:

$$\boldsymbol{\varrho} \sim \mathsf{N}(\mathbf{0}, 10I), \quad \psi \sim \mathsf{U}(-1, 1), \quad \sigma^2 \sim \mathsf{InvGamma}(3, 1) \ ,$$

where $\boldsymbol{\varrho} = (\varrho_0, \varrho_1)^{\top}$.

(a) Derive the posterior conditional densities $f(\boldsymbol{\varrho} \,|\, \mathbf{y}, \psi, \sigma^2)$, $f(\psi \,|\, \mathbf{y}, \boldsymbol{\varrho}, \sigma^2)$, and $f(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\varrho}, \psi)$.
(b) Fit the model using the data set `USCPI.csv`.
(c) Compute the posterior means of $\boldsymbol{\varrho}$, ψ, and σ^2, and compare them with their corresponding maximum likelihood estimates.

10.8. What is the "factor effects" representation of the one-factor random effects model in (10.15)?

10.9. Write the two-factor mixed ANOVA model in Example 10.12 as a Gaussian model. That is, arrange $\{Y_{ijk}\}$ as $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\top} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$, and determine $\boldsymbol{\mu}$ and Σ.

10.10. Determine the Gaussian graphical model for each of the following situations:

(a) For the random variables $Y_1, \ldots, Y_6$, Y_2 depends only on Y_1, and Y_t depends only on Y_{t-1} and Y_{t-2}, $t = 3, \ldots, 6$.
(b) A one-factor ANOVA model with $d = 3$, $n_1 = 2$, $n_2 = 3$, and $n_3 = 4$.
(c) A two-factor ANOVA model with $d_1 = 2$, $d_2 = 3$, and $n_{ij} = 1, i = 1, 2$, $j = 1, 2, 3$.

10.11. We wish to design a Gibbs sampler for estimating the one-factor random effects model. To that end, consider the following independent priors:

$$\mu \sim \mathsf{N}(\mu_0, V_\mu), \quad \sigma_\mu^2 \sim \mathsf{InvGamma}(\alpha_\mu, \lambda_\mu), \quad \sigma^2 \sim \mathsf{InvGamma}(\alpha, \lambda) .$$

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$. Derive the following conditional distributions:

(a) $f(\boldsymbol{\mu} \mid \mathbf{y}, \mu, \sigma_\mu^2, \sigma^2) = \prod_{i=1}^d f(\mu_i \mid \mathbf{y}, \mu, \sigma_\mu^2, \sigma^2)$.
(b) $f(\mu \mid \mathbf{y}, \boldsymbol{\mu}, \sigma_\mu^2, \sigma^2) = f(\mu \mid \boldsymbol{\mu}, \sigma_\mu^2)$.
(c) $f(\sigma_\mu^2 \mid \mathbf{y}, \boldsymbol{\mu}, \mu, \sigma^2) = f(\sigma_\mu^2 \mid \boldsymbol{\mu}, \mu)$.
(d) $f(\sigma^2 \mid \mathbf{y}, \boldsymbol{\mu}, \mu, \sigma_\mu^2) = f(\sigma^2 \mid \mathbf{y}, \boldsymbol{\mu})$.

10.12. Implement the Gibbs sampler developed in Problem 10.11 for the one-factor random effects model to fit the crop yield data in Example 10.10. Use the following independent priors:

$$\mu \sim \mathsf{N}(0, 100), \quad \sigma_\mu^2 \sim \mathsf{InvGamma}(3, 1), \quad \sigma^2 \sim \mathsf{InvGamma}(3, 1) .$$

Estimate the posterior means $\mathbb{E}(\mu \mid \mathbf{y})$, $\mathbb{E}(\sigma_\mu^2 \mid \mathbf{y})$, and $\mathbb{E}(\sigma^2 \mid \mathbf{y})$. What is the posterior probability that $\sigma_\mu^2 > 5\sigma^2$?

10.13. Show that the two-factor random effects model in (10.18) is a special case of the linear mixed model by writing the former in the form (10.20).

Chapter 11
State Space Models

In this chapter we discuss versatile generalizations of the basic time series models in Sect. 10.1, collectively known under the name **state space models**. These models not only can capture the *serial dependence* of the observations (i.e., the dependence across time), but also can describe the *persistence* and *volatility* of the measurements. That is, they can model continued periods of high or low measurements and time-varying amounts of random fluctuation. In contrast, the AR(p) model, for example, cannot capture these features, as the model parameters do not depend on time. Throughout this chapter we shall use *Bayesian* notation when specifying (conditional) densities, even when working in a non-Bayesian setting.

A state space model typically consists of two modeling levels: in the first level, observations are related to the latent or unobserved variables called **states** according to the **observation** or **measurement equation**. In the second level, the evolution of the states is modeled via the **state** or **transition equation.**

> **Definition 11.1. (State Space Model).** In a **state space model**, the **observations** $y_t, t = 1, 2, \ldots$ are drawn from a conditional pdf $f(y_t \mid x_t, y_{t-1}, \ldots, y_1, \boldsymbol{\theta})$, where x_t is the hidden **state** at time t. The states $x_t, t = 1, 2, \ldots$ evolve according to a Markov chain with transition density $f(x_t \mid x_{t-1}, \boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ denotes the vector of model parameters.

Typically one assumes that each observation y_t only depends on the latent state x_t and not on previous states or observations. In that case the state space model can be viewed as a *hidden Markov model*; see Problem 8.14. Note that the $\{x_t\}$ and $\{y_t\}$ may be vector-valued.

Example 11.1 (Kalman Filter). State space models originate from the analysis of *dynamical systems*. One of the most fundamental examples is the **linear Gaussian discrete-time state space model**:

$$\mathbf{x}_t = A\,\mathbf{x}_{t-1} + \boldsymbol{\delta}_t$$
$$\mathbf{y}_t = B\,\mathbf{x}_t + \boldsymbol{\varepsilon}_t\,, \qquad t = 1, 2, \ldots, \tag{11.1}$$

where $\mathbf{x}_t$ is an n-dimensional (hidden) **state** vector and $\mathbf{y}_t$ an m-dimensional **output** vector. A and B are fixed matrices, and $\boldsymbol{\delta}_t$ and $\boldsymbol{\varepsilon}_t$ are zero-mean normal random vectors with covariance matrices D and E, respectively. All $\{\boldsymbol{\delta}_t\}$ and $\{\boldsymbol{\varepsilon}_t\}$ are independent. The initial state $\mathbf{x}_0$ is assumed to be $\mathsf{N}(\boldsymbol{\mu}_0, \Sigma_0)$ distributed.

Define $\mathbf{y}_{1:t} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_t^\top)^\top$. Assuming the model parameters are known, two main objectives are to obtain the

- **Predictive distribution**, that is, the conditional distribution of $\mathbf{x}_t$ given $\mathbf{y}_{1:t-1}$ (the observations *before* time t).
- **Filtering distribution**, that is, the conditional distribution of $\mathbf{x}_t$ given $\mathbf{y}_{1:t}$ (the observations *up to* time t).

Since we are dealing only with affine transformations of Gaussian vectors, we have ☞ 85 by Theorem 3.6 that $(\mathbf{x}_t \mid \mathbf{y}_{1:t}) \sim \mathsf{N}(\boldsymbol{\mu}_t, \Sigma_t)$ for some mean vector $\boldsymbol{\mu}_t$ and covariance matrix Σ_t. Similarly, $(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}) \sim \mathsf{N}(\widetilde{\boldsymbol{\mu}}_t, \widetilde{\Sigma}_t)$ for some mean vector $\widetilde{\boldsymbol{\mu}}_t$ and covariance matrix $\widetilde{\Sigma}_t$. These mean vectors and covariance matrices can be computed sequentially. First, since $\mathbf{x}_t = A\,\mathbf{x}_{t-1} + \boldsymbol{\delta}_t$ and $(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \mathsf{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$, we have

$$(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}) \sim \mathsf{N}(A\,\boldsymbol{\mu}_{t-1},\ A\Sigma_{t-1}A^\top + D)\,.$$

Thus, the updating formulas for the predictive distribution are

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_t &= A\,\boldsymbol{\mu}_{t-1}\,, \\ \widetilde{\Sigma}_t &= A\Sigma_{t-1}A^\top + D\,. \end{aligned} \tag{11.2}$$

Next, we determine the joint pdf of $\mathbf{x}_t$ and $\mathbf{y}_t$, given $\mathbf{y}_{1:t-1}$. Decomposing $\widetilde{\Sigma}_t$ and E ☞ 83 as $\widetilde{\Sigma}_t = RR^\top$ and $E = QQ^\top$, respectively, we can write (using Definition 3.10 of the multivariate normal distribution)

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{pmatrix} \mathbf{y}_{1:t-1} = \begin{pmatrix} \widetilde{\boldsymbol{\mu}}_t \\ B\widetilde{\boldsymbol{\mu}}_t \end{pmatrix} + \begin{pmatrix} R & 0 \\ BR & Q \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

where, conditional on $\mathbf{y}_{1:t-1}$, $\mathbf{u}$ and $\mathbf{v}$ are independent standard normal random vectors. The corresponding covariance matrix is

$$\begin{pmatrix} R & 0 \\ BR & Q \end{pmatrix} \begin{pmatrix} R^\top & R^\top B^\top \\ 0 & Q^\top \end{pmatrix} = \begin{pmatrix} RR^\top & RR^\top B^\top \\ BRR^\top & BRR^\top B^\top + QQ^\top \end{pmatrix},$$

so that we have

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{pmatrix} \mathbf{y}_{1:t-1} \sim \mathsf{N}\left(\begin{pmatrix} \widetilde{\boldsymbol{\mu}}_t \\ B\widetilde{\boldsymbol{\mu}}_t \end{pmatrix}, \begin{pmatrix} \widetilde{\Sigma}_t & \widetilde{\Sigma}_t B^\top \\ B\widetilde{\Sigma}_t & B\widetilde{\Sigma}_t B^\top + E \end{pmatrix} \right). \tag{11.3}$$

A direct application of Theorem 3.8 yields that $\mathbf{x}_t$ given $\mathbf{y}_{1:t}$ has a $\mathsf{N}(\boldsymbol{\mu}_t, \Sigma_t)$ ☞ 86
distribution with

$$\begin{aligned}
\boldsymbol{\mu}_t &= \widetilde{\boldsymbol{\mu}}_t + \widetilde{\Sigma}_t B^\top (B \widetilde{\Sigma}_t B^\top + E)^{-1}(\mathbf{y}_t - B\widetilde{\boldsymbol{\mu}}_t) \,, \\
\Sigma_t &= \widetilde{\Sigma}_t - \widetilde{\Sigma}_t B^\top (B \widetilde{\Sigma}_t B^\top + E)^{-1} B \widetilde{\Sigma}_t \,.
\end{aligned} \tag{11.4}$$

We leave the details as an exercise; see Problem 11.1. Updating formulas (11.2)
and (11.4) form the (discrete-time) **Kalman filter**. Starting with some known $\boldsymbol{\mu}_0$
and Σ_0, one determines $\widetilde{\boldsymbol{\mu}}_1$ and $\widetilde{\Sigma}_1$, then $\boldsymbol{\mu}_1$ and Σ_1, and so on. Notice that $\widetilde{\Sigma}_t$ and
Σ_t do not depend on the observations $y_1, y_2, \ldots$ and can therefore be determined
off-line.

In the remainder of this chapter we will discuss various popular state space
models that fall within the framework defined above. From the definition it is
obvious that state space models are high-dimensional, often with more latent
variables and parameters than observations. Instead of using generalizations of the
Kalman filter, we will discuss the precision-based approach to estimating state space
models, where the derivation is relatively simple and transparent.

11.1 Unobserved Components Model

An important state space model is the **unobserved components model**. In the first
level, the (real-valued) observable y_t at time t is modeled to depend on the state or
unobserved component τ_t as follows:

$$y_t = \tau_t + \varepsilon_t \,, \tag{11.5}$$

where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$. That is, the observable y_t is modeled as the sum of the
unobserved component τ_t and the error term ε_t. As we shall see shortly, this is a
popular specification for modeling the evolution of univariate time series such as
inflation rate. For example, in the context of inflation modeling, the unobserved
component τ_t can be interpreted as the stochastic trend or underlying inflation.

Since for every y_t we have an associated latent variable τ_t, there are more
latent variables and parameters (i.e., $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$ and σ^2) than the number
of observations. As such, if we have the measurement equation only, the maxi-
mum likelihood estimator for $(\boldsymbol{\tau}, \sigma^2)$ is not defined. Specifically, the likelihood
function is unbounded in $(\boldsymbol{\tau}, \sigma^2)$, and therefore the maximum does not exist (see
Problem 11.3).

The fundamental problem is that if the unobserved components are unrestricted,
then we have the extreme situation where we can fit the data perfectly (e.g., by
choosing $\tau_t = y_t$). One way to get around this problem is to impose some structure
on the model to make estimation feasible. Since we are dealing with time series data,
it seems reasonable to assume that the unobserved component evolves gradually

over time. In the inflation example, consecutive inflation trends are likely to be
"close." More precisely, consider the following **random walk** specification

$$\tau_t = \tau_{t-1} + u_t , \tag{11.6}$$

for $t = 2, \ldots, T$, where $\{u_t\} \sim_{\text{iid}} N(0, \omega^2)$. That is, the conditional distribution of
τ_t given τ_{t-1} and ω^2 is $N(\tau_{t-1}, \omega^2)$: the current state τ_t centers around the previous
one τ_{t-1}, while ω^2 controls how close the two terms are on average.

The smoothness parameter ω^2 can either be fixed in advance to some "reasonable
value" or treated as a parameter to be estimated from the data. Should it be fixed as
a constant, its choice should reflect the desired smoothness of the evolution of the
states: large values for ω^2 allow τ_t to evolve quickly, whereas for small values the
transition of τ_t becomes more gradual. In a Bayesian framework one often assumes
a hierarchical prior distribution for ω^2 that reflects the desired smoothness of the
transition equation.

Note that (11.6) does not explicitly provide a distribution for τ_1. To complete
the model specification, one typically assumes that the process is initialized with
$\tau_1 \sim N(\tau_0, \omega_0^2)$ for some known constants τ_0 and ω_0^2. This is referred to as the **initial
condition**.

We summarize the unobserved components model as follows:

Definition 11.2. (Unobserved Components Model). In the **unobserved
components model**, the measurement equation is given by

$$y_t = \tau_t + \varepsilon_t ,$$

where $\{\varepsilon_t\} \sim_{\text{iid}} N(0, \sigma^2)$. The states, in turn, are initialized with $\tau_1 \sim$
$N(\tau_0, \omega_0^2)$ for some known constants τ_0 and ω_0^2 and evolve according to the
transition equation

$$\tau_t = \tau_{t-1} + u_t$$

for $t = 2, \ldots, T$, where $\{u_t\} \sim_{\text{iid}} N(0, \omega^2)$.

It is obvious that the unobserved components model falls within the family of state
space models. In fact, in the notation of Definition 11.1, the state x_t in this case is the
univariate unobserved component τ_t and $\boldsymbol{\theta} = (\sigma^2, \omega^2)$. The conditional distribution
of y_t given τ_t is $N(\tau_t, \sigma^2)$, whereas the transition density $f(\tau_t \mid \tau_{t-1}, \boldsymbol{\theta})$ corresponds
to the pdf of the $N(\tau_{t-1}, \omega^2)$ distribution. Furthermore, since both the measurement
and transition equations are linear in the states with Gaussian errors, the unobserved
components model is an example of the linear Gaussian state space model discussed
in Example 11.1.

11.1.1   Classical Estimation

Let $\mathbf{y} = (y_1, \ldots, y_T)^\top$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$ be the vector of observations and latent variables, respectively. Throughout this section we fix ω^2, and the only *parameter* in the model is σ^2. To obtain the maximum likelihood estimate of σ^2, which we denote as $\widehat{\sigma^2}$, in principle we can maximize the likelihood function

$$L(\sigma^2; \mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) f(\boldsymbol{\tau} \mid \omega^2) \, \mathrm{d}\boldsymbol{\tau} \qquad (11.7)$$

with respect to σ^2, where the densities $f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2)$ and $f(\boldsymbol{\tau} \mid \omega^2)$ follow from (11.5) and (11.6), respectively (their exact expressions are given below). In practice, however, evaluating the above integral directly is often time-consuming as it involves high-dimensional integration (see Problem 11.5 for an alternative method).

Instead, we will obtain $\widehat{\sigma^2}$ using the EM algorithm introduced in Chap. 6.6. To this end we first write the system (11.5)–(11.6) in matrix form and derive explicit expressions for $\ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2)$ and $\ln f(\boldsymbol{\tau} \mid \omega^2)$.

Defining $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top$, we can rewrite (11.5) as

$$\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I), \qquad (11.8)$$

where $\mathbf{0}$ is a $T \times 1$ column of zeros and I is the $T \times T$ identity matrix. From (11.8) we see that

$$\ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\tau})^\top (\mathbf{y} - \boldsymbol{\tau}). \qquad (11.9)$$

Next, we derive an expression for $\ln f(\boldsymbol{\tau} \mid \omega^2)$. For simplicity, we assume $\tau_0 = 0$; the general case follows similarly (see Problem 11.7). Note that we can rewrite the transition equation (11.6) as

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}}_{H} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \vdots \\ \tau_T \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_T \end{pmatrix}, \qquad (11.10)$$

i.e., $H\boldsymbol{\tau} = \mathbf{u}$, $\mathbf{u} \sim \mathsf{N}(\mathbf{0}, \Omega)$, where $\Omega = \mathrm{diag}(\omega_0^2, \omega^2, \ldots, \omega^2)$ is a diagonal matrix. Noting that $|H| \stackrel{\mathrm{def}}{=} |\det(H)| = 1$ and hence H is invertible, we have

$$\boldsymbol{\tau} = H^{-1}\mathbf{u} \sim \mathsf{N}(\mathbf{0}, (H^\top \Omega^{-1} H)^{-1}),$$

where $\Omega^{-1} = \text{diag}(\omega_0^{-2}, \omega^{-2}, \ldots, \omega^{-2})$ is again a diagonal matrix. It follows that

$$\ln f(\boldsymbol{\tau} \mid \omega^2) = -\frac{1}{2} \ln((2\pi)^T |(H^\top \Omega^{-1} H)^{-1}|) - \frac{1}{2}\boldsymbol{\tau}^\top (H^\top \Omega^{-1} H)\boldsymbol{\tau}$$

$$= -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \ln \omega_0^2 - \frac{T-1}{2} \ln \omega^2 - \frac{1}{2}\boldsymbol{\tau}^\top (H^\top \Omega^{-1} H)\boldsymbol{\tau} \ . \qquad (11.11)$$

To implement the *E-step*, we need to derive the conditional density of the states given the data

$$g_i(\boldsymbol{\tau}) = f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma_{i-1}^2, \omega^2) \ ,$$

where σ_{i-1}^2 is the current value for σ^2 in iteration i. We first show that $(\boldsymbol{\tau} \mid \mathbf{y}, \sigma_{i-1}^2, \omega^2)$ has a multivariate normal density of dimension T. Then we discuss how one can evaluate this typically high-dimensional density efficiently.

Using (11.9) and (11.11), while ignoring constant terms not involving $\boldsymbol{\tau}$, we have

$$\ln f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma_{i-1}^2, \omega^2) = \ln f(\mathbf{y}, \boldsymbol{\tau} \mid \sigma_{i-1}^2, \omega^2) + \text{const}$$

$$= \ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma_{i-1}^2) + \ln f(\boldsymbol{\tau} \mid \omega^2) + \text{const}$$

$$= -\frac{1}{2}\left(\frac{(\mathbf{y}-\boldsymbol{\tau})^\top (\mathbf{y}-\boldsymbol{\tau})}{\sigma_{i-1}^2} + \boldsymbol{\tau}^\top (H^\top \Omega^{-1} H)\boldsymbol{\tau} \right) + \text{const}$$

$$= -\frac{1}{2}\left(\boldsymbol{\tau}^\top K_i\, \boldsymbol{\tau} - \frac{2}{\sigma_{i-1}^2}\mathbf{y}^\top \boldsymbol{\tau} \right) + \text{const} \ ,$$

where $K_i = H^\top \Omega^{-1} H + \sigma_{i-1}^{-2} I$. Note that the expression above defines the pdf of a normal distribution, and we only need to determine the mean vector and the covariance matrix. By *completing the squares* as in Theorem 8.1, we see that

$$(\boldsymbol{\tau} \mid \mathbf{y}, \sigma_{i-1}^2, \omega^2) \sim \mathsf{N}(\widehat{\boldsymbol{\tau}}_i,\ K_i^{-1}) \ , \qquad (11.12)$$

where $\widehat{\boldsymbol{\tau}}_i = \sigma_{i-1}^{-2} K_i^{-1}\mathbf{y}$.

Next, we compute the expectation

$$Q_i(\sigma^2) = \mathbb{E}_{g_i} \ln f(\mathbf{y}, \boldsymbol{\tau} \mid \sigma^2, \omega^2) \ .$$

To simplify the computation, we ignore all the terms not involving σ^2, as they will eventually drop out when we maximize $Q_i(\sigma^2)$ with respect to σ^2. Hence, we have, from (11.9),

$$Q_i(\sigma^2) = \mathbb{E}_{g_i} \ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) + \text{const}$$

$$= -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_{g_i}(\mathbf{y}-\boldsymbol{\tau})^\top (\mathbf{y}-\boldsymbol{\tau}) + \text{const}$$

$$= -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\text{tr}(K_i^{-1}) + (\mathbf{y}-\widehat{\boldsymbol{\tau}}_i)^\top (\mathbf{y}-\widehat{\boldsymbol{\tau}}_i) \right] + \text{const} \ .$$

Note that we used the fact that for any random vector $\mathbf{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, we have $\mathbb{E}(\mathbf{x}^\top\mathbf{x}) = \text{tr}(\Sigma) + \boldsymbol{\mu}^\top\boldsymbol{\mu}$ (see Problem 11.2).

Finally, to implement the *M-step*, we differentiate $Q_i(\sigma^2)$ with respect to σ^2 and solve for the maximizer:

$$\sigma_i^2 = \underset{\sigma^2}{\text{argmax}}\, Q_i(\sigma^2) = \frac{1}{T}\left[\text{tr}(K_i^{-1}) + (\mathbf{y} - \widehat{\boldsymbol{\tau}}_i)^\top(\mathbf{y} - \widehat{\boldsymbol{\tau}}_i)\right] . \qquad (11.13)$$

We summarize the EM algorithm as follows: given a starting value σ_0^2, iterate the following steps until convergence:

- **E-Step.** Given the current value σ_{i-1}^2, compute

$$K_i = H^\top\Omega^{-1}H + \sigma_{i-1}^{-2}I \qquad \text{and} \qquad \widehat{\boldsymbol{\tau}}_i = K_i^{-1}\mathbf{y}/\sigma_{i-1}^2 .$$

- **M-Step.** Given K_i and $\widehat{\boldsymbol{\tau}}_i$ from the E-step, update the value for σ^2 using (11.13).

Although the estimation procedures presented above are relatively straightforward, one thing to notice is that the computations involve various large matrices. For example, K_i^{-1} is a full $T \times T$ matrix. In typical applications the sample size T could be as large as several hundred or a few thousand, and computing the inverse K_i^{-1} is very time-consuming. However, note that the matrix K_i, as well as H and Ω^{-1}, is *sparse*, and computations involving sparse matrices (multiplication, Cholesky decomposition, etc.) are generally very fast. See also Appendix A for some useful MATLAB built-in routines for handling sparse matrices.

Also, for computing $\widehat{\boldsymbol{\tau}}_i = K_i^{-1}\mathbf{y}/\sigma_{i-1}^2$, one need not obtain the inverse K_i^{-1}, which is a time-consuming matrix operation. Instead, we solve the linear system $K_i\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, the solution of which is $K_i^{-1}\mathbf{y}$. In contrast to inverting large matrices, the latter operation can be done much more quickly and accurately.

Finally, to compute $\text{tr}(K_i^{-1})$ in (11.13) without obtaining the inverse K_i^{-1}, we use the following result:

$$\text{tr}(K_i^{-1}) = \sum_{j=1}^{T}\lambda_j^{-1} ,$$

where $\lambda_1, \ldots, \lambda_T$ are the eigenvalues of the sparse matrix K_i.

At the end of the EM iterations, we obtain the estimate $\widehat{\sigma^2}$. However, it is typically not the quantity of interest in the analysis—what we are really after is the expected value of the underlying inflation, $\mathbb{E}(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2, \omega^2)$. We can estimate this quantity using the "plug-in" estimate $\mathbb{E}(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2 = \widehat{\sigma^2}, \omega^2)$, which is $\widehat{K}^{-1}\mathbf{y}/\widehat{\sigma^2}$, where $\widehat{K}$ is the K_i matrix evaluated at the final iteration of the EM algorithm.

We illustrate the EM algorithm using the following empirical example that involves fitting the US inflation with the unobserved components model.

Example 11.2 (Modeling Inflation with Unobserved Components Model). In
☞ 300 Example 10.6 we first modeled the US inflation data with an integrated MA(1)
model. Later we continued our analysis with a more general ARMA(1,1) model in
☞ 304 Example 10.7. In this empirical example, we consider the unobserved components
model for the same data. The unobserved components model may be viewed as a
convenient way to allow for a stochastic trend (see Problem 11.4). As such, it is
highly flexible and is capable of modeling a variety of features. In addition, using
the state space framework makes it easy to consider further extensions with richer
dynamics (e.g., see Problem 11.8).

Recall that the quarterly inflation rate is computed from the consumer price index
(CPI). Specifically, given z_t, the CPI at time t, we compute the (annualized) inflation
rate as $y_t = 400 \ln(z_t / z_{t-1})$.

In what follows, we fit the unobserved components model (11.5) and (11.6) with
the US CPI inflation data. In order to proceed, we first need to set the values for ω_0^2
and ω^2. Recall that ω_0^2 is the variance of the initial condition (i.e., $\tau_1 \sim N(0, \omega_0^2)$).
We set $\omega_0^2 = 9$. What this means is that the initial τ_1 is between -6 and 6 with a
probability approximately equal to 95 %. As for the smoothness parameter ω^2, we
consider two cases: $\omega^2 = 1^2$ and $\omega^2 = 0.5^2$. The value for ω^2 reflects the desired
smoothness of the transition for τ_t. For example, if $\omega^2 = 0.5^2$, then with high
probability the difference between consecutive unobserved components, $\tau_t - \tau_{t-1}$,
is between -1 and 1.

Recall that the estimation consists of two steps. First, given the prefixed values
for ω_0^2 and ω^2, we iterate the E- and M-steps until the sequence of σ_t^2 converges.
Then, given the maximum likelihood estimate $\widehat{\sigma^2}$, we once again use the E-step
to obtain $\mathbb{E}(\boldsymbol{\tau} \,|\, \sigma^2 = \widehat{\sigma^2}, \omega^2)$. The following MATLAB script performs these two
tasks.

```
%UC_EM.m
load 'USCPI.csv';
y = USCPI;
T = length(y);
omega2_0 = 9;      % initial condition
omega = .5^2;      % fix omega
H = speye(T) - sparse(2:T,1:(T-1),ones(1,T-1),T,T);
invOmega = sparse(1:T,1:T,[1/omega2_0 ...
    1/omega*ones(1,T-1)],T,T);
HinvOmegaH = H'*invOmega*H;
sigma2t = var(y);  % initial guess
err = 1;
while err> 10^(-4)
        % E-step
    Kt = HinvOmegaH + speye(T)/sigma2t;
    taut = Kt\(y/sigma2t);
        % M-step
    lam =  eig(Kt);
```

```
    newsigma2t = (sum(1./lam) + (y-taut)'*(y-taut))/T;
        % update
    err = abs(sigma2t-newsigma2t);
    sigma2t = newsigma2t;
end
Kt = HinvOmegaH + speye(T)/sigma2t;
taut = Kt\(y/sigma2t);
```

We used the above code to obtain the maximum likelihood estimates for σ^2 with $\omega^2 = 1^2$ and $\omega^2 = 0.5^2$. The plug-in estimates for τ are plotted in Fig. 11.1. It can be seen that both curves fit the data reasonably well, without fitting the observed series too closely (otherwise we might run into overfitting problems). In particular, both seem to be able to capture the high inflation periods in the 1970s and 1980s, whereas the estimated trend remains low and stable since the 1990s until the last credit crisis. But as expected, when ω^2 is larger, the estimated τ fit the data better. But we emphasize that if one sets ω^2 to be too large, one might run into overfitting problems.
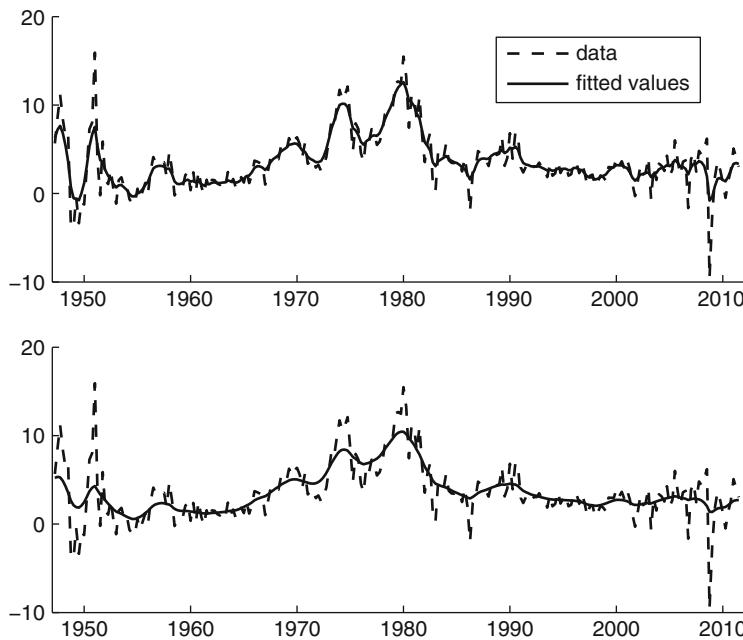
Fig. 11.1 Fitted values for τ under the unobserved components model with $\omega^2 = 1^2$ (*top panel*) and $\omega^2 = 0.5^2$ (*bottom panel*)

11.1.2   Bayesian Estimation

The unobserved components model, and state space models in general, may be viewed as a Bayesian hierarchical model, where the measurement equation provides the likelihood function and the transition equation specifies a prior for the states. For the remaining parameters, namely, σ^2 and ω^2 (note that ω^2 can be estimated if we specify a proper prior for ω^2), we assume the independent priors:

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_{\sigma^2}, \lambda_{\sigma^2}), \quad \omega^2 \sim \mathsf{InvGamma}(\alpha_{\omega^2}, \lambda_{\omega^2}) , \tag{11.14}$$

where $\alpha_{\sigma^2}, \lambda_{\sigma^2}, \alpha_{\omega^2}$, are λ_{ω^2} constants specified by the user. Typically we set the shape parameters α_{σ^2} and α_{ω^2} to be some small numbers, so that the priors are relatively non-informative. We then choose the scale parameters λ_{σ^2} and λ_{ω^2} such that the prior means for σ^2 and ω^2 have the desired values.

Given the measurement and state equations (11.5) and (11.6), as well as the prior for σ^2 and ω^2 in (11.14), we have the following joint posterior density:

$$f(\boldsymbol{\tau}, \sigma^2, \omega^2 \,|\, \mathbf{y}) \propto f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) f(\boldsymbol{\tau} \,|\, \omega^2) f(\sigma^2) f(\omega^2) , \tag{11.15}$$

where $f(\sigma^2)$ and $f(\omega^2)$ are the inverse-gamma priors. We can then obtain posterior draws via the following two-step Gibbs sampler: alternatively draw from $f(\boldsymbol{\tau} \,|\, \mathbf{y}, \sigma^2, \omega^2)$ and $f(\sigma^2, \omega^2 \,|\, \mathbf{y}, \boldsymbol{\tau})$.

We first show that $f(\boldsymbol{\tau} \,|\, \mathbf{y}, \sigma^2, \omega^2)$ is a normal density and then discuss how one can sample from it efficiently. To this end, note that from (11.15) we have

$$\begin{aligned}
\ln f(\boldsymbol{\tau} \,|\, \mathbf{y}, \sigma^2, \omega^2) &= \ln f(\boldsymbol{\tau}, \sigma^2, \omega^2 \,|\, \mathbf{y}) + \text{const} \\
&= \ln f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) + \ln f(\boldsymbol{\tau} \,|\, \omega^2) + \text{const} \\
&= -\frac{1}{2}\left(\frac{(\mathbf{y} - \boldsymbol{\tau})^\top (\mathbf{y} - \boldsymbol{\tau})}{\sigma^2} + \boldsymbol{\tau}^\top (H^\top \Omega^{-1} H) \boldsymbol{\tau} \right) + \text{const} \\
&= -\frac{1}{2}\left(\boldsymbol{\tau}^\top K \boldsymbol{\tau} - \frac{2}{\sigma^2} \mathbf{y}^\top \boldsymbol{\tau} \right) + \text{const}.
\end{aligned}$$

It follows, similar to the derivation of (11.12), that

$$(\boldsymbol{\tau} \,|\, \mathbf{y}, \sigma^2, \omega^2) \sim \mathsf{N}(\widehat{\boldsymbol{\tau}}, K^{-1}) ,$$

where $K = H^\top \Omega^{-1} H + \sigma^{-2} I$ and $\widehat{\boldsymbol{\tau}} = \sigma^{-2} K^{-1} \mathbf{y}$.

Since the covariance matrix K^{-1} is a full matrix and is typically of very high dimension, drawing $\mathsf{N}(\widehat{\boldsymbol{\tau}}, K^{-1})$ the usual way (i.e., via Algorithm 3.3) is time-consuming. Instead, we exploit the special structure of the *precision* matrix K, namely, that it is sparse (see the discussion on page 329). As such, a Cholesky decomposition of the precision matrix $K = CC^\top$ can be obtained quickly. Then,

we can use Algorithm 10.1 to quickly sample from $\mathsf{N}(\widehat{\boldsymbol{\tau}}, K^{-1})$. Specifically, if we
let $\mathbf{x} = (C^{\top})^{-1}\mathbf{z}$, where $\mathbf{z} \sim \mathsf{N}(\mathbf{0}, I)$, then $\mathbf{x} \sim \mathsf{N}(\mathbf{0}, K^{-1})$. Recall that one can
obtain $\widehat{\boldsymbol{\tau}}$ efficiently by solving $K\widehat{\boldsymbol{\tau}} = \sigma^{-2}\mathbf{y}$. Finally, $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}} + (C^{\top})^{-1}\mathbf{z}$ has the
desired distribution.

Next, we derive the conditional density $f(\sigma^2, \omega^2 \,|\, \mathbf{y}, \boldsymbol{\tau})$. From (11.15) we have

$$f(\sigma^2, \omega^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) \propto f(\boldsymbol{\tau}, \sigma^2, \omega^2 \,|\, \mathbf{y})$$
$$\propto f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) f(\sigma^2) \times f(\boldsymbol{\tau} \,|\, \omega^2) f(\omega^2) .$$

In other words, σ^2 and ω^2 are conditionally independent given $\mathbf{y}$ and $\boldsymbol{\tau}$, with

$$f(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) \propto f(\mathbf{y} \,|\, \boldsymbol{\tau}, \sigma^2) f(\sigma^2) \qquad \text{and} \qquad f(\omega^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) \propto f(\boldsymbol{\tau} \,|\, \omega^2) f(\omega^2) .$$

In fact, one can show that both conditional densities are inverse-gamma densities.
Namely, by (11.9) and the prior pdf of σ^2, we have (up to a constant)

$$\ln f(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) = \frac{T}{2} \ln(\frac{1}{\sigma^2}) - \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\tau})^{\top}(\mathbf{y} - \boldsymbol{\tau}) + (1 + \alpha_{\sigma^2}) \ln(\frac{1}{\sigma^2}) - \frac{\lambda_{\sigma^2}}{\sigma^2} ,$$

which shows that

$$(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) \sim \mathsf{InvGamma}\left(\alpha_{\sigma^2} + \frac{T}{2}, \ \lambda_{\sigma^2} + \frac{1}{2}(\mathbf{y} - \boldsymbol{\tau})^{\top}(\mathbf{y} - \boldsymbol{\tau})\right). \qquad (11.16)$$

Using a similar reasoning, we find

$$(\omega^2 \,|\, \mathbf{y}, \boldsymbol{\tau}) \sim \mathsf{InvGamma}\left(\alpha_{\omega^2} + \frac{T-1}{2}, \ \lambda_{\omega^2} + \frac{1}{2}\sum_{t=2}^{T}(\tau_t - \tau_{t-1})^2\right). \qquad (11.17)$$

11.2   Time-Varying Parameter Model

The unobserved components model discussed in the last section may be viewed as
a linear regression model with only an intercept, where the intercept is allowed to
change over time. More generally, one can consider linear regression models where
all the regression coefficients are time-varying. As discussed in the introduction,
this is motivated by the empirical findings that typical macroeconomic and financial
variables exhibit time-varying persistence and dynamics. In this section we discuss a
particular type of time-varying parameter models, called **time-varying parameter
autoregressive models**. Consider again the autoregressive model introduced in
Definition 10.1. Instead of assuming constant autoregressive coefficients, we allow
them to evolve over time.

Definition 11.3. (Time-Varying Parameter Autoregressive Model). In the
pth-order time-varying parameter autoregressive model, or time-varying
parameter AR(p), the measurement equation is given by

$$y_t = \beta_{0t} + \beta_{1t} y_{t-1} + \cdots + \beta_{pt} y_{t-p} + \varepsilon_t \qquad (11.18)$$

for $t = 1, \ldots, T$, where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, and $y_0, \ldots, y_{1-p}$ are initial
observations. The autoregressive coefficients $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, \ldots, \beta_{pt})^\top$ in
turn evolve according to the following transition equation:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \mathbf{u}_t \ , \qquad (11.19)$$

for $t = 2, \ldots, T$, where $\{\mathbf{u}_t\} \sim_{\text{iid}} \mathsf{N}(\mathbf{0}, \Omega)$, and the transition equation is
initialized with $\boldsymbol{\beta}_1 \sim \mathsf{N}(\boldsymbol{\beta}_0, \Omega_0)$.

In the above definition, we treat the initial observations $y_0, \ldots, y_{1-p}$ as given, and
we do not model them separately. For T much greater than p this has little influence
on estimation and inference.

11.2.1   *Bayesian Estimation*

We begin by writing (11.18) in matrix notation:

$$y_t = \mathbf{x}_t^\top \boldsymbol{\beta}_t + \varepsilon_t \ ,$$

where $\mathbf{x}_t^\top = (1, y_{t-1}, \ldots, y_{t-p})$, $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, \ldots, \beta_{pt})^\top$, and $\varepsilon_t \sim \mathsf{N}(0, \sigma^2)$.
Now, stack the observations over all times t:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \qquad (11.20)$$

where $\mathbf{y} = (y_1, \ldots, y_T)^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_T^\top)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$,
and

$$X = \begin{pmatrix} \mathbf{x}_1^\top & 0 & \ldots & 0 \\ 0 & \mathbf{x}_2^\top & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{x}_T^\top \end{pmatrix}.$$

Thus, the joint density of $\mathbf{y}$ is given by (suppressing the dependence on the initial observations $y_0, \ldots, y_{1-p}$)

$$\ln f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) = -\frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) + \text{const} . \qquad (11.21)$$

Next, we stack the transition equation (11.19) over t. For simplicity we set $\boldsymbol{\beta}_0 = \mathbf{0}$ (the general case follows similarly). The transition equations can be written in matrix form as

$$H\boldsymbol{\beta} = \mathbf{u} ,$$

where $\mathbf{u} \sim \mathsf{N}(\mathbf{0}, S)$, $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_T^\top)^\top$, with

$$H = \begin{pmatrix} I & 0 & \ldots & 0 & 0 \\ -I & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & -I & I \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} \Omega_0 & 0 & \ldots & 0 \\ 0 & \Omega & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Omega \end{pmatrix} .$$

Note that $|H| = 1$ and $|S| = |\Omega_0| \, |\Omega|^{T-1}$. It follows that the joint density of $\boldsymbol{\beta}$ satisfies

$$\ln f(\boldsymbol{\beta} \mid \Omega) = -\frac{T-1}{2} \ln |\Omega| - \frac{1}{2} \boldsymbol{\beta}^\top H^\top S^{-1} H \boldsymbol{\beta} + \text{const} . \qquad (11.22)$$

Since Ω is a $(p+1) \times (p+1)$ symmetric matrix, it contains $(p+1)(p+2)/2$ distinct parameters. Even when p is small, say, $p = 4$, there are 15 distinct parameters. In typical empirical applications one cannot accurately estimate these many parameters. We can reduce the number of parameters by assuming that Ω is diagonal. We adopt this approach and let $\boldsymbol{\omega}^2 = (\omega_0^2, \omega_1^2, \ldots, \omega_p^2)^\top$ denote the vector of diagonal elements of Ω.

To derive the posterior density, it remains to specify the prior for σ^2 and $\boldsymbol{\omega}^2$ (note that $\boldsymbol{\omega}^2$ can be estimated from the data rather than fixed as a vector of constants if a proper prior is adopted). We assume an independent prior $f(\sigma^2, \boldsymbol{\omega}^2) = f(\sigma^2) f(\boldsymbol{\omega}^2)$, where

$$\sigma^2 \sim \mathsf{InvGamma}(\alpha_{\sigma^2}, \lambda_{\sigma^2}), \quad \omega_i^2 \sim \mathsf{InvGamma}(\alpha_{\omega_i^2}, \lambda_{\omega_i^2}) , \qquad (11.23)$$

and $\alpha_{\sigma^2}, \lambda_{\sigma^2}, \alpha_{\omega_i^2}$, and $\lambda_{\omega_i^2}, i = 0, \ldots, p$, are constants specified by the user.

Finally, the posterior density is given by

$$f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\omega}^2 \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta} \mid \boldsymbol{\omega}^2) f(\sigma^2) f(\boldsymbol{\omega}^2) , \qquad (11.24)$$

where $f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)$ and $f(\boldsymbol{\beta} \mid \Omega) = f(\boldsymbol{\beta} \mid \boldsymbol{\omega}^2)$ are provided in (11.21) and (11.22), respectively. Posterior draws can be obtained using the Gibbs sampler. Specifically, we sequentially draw from $f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\omega}^2)$ followed by a draw from $f(\sigma^2, \boldsymbol{\omega}^2 \mid \mathbf{y}, \boldsymbol{\beta})$.

For the first step, we note that $f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\omega}^2)$ is again a normal density. Hence, once we determine the mean vector and the precision matrix, we can apply

Algorithm 10.1 to obtain a draw from it efficiently. Using (11.21) and (11.22), we have

$$\ln f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\omega}^2) = \ln f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) + \ln f(\boldsymbol{\beta} \mid \boldsymbol{\omega}^2) + \text{const}$$

$$= -\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top H^\top S^{-1} H \boldsymbol{\beta} + \text{const}$$

$$= -\frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top K_\beta (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \text{const} ,$$

where

$$K_\beta = \frac{1}{\sigma^2} X^\top X + H^\top S^{-1} H \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}} = K_\beta^{-1}\left(\frac{1}{\sigma^2} X^\top \mathbf{y}\right).$$

In other words, $(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\omega}^2) \sim \mathsf{N}(\widehat{\boldsymbol{\beta}}, K_\beta^{-1})$.

Next, note that σ^2 and $\boldsymbol{\omega}^2$ are conditionally independent given $\mathbf{y}$ and $\boldsymbol{\beta}$. Namely, from (11.24) we have

$$f(\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}) \propto f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) f(\sigma^2) \qquad \text{and} \qquad f(\boldsymbol{\omega}^2 \mid \mathbf{y}, \boldsymbol{\beta}) \propto f(\boldsymbol{\beta} \mid \boldsymbol{\omega}^2) f(\boldsymbol{\omega}^2) .$$

Similar to (11.16) it follows from (11.21) and the prior $f(\sigma^2)$ that

$$(\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}) \sim \mathsf{InvGamma}\left(\alpha_{\sigma^2} + \frac{T}{2}, \lambda_{\sigma^2} + \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta})\right).$$

To find the distribution of $(\boldsymbol{\omega}^2 \mid \mathbf{y}, \boldsymbol{\beta})$, we use (11.22) and the assumption that $\Omega = \text{diag}(\boldsymbol{\omega}^2)$, to find

$$\ln f(\boldsymbol{\omega}^2 \mid \mathbf{y}, \boldsymbol{\beta}) = -\frac{T-1}{2} \sum_{i=0}^{p} \ln \omega_i^2 - \frac{1}{2} \sum_{i=0}^{p} \frac{1}{\omega_i^2} \sum_{t=2}^{T} (\beta_{it} - \beta_{i,t-1})^2 + \text{const} .$$

From this we can deduce that conditional on $\mathbf{y}$ and $\boldsymbol{\beta}$ the components of $\boldsymbol{\omega}^2$ are independent of each other and each has an inverse-gamma distribution:

$$(\omega_i^2 \mid \mathbf{y}, \boldsymbol{\beta}) \overset{\text{ind}}{\sim} \mathsf{InvGamma}\left(\alpha_{\omega_i^2} + \frac{T-1}{2}, \lambda_{\omega_i^2} + \frac{1}{2} \sum_{t=2}^{T} (\beta_{it} - \beta_{i,t-1})^2\right)$$

for $i = 0, \ldots, p$.

Example 11.3 (Modeling Inflation with Time-Varying Parameter AR Model).
In Example 11.2 we used the unobserved components model to fit the US quarterly

CPI inflation rate from 1947 to 2011. Here we illustrate Bayesian estimation in the more general time-varying parameter AR model. Specifically, we fit the time-varying parameter AR model in (11.18) and (11.19) using the inverse-gamma priors in (11.23). For simplicity, we fix p, the number of lags, to be 2. As for the hyperparameters in the prior, we choose relatively small values for the shape parameters so that the prior is relatively non-informative (e.g., large prior variances): $\alpha_{\sigma^2} = \alpha_{\omega_i^2} = 5, i = 0, \ldots, p$. Next, we set $\lambda_{\sigma^2} = (\alpha_{\sigma^2} - 1)$, $\lambda_{\omega_0^2} = 0.5^2(\alpha_{\omega_0^2} - 1)$, and $\lambda_{\omega_i^2} = 0.1^2(\alpha_{\omega_i^2} - 1)$ for $i = 1, \ldots, p$. These values imply $\mathbb{E}\sigma^2 = 1$, $\mathbb{E}\omega_0^2 = 0.5^2$, and $\mathbb{E}\omega_i^2 = 0.1^2$ for $i = 1, \ldots, p$. The covariance matrix Ω_0 is set to be diagonal with diagonal elements 5.

Before we discuss the main Gibbs sampler, we need a fast routine to build an appropriate sparse matrix. Recall that we want to write the measurement equation in matrix notation $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ [see (11.20)].

The following function SURform takes the $T \times (p + 1)$ matrix

$$\begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_T^\top \end{pmatrix}$$

and produces the sparse matrix X, which is of dimension $T \times T(p + 1)$.

```
function Xout = SURform(X)
[r c] = size( X );
idi = kron((1:r)',ones(c,1));
idj = (1:r*c)';
Xout = sparse(idi,idj,reshape(X',r*c,1));
```

The function to implement the Gibbs sampler is given below. The number of iterations in the main Gibbs run is 11000, where the first 1000 is a burn-in sample. Note that the Statistics Toolbox function gamrnd.m is used instead of gamrand.m because the latter assumes scalar input parameters.

```
%TVPAR.m
load 'USCPI.csv';
nloop = 11000;
burnin = 1000;
p = 2;     % no of lags
y0 = USCPI(1:p);
y = USCPI(p+1:end);
T = length(y);
q = p+1;   % dim of states
```

```
Tq = T*q;
    %% prior
asigma2 = 5; lsigma2 = 1*(asigma2-1);
aomega2 = 5; lomega2 = (aomega2-1)*[0.5^2; 0.1^2*ones(p,1)];
invOmega0 = ones(q,1)/5;
    %% initialize
omega2 = .1*ones(q,1);
sigma2 = 1;
store_omega2 = zeros(nloop-burnin,q);
store_sigma2 = zeros(nloop-burnin,1);
store_beta = zeros(Tq,1);
    %% construct/compute a few things
X = [ones(T,1) [y0(end); y(1:end-1)] [y0; y(1:end-2)]];
bigX = SURform(X);
H = speye(Tq) - sparse(q+1:Tq,1:(T-1)*q, ...
    ones(1,(T-1)*q),Tq,Tq);
newaomega2 = aomega2 + T - 1;
newasigma2 = asigma2 + T ;
for loop = 1:nloop
        %% sample beta
    invS = sparse(1:Tq,1:Tq,[invOmega0' ...
        repmat(1./omega2',1,T-1)]);
    K = H'*invS*H + bigX'*bigX/sigma2;
    C = chol(K,'lower');
    betahat = K\(bigX'*y/sigma2);
    beta = betahat + C'\randn(Tq,1);
        %% sample omega2
    erromega2 = reshape(H*beta,q,T);
    newlomega2 = lomega2 + sum(erromega2(:,2:end).^2,2)/2;
    omega2 = 1./gamrnd(newaomega2, 1./newlomega2);
        %% sample sigma2
    newlsigma2 = lsigma2 + sum((y-bigX*beta).^2)/2;
    sigma2 = 1/gamrnd(newasigma2, 1./newlsigma2);
    if loop>burnin
        i=loop-burnin;
        store_beta = store_beta + beta;
        store_omega2(i,:) = omega2';
        store_sigma2(i,:) = sigma2;
    end
end
betahat = store_beta/(nloop-burnin);
sigma2hat = mean(store_sigma2);
omega2hat = mean(store_omega2);
```

The estimated posterior means for $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, \beta_{2t})^\top$ are reported in Fig. 11.2. It is evident from the plots that there is a lot of time variation in the regression coefficients, which suggests that a time-invariant autoregressive model might not be appropriate. For instance, the intercept β_{0t} is estimated to be about 1 % in the 1960s,
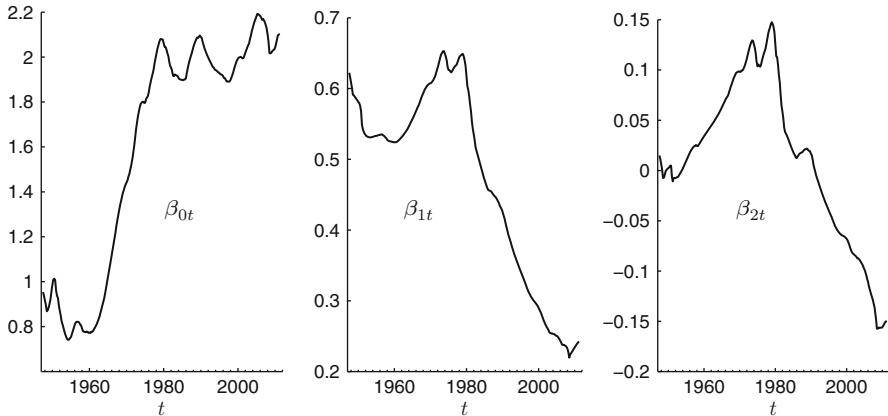
Fig. 11.2 Estimated posterior means for $\boldsymbol{\beta}_t$

while the estimate jumps to around 2 % in the 1980s. Moreover, the estimate for the lag-1 coefficient β_{1t} increases from around 0.5 in 1960 to about 0.65 in the 1980s, which then decreases gradually in the following two decades, and reaches a small value of 0.2 in 2010. Taken together, the hyperinflation in the 1970s–1980s may be viewed as a combination of a large shift in the level of the underlying inflation, together with an increase in persistence. After the 1980s, however, the underlying inflation stays at around 2 %, but since the persistence decreases substantially, the inflation rate remains at a relatively low level.

11.3   Stochastic Volatility Model

A prominent feature of many time series, particularly macroeconomic and financial data, is the so-called **volatility clustering**—the phenomenon that large changes in observations tend to be followed by large changes and small changes followed by small changes. For example, large movements in asset returns tend to cluster together (e.g., during crisis), whereas there might be little variation over long stretches of "normal periods." Models with constant variance obviously do not allow the volatility of the observations to change over time and hence cannot model volatility clustering. In this section we introduce a class of state space models that can accommodate time-varying volatility. To focus our discussion on modeling the variance of the time series, we assume for the moment that the observations $\{y_t\}$ have zero mean; one could add a suitable conditional mean process such as an AR(p) component later on.

Definition 11.4. (Stochastic Volatility Model). In the **stochastic volatility** model the observation at time t is given by

$$y_t = e^{h_t/2} \varepsilon_t \, , \tag{11.25}$$

where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, 1)$. Consequently, the **volatility** of y_t is $\mathrm{Var}(y_t) = e^{h_t}$. The states are initialized with $h_1 \sim \mathsf{N}(h_0, \sigma_0^2)$ for some known constants h_0 and σ_0^2 and evolve according to a random walk

$$h_t = h_{t-1} + v_t, \quad t = 2, \dots, T \, , \tag{11.26}$$

where $\{v_t\} \sim_{\text{iid}} \mathsf{N}(0, \omega^2)$. The state h_t is called the **log-volatility**.

The stochastic volatility model is an example of a *nonlinear* state space model where the measurement equation (11.25) is not linear in the state. One challenge of fitting this nonlinear model is that the joint conditional density of the states $\mathbf{h} = (h_1, \dots, h_T)^\top$ given $\mathbf{y}$ is nonstandard (in contrast to previous examples where the conditional densities of the states are all Gaussian). As such, Bayesian estimation using MCMC and classical estimation via EM both become more difficult.

11.3.1 Auxiliary Mixture Sampling Approach

A popular method for estimating the stochastic volatility model is **auxiliary mixture sampling**. The basic idea underlying this approach is as follows. First, we transform the observation y_t so that the measurement equation becomes linear in h_t. Specifically, we square both sides of the measurement equation (11.25) and take the (natural) logarithm:

$$y_t^* = h_t + \varepsilon_t^* \, , \tag{11.27}$$

where $y_t^* = \ln y_t^2$ and $\varepsilon_t^* = \ln \varepsilon_t^2$. In practice, it is often recommended to set $y_t^* = \ln(y_t^2 + c)$ for some small constant c, say, $c = 0.0001$, to avoid numerical problems when y_t is close to zero. Now, after the transformation, (11.27) and (11.26) define a linear state space model. However, the error ε_t^* no longer has a Gaussian distribution (in fact, it has a log-χ_1^2 distribution), and the estimation techniques for linear Gaussian state space models discussed earlier cannot be directly applied.

In view of this difficulty, the second ingredient of the auxiliary mixture sampling approach is to find a suitable Gaussian mixture that approximates the pdf of ε_t^*:

$$f(\varepsilon_t^*) \approx \sum_{i=1}^{n} p_i \, \varphi(\varepsilon_t^* \, ; \mu_i, \sigma_i^2) \, , \tag{11.28}$$

where $\varphi(x\,;\,\mu,\sigma^2)$ is the Gaussian density with mean μ and variance σ^2, p_i is the mixture probability for the ith component, and n is the number of components. The idea is to approximate the nonlinear stochastic volatility model using a mixture of linear Gaussian models, where the estimation of the latter models is standard. We can equivalently write (11.28) in terms of an auxiliary random variable $s_t \in \{1,\ldots,n\}$ that serves as the mixture component indicator (hence, the name of the approach):

$$(\varepsilon_t^* \,|\, s_t = i) \sim \mathsf{N}(\mu_i, \sigma_i^2) \,, \tag{11.29}$$

$$\mathbb{P}(s_t = i) = p_i \,. \tag{11.30}$$

Now, conditional on the component indicator s_t, we have a linear Gaussian model and the machinery for estimating such models can be applied.

It remains to select a suitable Gaussian mixture. By matching the moments of the log-χ_1^2 distribution, Kim et al. (1998) propose a seven-component Gaussian mixture

$$f(x) = \sum_{i=1}^{7} p_i \,\varphi(x\,;\,\mu_i - 1.2704, \sigma_i^2) \,,$$

where the values of the parameters are given in Table 11.1. It is important to note that since the log-χ_1^2 distribution does not involve any unknown parameters, neither does this Gaussian mixture. In fact, all the parameter values of the approximating density are known.

Table 11.1   A seven-component Gaussian mixture for approximating the log-χ_1^2 distribution

Component	p_i	μ_i	σ_i^2
1	0.00730	-10.12999	5.79596
2	0.10556	-3.97281	2.61369
3	0.00002	-8.56686	5.17950
4	0.04395	2.77786	0.16735
5	0.34001	0.61942	0.64009
6	0.24566	1.79518	0.34023
7	0.25750	-1.08819	1.26261

To summarize the model, define $\mathbf{s} = (s_1,\ldots,s_T)^\top$, $\mathbf{y}^* = (y_1^*,\ldots,y_T^*)^\top$, $\mathbf{h} = (h_1,\ldots,h_T)^\top$, $\mathbf{v} = (v_1,\ldots,v_T)^\top$, and $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*,\ldots,\varepsilon_T^*)^\top$. Let H be the same matrix as in (11.10). By (11.27) we can write

$$\mathbf{y}^* = \mathbf{h} + \boldsymbol{\varepsilon}^*,$$

where $(\boldsymbol{\varepsilon}^* \,|\, \mathbf{s}) \sim \mathsf{N}(\mathbf{d}, \Sigma_{\mathbf{y}^*})$, with $\mathbf{d} = (\mu_{s_1} - 1.2704, \ldots, \mu_{s_T} - 1.2704)^\top$ and $\Sigma_{\mathbf{y}^*} = \mathrm{diag}(\sigma_{s_1}^2, \ldots, \sigma_{s_T}^2)$. The (fixed) $\{\mu_i\}$ and $\{\sigma_i^2\}$ are given in Table 11.1. Consequently,

$$(\mathbf{y}^* \,|\, \mathbf{s}, \mathbf{h}) \sim \mathsf{N}(\mathbf{h} + \mathbf{d}, \Sigma_{\mathbf{y}^*}) \,. \tag{11.31}$$

Using an inverse-gamma prior for ω^2, the hierarchical Bayesian model is thus as follows:

1. $(\mathbf{y}^* \mid \mathbf{s}, \mathbf{h}) \sim \mathsf{N}(\mathbf{h} + \mathbf{d}, \ \Sigma_{\mathbf{y}^*})$.
2. It follows from (11.26) that the random vector $\mathbf{h}$ is of the form $\mathbf{h} = H^{-1}\mathbf{v}$, where $\mathbf{v} \sim \mathsf{N}(\mathbf{0}, \Omega_{\mathbf{v}})$, with $\Omega_{\mathbf{v}} = \mathrm{diag}(\omega_0^2, \omega^2, \dots, \omega^2)$.
3. The components $s_1, \dots, s_t$ of $\mathbf{s}$ are independent, with $\mathbb{P}(s_t = i) = p_i$, $\quad i = 1, \dots, 7$.
4. $\omega^2 \sim \mathsf{InvGamma}(\alpha_{\omega^2}, \lambda_{\omega^2})$.

In order to perform a Bayesian analysis we need to be able to sample from the posterior pdf $f(\mathbf{h}, \mathbf{s}, \omega^2 \mid \mathbf{y})$. We can do this via a Gibbs sampler, by sequentially sampling from (1) $f(\mathbf{h} \mid \mathbf{y}^*, \mathbf{s}, \omega^2)$; (2) $f(\mathbf{s} \mid \mathbf{y}^*, \mathbf{h}, \omega^2) = f(\mathbf{s} \mid \mathbf{y}^*, \mathbf{h})$; and (3) $f(\omega^2 \mid \mathbf{y}, \mathbf{h}, \mathbf{s}) = f(\omega^2 \mid \mathbf{h})$.

To implement (1), we write

$$\ln f(\mathbf{h} \mid \mathbf{y}^*, \mathbf{s}, \omega^2) = \ln f(\mathbf{y}^* \mid \mathbf{s}, \mathbf{h}) + \ln f(\mathbf{h} \mid \omega^2) + \text{const} ,$$

where $f(\mathbf{y}^* \mid \mathbf{s}, \mathbf{h})$ follows from (11.31) and $f(\mathbf{h} \mid \omega^2)$ follows from

$$(\mathbf{h} \mid \omega^2) \sim \mathsf{N}(\mathbf{0}, (H^\top \Omega_{\mathbf{v}}^{-1} H)^{-1}) .$$

Using a similar reasoning as in Sect. 11.1.2, we find

$$(\mathbf{h} \mid \mathbf{y}^*, \mathbf{s}, \omega^2) \sim \mathsf{N}(\widehat{\mathbf{h}}, K_{\mathbf{h}}^{-1}) ,$$

where

$$K_{\mathbf{h}} = \Sigma_{\mathbf{y}^*}^{-1} + H^\top \Omega_{\mathbf{v}}^{-1} H \qquad \text{and} \qquad \widehat{\mathbf{h}} = K_{\mathbf{h}}^{-1} \Sigma_{\mathbf{y}^*}^{-1} (\mathbf{y}^* - \mathbf{d}) .$$

A draw from the above Gaussian distribution can be efficiently obtained using Algorithm 10.1.

To implement (2), note that $f(\mathbf{s} \mid \mathbf{y}^*, \mathbf{h}) = \prod_{t=1}^{T} f(s_t \mid y_t^*, h_t)$, and therefore we can draw each s_t independently. Since s_t is a discrete random variable that follows a seven-point distribution, it can be easily sampled as long as we can compute $\mathbb{P}(s_t = i \mid y_t^*, h_t)$ for $i = 1, \dots, 7$. In fact, we have

$$\mathbb{P}(s_t = i \mid y_t^*, h_t) = \frac{1}{c_t} p_i \, \varphi(y_t^* ; h_t + \mu_i - 1.2704, \sigma_i^2) ,$$

where $c_t = \sum_{j=1}^{7} p_j \, \varphi(y_t^* ; h_t + \mu_j - 1.2704, \sigma_j^2)$ is the normalization constant.

Finally, implementation of (3) is similar to the derivation in (11.17). In particular, given $\mathbf{y}$ and $\mathbf{h}$, ω^2 has again an inverse-gamma distribution:

$$(\omega^2 \mid \mathbf{y}, \mathbf{h}) \sim \mathsf{InvGamma}\left(\alpha_{\omega^2} + \frac{T-1}{2}, \ \lambda_{\omega^2} + \frac{1}{2}\sum_{t=2}^{T}(h_t - h_{t-1})^2\right).$$

Example 11.4 (Modeling Inflation with Unobserved Components Stochastic Volatility Model). We have considered in Example 11.2 an unobserved components model with constant variance for modeling the US quarterly CPI inflation. In this example we extend the constant variance to include stochastic volatility in the measurement equation. Specifically, consider

$$y_t = \tau_t + e^{h_t/2}\varepsilon_t \ ,$$

$$\tau_t = \tau_{t-1} + u_t \ ,$$

$$h_t = h_{t-1} + v_t \ ,$$

where $\{\varepsilon_t\} \sim_{\text{iid}} N(0, 1)$, $\{u_t\} \sim_{\text{iid}} N(0, \omega_\tau^2)$, and $\{v_t\} \sim_{\text{iid}} N(0, \omega_h^2)$. The state equations are initialized with $\tau_1 \sim N(\tau_0, V_\tau)$ and $h_1 \sim N(h_0, V_h)$, where $\tau_0 = h_0 = 0$ and $V_\tau = V_h = 9$. Again we assume independent inverse-gamma priors for ω_τ^2 and ω_h^2:

$$\omega_\tau^2 \sim \text{InvGamma}(\alpha_\tau, \lambda_\tau) \qquad \text{and} \qquad \omega_h^2 \sim \text{InvGamma}(\alpha_h, \lambda_h) \ ,$$

where we set $\alpha_\tau = \alpha_h = 10$, $\lambda_\tau = 0.25^2(\alpha_\tau - 1)$, and $\lambda_h = 0.2^2(\alpha_h - 1)$. These values imply $\mathbb{E}\omega_\tau^2 = 0.25^2$ and $\mathbb{E}\omega_h^2 = 0.2^2$. By defining y_t^* appropriately, the results derived earlier in this section can be applied to construct a suitable Gibbs sampler. More precisely, let

$$y_t^* = \ln((y_t - \tau_t)^2 + 0.0001) \ .$$

Then, by using the auxiliary mixture sampling approach, we sequentially draw from (1) $f(\mathbf{s}\,|\,\mathbf{y}, \boldsymbol{\tau}, \mathbf{h}, \omega_\tau^2, \omega_h^2) = f(\mathbf{s}\,|\,\mathbf{y}^*, \boldsymbol{\tau}, \mathbf{h})$; (2) $f(\mathbf{h}\,|\,\mathbf{y}, \boldsymbol{\tau}, \mathbf{s}, \omega_\tau^2, \omega_h^2) = f(\mathbf{h}\,|\,\mathbf{y}^*, \boldsymbol{\tau}, \mathbf{s}, \omega_h^2)$; (3) $f(\boldsymbol{\tau}\,|\,\mathbf{y}, \mathbf{h}, \mathbf{s}, \omega_\tau^2, \omega_h^2) = f(\boldsymbol{\tau}\,|\,\mathbf{y}, \mathbf{h}, \omega_\tau^2)$; and (4) $f(\omega_\tau^2, \omega_h^2\,|\,\mathbf{y}, \boldsymbol{\tau}, \mathbf{h}, \mathbf{s}) = f(\omega_\tau^2, \omega_h^2\,|\,\boldsymbol{\tau}, \mathbf{h})$.

For Steps (1) and (2), we can use the following MATLAB function SVRW to draw from the full conditional densities for **s** and **h**:

```
function [h s] = SVRW(ystar,h,omega2h,Vh)
T = length(h);
  %% parameters for the Gaussian mixture
pi = [0.0073 .10556 .00002 .04395 .34001 .24566 .2575];
mui = [-10.12999 -3.97281 -8.56686 2.77786 .61942 1.79518 ...
       -1.08819] - 1.2704;
sig2i =[5.79596 2.61369 5.17950 .16735 .64009 .34023 1.26261];
sigi = sqrt(sig2i);
  %% sample s from a 7-point discrete distribution
temprand = rand(T,1);
q = repmat(pi,T,1).*normpdf(repmat(ystar,1,7),repmat(h,1,7)...
    +repmat(mui,T,1),repmat(sigi,T,1));
q = q./repmat(sum(q,2),1,7);
s = 7 - sum(repmat(temprand,1,7)<cumsum(q,2),2)+1;
```

```
  %% sample h
H = speye(T) - sparse(2:T,1:(T-1),ones(1,T-1),T,T);
invOmegah = spdiags([1/Vh; 1/omega2h*ones(T-1,1)],0,T,T);
d = mui(s)'; invSigystar = spdiags(1./sig2i(s)',0,T,T);
Kh = H'*invOmegah*H + invSigystar;
Ch = chol(Kh,'lower');
hhat = Kh\(invSigystar*(ystar-d));
h = hhat + Ch'\randn(T,1);
```

Next, using a similar derivation as on page , one can show that

$$(\boldsymbol{\tau} \mid \mathbf{y}, \mathbf{h}, \omega_\tau^2) \sim \mathsf{N}(\widehat{\boldsymbol{\tau}}, K_{\boldsymbol{\tau}}^{-1}) \,,$$

where $\Sigma_{\mathbf{y}}^{-1} = \mathrm{diag}(\mathrm{e}^{-h_1}, \ldots, \mathrm{e}^{-h_T})$,

$$K_{\boldsymbol{\tau}} = H^\top \Omega_{\boldsymbol{\tau}}^{-1} H + \Sigma_{\mathbf{y}}^{-1}, \qquad \text{and} \qquad \widehat{\boldsymbol{\tau}} = K_{\boldsymbol{\tau}}^{-1} \Sigma_{\mathbf{y}}^{-1} \mathbf{y} \,.$$

Hence, Step 3 can be implemented easily. Lastly, to complete Step 4, note that ω_τ^2 and ω_h^2 are conditionally independent given the states. Moreover,

$$(\omega_\tau^2 \mid \boldsymbol{\tau}, \mathbf{h}) \sim \mathsf{InvGamma}\left(\alpha_\tau + \frac{T-1}{2}, \ \lambda_\tau + \frac{1}{2}\sum_{t=2}^{T}(\tau_t - \tau_{t-1})^2\right) \,,$$

$$(\omega_h^2 \mid \boldsymbol{\tau}, \mathbf{h}) \sim \mathsf{InvGamma}\left(\alpha_h + \frac{T-1}{2}, \ \lambda_h + \frac{1}{2}\sum_{t=2}^{T}(h_t - h_{t-1})^2\right) \,.$$

The main script below fits the unobserved components model with stochastic volatility using the auxiliary mixture sampling approach.

```
%UCSV.m
load 'USCPI.csv';
y = USCPI;
T = length(y);
nloop = 11000;
burnin = 1000;
  %% prior
Vtau = 9; Vh = 9;
atau = 10; ltau = .25^2*(atau-1);
ah = 10; lh = .2^2*(ah-1);
  %% initialize the Markov chain
omega2tau = .25^2;
omega2h = .2^2;
```

```
h = log(var(y)*.8)*ones(T,1);
H = speye(T) - sparse(2:T,1:(T-1),ones(1,T-1),T,T);
  %% initialize for storage
store_omega2tau = zeros(nloop - burnin,1);
store_omega2h = zeros(nloop - burnin,1);
store_tau = zeros(nloop - burnin,T);
store_h = zeros(nloop - burnin,T);
  %% compute a few things
newatau = (T-1)/2 + atau;
newah = (T-1)/2 + ah;
for loop = 1:nloop
      %% sample tau
    invOmegatau = sparse(1:T,1:T, ...
        [1/Vtau 1/omega2tau*ones(1,T-1)]);
    invSigy = sparse(1:T,1:T,exp(-h));
    Ktau = H'*invOmegatau*H + invSigy;
    Ctau = chol(Ktau,'lower');
    tauhat = Ktau\(invSigy*y);
    tau = tauhat + Ctau'\randn(T,1);
      %% sample h
    ystar = log((y-tau).^2 + .0001 );
    h = SVRW(ystar,h,omega2h,Vh);
      %% sample omega2tau
    newltau = ltau + sum((tau(2:end)-tau(1:end-1)).^2)/2;
    omega2tau = 1/gamrnd(newatau, 1./newltau);
      %% sample omega2h
    newlh = lh + sum((h(2:end)-h(1:end-1)).^2)/2;
    omega2h = 1/gamrnd(newah, 1./newlh);
    if loop>burnin
        i = loop-burnin;
        store_tau(i,:) = tau';
        store_h(i,:) = h';
        store_omega2tau(i,:) = omega2tau;
        store_omega2h(i,:) = omega2h;
    end
end
tauhat = mean(store_tau)';
hhat = mean(store_h)';
```

We use the above code to obtain 10000 posterior draws after a period of 1000 draws as burn-in. We present in Fig. 11.3 the estimated posterior means for the underlying inflation τ and the log-volatilities **h**.

Compared to the results obtained under the constant variance unobserved components model, the estimated underlying inflation exhibits a similar pattern, but is seemingly more variable. In addition, the estimated log-volatilities show that there is substantial time variation in the variance in the measurement equation, highlighting the relevance of the stochastic volatility model.
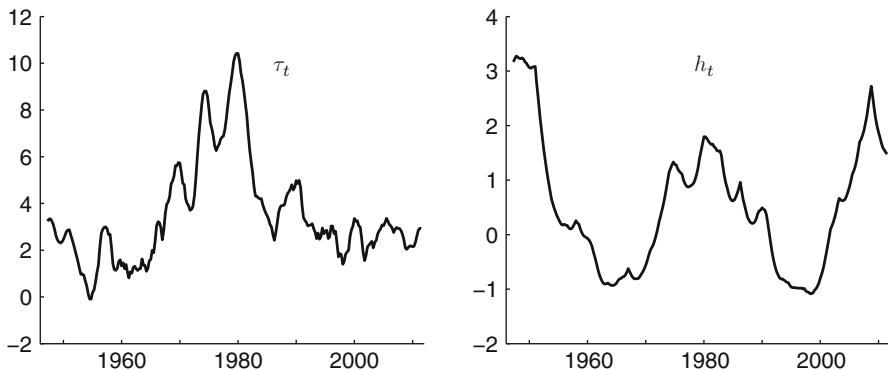
Fig. 11.3 Estimated posterior means for $\boldsymbol{\tau}$ (*left panel*) and $\mathbf{h}$ (*right panel*)

11.4 Problems

11.1. Prove the updating formulas in (11.4) by using the joint distribution in (11.3) and Theorem 3.8.

11.2. The **trace** of a square matrix $A = (a_{ij})$ is the sum of the diagonal elements: $\mathrm{tr}(A) = \sum_i a_{ii}$.

(a) Let A and B be matrices (not necessarily square) such that AB and BA are square matrices (not necessarily of the same dimension). Show that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

(b) Let A be a square matrix and let $\mathbf{x}$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Using (a) and the fact that $\mathrm{tr}(\mathbb{E}Z) = \mathbb{E}\,\mathrm{tr}(Z)$ for a random square matrix Z, show that

$$\mathbb{E}(\mathbf{x}^\top A \mathbf{x}) = \mathrm{tr}(A\Sigma) + \boldsymbol{\mu}^\top A \boldsymbol{\mu} \ .$$

11.3. Show that for the measurement equation in (11.5), if one fixes $\tau_t = y_t$, then the function

$$g(\sigma^2) = \ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2)$$

is unbounded in σ^2.

11.4. Another interpretation of the unobserved components model is to view it as a way to specify *stochastic trends*. Using the transition equation (11.6) and recursive substitution, show that

$$\mathrm{Var}(\tau_t \mid \tau_1) = (t-1)\omega^2 \ ,$$

i.e., the stochastic trend τ_t has variance that is increasing with time, which implies that τ_t can wander over an increasing range of values as time increases.

11.5. For the unobserved components model (and more generally linear Gaussian state space models), it is possible to evaluate the likelihood function $L(\sigma^2; \mathbf{y})$ without computing the high-dimensional integral in (11.7). More specifically, by Bayes' theorem, the likelihood function can be written as (recall that ω^2 is a fixed constant)

$$L(\sigma^2; \mathbf{y}) = f(\mathbf{y} \mid \sigma^2, \omega^2) = \frac{f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) f(\boldsymbol{\tau} \mid \omega^2)}{f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2, \omega^2)} \, ,$$

where the densities $f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2)$, $f(\boldsymbol{\tau} \mid \omega^2)$, and $f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2, \omega^2)$ are all normal, and can be evaluated quickly. Since the second equality holds for all $\boldsymbol{\tau}$, one can simply choose some convenient values, say, $\boldsymbol{\tau} = \mathbf{0}$.

Redo Example 11.2 by directly maximizing the log-likelihood function $l(\sigma^2; \mathbf{y})$. Specifically, plot $l(\sigma^2; \mathbf{y})$ as a function of σ^2. Moreover, find the maximum likelihood estimate for σ^2.

11.6. In this exercise we generalize the unobserved components model to allow for an additional channel for persistence. Specifically, consider

$$y_t = \tau_t + \beta(y_{t-1} - \tau_{t-1}) + \varepsilon_t \, ,$$
$$\tau_t = \tau_{t-1} + u_t \, ,$$

where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, $\{u_t\} \sim_{\text{iid}} \mathsf{N}(0, \omega^2)$, and $\omega^2 = 1$. The underlying trend is initialized with $\tau_1 \sim \mathsf{N}(0, 5)$ and $\tau_0 = 0$. It is obvious that if $\beta = 0$ it reduces to the standard unobserved components model.

(a) Derive the log-density $\ln f(\mathbf{y} \mid y_0, \boldsymbol{\tau}, \beta, \sigma^2)$.
(b) Show that the conditional density $f(\boldsymbol{\tau} \mid \mathbf{y}, y_0, \beta, \sigma^2, \omega^2)$ is normal, and derive its parameters.
(c) Describe how one can estimate the model parameters using classical and Bayesian methods.

11.7. Under the unobserved components model, suppose the state equation is given by

$$\tau_t = \beta \tau_{t-1} + u_t, \quad u_t \sim \mathsf{N}(0, \omega^2)$$

for $t = 2, \ldots, T$, with $\tau_1 \sim \mathsf{N}(\tau_0, \omega_0^2)$. Derive the joint density $f(\boldsymbol{\tau} \mid \beta, \omega^2)$.

11.8. Consider the following unobserved components model with AR(1) transition equation:

$$y_t = \tau_t + \varepsilon_t \, ,$$
$$\tau_t = \beta \tau_{t-1} + u_t \, ,$$

where $\{\varepsilon_t\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, $\{u_t\} \sim_{\text{iid}} \mathsf{N}(0, \omega^2)$, and the underlying trend is initialized with $\tau_1 \sim \mathsf{N}(0, 5)$. Suppose we assume the priors: $\beta \sim \mathsf{N}(0, 1)$, $\sigma_\tau^2 \sim$ InvGamma$(10, 9)$, and $\omega^2 \sim$ InvGamma$(10, 9)$. Derive all the full conditional distributions. Fit this model with the US CPI data. In particular, use the kde.m program to plot a kernel density estimate of posterior distribution of β.

Appendix A
Matlab Primer

MATLAB, a portmanteau of MATrix LABoratory, is an interactive matrix-based program for numerical computation. It is a very easy to use high-level language that requires minimal programming skills. The purpose of this appendix is to introduce the reader to some basic MATLAB functions that are used in the main text. For more detailed information and full documentation, please visit the official documentation site

<p style="text-align:center">http://www.mathworks.com/help/techdoc/.</p>

In addition, the command `help` *function_name* gives information about the function *function_name*. Alternatively, select `Help -> Product Help` in the toolbar in the MATLAB command window.

A.1 Matrices and Matrix Operations

The most fundamental objects in MATLAB are, not surprisingly, matrices. For instance, to create a 1×3 matrix (i.e., row vector) **a**, enter into the MATLAB command window:

```
a = [1 2 3]
```

MATLAB returns

```
a =
     1 2 3
```

To create a matrix with more than one row, use semicolons to separate the rows. For example, the line

```
A = [1 2 3; 4 5 6; 7 8 9];
```

creates a 3×3 matrix A. It is worth noting that MATLAB is case sensitive for variable names and built-in functions. That means MATLAB treats a and A as different objects. To display the ith element in a vector **x**, just type x(i). For example,

```
a(2)
```

refers to the second element of **a**. Similarly, one can access a particular element of A by specifying its row and column number (row first followed by column). For instance,

```
A(2,3)
```

displays the $(2, 3)$ entry of the matrix A. To display multiple elements in the matrix, one can use expressions involving colons. For example,

```
A(1,1:2)
```

displays the first and second elements in the first row, whereas

```
A(:,2)
```

displays all the elements in the second column.

To perform numerical computation, one needs some basic matrix operations. In MATLAB, the following matrix operations, among many others, are available:

+	addition	/	right division
−	subtraction	^	power
*	multiplication	'	transpose
\	left division		

For example,

```
a'
```

returns the transpose of **a**:

```
ans =

     1
     2
     3
```

whereas

```
a*A
```

gives the product of **a** and A:

```
ans =

    30    36    42
```

Other operations are obvious, except for the matrix divisions $\backslash$ and $/$. If A is an invertible square matrix and **a** is a compatible vector, then $\mathbf{x} = A \backslash \mathbf{a}$ is the solution of $A\mathbf{x} = \mathbf{a}$ and $\mathbf{x} = \mathbf{a}/A$ is the solution of $\mathbf{x} A = \mathbf{a}$. In other words, $A \backslash \mathbf{a}$ gives the same result (in principle) as $A^{-1} \mathbf{a}$, though they compute their results in different ways. Specifically, the former solves the linear system $A\mathbf{x} = \mathbf{a}$ for $\mathbf{x}$ by Gaussian elimination, whereas the latter first computes the inverse A^{-1}, then multiplies it by **a**. As such, the second method is in general slower as computing the inverse of a matrix is time-consuming (and inaccurate).

It is important to note that although addition and subtraction are element-wise operations, the other operations listed above are not—they are matrix operations. For example, $A\hat{~}2$ gives the square of the matrix A, not a matrix whose entries are the squares of those in A. One can make the operations $*$, $\backslash$, $/$, and $\hat{~}$ to operate element-wise by preceding them by a full stop. For example,

```
A^2
```

returns the square of the matrix A:

```
ans =

    30     36     42
    66     81     96
   102    126    150
```

On the other hand,

```
A.^2
```

computes the squares element-wise:

```
ans =

     1      4      9
    16     25     36
    49     64     81
```

A.2   Some Useful Built-In Functions

In this section we list some common built-in functions which are used throughout the main text. One can learn more about a specific function, say, eye, by typing help eye in the command window. Here are some useful matrix-building functions:

eye	create an identity matrix
zeros	create a matrix of zeros
ones	create a matrix of ones
diag	create a diagonal matrix or extract the diagonal from a matrix
rand	generate $U(0, 1)$ random variables
randn	generate $N(0, 1)$ random variables

For example, eye(n) creates an $n \times n$ identity matrix, and ones(m,n) produces an $m \times n$ matrix of ones. Given the 3×3 matrix A,

```
diag(A)
```

extracts the diagonal of the matrix A:

```
ans =

     1
     5
     9
```

But for the 3×1 vector **a**, the same command

```
diag(a)
```

builds a diagonal matrix whose main diagonal is **a**:

```
ans =

     1     0     0
     0     2     0
     0     0     3
```

Some other useful vector and matrix functions:

exp	exponential	log	natural log
sqrt	square root	abs	absolute value
sin	sine	cos	cosine
sum	sum	prod	product
max	maximum	min	minimum
chol	Cholesky factorization	inv	inverse
det	determinant	size	size

If **x** is a vector, $\mathtt{sum(x)}$ returns the sum of the elements in **x**. For a matrix X, $\mathtt{sum(X)}$ returns a row vector consisting of sums of each column, while $\mathtt{sum(X,2)}$ returns a column vector of sums of each row. For example,

```
sum(A)
```

returns

```
ans =

    12    15    18
```

whereas

```
sum(A,2)
```

gives

```
ans =

     6
    15
    24
```

For a positive definite matrix C, $\mathtt{chol(C,'lower')}$ returns the lower Cholesky factorization B such that $BB^\top = C$. For example,

```
B = [ 1 0 0; 2 3 0; 4 5 6];
C = B*B';
chol(C,'lower')
```

returns the lower Cholesky factor of $BB^\top$, which is, of course, B.

A.3   Flow Control

MATLAB has the usual control flow statements such as if-then-else, while, and for. For instance, the general form of a simple if statement is

```
if condition
    statements
end
```

The statements will be executed if the condition is true. Multiple branching is done by using elseif and else. For example, the following code simulates rolling a four-sided die:

```
u = rand;
if u <= .25
    disp('1');
elseif u <= .5
    disp('2');
elseif u <= .75
    disp('3');
else
    disp('4');
end
```

The general form of a while loop is

```
while condition
    statements
end
```

The statements will be repeatedly executed while the condition remains true. To illustrate the while loop syntax, suppose we wish to generate a positive normal
☞ 56   random variable [with pdf given in (2.25)]. We can do that using the following simple while loop:

```
u = randn;
while u <= 0
    u = randn;
end
```

Another useful control flow statement is the `for` loop, whose general form is

```
for count
    statements
end
```

Unlike a `while` loop, the `for` loop executes the statements for a fixed number of times. As an example, the following code generates five draws from the positive normal distribution.

```
x = zeros(1,5);   %% create a storage vector
for i=1:5
    u = randn;
    while u <= 0
        u = randn;
    end
    x(i)=u;
end
```

A.4   Function Handles and Function Files

In previous sections we have introduced some built-in functions in MATLAB. For instance, `sqrt` is a function that takes an argument and returns an output (its square root). Later on we will need to create our own functions that take one or more input arguments, operate on them, and return the results. One way to create new functions is through function handles. For example,

```
f = @(x) x.^2 + 5*x - 10 ;   % Note the use of the dot
```

creates the function $f(x) = x^2 + 5x - 10$ with the function handle `f`. The function handle gives you a means of invoking the function. To evaluate, say, $f(10)$, we can type `feval(f,10)`, or simply, `f(10)`.

Function handles can be passed to other functions as inputs. For instance, if we want to find the minimum point of $f(x)$ in the interval $(-10, 10)$, we can use the built-in function `fminbnd` (see Sect. A.6 for a more detailed discussion on optimization routines):

```
[xmin fmin] = fminbnd(f,-10, 10)
```

Note that `fminbnd` takes three inputs (a function handle and the two end-points of the interval) and returns two outputs (the minimizer and the minimal value). In our example, the minimizer of $f(x)$ in $(-10, 10)$ is -2.5, and the corresponding

functional value is -16.25. To create a function that takes more than one input is just as easy. For example,

```
g = @(x,y) x.^2 + y.^3 + x.*y
```

defines the two-variable function $g(x, y) = x^2 + y^3 + xy$.

For more complex functions that involve multiple lines and intermediate variables, we need the command `function`. For example, the following code takes a column vector of data and computes its mean and standard deviation:

```
function [meanx, stdevx] = stat(x)
n = length(x);
meanx = sum(x)/n;
stdevx = sqrt(sum(x.^2)/n - meanx.^2);
```

It is important to note that all the code must be written and saved in a separate m-file. Also, the name of the file should coincide with the name of the function; in this case, the file must be called `stat.m`. After saving the file, it can be used the same way as other built-in functions, for example:

```
[meanx stdx] = stat(randn(100,1))
```

returns

```
meanx =

    0.0530


stdx =

    0.9902
```

A.5   Graphics

MATLAB has several high-level graphical routines and very extensive plotting capabilities. It allows users to create various graphical objects including two- and three-dimensional graphs. One can also have a title on a graph, add a legend, change

the font and font size, label the axis, etc. For more information, in the command window, click on `Help` and next select `Demos`. Then choose `Graphics` followed by `2D Plots`.

In MATLAB the most basic function used to create 2D graphs is `plot`. For example, to make a graph of $y = \sin(x)$ on the interval from $x = 0$ to $x = 2\pi$, we use the following code:

```
x = 0:.01:2*pi;
y = sin(x);
plot(x,y);
```

Fig. A.1 A plot of the graph $y = \sin(x)$ from 0 to 2π
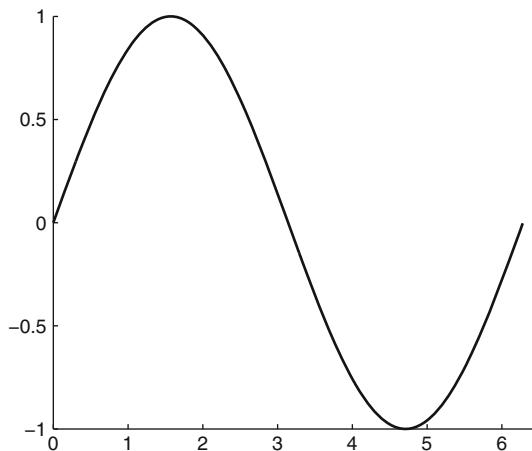
The graph produced is given in Fig. A.1. Note that the command `x = 0:.01:2*pi;` creates a vector whose components range from 0 to 2π in steps of 0.01. Another useful command to create a grid is `linspace` (use `help linspace` to learn more about this function).

Another useful function is `hist`, which allows us to plot histograms. For example,

```
hist(randn(1000,1),50);
```

creates a histogram where the 1000 standard normal draws are put into 50 equally spaced bins (see Fig. A.2).

Fig. A.2 A histogram of
1000 standard normal draws
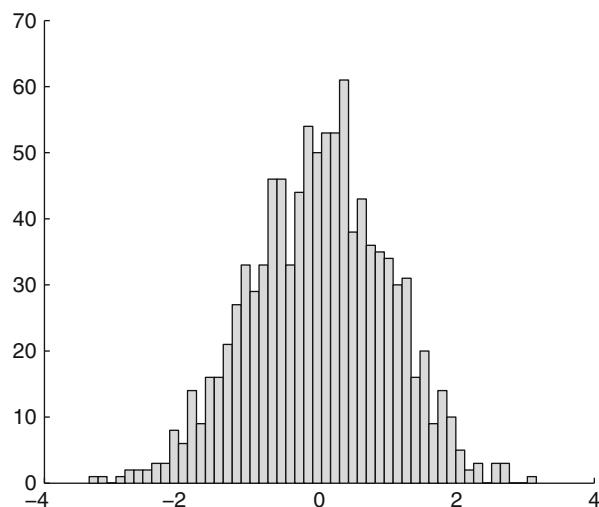
Instead of a histogram, it is often more useful to have a density estimate. One fast and reliable Gaussian kernel density estimator is the **theta KDE** of Botev et al. (2010). The MATLAB function kde.m can be downloaded from http://www.mathworks.com/matlabcentral/fileexchange/14034-kernel-density-estimator. See also Example 7.4 for an illustration.

☞ 202

It is often desirable to plot several graphs in the same figure window. For this purpose we need the function subplot. The function subplot(i,j,k) takes three arguments: the first two tells MATLAB that an $i \times j$ array of plots will be created, and the third is the running index that indicates the kth subplot is currently generated. Suppose we wish to plot the functions $y = \sin(x^2/2)$ and $y = \sin(2x)$ in the same figure window. A little modification of the above code accomplishes this goal:
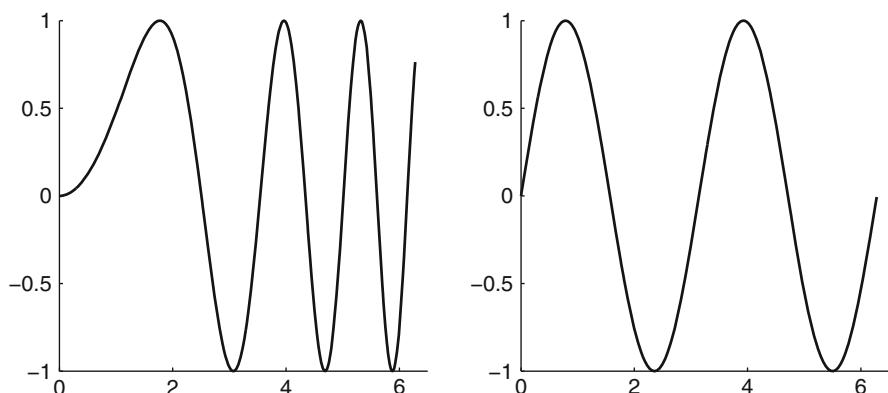
Fig. A.3 Plots of the graphs $y = \sin(x^2/2)$ and $y = \sin(2x)$ from 0 to 2π

```
x = 0:.01:2*pi;
y1 = sin(x.^2/2);     y2 = sin(2*x);
subplot(1,2,1); plot(x,y1);
subplot(1,2,2); plot(x,y2);
```

In addition, one can also easily produce 3D graphical objects in MATLAB. To illustrate various useful routines, suppose we want to plot the density function of the bivariate normal distribution (see Sect. 3.6) given by

$$f(x, y; \varrho) = \frac{1}{2\pi \sqrt{1 - \varrho^2}} \, e^{-\frac{1}{2(1-\varrho^2)}(x^2 - 2\varrho xy + y^2)} \ .$$

As in plotting a 2D graph, we first need to build a grid, and this can be done with the function meshgrid. After computing the values of the function at each point on the grid, we can plot the 3D graph using mesh. For example, we use the following code

```
rho = .6;
[x y] = meshgrid(-2:.1:2, -2:.1:2); %% build a 2D grid
z = 1/(2*pi*sqrt(1-rho^2)) ...
    * exp(-(x.^2 -2*rho*x.*y + y.^2)/(2*(1-rho^2)));
mesh(x,y,z);
```

to plot the bivariate normal density function with $\varrho = 0.6$ in Fig. A.4. The ellipsis (...) is used to break up a long line into multiple lines.

Fig. A.4 The density function of the bivariate normal distribution with $\varrho = 0.6$
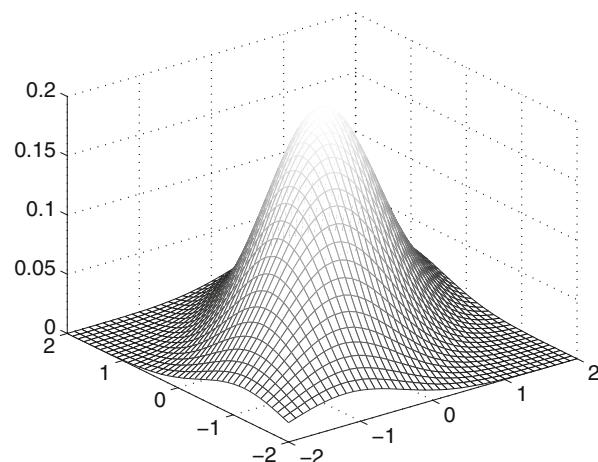
We can produce a contour plot by using the function `contour`:

```
contour(x,y,z);
```

The result is shown in Fig. A.5.

Fig. A.5 A contour plot of the bivariate normal density function with $\varrho = 0.6$
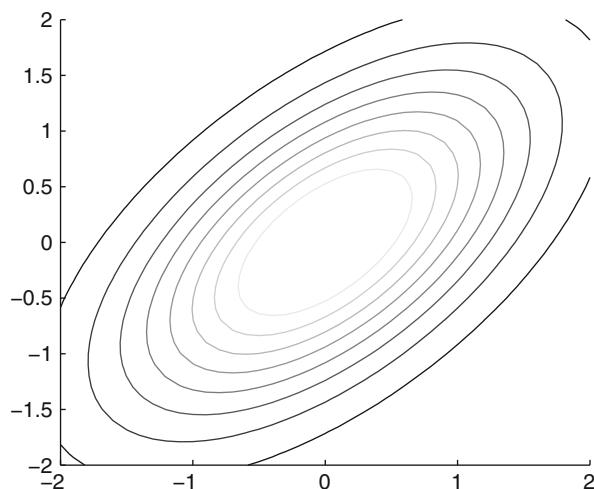
A.6   Optimization Routines

MATLAB provides various built-in optimization routines. In this section we discuss some of them that are used in the main text. Note that all the optimization routines in MATLAB are framed in terms of minimization. In order to perform maximization, some minor changes to the objective function are required. More precisely, suppose we want to maximize the function $f(\mathbf{x})$ and find a maximizer $\mathbf{x}_{\max} = \text{argmax}_{\mathbf{x}} f(\mathbf{x})$. Instead of the original maximization problem, consider minimizing $-f(\mathbf{x})$ and noting that

$$\mathbf{x}_{\max} = \underset{\mathbf{x}}{\text{argmax}}\, f(\mathbf{x}) = \underset{\mathbf{x}}{\text{argmin}}\, -f(\mathbf{x})\ .$$

Hence, without loss of generality, we will focus on minimization routines. One basic minimization function is `fminbnd`, which finds the minimum of a single-variable function on a fixed interval. To illustrate its usage, suppose we wish to minimize the function $f(x) = \sin(x^2)$ over the interval $[0, 3]$ (see Fig. A.6).

Fig. A.6 A plot of
$f(x) = \sin(x^2)$ from 0 to 3
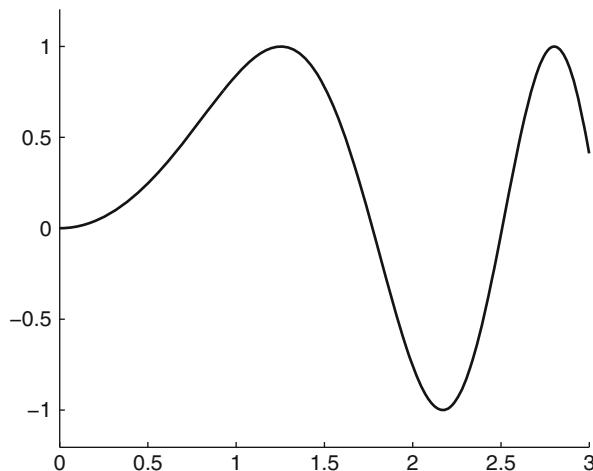
After defining the function $f(x) = \sin(x^2)$ using the command

```
f = @(x) sin(x.^2);
```

we pass f to fminbnd, which takes three inputs (the function handle, lower and upper bounds of the interval) and gives two outputs (the minimizer and value of the function evaluated at the minimizer):

```
[xmin fmin] = fminbnd(f,0,3);
```

For this example, we have

```
[xmin fmin]

ans =

    2.1708    -1.0000
```

The function fminbnd can only be used to minimize univariate functions on a closed interval. For multivariate minimization, one very useful function is fminsearch that finds the unconstrained minimum of a function of several variables. fminsearch takes two inputs, namely, the function handle and a starting value. Like fminbnd, fminsearch gives two outputs: the minimizer and the minimum (the value of the function evaluated at the minimizer). As an example, suppose we wish to maximize the bivariate normal pdf

$$f(x_1, x_2; \varrho) = \frac{1}{2\pi \sqrt{1 - \varrho^2}} \, e^{-\frac{1}{2(1-\varrho^2)}(x_1^2 - 2\varrho x_1 x_2 + y_2^2)}$$

with respect to $\mathbf{x} = (x_1, x_2)$ with $\varrho = 0.6$. To this end, first define $g(x_1, x_2) = -f(x_1, x_2; \varrho)$:

```
rho = .6;
g = @(x) -1/(2*pi*sqrt(1-rho^2)) ...
    *exp(-(x(1).^2 -2*rho*x(1).*x(2) +x(2).^2)
    /(2*(1-rho^2)));
```

Note that the variable x is a 1×2 vector. Then, we pass g to fminsearch with starting values, say, $[1, -1]$:

```
[xmin gmin] = fminsearch(g, [1 -1]);
```

For this example, we have

```
[xmin gmin]

ans =

    0.0000     0.0000    -0.1989
```

That is, the mode of $f(x_1, x_2; \varrho = 0.6)$ is $\mathbf{x} = (0, 0)$, and $f(0, 0) = 0.1989$.

A.7   Handling Sparse Matrices

A **sparse matrix** is simply a matrix that contains a large proportion of zeros. Computation for sparse matrices can typically be done much faster than for full matrices. In addition, as most of the elements in a sparse matrix are zeros, the storage cost of a sparse matrix is also small. In statistics we often need to deal with large sparse matrices. Thus it is useful to learn how to handle them in MATLAB.

A basic function for creating sparse matrices is sparse. For example, suppose the matrix

$$W = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 \end{pmatrix}$$

is stored as a full matrix in MATLAB. The command `sparse(W)` converts W to sparse form by squeezing out any zero elements and returns:

```
ans =

   (1,1)        1
   (2,2)        1
   (3,3)        2
   (4,4)        3
   (4,5)        1
```

Notice that only the nonzero elements in W are stored. In general, we can create a matrix S by the command `S = sparse(i,j,s,m,n)`, which uses vectors **i**, **j**, and **s** to generate an $m \times n$ sparse matrix such that $S(\mathbf{i}(k), \mathbf{j}(k)) = \mathbf{s}(k)$. For example, to create the matrix W above, we first need to build a vector **s** that stores all the nonzero elements:

```
s = [1 1 2 3 1]';
```

Next, we create a vector **i** that stores the row position for each element in **s**. For example, the first element in **s** should be in the first row, the second element in second row, and so on. We then do the same thing for the column positions and store them in the vector **j**:

```
i = [1 2 3 4 4]';
j = [1 2 3 4 5]';
```

Finally,

```
W = sparse(i,j,s,4,5);
```

creates the 4×5 matrix W above.

There are several useful built-in functions for creating special sparse matrices. For example,

```
I = speye(100);
```

creates the 100×100 sparse identity matrix. Of course we can accomplish the same goal by using

```
I = sparse(1:100,1:100,ones(1,100));
```

though the latter is more clumsy. Another useful function is `spdiags`, the sparse version of `diag`, which can be used to extract and create sparse diagonal matrices. Use `help spdiags` to learn more about this function.

As mentioned earlier, one main advantage of working with sparse rather than full matrices is that computations involving sparse matrices are usually much quicker. For instance, it takes about 2.7 seconds to obtain the Cholesky decomposition of the full 5000×5000 identity matrix:

```
tic; chol(eye(5000)); toc;
Elapsed time is 2.728245 seconds.
```

whereas the same operation takes only 0.015 second for a sparse 5000×5000 identity matrix:

```
tic; chol(speye(5000)); toc;
Elapsed time is 0.014867 seconds.
```

A.8   Gamma and Dirichlet Generator

The following MATLAB program `gamrand` implements the method developed in (Marsaglia and Tsang 2000) to generate samples from a Gamma(α, λ) distribution. If the *Statistics Toolbox* is available, the function `gamrnd` can be used instead; but note that `gamrnd(a,b)` generates random variables from a Gamma$(a, 1/b)$ distribution.

```
function x=gamrand(alpha,lambda)
if alpha>1
    d=alpha-1/3; c=1/sqrt(9*d); flag=1;
    while flag
        Z=randn;
        if Z>-1/c
            V=(1+c*Z)^3; U=rand;
            flag=log(U)>(0.5*Z^2+d-d*V+d*log(V));
        end
    end
```

```
    x=d*V/lambda;
else
    x=gamrand(alpha+1,lambda);
    x=x*rand^(1/alpha);
end
```

As a direct consequence of Theorem 8.2, the following MATLAB program ☞ 241
dirichrnd generates samples from a Dirichlet(α) distribution. Draws from a
Beta(α, β) are obtained by taking $\alpha = (\alpha, \beta)$.

```
function x=dirichrnd(alpha)
n=length(alpha)-1;
Y=nan(1,n+1);
for k=1:n+1
    Y(k)=gamrand(alpha(k),1);
end
x=Y(1:n)/sum(Y);
```

A.9   Cdfs and Inverse Cdfs

The following MATLAB program cumdf evaluates the cdfs of normal, Student's t,
gamma, chi-squared, and F distributions. If the *Statistics Toolbox* is available, the
function cdf can be used instead.

```
function y = cumdf(dist,x,varargin)
switch dist
    case 'norm'
        mu = varargin{1}; sigma = varargin{2};
        y = (erf((  (x - mu)/sigma )/sqrt(2)) + 1)/2;
    case 't'
        nu = varargin{1};
        y = 1-0.5*betainc(nu/(nu+x.^2),nu/2,1/2);
    case 'gamma'
        alpha = varargin{1}; lambda = varargin{2};
            % different from Stats toolbox
        y = gammainc(lambda*x,alpha);
    case 'chi2'
        n = varargin{1};
        y = gammainc(x/2,n/2);
    case 'F'
        m = varargin{1}; n = varargin{2};
        y = 1 - betainc(n/(n+m*x),n/2,m/2);
end
```

The following MATLAB program `icumdf` evaluates the inverse cdfs of normal, Student's t, gamma, chi-squared, and F distributions. The corresponding built-in function in the *Statistics Toolbox* is `icdf`.

```
function x = icumdf(dist,y,varargin)
switch dist
    case 'norm'
        mu = varargin{1}; sigma = varargin{2};
        x = mu  + sigma*sqrt(2)*erfinv(2*y -1);
    case 't'
        nu = varargin{1};
        x = sqrt(nu/betaincinv(2*(1-y),nu/2,1/2) - nu);
    case 'gamma'
        alpha = varargin{1}; lambda = varargin{2};
            % different from Stats toolbox
        x = gammaincinv(y,alpha)/lambda;
    case 'chi2'
        n = varargin{1};
        x = gammaincinv(y,n/2)*2;
    case 'F'
        m = varargin{1}; n = varargin{2};
        x = n/m/betaincinv(1-y,n/2,m/2) - n/m;
end
```

A.10   Further Reading and References

The official MATLAB documentation site is

http://www.mathworks.com/help/techdoc/

A good place to learn more about the major functionality in MATLAB is the MATLAB *Getting Started Guide* available at

http://www.mathworks.com/help/pdf_doc/matlab/getstart.pdf

MATLAB programs for generating random variables from a wide range of distributions may be found on the homepage of the *Handbook of Monte Carlo Methods* (Kroese et al. 2011):

http://www.montecarlohandbook.org

Finally, all programs and (large) data files in this book may be downloaded from the homepage

http://www.statmodcomp.org

To accommodate the users of the statistical programming language R, we have mirrored each MATLAB program with its equivalent in R.

Appendix B
Mathematical Supplement

B.1 Multivariate Differentiation

For a real-valued multivariate function $f(x_1, \ldots, x_n)$ the **partial derivative** with respect to x_i, denoted $\frac{\partial f}{\partial x_i}$ or simply $\partial_i f$, is the derivative taken with respect to x_i while all other variables are held constant. The partial derivative of $\partial_i f$ with respect to x_j is denoted $\frac{\partial^2 f}{\partial x_i \, \partial x_j}$ or simply $\partial_{ij} f$.

Let **f** be a multivariate function taking values in $\mathbb{R}^m$, defined by

$$
\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} = \mathbf{f}(\mathbf{x}) \ .
$$

The **derivative** of **f** at **x** is defined as the matrix of partial derivatives,

$$
J_{\mathbf{f}}(\mathbf{x}) = \begin{pmatrix} \partial_1 f_1(\mathbf{x}) & \cdots & \partial_n f_1(\mathbf{x}) \\ \vdots & \ldots & \vdots \\ \partial_1 f_m(\mathbf{x}) & \cdots & \partial_n f_m(\mathbf{x}) \end{pmatrix} , \tag{B.1}
$$

and is called the **matrix of Jacobi** of **f** at **x**, sometimes written as $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x})$.

Example B.1 (Differentiating a Linear Function). Let $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ for some $m \times n$ constant matrix A. Then,

$$
\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = A \ . \tag{B.2}
$$

To see this, let a_{ij} denote the (i, j)th element of A, so that

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k} x_k \\ \vdots \\ \sum_{k=1}^{n} a_{mk} x_k \end{pmatrix}.$$

To find the (i, j)th element of the $m \times n$ Jacobian matrix $\mathbf{J_f}$, we differentiate the ith element of $\mathbf{f}$ with respect to x_j:

$$\frac{\partial f_i(\mathbf{x})}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} a_{ik} x_k = a_{ij} .$$

In other words, the (i, j)th element of $\mathbf{J_f}$ is a_{ij}, the (i, j)th element of A.

For a real-valued multivariate function, that is, $f : \mathbb{R}^n \to \mathbb{R}$, the **gradient** of f is the transpose of the Jacobian matrix, that is, the *column* vector

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \partial_1 f(\mathbf{x}) \\ \vdots \\ \partial_n f(\mathbf{x}) \end{pmatrix} . \tag{B.3}$$

The derivative of the function $\mathbf{x} \mapsto \nabla f(\mathbf{x})$ is called the **Hessian matrix** of f, denoted $H_f(\mathbf{x})$ or $\nabla^2 f(\mathbf{x})$. In other words, the Hessian is the matrix of second derivatives:

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \partial_{11} f(\mathbf{x}) & \cdots & \partial_{1n} f(\mathbf{x}) \\ \vdots & \cdots & \vdots \\ \partial_{n1} f(\mathbf{x}) & \cdots & \partial_{nn} f(\mathbf{x}) \end{pmatrix} . \tag{B.4}$$

If the partial derivatives are *continuous* in a region around $\mathbf{x}$, then $\partial_{ij} f(\mathbf{x}) = \partial_{ji} f(\mathbf{x})$ and, hence, the Hessian matrix $H_f(\mathbf{x})$ is *symmetric*.

Example B.2 (Differentiating a Quadratic Function). Let $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ for some $n \times n$ constant matrix A. Then,

$$\nabla f(\mathbf{x}) = (A + A^\top)\mathbf{x} . \tag{B.5}$$

It follows immediately that if A is *symmetric*, i.e., $A = A^\top$, then $\nabla(\mathbf{x}^\top A \mathbf{x}) = 2A\mathbf{x}$ and $\nabla^2 (\mathbf{x}^\top A \mathbf{x}) = 2A$.

To prove (B.5), first, observe that the quadratic function $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ is real-valued, and therefore the Jacobian $\mathbf{J}_f$ is a $1 \times n$ vector (and its transpose is the gradient). Specifically,

$$f(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\, x_i x_j \ ,$$

and the kth element of $\mathbf{J}_f$ is obtained by differentiating $f(\mathbf{x})$ with respect to x_k:

$$\frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i \ .$$

The first term on the right-hand side is equal to the kth element of $A\mathbf{x}$, whereas the second term equals the kth element of $\mathbf{x}^{\top} A$, or equivalently the kth element of $A^{\top}\mathbf{x}$.

Gradients and Hessian matrices feature prominently in multidimensional Taylor expansions.

Theorem B.1. (Multidimensional Taylor Expansions). Let $\mathscr{X}$ be an open subset of $\mathbb{R}^n$ and let $\mathbf{a} \in \mathscr{X}$. If $f : \mathscr{X} \to \mathbb{R}$ is a continuously twice differentiable function with gradient $\nabla f(\mathbf{x})$ and Hessian matrix $H_f(\mathbf{x})$, then for every $\mathbf{x} \in \mathscr{X}$ we have the following first- and second-order Taylor expansions:

$$f(\mathbf{x}) = f(\mathbf{a}) + [\nabla f(\mathbf{a})]^{\top}(\mathbf{x} - \mathbf{a}) + \mathcal{O}(\|\mathbf{x} - \mathbf{a}\|^2)$$

and

$$f(\mathbf{x}) = f(\mathbf{a}) + [\nabla f(\mathbf{a})]^{\top}(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^{\top} H_f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathcal{O}(\|\mathbf{x} - \mathbf{a}\|^3)$$

as $\|\mathbf{x} - \mathbf{a}\| \to 0$. By dropping the $\mathcal{O}$ remainder terms, one obtains the corresponding Taylor approximations.

B.2   Proof of Theorem 2.6 and Corollary 2.2

The proof makes use of two fundamental properties of the expectation $\mathbb{E}$: the *monotone convergence theorem* and the *dominated convergence theorem*. The first states that if $X_1 \leqslant X_2 \leqslant X_3 \leqslant \ldots$ is a sequence of positive random variables that increases to a random variable X, then the corresponding expectations $\mathbb{E}X_1 \leqslant \mathbb{E}X_2 \leqslant \mathbb{E}X_3 \ldots$ converge to $\mathbb{E}X$. The second theorem states that the same holds true for any positive sequence $X_1, X_2, \ldots$ converging to X, if there exists a Y with $\mathbb{E}Y < \infty$ such that $X_n \leqslant Y$ for all n. An accessible account of these theorems may be found, for example, in (Williams 1991).

We prove Theorem 2.6 for the case $k = 1$ only. Let $G(z) = \mathbb{E}z^X$. Take a fixed z with $|z| < R$ and any $r < R$ such that $r < |z| < R$. Let (h_n) be any sequence converging to 0 such that $|z + h_n| < r$. By definition, the derivative of G at z is $\lim_{n\to\infty} \mathbb{E}C_n$, where $C_n = h_n^{-1}[(z + h_n)^X - z^X]$. Observe that

1. $|C_n|$ is dominated by $X r^{X-1}$,
2. $\mathbb{E}X r^{X-1} < \infty$, because the power series $\sum_{x=0}^{\infty} xz^{x-1} f(x)$ has again radius of convergence R,
3. $\lim_{n\to\infty} C_n = Xz^{X-1}$.

It follows by the dominated convergence theorem that

$$\lim_{n\to\infty} \mathbb{E}C_n = \mathbb{E} \lim_{n\to\infty} C_n = \mathbb{E}X z^{X-1} \ .$$

Next, let (z_n) be a sequence of real numbers that is converging to 1, where $|z_n| < 1$ for all n. The sequence of random variables (Y_n) defined by $Y_n = X(X - 1)\cdots(X - k + 1)z_n^k$ is increasing to $Y = X(X - 1)\cdots(X - k + 1)$. Hence, by the monotone convergence theorem $\lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}Y$. This shows (2.10). The second statement of the corollary is left as an exercise.

B.3   Proof of Theorem 2.7

If the moment generating function of a random variable X is finite in an open interval containing 0, then for all $n = 0, 1, \ldots$,

$$\mathbb{E}X^n = M^{(n)}(0) \ ,$$

where $M^{(n)}$ is the nth derivative of the MGF M evaluated at 0.

Proof. Let $R > 0$ be such that $M(s) < \infty$ for all $|s| < R$. Choose any numbers r and s such that $0 < r < R$ and $|s| < r$. Let (h_n) be a sequence converging to 0 satisfying $|h_n| < \varepsilon$ and $|s+h_n| < r$ for some $\varepsilon > 0$. Let $C_n = h_n^{-1}[e^{(s+h_n)X} - e^{sX}] = e^{sX}(e^{h_n X} - 1)/h_n$, which converges to Xe^{sX}. Also, $|C_n| \leqslant H(X) \overset{\text{def}}{=} e^{(|s|+\varepsilon)|X|}|X|$, because $0 \leqslant (e^t - 1)/t \leqslant e^{|t|}$ for all t. Moreover, because $|s| + \varepsilon < r$ and x grows at a lesser rate than e^{ax} for any $a > 0$, there must exist an $M > 0$ such that for all $|x| > M$, $H(x) < e^{r|x|}$. It follows that

$$\mathbb{E}H(X) \leqslant \mathbb{E}H(X)\,\mathrm{I}_{\{|X|>M\}} + \mathbb{E}H(X)\,\mathrm{I}_{\{|X|\leqslant M\}}$$

$$\leqslant \mathbb{E}e^{r|X|} + \max_{|x|\leqslant M} H(x) < \infty \ .$$

By the dominated convergence theorem we have $M'(s) = \lim_{n\to\infty} \mathbb{E}C_n = \mathbb{E} \lim_{n\to\infty} C_n = \mathbb{E}[Xe^{sX}]$. Finally, take a monotone sequence (s_n) converging to 0 and apply the monotone convergence theorem to the sequence $(Xe^{s_n X})$ to find $M'(0) = \mathbb{E}X$. The proof for higher moments is similar. $\square$

B.4   Proof of Theorem 3.10

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ spans $\mathcal{V}_k$ and $\mathbf{v}_1, \ldots, \mathbf{v}_m$ spans $\mathcal{V}_m$. We can write the orthogonal projection matrices onto $\mathcal{V}_j$, as $P_j = \sum_{i=1}^{j} \mathbf{v}_i \mathbf{v}_i^\top$, $j = k, m, n$, where $\mathcal{V}_n$ is defined as $\mathbb{R}^n$. Note that P_n is simply the identity matrix. Let $V = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ and define $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top = V^\top \mathbf{X}$. Recall that any orthogonal transformation such as $\mathbf{z} = V^\top \mathbf{x}$ is *length preserving*; that is, $\|\mathbf{z}\| = \|\mathbf{x}\|$.

To prove the first statement of the theorem, note that $V^\top \mathbf{X}_j = V^\top P_j \mathbf{X} = (Z_1, \ldots, Z_j, 0, \ldots, 0)^\top$, $j = k, m$. It follows that $V^\top (\mathbf{X}_m - \mathbf{X}_k) = (0, \ldots, 0, Z_{k+1}, \ldots, Z_m, 0, \ldots, 0)^\top$ and $V^\top (\mathbf{X} - \mathbf{X}_m) = (0, \ldots, 0, Z_{m+1}, \ldots, Z_n)^\top$. Moreover, being a linear transformation of a normal random vector, $\mathbf{Z}$ is also normal, with covariance matrix $V^\top V = I$; see also Problem 3.13. In particular, the $\{Z_i\}$ are *independent*. This shows that $\mathbf{X}_k$, $\mathbf{X}_m - \mathbf{X}_k$, and $\mathbf{X} - \mathbf{X}_m$ are independent as well. ☞ 95

Next, observe that $\|\mathbf{X}_k\| = \|V^\top \mathbf{X}_k\| = \|\mathbf{Z}_k\|$, where $\mathbf{Z}_k = (Z_1, \ldots, Z_k)^\top$. The latter vector has independent components with variances 1, and its squared norm has therefore (by definition) a $\chi_k^2(\theta)$ distribution. The noncentrality parameter is $\theta = \|\mathbb{E}\mathbf{Z}_k\| = \|\mathbb{E}\mathbf{X}_k\| = \|\boldsymbol{\mu}_k\|$, again by the length-preserving property of orthogonal transformations. This shows that $\|\mathbf{X}_k\|^2 \sim \chi_k^2(\|\boldsymbol{\mu}_k\|)$. The distributions of $\|\mathbf{X}_m - \mathbf{X}_k\|^2$ and $\|\mathbf{X} - \mathbf{X}_m\|^2$ follow by analogy. □

B.5   Proof of Theorem 5.2

First, observe that, by Theorem 5.1, ☞ 131

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi_{m-1}^2 \quad \text{and} \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi_{n-1}^2 .$$

Because these random variables are independent of each other, their sum, V, say, can be written as the sum of $m + n$ independent squared standard normal random variables and has therefore a χ_{m+n-2}^2 distribution. Thus,

$$V = \frac{(m+n-2)S_p^2}{\sigma^2} \sim \chi_{m+n-2}^2 .$$

Second, let

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma / \sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Then, $Z \sim \mathsf{N}(0, 1)$ and the square of the pivot T in Theorem 5.2 can be written as

$$T^2 = \frac{Z^2}{V/(m + n - 2)} \, ,$$

where Z and V are independent, because $\bar{X}$ and $\bar{Y}$ are independent of each other, and are both independent of S_X^2 and S_Y^2; see Theorem 5.1. The random variable T^2 is thus the independent quotient of a χ_1^2 and a χ_{m+n-2}^2 random variable. Hence, by Theorem 3.11, $T^2 \sim \mathsf{F}(1, m + n - 2)$. It follows now from Theorem 2.19 (and the fact that the pdf of T is symmetric around 0) that $T \sim \mathsf{t}_{m+n-2}$.   □

References

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc., Secaucus, NJ.

Botev, Z. I., J. F. Grotowski, & D. P. Kroese 2010. Kernel density estimation via diffusion. Annals of Statistics, 38(5):2916–2957.

Chib, S. 1995. Marginal Likelihood from the Gibbs Output. Journal of the American Statistical Association, 90:1313–1321.

Chib, S., & I. Jeliazkov 2001. Marginal Likelihood from the Metropolis-Hastings Output. Journal of the American Statistical Association, 96:270–281.

Fair, R. C. 1978. A Theory of Extramarital Affairs. Journal of Political Economy, 86:45–61.

Feller, W. 1970. An Introduction to Probability Theory and Its Applications, volume I. John Wiley & Sons, New York, second edition.

Gelfand, A. E., S. Hills, A. Racine-Poon, & A. F. M. Smith 1990. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. Journal of American Statistical Association, 85:972–985.

Kim, S., N. Shepherd, & S. Chib 1998. Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. Review of Economic Studies, 65(3):361–393.

Koop, G., D. J. Poirier, J. L., & Tobias 2007. Bayesian Econometric Methods. Cambridge University Press.

Kroese, D. P., T. Taimre, & Z. I. Botev 2011. Handbook of Monte Carlo Methods. John Wiley & Sons, New York.

L'Ecuyer, P. 1999. Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators. Operations Research, 47(1):159 – 164.

Marsaglia, G., & W. Tsang 2000. A Simple Method for Generating Gamma Variables. ACM Transactions on Mathematical Software, 26(3):363–372.

McLachlan, G. J., & T. Krishnan 2008. The EM Algorithm and Extensions. John Wiley & Sons, Hoboken, NJ, second edition.

Metropolis, M., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, & E. Teller 1953. Equations of state calculations by fast computing machines. J. of Chemical Physics, 21:1087–1092.

Verdinelli, I., & L. Wasserman 1995. Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio. Journal of the American Statistical Association, 90(430): 614–618.

Williams, D. 1991. Probability with Martingales. Cambridge University Press, Cambridge.

Solutions

Selected Problems of Chap. 1

1.1 (a) $\Omega = \{1, 2, 3, 4, 5, 6\}$; (b) $\Omega = \mathbb{R}_+$; (c) $\Omega = \{0, 1, \dots \}$; (d) $\Omega = \{0, 1, \dots, 50\}$; (e) $\Omega = \{(x_1, \dots, x_{10}) : x_i \geqslant 0, i = 1, \dots, 10\} = \mathbb{R}_+^{10}$.

1.4 (a) $1/5$; (b) $5/36$.

1.5 (a) Ω is the set of all 6! permutations of $(1, \dots, 6)$; (b) $\mathbb{P}(A) = |A|/720$; (c) $15/720$.

1.8 (a) $\Omega = \{(1, 2, 3), \dots, (52, 51, 50)\}$. Each elementary event is equally likely; (b) $\frac{4 \times 3 \times 2}{52 \times 51 \times 50} = \frac{3}{16575}$; (c) $\frac{6 \times 4^4}{52 \times 51 \times 50} = 64/5525$; (d) $\frac{36 \times 35 \times 34}{52 \times 51 \times 50} = \frac{1071}{3315}$.

1.9 $\frac{\binom{17}{7}}{\binom{20}{10}} = \frac{2}{19}$.

1.10 (a) $\frac{1}{6}$; (b) $\frac{1}{3}$.

1.13 $\frac{1}{365^2}$.

1.14 $\binom{10}{4} \times 0.4^4 \times 0.6^6 = 0.2508$.

1.15 (a) 0.0791; (b) 0.1239.

1.17 0.01.

1.19 (a) $\frac{1}{36} \times \left(\frac{35}{36}\right)^9 = 0.0216$; (b) $1 - (35/36)^{100} = 0.94022$.

1.21 `ceil(6*rand(1,100))`.

Selected Problems of Chap. 2

2.1 (a)

x	1	2	3	4	5	6
$f(x)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

;   (b) $\frac{4}{9}$; (c) $\mathbb{E}M = \frac{91}{36}$, $\mathrm{Var}(M) = \frac{2555}{1296}$.

2.2 (b): (i) 4/5; (ii) 3/5; (iii) 3/5; (iv) 19/20.

2.6 $X \sim \mathsf{Bin}(100, 0.12)$; $\mathbb{P}(X \leqslant 7) = \sum_{k=0}^{7} \binom{100}{k} 0.12^k (1 - 0.12)^{100-k} = 0.0761$.

2.8 $M(s) = \frac{e^{bs} - e^{as}}{s(b-a)}$, $s \in \mathbb{R}$.

2.11 (a) $1 - e^{-2}$; (b) e^{-8}; (c) e^{-4}; (d) $1/2$.

2.12 (a) The expectation does not exist ($\infty - \infty$ is ill-defined); (b) the expectation is ∞.

2.16 In the first model $X \sim \mathsf{Exp}(1/3)$ and $\mathbb{P}(X > 4.5 \mid X > 4) = 0.8465$. In the second model $X \sim \mathsf{N}(3, 9)$ and $\mathbb{P}(X > 4.5 \mid X > 4) = 0.8351$.

2.18 (b): (i) $\Phi(-1/3)$, (ii) $1 - \Phi(0) = 1/2$, (iii) $\Phi(1/3) - \Phi(-5/3)$; (c) 9; (d) 25.

2.20 (a) `2+rand`; (b) `3+3*randn`; (c) `-log(rand)/4`; (d) `sum(rand(10,1)<0.5)`; (e) `ceil(log(rand)/(1-1/6))`.

2.21 Use `X = sqrt(-log(rand(1000,1)));  hist(X)`.

Selected Problems of Chap. 3

3.1

x / y

	-2	0	2	
-1	0	$\frac{1}{2}$	0	$\frac{1}{2}$
1	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

X and Y are not independent since, for example, $\mathbb{P}(X = -1, Y = -2) = 0 \neq \frac{1}{2} \times \frac{1}{4} = \mathbb{P}(X = -1)\mathbb{P}(Y = -2)$.

3.3 (a) $f_X(x) = \frac{1}{3}$ for $x = 1, 2, 3$; (b) $f_{Y|X}(y \mid 1) = \frac{1}{2}$ for $y = 0, 1$, $f_{Y|X}(y \mid 2) = \frac{4-3(y-1)^2}{6}$ for $y = 0, 1, 2$, $f_{Y|X}(y \mid 3) = \frac{1}{2}$ for $y = 1, 2$;

(c) x / y

	0	1	2	
1	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{3}$
2	$\frac{1}{18}$	$\frac{2}{9}$	$\frac{1}{18}$	$\frac{1}{3}$
3	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{2}{9}$	1

; (d) $f_Y(0) = f_Y(2) = 2/9$ and $f_Y(1) = 5/9$; (e)

$$f_{X|Y}(x \mid 0) = \begin{cases} \frac{3}{4}, & x = 1 \\ \frac{1}{4}, & x = 2, \end{cases}, \quad f_{X|Y}(x \mid 1) = \begin{cases} \frac{3}{10}, & x = 1 \\ \frac{2}{5}, & x = 2 \\ \frac{3}{10}, & x = 3, \end{cases}$$

$$f_{X|Y}(x \mid 2) = \begin{cases} \frac{1}{4}, & x = 2 \\ \frac{3}{4}, & x = 3 \end{cases}.$$

3.4 $f_{X|Y}(x \mid 1) = \frac{1/(6x)}{147/360} = \frac{60}{147x}$ for $x = 1, \ldots, 6$, and $\mathbb{E}[X \mid Y = 1] = \sum_{x=1}^{6} x \times \frac{60}{147x} = \frac{360}{147}$.

3.6 (a) $f(x, y) = f_X(x) f_Y(y) = e^{-y}$ for $0 \leqslant x \leqslant 1$, and $y \geqslant 0$; (b) $1 - e^{-1}$; (c) e^{-1}.

3.8 (a) $f(x, y) = f_X(x) f_{Y|X}(y \mid x) = e^{-x} \times x e^{-xy} = x e^{-x(y+1)}$ for $x > 0, y > 0$; (b) $f_Y(y) = \frac{1}{(y+1)^2}, y \geqslant 0$.

3.9 Since $X \sim \mathsf{U}[-\pi/2, \pi/2]$, $f_X(x) = 1/\pi$, $x \in (-\pi/2, \pi/2)$. Let $Y = \tan(X)$. Then, the inverse transformation is $g^{-1}(y) = \arctan(x)$ and the associated matrix of Jacobi is $J_{g^{-1}}(y) = 1/(1 + y^2)$. Hence, $f_Y(y) = \frac{1}{\pi(1+y^2)}$, which is the pdf of the Cauchy distribution.

3.11 (a) $f_{S_2}(x) = x$ for $0 \leqslant x \leqslant 1$, $f_{S_2}(x) = 2 - x$ for $1 < x \leqslant 2$, and 0 otherwise; (b) $\mathsf{N}(10, 5/3)$; (c) 0.0607.

3.12 (a) $\mathbb{E}T = 5, \mathrm{Var}(T) = \frac{5}{2}$; (b) 0.2635; (c) $T \sim \mathsf{Gamma}(10, 2)$.

3.15 (a) $\left(\frac{7}{8}\right)^6$; (b) $1 - \left(\frac{7}{8}\right)^6$.

3.18 (b) $\frac{(n-1)}{n} \sigma^2$.

3.20 W is the sum of the weights of 6 randomly chosen people and has a $\mathsf{N}(600, 600)$ distribution. $6X_1$ is 6 times the weight of the first chosen person and has a $\mathsf{N}(600, 3600)$ distribution.

3.21 Recall that if $Z \sim \chi_\nu^2$, then its moment generating function is $M_Z(s) = 1/(1 - 2s)^{\nu/2} = (1 - 2s)^{-\nu/2}$. Now, $M_{X+Y}(s) = \mathbb{E}e^{s(X+Y)} = \mathbb{E}e^{sX} \mathbb{E}e^{sY} = (1 - 2s)^{-m/2}(1 - 2s)^{-n/2} = (1 - 2s)^{-(m+n)/2}$, which is the moment generating function of χ_{m+n}^2.

3.23 0.9867.

3.24 (a) 1.0524; (b) 0.0838.

Selected Problems of Chap. 4

4.1

(a) $X_1, \ldots, X_9 \sim_{\text{iid}} N(\mu, \sigma^2)$, where X_i is the volume of paint in the ith tin. The primary interest is to determine if μ is less than 20.

(b) $X_1, \ldots, X_{12} \sim_{\text{iid}} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{12} \sim_{\text{iid}} N(\mu_2, \sigma_2^2)$ independently. Here, X_i (Y_i) is the time of completion for the ith man (woman). The primary interest is to determine if $\mu_1 - \mu_2$ is significantly different from 0 or not.

(c) $Z_1, \ldots, Z_{12} \sim_{\text{iid}} N(\mu, \sigma^2)$, where Z_i is the difference in marks for the ith exam as marked by lecturers A and B. The primary interest is to determine if μ is significantly different from 0 or not.

(d) $X_1, \ldots, X_{500} \sim_{\text{iid}} \text{Ber}(p)$, where $X_i = 1$ if the ith coin toss is Heads and $X_i = 0$ otherwise. The primary interest is to determine if p equals 1/2.

4.2

(a) Let X_i be the *average* weight of 5 randomly selected packets from the packaging line at hour i, $i = 1, \ldots, 24$. A possible model is to assume that the $\{X_i\}$ are independent and that each $X_i \sim N(\mu_i, \sigma_i^2)$ for some unknown μ_i and σ_i^2. Typical questions of interest are whether the $\{\mu_i\}$ and $\{\sigma_i^2\}$ lie within an acceptable range.

4.7

(a) If one expects that shipping cost is a linear or quadratic function in distance, then a possible model is $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $\{\varepsilon_i\} \sim_{\text{iid}} N(0, \sigma^2)$, $i = 1, \ldots, 9$, where Y_i is the shipping cost of the ith air freight and x_i the corresponding distance traveled.

(b) Single-factor ANOVA model $Y_{ik} = \mu_i + \varepsilon_{ik}$, $\{\varepsilon_{ik}\} \sim_{\text{iid}} N(0, \sigma^2)$, $k = 1, \ldots, 20$, $i = 1, 2, 3$, where Y_{ik} is the average fuel consumption for the kth car of brand i, $k = 1, \ldots, 20$, $i = 1, 2, 3$.

(d) Simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\{\varepsilon_i\} \sim_{\text{iid}} N(0, \sigma^2)$, $i = 1, \ldots, 10$, where Y_i is the military expenditure of the country in year i, and x_i is the gross national product in that year.

4.9 The $n \times 6$ design matrix is given by

$$
\begin{pmatrix}
1 & x_{11} & x_{21} & x_{11}^2 & x_{21}^2 & x_{11}x_{21} \\
1 & x_{12} & x_{22} & x_{12}^2 & x_{22}^2 & x_{12}x_{22} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{1n} & x_{2n} & x_{1n}^2 & x_{2n}^2 & x_{1n}x_{2n}
\end{pmatrix}.
$$

4.11 The following MATLAB script generates realizations from the ANOVA model.

```
%p4_11.m
beta= [-1 0 1]; % 2 free parameters
alpha = [-2 2]; % 1 free parameter
```

```
mu = 6; % 1 parameter
gamma = [0.2 -1 0.8; -0.2 1 -0.8]; %2 free parameters
eps = 0.1*randn(2,3,3);
y = zeros(2,3,3);
for i=1:2 % generate the responses
 for j=1:3
  y(i,j,:) = mu + alpha(i) + beta(j) + gamma(i,j) + eps(i,j,:);
 end
end
for i=1:2 % print the table
  for j=1:3
    for k=1:3
        fprintf('%3.2f, ',y(i,j,k))
    end
    fprintf('\n');
  end
end
```

Selected Problems of Chap. 5

5.1 For $\mathsf{Geom}(p)$, $\widehat{p} = 1/\bar{x}$; for $\mathsf{Poi}(\lambda)$, $\widehat{\lambda} = \bar{x}$; for $\mathsf{Gamma}(\alpha, \lambda)$, $\widehat{\alpha} = \bar{x}^2/v^2$ and $\widehat{\lambda} = \bar{x}/v^2$, where $v^2 = n^{-1} \sum_{i=1}^{n}(x_i - \bar{x})^2$.

5.5 (a) $\widehat{\lambda} = \bar{X}$; (b) from the central limit theorem, $\bar{X}$ has approximately a normal distribution with expectation $\mathbb{E}\bar{X} = 1/\lambda$ and variance $\mathrm{Var}(\bar{X}) = 1/(\lambda^2 n)$, so that

$$\mathbb{P}\left(\frac{1 - z_{1-\alpha/2}/\sqrt{n}}{\bar{X}} \leqslant \lambda \leqslant \frac{1 + z_{1-\alpha/2}/\sqrt{n}}{\bar{X}}\right) \approx 1 - \alpha,$$

from which the $1 - \alpha$ approximate confidence interval for λ follows.

5.7 $(0.045, 0.055)$ ml.

5.8 The confidence interval $(1.015, 1.810)$ does not contain the value 1, so there is reasonable evidence to suspect that the claim on the packet is not true.

5.10 Evaluating $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$ gives

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

$$= \frac{1}{n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

$$= \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 n\bar{Y} - n\bar{x} \sum_{i=1}^{n} x_i Y_i \\ -n^2 \bar{x}\bar{Y} + n \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \frac{1}{S_{xx}} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \bar{Y} - \bar{x} \sum_{i=1}^{n} x_i Y_i \\ S_{xY} \end{pmatrix}$$

$$= \frac{1}{S_{xx}} \begin{pmatrix} \bar{Y}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) + \bar{x}(\bar{Y}n\bar{x} - \sum_{i=1} x_i Y_i) \\ S_{xY} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{x} S_{xY}/S_{xx} \\ S_{xY}/S_{xx} \end{pmatrix}.$$

5.12 Model: $Y_i = a\sqrt{h_i} + \varepsilon_i$, $i = 1, \ldots, 4$, where the $\{\varepsilon_i\}$ are iid and $\mathsf{N}(0, \sigma^2)$ distributed. The least-squares estimate of a is $\widehat{a} = 0.452$.

5.13 Let X be the number of low fat milk sales out of 1500. The model is $X \sim$ Bin$(1500, p)$ for some unknown p. We wish to test $H_0 : p = 0.3$ versus $H_1 : p < 0.3$. The outcome of X is $x = 400$. The corresponding p-value is $\mathbb{P}_{H_0}(X \leqslant 400) \approx 0.00243$. There is thus very strong evidence that the true proportion p is less than 0.3, indicating a move toward low fat milk.

5.17 Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, 16)$ be the PFC amounts.

(a) The average PFC is $\bar{X}$, which has a $\mathsf{N}(\mu, 16/n) = \mathsf{N}(\mu, 4)$ distribution.
(b) $\mathbb{P}_{\mu=38.5}(\bar{X} < 39) = \Phi(0.025) = 0.51$.
(c) Find n such that $\mathbb{P}_{\mu=38}(\bar{X} < 39) \geqslant 0.9$. The smallest such n is 27.

5.23 Let Y_{ik} be the walking age of the kth baby in group $i = 1, 2, 3, 4$ (corresponding to A,B,C,D). Consider the 1-factor ANOVA model

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \quad i = 1, \ldots, 4, \quad k = 1, \ldots, 6,$$

with $\{\varepsilon_{ik}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$, and $\sum_{j=1}^{4} \alpha_i = 0$. To test the hypothesis $\alpha_1 = \ldots = \alpha_4 = 0$ we use the test statistic $T = \frac{\text{MS}_{\text{treatment}}}{\text{MS}_{\text{error}}}$, which under H_0 has an $\mathsf{F}(3, 20)$ distribution.

☞ 143 By changing the data matrix yy in the first MATLAB program in Example 5.15 we find the outcome 2.1370 for the test statistic, which gives a p-value of 0.1275. Since this is not very small, we accept the null hypothesis that there is no difference in expected walking age between the groups.

To compute the 95 % confidence intervals for the expected walking ages $\mu_i =$
☞ 138 $\mu + \alpha_i$, $i = 1, \ldots, 4$, we apply Theorem 5.3 using specific vectors **a**. For example, to find a confidence interval for μ_1, take $\mathbf{a} = (1, 1, 0, 0)^\top$. Similarly, for μ_4, take $\mathbf{a} = (1, -1, -1, -1)^\top$. By modifying the MATLAB program linregestconf.m
☞ 139 in Example 5.12 (see the code below), we find the following 95 % numerical confidence intervals, $(8.85, 11.39)$, $(10.10, 12.64)$, $(10.44, 12.97)$, and $(10.94, 13.48)$, which clearly overlap, corroborating our finding that there is no evidence for a difference in expected walking age.

```
%p5_22.m
tquant = icdf('t',0.975,n-m) % 0.975 quantile
a = [1 1 0 0]'; %change for other mu's
ucl = a'*betahat + tquant*norm(y - X*betahat)* ...
sqrt(a'*inv(X'*X)*a)/sqrt(n-m);
lcl = a'*betahat - tquant*norm(y - X*betahat)* ...
sqrt(a'*inv(X'*X)*a)/sqrt(n-m);
display([lcl,ucl])
```

Selected Problems of Chap. 6

6.2 The derivative of the log-likelihood for θ is $\frac{dl}{d\theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}x_i^2 - \frac{n}{2}$. Solving the likelihood equation $\frac{dl}{d\theta} = 0$, one obtains the maximum likelihood estimate $\widehat{\theta} = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{n}\sum_{i=1}^{n}x_i^2}$. Substitute X_i for x_i to obtain the estimator.

6.5 (a) $\widehat{\theta}_M = (2\bar{X} - 1)/(1 - \bar{X})$; (b) $\widehat{\theta} = -1 - n/\sum_{i=1}^{n}\ln X_i$.

6.6 $\mathbb{P}(X > 68.5) = \mathbb{P}((X - \mu)/\sigma > (68.5 - \mu)/\sigma) = 1 - \Phi((68.5 - \mu)/\sigma)$. Hence, by Theorem 6.6, the MLE is $1 - \Phi((68.5 - 56.3)/7.6) = 0.0542$.   ☞ 175

6.9 $l(\lambda; \mathbf{x}) = n\ln(\lambda/2) - \lambda\sum_{i=1}^{n}|x_i|$. Setting $l'(\lambda; \mathbf{x}) = 0$ gives $\widehat{\lambda} = n/\sum_{i=1}^{n}|x_i| = 0.5893$.

6.12 $l(p; \mathbf{x}) = \ln p\sum_{i=1}^{n}x_i + \ln(1 - p)\sum_{i=1}^{n}(k - x_i) + \text{const}$, so that $l'(p; \mathbf{X}) = \sum_{i=1}^{n}X_i/p - \sum_{i=1}^{n}(k - X_i)/(1 - p)$. Setting $l'(p; \mathbf{X}) = 0$ gives the maximum likelihood estimator $\widehat{p} = \bar{X}/k$. The information number $I(p)$ is, by Theorem 6.4, equal to $n\mathring{I}(p)$, where $\mathring{I}(p)$ is the information number for $X \sim \text{Bin}(k, p)$. By (6.11) $\mathring{I}(p) = k/(p(1 - p))$. Also, $\text{Var}(\widehat{p}) = \text{Var}(\bar{X}/k) = \text{Var}(X_1)/(nk^2) = kp(1 - p)/(nk^2) = p(1 - p)/(nk) = I^{-1}(p)$. Hence, $\widehat{p}$ attains the Cramér–Rao lower bound.   ☞ 170

6.14

(a) The log-likelihood corresponding to $X \sim \text{Exp}(1/v)$ is $\mathring{l}(v; x) = -\ln(v) - x/v$, with score function $\mathring{S}(v; x) = -1/v + x/v^2$. The score function corresponding to $X_1, \ldots, X_n \sim_{\text{iid}} \text{Exp}(1/v)$ is $S(v; \mathbf{x}) = \sum_{i=1}^{n}\mathring{S}(v; x_i) = -n/v + \sum_{i=1}^{n}x_i/v^2$.

(b) The Fisher information for $X \sim \text{Exp}(1/v)$ is $\mathring{I}(v) = \text{Var}(\mathring{S}(v, \mathbf{X})) = \text{Var}(X)/v^4 = v^2/v^4 = 1/v^2$. The Fisher information corresponding to $X_1, \ldots, X_n \sim_{\text{iid}} \text{Exp}(1/v)$ is $I(v) = n\mathring{I}(v) = n/v^2$.

(c) Setting $S(v; \mathbf{X}) = 0$, we find the MLE $\widehat{v} = \bar{X}$.

(d) $\widehat{\sin(v)} = \sin(\widehat{v}) = \sin(\bar{X})$.

(e) $\text{Var}(\widehat{v}) = \text{Var}(\bar{X}) = \text{Var}(X_1)/n = v^2/n = I^{-1}(v)$.

6.18 The score function is $n/\theta - \sum_{i=1}^{n}X_i$ and the Fisher information is $n\text{Var}(\theta - X_1) = n/\theta^2$. The $1 - \alpha$ stochastic confidence set is thus

$$\left\{\theta : -z_{1-\alpha/2} < \frac{n/\theta - \sum_{i=1}^{n}X_i}{\sqrt{n/\theta^2}} < z_{1-\alpha/2}\right\}$$

$$= \left\{\frac{n - z_{1-\alpha/2}\sqrt{n}}{\sum_{i=1}^{n}X_i} < \theta < \frac{n + z_{1-\alpha/2}\sqrt{n}}{\sum_{i=1}^{n}X_i}\right\},$$

which is an interval. Taking $z_{1-\alpha/2} = z_{0.95} = 1.645$, we find the numeric 90% confidence interval $(0.480, 1.520)$.

6.19 Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from $\mathsf{Exp}(\lambda)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ an iid sample from $\mathsf{Exp}(\mu)$.

(a) The MLEs of λ and μ are, respectively, $\widehat{\lambda} = 1/\bar{X}$ and $1/\bar{Y}$.

(b) Under H_0, the MLE of θ is $\widehat{\theta} = \frac{2n}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}$.

(c) The likelihood ratio is given by

$$\Lambda = \frac{(2n)^{2n}}{\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right)^{2n}} \times \frac{\left(\sum_{i=1}^{n} x_i\right)^{n} \left(\sum_{i=1}^{n} y_i\right)^{n}}{n^{2n}}$$

$$= \frac{2^{2n}}{\left(1 + \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right)^{2n}} \times \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right)^{n}$$

$$= 2^{2n}(1 + T)^{-2n} T^n = 2^{2n} \left(\frac{T}{(1 + T)^2}\right)^{n}.$$

Hence, we can use $T/(1 + T)^2$ as a test statistic and we reject H_0 when the likelihood ratio is "too small," i.e., when $\frac{T}{(1+T)^2} < \alpha$ for some critical value α.

(d) T has approximately a $\mathsf{N}(\lambda/\mu, \sigma^2/n)$ distribution, with

$$\sigma^2 = J \Sigma J^\top = \left(-\frac{\lambda^2}{\mu}, \lambda\right) \begin{pmatrix} \lambda^{-2} & 0 \\ 0 & \mu^{-2} \end{pmatrix} \begin{pmatrix} -\frac{\lambda^2}{\mu} \\ \lambda \end{pmatrix} = \frac{2\lambda^2}{\mu^2}.$$

6.24 Initial guesses for α and λ are obtained via the method of moments: $\alpha_0 = \frac{\bar{x}^2}{s^2}, \lambda_0 = \frac{\bar{x}}{s^2}$. The following MATLAB program implements the Newton–Raphson scheme to find the MLE of α and λ, which are estimated to be $\widehat{\alpha} = 3.9853$ and $\widehat{\lambda} = 0.0696$.

```
%p6_24.m
x = [
29.7679    12.8406   105.3225    46.6101    75.7135    72.0340...
64.1004    33.9008    35.2510    50.9201    29.8086    32.6963...
131.5229   65.3381    29.1369    61.8774    31.0650    54.4877...
103.6889   68.0230    89.6879    30.1994    48.3140    54.4447...
29.2253    27.0242   102.5929    63.7344    43.0354    96.5552]';
n = length(x);
sumlogx = sum(log(x)); sumx = sum(x);
alpt = mean(x)^2/var(x); lamt = mean(x)/var(x); %intl. guess
thetat = [alpt; lamt]
for i=1:5 %just repeat the NR step 5 times
    S = [ n*(log(lamt) - psi(alpt)) + sumlogx; ...
          n*alpt/lamt - sumx ];
    I = n * [psi(1,alpt) -1/lamt; -1/lamt alpt/lamt^2 ];
    thetat = thetat + inv(I)*S % using inv is OK (dim =2)
    alpt = thetat(1); lamt = thetat(2);
end
```

Selected Problems of Chap. 7

7.1

(a) The distribution of T does not depend on σ.

(b) Let $Y_i = X_i/\sigma$. Then $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathsf{N}(0, 1)$ under H_0, and $T = \sqrt{n}\bar{X}/S_X = \sqrt{n}\bar{Y}/S_Y$, which is exactly of the form (5.16). The true p-value is 0.0197 (using
    $\mathtt{cdf('t',-3.5,3)}$ in MATLAB).

(c) The following code uses vectorization.

```
%p7_1.m
xbar_obs = -0.7; s_obs = 0.4; t_obs = 2*xbar_obs/s_obs;
N = 10^5;
x = randn(4,N);
xbar = mean(x,1);
s = std(x,0,1);
t = 2*xbar./s;
count = sum(t <= t_obs);
phat = count/N % estimated p-value
```

7.3 X takes values $x_1, \ldots, x_N$ with probability $1/N$, so the expectation is $\mathbb{E}X = x_1/N + x_2/N + \cdots + x_N/N = \bar{x}$, and the second moment is $\mathbb{E}X^2 = N^{-1}\sum_{i=1}^{N} x_i^2$. The variance is therefore $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = N^{-1}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 = N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$.

7.4

(a) By the product rule, the joint pdf of X and J is $f_{X,J}(x, j) = f_J(j)f_{X|J}(x \mid j) = w_j f_j(x)$ and the marginal pdf of X is $f(x) = \sum_{j=1}^{k} f_{X,J}(x, j) = \sum_{j=1}^{k} w_j f_j(x)$.

(b) First, draw $J \in \{1, \ldots, k\}$ with probabilities $w_1, \ldots, w_k$. Then, given $J = j$, draw X from the pdf f_j.

(c) $\mathbb{E}X = \sum_{j=1}^{k} w_j \mu_j$ and $\mathrm{Var}(X) = \sum_{j=1}^{k} w_j(\sigma_j^2 + \mu_j^2) - (\sum_{j=1}^{k} w_j \mu_j)^2$.

7.8

(a) Solving $\mathbb{P}(X \geqslant m) = \mathrm{e}^{-\lambda m} = 1/2$ gives $m = \ln(2)/\lambda$.

(b) The sample median is $\tilde{x} = 1.4073$. Hence, the estimate is $\tilde{\lambda} = \ln(2)/\tilde{x} = 0.4925$. The maximum likelihood estimate is $1/\bar{x} = 0.2773$.

(c) The following code produces Fig. 1. The pdf of the "median" estimator (solid line) is bimodal and much more spread-out than the pdf of the maximum likelihood estimator.

```
%p7_8.m
xorg = [1.4066, 1.2917,  1.408, 4.2801, 1.2136, 2.7461,...
    11.1076,  0.9247,  5.8833, 10.2513, 3.8285, 3.2116, ...
    0.5451, 0.9896, 1.1602, 7.7723, 0.1702, 0.8907,...
    0.2276,  3.1197, 11.4909,  0.6475, 11.2279, 0.7639]
```

```
n = numel(xorg);
est1org = log(2)/median(xorg)
est2org = 1/mean(xorg)
K= 10000;
for i=1:K
    ind = ceil(n*rand(1,n)); % draw random indices
    x = xorg(ind); % resampled   data
    est1(i) =log(2)/median(x);
    est2(i) =1/mean(x);
end
hold on
kde(est1,2^4)  %choose here a low number of meshpoint
kde(est2,2^8)
```
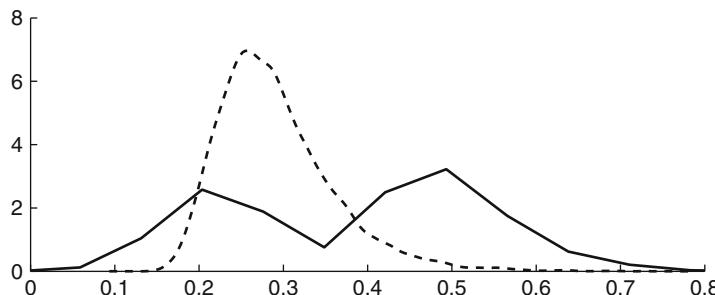

Fig. 1 KDEs for the pdfs of $\ln(2)/\widetilde{X}$ (*solid line*) and $1/\bar{X}$ (*dashed line*)

7.11 All the $x_1, \ldots, x_n$ are smaller than θ, and so are the $M_i^*, k = i, \ldots, K$. Hence, θ is not contained in any $1 - \alpha$ bootstrap confidence interval.

7.13

(a) With positive probability it is possible to reach each state from another state, in at most 4 steps. So the chain is irreducible. However, to return to a starting state, it always requires a multiple of 2 steps. Hence, the chain is periodic with period 2.
(b) The local balance equations hold, because the system is reversible. By symmetry $f(1) = f(2) = f(5) = f(6)$ and $f(3) = f(4)$. By local balance, $f(1)/2 = f(3)/3$. Hence, $4f(1) + 3f(1) = 1$, so that $f(1) = 1/7$ and $f(3) = 3/14$.
(c) For example, for odd t the probability $\mathbb{P}(X_t = 1) = 0$, because, starting from 1 at time 0, it requires an even number of steps to return to 1. On the other hand, the probability $\mathbb{P}(X_{2t} = 1)$ converges to $2/7$ as $t \to \infty$. Hence, the sequence $\mathbb{P}(X_t = 1), t = 0, 1, 2, \ldots$ does not converge. In this case the *stationary* probability is not equal to the *limiting* probability.

7.16 $Z = -\ln U_1$ has an Exp(1) distribution by the inverse-transform method and $R = 2I_{\{U_2 \leqslant 1/2\}} - 1$ takes values -1 and 1 with equal probability. Hence, X is obtained by first generating Z and then flipping its sign with probability $1/2$. The cdf of X is therefore given by

$$\mathbb{P}(X \leqslant x) = \begin{cases} 1 - \mathbb{P}(X > x) = 1 - \mathbb{P}(Z > x)/2 = 1 - e^{-x}/2 & \text{for } x \geqslant 0 \\ \mathbb{P}(Z > -x)/2 = e^x/2 & \text{for } x \leqslant 0. \end{cases}$$

By differentiating the cdf we obtain the pdf $g(x) = e^{-|x|}/2$ for all x.

Selected Problems of Chap. 8

8.2

(a) The posterior pdf is $f(\lambda \mid \mathbf{x}) \propto f(\lambda) \times f(\mathbf{x} \mid \lambda) = (1/\lambda) \times \lambda^5 \exp(-\lambda \sum_{i=1}^5 x_i)$. This is the pdf of the Gamma$(5, \sum_{i=1}^5 x_i)$ distribution, where $\sum_{i=1}^5 x_i = 15.7487$.
(b) The expectation is $5/\sum_{i=1}^5 x_i = 0.317487$.

8.3 The posterior pdf is $f(\lambda \mid x) \propto f(\lambda) \times f(x \mid \lambda) \propto \lambda^{a-1} \exp(-b\lambda) \times \lambda^x \exp(-\lambda) = \lambda^{a+x-1} \exp(-\lambda(1+b))$. This is the pdf of the Gamma$(a + x, b + 1)$ distribution.

8.6 Let $Y_1, \ldots, Y_{m+1}$ with $Y_i \sim$ Gamma$(\alpha_i, 1), i = 1, \ldots, m + 1$ be independent random variables. By Theorem 8.2, $Z_i = Y_i/(Y_i + Y)$, with $Y = \sum_{j \neq i} Y_j$, has the same distribution as the ith coordinate of $\mathbf{Z} = (Z_1, \ldots, Z_m) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_{m+1})$. Moreover, Y_i and Y are independent, and $Y \sim$ Gamma $(\sum_{j \neq i} \alpha_j, 1)$, because its moment generating function is $(1/(1-s))^{\sum_{j \neq i} \alpha_j}$. Hence, again by Theorem 8.2, $Z_i \sim$ Dirichlet$(\alpha_i, \sum_{j \neq i} \alpha_j)$.

8.7

(a) The prior pdf is $f(p \mid x) \propto f(x \mid p) = p(1 - p)^{x-1}$, which is a Beta$(2, x)$ distribution.
(b) The posterior mode is $1/x$; see Problem 8.6.
(c) The posterior expectation is $2/(2 + x)$; see Problem 8.6.

8.13 We have $f(x_{i+1} \mid x_i) = f(x_i, x_{i+1})/f(x_i) = f(x_i \mid x_{i+1})f(x_{i+1})/f(x_i)$. Hence,

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1)f(x_2 \mid x_1) \cdots f(x_n \mid x_{n-1}) \\ &= f(x_1)f(x_1 \mid x_2)\frac{f(x_2)}{f(x_1)} f(x_2 \mid x_3)\frac{f(x_3)}{f(x_2)} \cdots f(x_{n-1} \mid x_n)\frac{f(x_n)}{f(x_{n-1})} \\ &= f(x_n)f(x_{n-1} \mid x_n) \cdots f(x_1 \mid x_2) . \end{aligned}$$

8.14

(a) By the law of total probability $f(x_t, \mathbf{y}_{1:t}) = \sum_{x_{t-1}} f(x_t, x_{t-1}, y_t, \mathbf{y}_{1:t-1})$.
   By conditioning $f(x_t, x_{t-1}, y_t, \mathbf{y}_{1:t-1})$ on x_{t-1} and $\mathbf{y}_{1:t-1}$ we find $f(x_t, \mathbf{y}_{1:t})$
   $= \sum_{x_{t-1}} f(x_t, y_t \mid x_{t-1}, \mathbf{y}_{1:t-1}) f(x_{t-1}, \mathbf{y}_{1:t-1})$.

(b) $f(x_t, y_t \mid x_{t-1}, \mathbf{y}_{1:t-1}) = f(x_t \mid x_{t-1}, \mathbf{y}_{1:t-1}) f(y_t \mid x_t, x_{t-1}, \mathbf{y}_{1:t-1})$. Because of
   the structure of the Bayesian network, x_t given x_{t-1} is independent of $\mathbf{y}_{1:t-1}$,
   and y_t given x_t is independent of x_{t-1} and $\mathbf{y}_{1:t-1}$. Hence, $f(x_t \mid x_{t-1}, \mathbf{y}_{1:t-1})$
   $= f(x_t \mid x_{t-1})$ and $f(y_t \mid x_t, x_{t-1}, \mathbf{y}_{1:t-1}) = f(y_t \mid x_t)$.

(c) $f(x_1, y_1) = f(y_1 \mid x_1) f(x_1)$, where both $f(x_1)$ and $f(y_1 \mid x_1)$ are known.
   Hence, $f(x_1, y_1)$ can be evaluated. Next, $f(x_2 \mid y_1, y_2)$ can be evaluated via
   (8.32) because both factors in the sum are known. The first one is known via
   (8.33) and the second as part of the recursion for $t = 1$. Repeating this, we
   see that $f(x_t, \mathbf{y}_{1:t})$ can be evaluated for any t. The posterior pdf $f(x_t \mid \mathbf{y}_{1:t}) \propto$
   $f(x_t, \mathbf{y}_{1:t})$ follows simply by normalization.

8.18

(a) The Bayesian network is given in Fig. 2.

Fig. 2 The Bayesian network for the bag-of-words model
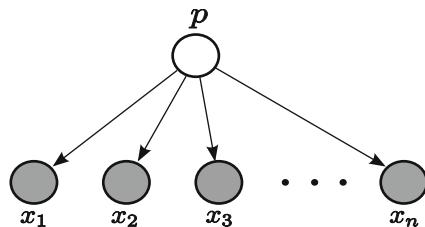
(b) The posterior pdf is

$$f(p \mid \mathbf{x}) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_{pi})^2}{\sigma^2}\right) = \exp\left(-\frac{1}{2} \frac{\|\mathbf{x} - \boldsymbol{\mu}_p\|^2}{\sigma^2}\right).$$

This is maximal when $\|\mathbf{x} - \boldsymbol{\mu}_p\|$ is minimal. Thus p^* maximizes the posterior
pdf.

(c) The posterior pdf is

$$f(p \mid \mathbf{x}) \propto (\sigma_{p1} \cdots \sigma_{pn})^{-1} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_{pi})^2}{\sigma_{pi}^2}\right).$$

The (unscaled) values for $f(p)$, $p = 1, \ldots, 4$ are $53, 0.24, 8.36$, and 3.5×10^{-6}.
Hence the object should be classified as 1. The following code was used.

```
%p8_17c.m
x = [1.67,2,4.23]
mu = [1.6, 2.4, 4.3; 1.5, 2.9, 6.1; ...
```

```
        1.8, 2.5, 4.2; 1.1, 3.1, 5.6];
sig = [0.1, 0.5, 0.2; 0.2, 0.6, 0.9;...
        0.3, 0.3, 0.3; 0.2, 0.7, 0.3];
f =@(p) prod(sig(p,:))^(-1)*...
        exp(-0.5*sum((x-mu(p,:)).^2./sig(p,:).^2));
f(1), f(2), f(3), f(4)
```

Selected Problems of Chap. 9

9.4 We are interested in the pdf of $\Phi(\mathbf{x}^\top\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is distributed according to the posterior pdf and $\mathbf{x}^\top = (1, 1, 10, 1, 0, 16, 1)$. The following MATLAB script uses the posterior draws for $\boldsymbol{\beta}$ (stored in store_beta) to draw from the corresponding posterior pdf.

```
%p9_4.m
N = length(store_beta);
store_prob = zeros(N,1);
x = [1 1 10 1 0 16 1];
for loop=1:N
    store_prob(loop) = normcdf(x*store_beta(loop,:)');
end
```

The expected value of the posterior probability is estimated to be 0.324. A KDE of the posterior probability is plotted in Fig. 3.

Fig. 3 A KDE of the posterior probability that a subject with certain characteristics will have an extramarital affair
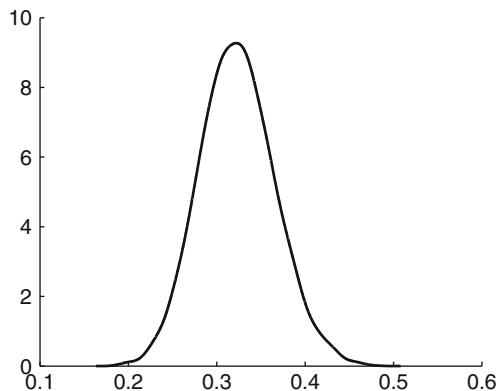
9.6 (c) $1/(\nu - 2)$.

9.8 First recall that $Q_t(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{(\mathbf{x}_i^\top\boldsymbol{\beta})^2 - 2v_i\,\mathbf{x}_i^\top\boldsymbol{\beta}\right\} + \text{const.}$ Noting that $(\mathbf{x}_i^\top\boldsymbol{\beta})^2 = \boldsymbol{\beta}^\top\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta}$ and using the formulas for multivariate differentiation

in Appendix B.1, we have

$$\nabla\, Q_t(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(2\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta} - 2v_i\,\mathbf{x}_i\right)\ .$$

Now solve $\nabla\, Q_t(\boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{\beta}$ to find $\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta} = \sum_{i=1}^{n}v_i\,\mathbf{x}_i$.

Selected Problems of Chap. 10

10.1 The lag-1, lag-2, and lag-3 autocorrelations are, respectively, 0.830, 0.618, and 0.448.

10.3 Let $\boldsymbol{\varepsilon} = (\varepsilon_{1-q},\ldots,\varepsilon_T)^\top$ be the vector of error terms, and let $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_q)^\top$ denote the vector of MA coefficients. Then we can write the MA(q) model as

$$\mathbf{Y} = H\,\boldsymbol{\varepsilon}\ ,$$

where H is a $T \times (T + q)$ **circulant** matrix, where each row vector is rotated one element to the right relative to the previous row vector. In particular, the first row is $(\psi_q, \psi_{q-1}, \ldots, \psi_1, 0, \ldots, 0)$. Since $\boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2\, I)$, the log-likelihood function is given by

$$l(\boldsymbol{\psi}, \sigma^2; \mathbf{y}) = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2}|HH^\top| - \frac{1}{2\sigma^2}\mathbf{y}^\top(HH^\top)^{-1}\mathbf{y}\ .$$

10.7 (a) To derive the full conditional distribution, we first write the ARMA(1,1) as

$$\mathbf{y} = X\boldsymbol{\varrho} + H\boldsymbol{\varepsilon}\ ,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top \sim \mathsf{N}(\mathbf{0}, \sigma^2 I)$, $\boldsymbol{\varrho} = (\varrho_0, \varrho_1)^\top$,

$$X = \begin{pmatrix} 1 & y_0 \\ 1 & y_1 \\ \vdots & \vdots \\ 1 & y_{T-1} \end{pmatrix}, \qquad H = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \psi & 1 & 0 & \cdots & 0 \\ 0 & \psi & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & \psi & 1 \end{pmatrix}\ .$$

The likelihood function is given by

$$f(\mathbf{y} \mid \varrho, \psi, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\varrho)^\top (HH^\top)^{-1}(\mathbf{y}-X\varrho)} .$$

Since this has the form of a linear regression model with covariance matrix $\sigma^2 HH^\top$, it follows from Corollary 8.1 that   ☞ 239

$$(\varrho \mid \mathbf{y}, \psi, \sigma^2) \sim \mathsf{N}(\widehat{\varrho}, D_\varrho) ,$$

where

$$D_\varrho = \left(\frac{1}{10} I + \frac{1}{\sigma^2} X^\top (HH^\top)^{-1} X \right)^{-1} , \qquad \widehat{\varrho} = D_\varrho \left(\frac{1}{\sigma^2} X^\top (HH^\top)^{-1} \mathbf{y} \right).$$

Next, using the likelihood function given above, it can be easily checked that

$$(\sigma^2 \mid \mathbf{y}, \varrho, \psi) \sim \mathsf{InvGamma} \left(3 + \frac{T}{2}, \lambda \right) ,$$

where $\lambda = 1 + (\mathbf{y} - X\varrho)^\top (HH^\top)^{-1}(\mathbf{y} - X\varrho)/2$.

Lastly, given the uniform prior $\psi \sim \mathsf{U}[-1, 1]$, the full conditional posterior distribution for ψ is simply

$$f(\psi \mid \mathbf{y}, \varrho, \sigma^2) \propto f(\mathbf{y}, \varrho, \psi, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\varrho)^\top (HH^\top)^{-1}(\mathbf{y}-X\varrho)}$$

for $-1 < \psi < 1$ and 0 otherwise.

10.9 We first write the two-factor mixed model in matrix form. To that end, let

$$\mathbf{Y} = (Y_{111}, Y_{112}, Y_{113}, Y_{121}, Y_{122}, Y_{123}, \ldots, Y_{3,10,1}, Y_{3,10,2}, Y_{3,10,3})^\top,$$

and define $\boldsymbol{\varepsilon}$ accordingly. In addition, stack $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})^\top$, and $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{39}, \gamma_{3,10})^\top$. Then,

$$\mathbf{Y} = \mu \mathbf{1}_{90} + X_{\boldsymbol{\alpha}} \boldsymbol{\alpha} + X_{\boldsymbol{\beta}} \boldsymbol{\beta} + X_{\boldsymbol{\gamma}} \boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $X_{\boldsymbol{\alpha}} = I_3 \otimes \mathbf{1}_{30}$, $X_{\boldsymbol{\beta}} = \mathbf{1}_3 \otimes A$, $A = I_{10} \otimes \mathbf{1}_3$, $X_{\boldsymbol{\gamma}} = I_{30} \otimes \mathbf{1}_3$, $\otimes$ is the Kronecker product, $\mathbf{1}_p$ is a $p \times 1$ vector of ones, and I_q is the q-dimensional identity   ☞ 116
matrix.

Since $\mathbf{Y}$ is an affine transformation of normal random variables, it has a normal distribution. Its expected value is $\mathbb{E}\mathbf{Y} = \mu \mathbf{1}_{90}$ and its covariance matrix is given by

$$\Sigma = \sigma_{\boldsymbol{\alpha}}^2 X_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}^\top + \sigma_{\boldsymbol{\beta}}^2 X_{\boldsymbol{\beta}} X_{\boldsymbol{\beta}}^\top + \sigma_{\boldsymbol{\gamma}}^2 X_{\boldsymbol{\gamma}} X_{\boldsymbol{\gamma}}^\top + \sigma^2 I_{90} .$$

10.11 In deriving the full conditional distributions, we will repeatedly make use

Theorem 8.1. First, the one-factor random effects model can be written as

$$\mathbf{Y}_i = \mu_i \mathbf{1}_{n_i} + \boldsymbol{\varepsilon}_i \ ,$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_1, \ldots, \varepsilon_{n_i})^\top \sim_{\text{ind}} \mathsf{N}(\mathbf{0}, \sigma^2 I_{n_i})$. From this and the assumption $\{\mu_i\} \sim_{\text{iid}} \mathsf{N}(\mu, \sigma_\mu^2)$, the random effects $\mu_1, \ldots, \mu_d$ are conditionally independent given $\mathbf{y}$, μ, σ_μ^2, and σ^2. In fact, using Theorem 8.1, we have

$$(\mu_i \mid \mathbf{y}, \mu, \sigma_\mu^2, \sigma^2) \stackrel{\text{ind}}{\sim} \mathsf{N}(\widehat{\mu}_i, D_{\mu_i}) \ ,$$

where $D_{\mu_i} = (1/\sigma_\mu^2 + n_i/\sigma^2)^{-1}$ and $\widehat{\mu}_i = D_{\mu_i}(\mu/\sigma_\mu^2 + \mathbf{1}_{n_i}^\top \mathbf{y}_i/\sigma^2)$.

Next, to derive $f(\mu \mid \mathbf{y}, \boldsymbol{\mu}, \sigma_\mu^2, \sigma^2)$, the relevant distributions are the prior for μ and $\{\mu_i\} \sim_{\text{iid}} \mathsf{N}(\mu, \sigma_\mu^2)$. It is then clear that given $\boldsymbol{\mu}$ and σ_μ^2, μ is conditionally independent of $\mathbf{y}$ and σ^2. Again, using Theorem 8.1,

$$(\mu \mid \boldsymbol{\mu}, \sigma_\mu^2) \sim \mathsf{N}(\widehat{\mu}, D_\mu) \ ,$$

where $D_\mu = (1/V_\mu + d/\sigma_\mu^2)^{-1}$ and $\widehat{\mu} = D_\mu(\mu_0/V_\mu + \sum_{i=1}^d \mu_i/\sigma_\mu^2)$.

Similarly, σ_μ^2 is conditionally independent of $\mathbf{y}$ and σ^2 given $\boldsymbol{\mu}$ and μ. In fact, using Theorem 8.1, we have

$$(\sigma_\mu^2 \mid \boldsymbol{\mu}, \mu) \sim \mathsf{InvGamma}\left(\alpha_\mu + \frac{d}{2}, \lambda_\mu + \frac{\sum_{i=1}^d (\mu_i - \mu)^2}{2}\right) \ .$$

Finally, following a similar reasoning,

$$(\sigma^2 \mid \mathbf{y}, \boldsymbol{\mu}) \sim \mathsf{InvGamma}\left(\alpha + \frac{n}{2}, \lambda + \frac{\sum_{i=1}^d (\mathbf{y}_i - \mu_i \mathbf{1}_{n_i})^\top (\mathbf{y}_i - \mu_i \mathbf{1}_{n_i})}{2}\right) \ .$$

Selected Problems of Chap. 11

11.3 Recall that the log-density $\ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2)$ is given by

$$\ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) = -\frac{T}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\tau})^\top (\mathbf{y} - \boldsymbol{\tau}) \ .$$

Hence, setting $\boldsymbol{\tau} = \mathbf{y}$, we obtain $\ln f(\mathbf{y} \mid \boldsymbol{\tau} = \mathbf{y}, \sigma^2) = -\frac{T}{2} \ln(2\pi\sigma^2)$, which approaches infinity as σ^2 approaches 0.

11.4 By recursive substitution using the transition equation, we have $\tau_t = \tau_1 + \sum_{s=2}^{t} u_s$. Hence,

$$\text{Var}(\tau_t \mid \tau_1) = \text{Var}\left(\sum_{s=2}^{t} u_s\right) = \sum_{s=2}^{t} \text{Var}(u_s) = (t-1)\omega^2 .$$

11.5 We first write a MATLAB script to evaluate the log-likelihood function via

$$l(\sigma^2; \mathbf{y}) = \ln f(\mathbf{y} \mid \boldsymbol{\tau}, \sigma^2) + \ln f(\boldsymbol{\tau} \mid \omega^2) - \ln f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2, \omega^2) .$$

Since the equality holds for any $\boldsymbol{\tau}$, we choose $\boldsymbol{\tau} = \mathbf{0}$ to reduce the number of computations involved.

```
function l = loglike_UC(sigma2,omega2,omega2_0,y)
T = length(y);
H = speye(T) - sparse(2:T,1:(T-1),ones(1,T-1),T,T);
invOmega = sparse(1:T,1:T, ...
    [1/omega2_0 1/omega2*ones(1,T-1)],T,T);
HinvOmegaH = H'*invOmega*H;
K = HinvOmegaH + speye(T)/sigma2;
tauhat = K\y/sigma2;
C = chol(K,'lower');
logfy = -T/2*log(2*pi*sigma2) - .5/sigma2*(y'*y);
logftau_pri = -T/2*log(2*pi) - .5*log(omega2_0) ...
    - (T-1)/2*log(omega2);
logftau_post = -T/2*log(2*pi) + sum(log(diag(C))) ...
    -.5*tauhat'*K*tauhat;
l = logfy + logftau_pri - logftau_post;
```

Note that evaluating the log-density $\ln f(\boldsymbol{\tau} \mid \mathbf{y}, \sigma^2, \omega^2)$ involves the term $-\frac{1}{2} \ln |K^{-1}|$, where $K = H\Omega^{-1}H + I/\sigma^2$. To speed up computation, we use the fact that if C is the Cholesky factor of K such that $K = CC^\top$, then

$$-\frac{1}{2} \ln |K^{-1}| = \frac{1}{2} \ln |K| = \ln |C| = \sum_{i} \ln c_{ii} ,$$

where c_{ii} is the ith diagonal element of C. The last equality holds because C is lower triangular.

Next, in the main script, we build a grid and use the function `loglike_UC` to evaluate the log-likelihood function at every point on the grid. A plot of $l(\sigma^2; \mathbf{y})$ is given in Fig. 4. The maximum likelihood estimate of σ^2 computed using this grid search is about 4.401, compared to the value 4.405 obtained by the numerical maximization.

```
%p11_5.m
load 'USCPI.csv';
y = USCPI;
T = length(y);
omega2_0 = 9;    % initial condition
omega2 = .5^2;   % fix omega2
ngrid = 300;     % # of grid points
sigma2grid = linspace(1,10,ngrid)';
l = zeros(ngrid,1);
for i=1:ngrid
    l(i) = loglike_UC(sigma2grid(i),omega2,omega2_0,y);
end
plot(sigma2grid,l,'k','LineWidth',2); box off;
[maxl maxid] = max(l);
sigma2hat = sigma2grid(maxid);
```

Fig. 4 The log-likelihood function $l(\sigma^2; \mathbf{y})$ under the unobserved components model
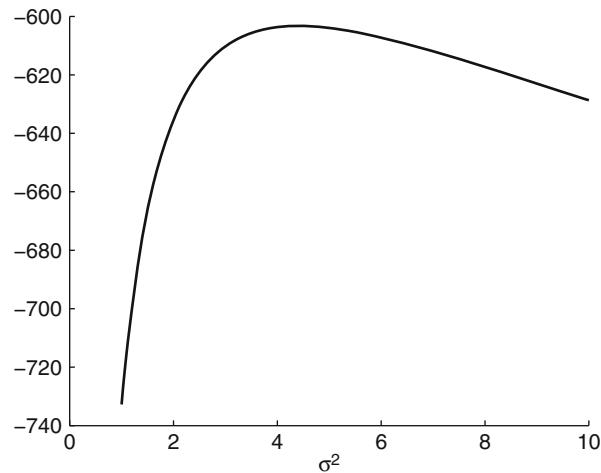
11.6

(a) First write the model in matrix form: $H_\beta \mathbf{y} = H_\beta \boldsymbol{\tau} + \widetilde{\boldsymbol{\alpha}} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = (y_1, \ldots, y_T)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^\top$, $\widetilde{\boldsymbol{\alpha}} = (\beta y_0, 0, \ldots, 0)^\top$, and

$$
H_\beta = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
-\beta & 1 & 0 & \cdots & 0 \\
0 & -\beta & 1 & \cdots & 0 \\
\vdots & & & \ddots & \vdots \\
0 & 0 & \cdots & -\beta & 1
\end{pmatrix}.
$$

Since the determinant of H_β is 1 for any β, H_β is invertible. Then, $\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\alpha} + H_\beta^{-1}\boldsymbol{\varepsilon}$, where $\boldsymbol{\alpha} = H_\beta^{-1}\widetilde{\boldsymbol{\alpha}}$. In other words,

$$(\mathbf{y} \mid y_0, \boldsymbol{\tau}, \beta, \sigma^2) \sim \mathsf{N}(\boldsymbol{\tau} + \boldsymbol{\alpha}, \sigma^2 (H_\beta^\top H_\beta)^{-1}) \,,$$

and the joint log-density of $\mathbf{y}$ is

$$\ln f(\mathbf{y} \mid y_0, \boldsymbol{\tau}, \beta, \sigma^2) = -\frac{T}{2} \ln(2\pi\sigma^2)$$

$$- \frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\tau} - \boldsymbol{\alpha})^\top H_\beta^\top H_\beta (\mathbf{y} - \boldsymbol{\tau} - \boldsymbol{\alpha}) \,.$$

(b) The derivation of $f(\boldsymbol{\tau} \mid \mathbf{y}, y_0, \beta, \sigma^2, \omega^2)$ follows closely the discussion in Sect. 11.1.2. More specifically, since the transition equation is exactly the same as before, we have ☞ 332

$$\ln f(\boldsymbol{\tau} \mid \omega^2) = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \ln \omega_0^2 - \frac{T-1}{2} \ln \omega^2 - \frac{1}{2} \boldsymbol{\tau}^\top (H^\top \Omega^{-1} H) \boldsymbol{\tau} \,,$$

where H is the usual first difference matrix, $\Omega = \mathrm{diag}(\omega_0^2, \omega^2, \dots, \omega^2)$, and $\omega_0^2 = 5$. Then, using the expression for $\ln f(\mathbf{y} \mid y_0, \boldsymbol{\tau}, \beta, \sigma^2)$ given above, we "complete the squares" to obtain ☞ 238

$$(\boldsymbol{\tau} \mid \mathbf{y}, y_0, \beta, \sigma^2, \omega^2) \sim \mathsf{N}(\widehat{\boldsymbol{\tau}}, K^{-1}) \,,$$

where $K = H^\top \Omega^{-1} H + H_\beta^\top H_\beta / \sigma^2$ and $\widehat{\boldsymbol{\tau}} = K^{-1} H_\beta^\top H_\beta (\mathbf{y} - \boldsymbol{\alpha}) / \sigma^2$.

(c) For classical estimation, we can maximize the log-likelihood function numerically using the method of direct likelihood evaluation described in Problem 11.5. Since the number of parameters is only two (β and σ^2), this approach is computationally feasible.

For Bayesian estimation, if we assume conjugate priors for β and σ^2, we can implement a 3-block Gibbs sampler for posterior analysis. The full conditional distribution for $\boldsymbol{\tau}$ is normal as given above. The full conditional distributions for β and σ^2 are normal and inverse-gamma, respectively, which can be sampled from easily.

Index