Nonparametric Curve Estimation: Methods, Theory, and Applications

Sam Efromovich

Springer

Springer Series in Statistics

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

Springer Series in Statistics

Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Andrews/Herzberg: Data: A Collection of Problems from Many Fields for the Student and Research Worker.
Anscombe: Computing in Statistical Science through APL.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Bolfarine/Zacks: Prediction Theory for Finite Populations.
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications
Brémaud: Point Processes and Queues: Martingale Dynamics.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Daley/Vere-Jones: An Introduction to the Theory of Point Processes.
Dzhaparidze: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models.
Farebrother: Fitting Linear Relationships: A History of the Calculus of Observations 1750 - 1900.
Farrell: Multivariate Calculation.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
Fisher/Sen: The Collected Works of Wassily Hoeffding.
Good: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.
Goodman/Kruskal: Measures of Association for Cross Classifications.
Gouriéroux: ARCH Models and Financial Applications.
Grandell: Aspects of Risk Theory.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Hart: Nonparametric Smoothing and Lack-of-Fit Tests.
Hartigan: Bayes Theory.
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications.
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.
Heyer: Theory of Statistical Experiments.
Huet/Bouvier/Gruet/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.
Jolliffe: Principal Component Analysis.
Kolen/Brennan: Test Equating: Methods and Practices.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.

(continued after index)

Springer Series in Statistics

(continued from p. ii)

Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume III.
Kres: Statistical Tables for Multivariate Analysis.
Küchler/Sørensen: Exponential Families of Stochastic Processes.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.
Longford: Models for Uncertainty in Educational Testing.
Manoukian: Modern Concepts and Theorems of Mathematical Statistics.
Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.
Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of the Federalist Papers.
Parzen/Tanabe/Kitagawa: Selected Papers of Hirotugu Akaike.
Politis/Romano/Wolf: Subsampling.
Pollard: Convergence of Stochastic Processes.
Pratt/Gibbons: Concepts of Nonparametric Theory.
Ramsay/Silverman: Functional Data Analysis.
Rao/Toutenburg: Linear Models: Least Squares and Alternatives.
Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.
Reinsel: Elements of Multivariate Time Series Analysis, 2nd edition.
Reiss: A Course on Point Processes.
Reiss: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
Rieder: Robust Asymptotic Statistics.
Rosenbaum: Observational Studies.
Ross: Nonlinear Estimation.
Sachs: Applied Statistics: A Handbook of Techniques, 2nd edition.
Särndal/Swensson/Wretman: Model Assisted Survey Sampling.
Schervish: Theory of Statistics.
Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.
Shao/Tu: The Jackknife and Bootstrap.
Siegmund: Sequential Analysis: Tests and Confidence Intervals.
Simonoff: Smoothing Methods in Statistics.
Singpurwalla and Wilson: Statistical Methods in Software Engineering: Reliability and Risk.
Small: The Statistical Theory of Shape.
Stein: Interpolation of Spatial Data: Some Theory for Kriging
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.
Tong: The Multivariate Normal Distribution.
van der Vaart/Wellner: Weak Convergence and Empirical Processes: With Applications to Statistics.
Vapnik: Estimation of Dependences Based on Empirical Data.
Weerahandi: Exact Statistical Methods for Data Analysis.
West/Harrison: Bayesian Forecasting and Dynamic Models, 2nd edition.
Wolter: Introduction to Variance Estimation.
Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.

Sam Efromovich

Nonparametric Curve Estimation

Methods, Theory, and Applications

With 130 Figures

Springer

Sam Efromovich
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131-1141
USA

Library of Congress Cataloging-in-Publication Data
Efromovich, Sam.
Nonparametric curve estimation : methods, theory, and applications
/ Sam Efromovich.
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-98740-1 (hardcover)
1. Nonparametric statistics. 2. Estimation theory. I. Title.
II. Series.
QA278.8.E35 1999
519.5—dc21 99-13253

ISBN 0-387-98740-1 Springer-Verlag New York Berlin Heidelberg SPIN 10709119

To my parents

Preface

Appropriate for a one-semester course, this self-contained book is an introduction to nonparametric curve estimation theory. It may be used for teaching graduate students in statistics (in this case an intermediate course in statistical inference, on the level of the book by Casella and Berger (1990), is the prerequisite) as well as for diverse classes with students from other sciences including engineering, business, social, medical, and biological among others (in this case a traditional intermediate calculus course plus an introductory course in probability, on the level of the book by Ross (1997), are the prerequisites).

There are several distinguishing features of this book that should be highlighted:

- All basic statistical models, including probability density estimation, nonparametric regression, time series analysis including spectral analysis, and filtering of time-continuous signals, are considered as one general problem. As a result, universal methods of estimation are discussed, and students become familiar with a wide spectrum of applications of nonparametric methods.
- Main emphasis is placed on the case of small sample sizes and data-driven orthogonal series estimates (Chapters 1–6). Chapter 7 discusses (with proofs) modern asymptotic results, and Chapter 8 is devoted to a thorough discussion of nonseries methods.
- The companion software package (available over the World Wide Web) allows students to produce and modify almost all figures of the book as well as to analyze a broad spectrum of simulated and real data sets. Based on the S–PLUS environment, this package requires no knowledge of S–PLUS

and is elementary to use. Appendix B explains how to install and use this package; it also contains information about the affordable S–PLUS Student Edition for PC.
- "Practical Seminar" sections are devoted to applying the methods studied to the analysis and presentation of real data sets. The software for these sections allows students to analyze any data set that exists in the S–PLUS environment.
- "Case Study" sections allow students to explore applications of basic methods to more complicated practical problems. These sections together with "Special Topic" sections give the instructor some flexibility in choosing additional material beyond the core.
- Plenty of exercises with different levels of difficulty will allow the instructor to keep students with different mathematical and statistical backgrounds out of trouble!
- "Notes" sections at the end of each chapter are primarily devoted to books for further reading. They also capture some bibliographic comments, side issues, etc.
- Appendix A contains a brief review of fundamentals of statistical inference. All the related notions and notations used in the book may be found there. It is highly recommended to review these fundamentals prior to studying Chapters 3–8. Also, exercises for Appendix A may be used as a first quiz or homework.

A bit of advice to the reader who would like to use this book for self-study and who is venturing for the first time into this area. You can definitely just read this book as any other text without using the companion software. There are plenty of figures (more than a hundred), which will guide you through the text. However, if you have decided to study nonparametrics, then you are probably interested in data analysis. I cannot stress too strongly the importance of combining reading with analyzing both simulated and real data sets. This is the kind of experience that you can gain only via repeated exercises, and here the software can make this process dramatically quicker and less painful. Using the software will allow you to check virtually every claim and development mentioned in the book and make the material fully transparent. Also, please review the fundamentals outlined in Appendix A prior to studying Chapters 3–8.

All further developments related to this book will be posted on the WWW page http://www.math.unm.edu/∼efrom/book1, and the author may be contacted by electronic mail as efrom@math.unm.edu.

Acknowledgments

I would like to thank everyone who in various ways has had influence on this book. My biggest thanks go to Mark Pinsker. Alex Samarov graciously read and gave comments on a draft of the book. John Kimmel provided

invaluable assistance through the publishing process. Three "generations" of my students who took the class on nonparametric curve estimation based on this book shared with me their thoughts, comments, and suggestions. I thank all of you.

Sam Efromovich
Albuquerque, USA, 1999

Contents

1
Introduction

Methods of nonparametric curve estimation allow one to analyze and present data at hand without any prior opinion about the data. In this chapter we discuss several basic applications of this approach via analyzing real data sets. These data sets are interesting, give insight into the nature of nonparametric curve estimation, and raise important practical questions that will be answered in the following chapters.

1.1 Density Estimation in the Exploration and Presentation of Data

Density estimation is one of the most fundamental problems in statistics. The simplest problem is formulated mathematically as follows. Let us consider a univariate continuous random variable X distributed according to a probability density f. This phrase means that for any practically interesting set B of real numbers we can find the probability (likelihood) that X belongs to this set by the formula

$$P(X \in B) = \int_B f(x)dx. \tag{1.1.1}$$

For instance, for $B = [a, b]$ we get that $P(a \le X \le b) = \int_a^b f(x)dx$.

Then one observes n independent realizations $X_1, X_2, \ldots, X_n$ of X, and the aim is to find an estimate $\tilde{f}(x)$, based on these observations, that fits the underlying $f(x)$.

A customary and very natural use of density estimates is to present a data set at hand, as well as to make a kind of informal investigation of properties of the data. This is exactly what we would like to do in this section, leaving the more rigorous analysis until Chapters 3 and 7.

As an example, consider a particular data set of a daily numbers game (lottery) run by a state government. This type of lottery is an excellent tool for understanding the problem because lotteries are a common feature of our life, and their simplicity makes it attractive for millions of people; the foundations of probability and statistics are based on games of chance; and lotteries raise many unanswered questions ranging from "Is the lottery fair?" to "Can we win, and if so, then how?"

We begin with a specific data set for the New Jersey Pick-It lottery, a daily numbers game run by the state of New Jersey. (In this book we use data sets available as a part of the standard S–PLUS package.) The data set is **lottery** and it is for 254 drawings just after the lottery was started, from May 1975 to March 1976.

The rules of the game are as follows. At the time of purchase, a player chooses a three-digit number ranging from 000 to 999. Pick-It lottery is a pari-mutuel game where the winners share a fraction of the money taken in for a particular drawing. Half of the money bet during the day is placed into a prize pool (the state takes the other half), and anyone who picked the winning number shares equally in the pool. The lottery allows a player to make a combinations bet on three different digits. For example, if the player picks the number 791, then the player wins if in a winning number the digits appear in any order. Thus, the ticket 791 wins on 179, 197, etc. Payoffs for the numbers with duplicate digits are not shared with combination betters, and thus may be higher. In short, to win big, pick numbers like 001 or 998, whereas to win often, pick numbers like 012 or 987.

The available data set consists of the winning number and the payoff for a winning ticket for each drawing. Let us begin with the statistical analysis of winning numbers. The stream of these numbers is shown in Figure 1.1a, and the first number was 810 and the last 479. Such a diagram nicely shows the "wild" dynamics of the winning numbers, but it is very difficult to assess this data set in such a representation. Unfortunately, this is how such data sets are typically presented to the public.

Thus, our first aim is to understand how to present such a data set better. Before doing this, let us pause for a second and try to understand what we would like to realize and gain from the analysis of the winning numbers. Let us look at one of the possible reasons to look more closely at this data set. It is assumed that the game is fair, so the chance for a winning number to fall within an interval, say $[200, 240]$ should be the same for any interval of the same length. Thus, we may divide the interval $[0, 999]$ into equal subintervals and then count the winning numbers that fall into each of the subintervals. If the lottery is fair, then such a frequency distribution of winning numbers should look "reasonably" flat. If this is not the case,

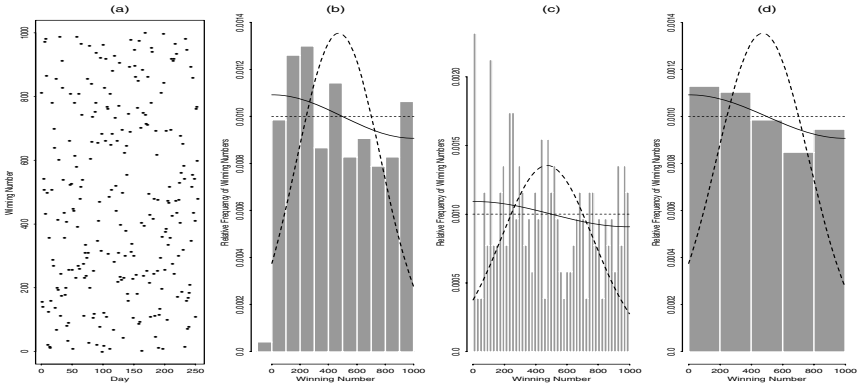

FIGURE 1.1. (a) Daily winning numbers from the 254 New Jersey Pick-It lottery drawings from May 1975 to March 1976. (b) Default S–PLUS histogram for this data set. (c) A histogram with 50 bins. (d) A histogram with 5 bins. All the histograms are overlaid by the parametric normal density estimate (dashed line), the universal nonparametric density estimate (solid line), and the ideal uniform density (dotted line), which corresponds to a fair drawing procedure.

then the lottery is not fair, and one can use this lack of fairness to one's advantage in picking more favorable numbers or, nowadays, do even better by visiting a trial lawyer.

In Figure 1.1b a default S–PLUS *histogram*, which is the most common format for representing relative frequencies of grouped data, is shown. Such a default histogram is just a realization of the above-discussed idea about how to look at the data in the frequency domain. A histogram is also a favorite tool of data analysts—they use it as a first look at the data—so let us briefly explain how it is created (a more rigorous discussion of a histogram may be found in Section 8.1). A histogram is formed by dividing the real line into equally sized intervals (called *bins*); it is a step function that begins at the *origin* and whose heights are proportional to the number of sample points (here the winning numbers) contained in each bin. The simplicity of a histogram explains why it is the oldest and most widely used density estimate. The bin width and origin must be chosen to show features of the data, and in Figure 1.1b they are chosen by the S–PLUS function **hist**.

Now that we have seen what professional data analysts like to observe, let us find out what they think about this particular data set and this particular histogram. (Recall that the "ideal" histogram, corresponding to a fair drawing, would look like the dotted line.) A typical subjective conclusion based on visualization of this histogram sounds like this one: "The histogram looks fairly flat—no need to inform a grand jury." This particular quotation is taken from the book by Becker, Chambers, and Wilks (1988). Clearly, a data analyst should be very experienced and well

trained to make such a conclusion from just visualizing this histogram, because, at first glance, it is absolutely not flat, and should you show it to players at the wrong place and time, expect a riot.

We cannot judge this conclusion right now (a lot of training in analyzing similar diagrams awaits us). Instead, let us return to a more thorough discussion of how the S–PLUS function **hist** creates default histograms.

The procedure of creating a default histogram is based on the assumption that an underlying density is the familiar bell-shaped normal density (the definition of a normal density may be found in Appendix A, but it is not needed at this point). In Figure 1.1b the dashed line is the estimated normal density shown over the interval $[0, 999]$. Note that this normal curve does not resemble the histogram at all, and it is also far from the flat uniform distribution that should be expected for a fair game. As a reflection of this inconsistency, the default histogram shows us the peculiar small bar at the left tail that represents the single time that 000 was the winning number. This is in part because the normal curve (the dashed line) is skewed, and its mode (mean) is apparently less than 500.

Thus, we can reasonably suspect that the default histogram does not tell us the whole story about the winning numbers and even may misrepresent them because in no way does the normal density (the dashed line) resemble the default histogram.

Finally, on the top of everything, Figure 1.1b shows us the universal nonparametric estimate (the solid line), which we shall thoroughly study in this book. This estimate tells us an absolutely unbelievable story from the point of view of both the normal density estimate and the default histogram: smaller numbers were more likely than larger ones.

Probably, we are now too overwhelmed by all this information, so let us again pause for a second and summarize our preliminary conclusions. We now understand that the default histogram is an estimate and it may be imperfect especially if an underlying density does not resemble a normal (bell-shaped) density. Second, both the nonparametric estimate and the parametric normal estimate tell us that the data are skewed. Thus, thinking logically, our next step should be to look more closely at the underlying winning numbers. After all, these numbers and only these numbers may tell as the full story. We can easily do this using a histogram with a larger number of bins (smaller bin width).

In Figure 1.1c we see the "zoomed-in" histogram with 50 bins and a correctly chosen origin at the number 000. This histogram is also overlaid by the above-discussed estimates. Here each bin contains winning numbers within the range of 20 numbers, so it gives us a rather detailed impression about relative frequencies of the winning numbers. But does such a detailed histogram help us to answer the questions raised? The answer is "no" because this particular histogram is even more confusing. How many local modes do you see? What is the strange situation with the smallest winning

numbers; is it a fraud? In short, the detailed histogram only increases our confusion.

If our attempt at zooming-in failed, let us try to zoom-out. After all, if the game is fair, then the numbers of winning numbers within wide bins should be approximately the same due to all the basic limit theorems of probability theory (they are discussed in Appendix A).

Figure 1.1d shows us the histogram with only 5 bins (that is, winning numbers are combined in groups within every 200 of numbers, and we may expect about fifty winning numbers within each bin). As soon as you see this histogram, you may cry "BINGO." Now we clearly see what was happening during the first 11 months of the lottery. Indeed, the smaller numbers were more generous to the players, and this coincides with the main message of the nonparametric estimate. There are some disagreements about the largest numbers, but at least in general the situation has been cleared up.

What we have done with you so far reflects the main issue of modern nonparametric estimation—the science and art of smoothing a data set. We have seen that improper smoothing may make a problem extremely complicated. It takes some time and training to understand how to smooth a particular data set, and it takes some time and training to understand the messages of nonparametric estimates.

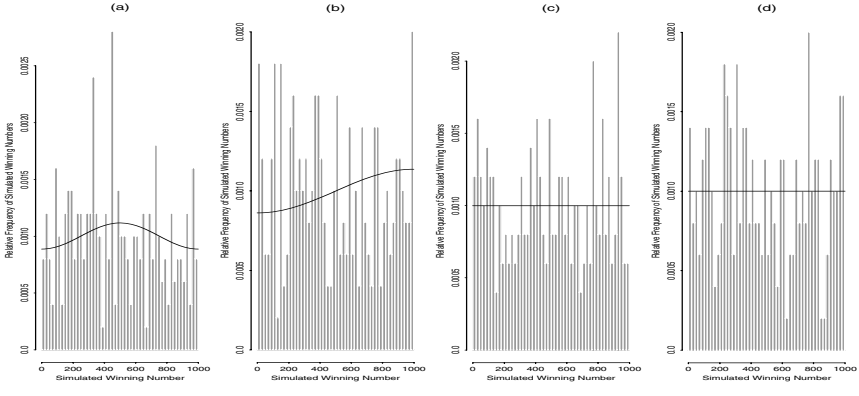

FIGURE 1.2. Zoomed–in histograms and the universal nonparametric density estimates for 4 fairly (uniformly) simulated sets of 250 winning lottery numbers. {Information in curly brackets contains comments on how to use the software to repeat or modify a particular figure, whereas information in square brackets is about arguments of the S-function. As is indicated by the square brackets in this caption, another simulation can be performed and then visualized on the screen of the monitor. The argument l, shown in the square brackets, allows one to control the number of bins (or respectively the bin width, which will be $1000/l$). For instance, to get a histogram with just 5 bins, like the one shown in Figure 1.1d, type > **ch1(f=2,**$l = 5$**)**. More about using this software may be found in Appendix B.} [$l = 50$]

Now let us return to one of the questions raised at the beginning of this section. Namely, we have this particular data set at hand. What can be said about the fairness of the drawing? Does the nonparametric estimate indicate an unfair game, or is this just a natural deviation due to the stochastic nature of the drawing? The scientific answer, based on the theory of hypothesis testing, will be given in Section 3.8. Here, instead of invoking that powerful theory, we shall use a very simple and convincing approach based on Monte Carlo simulations of winning numbers ranging from 000 to 999. Modern computers allow us to simulate such numbers with great confidence in their fairness; after all, let us hope that computers are not influenced by a local government. Another reason to look at such generators is the fact that nowadays many players use a computer (called an electronic advisor) to pick a number.

In Figure 1.2 four fairly (according to a uniform distribution) generated sets of 250 winning numbers are shown via detailed histograms and the nonparametric estimates. As we see, both the particular data sets and the estimates may have peculiar forms. At least formally, these simulations support the conclusion of the experts cited earlier "... no need to inform a grand jury."

Note that as in our previous discussion, the detailed histograms do not clarify the situation and do not help to visualize the underlying uniform density. Here again a zoomed-out histogram is a better way to analyze a data set.

Another lesson from Figure 1.2 is that you cannot judge one estimate (for instance the nonparametric one) via analyzing another nonparametric estimate (in this case the histogram).

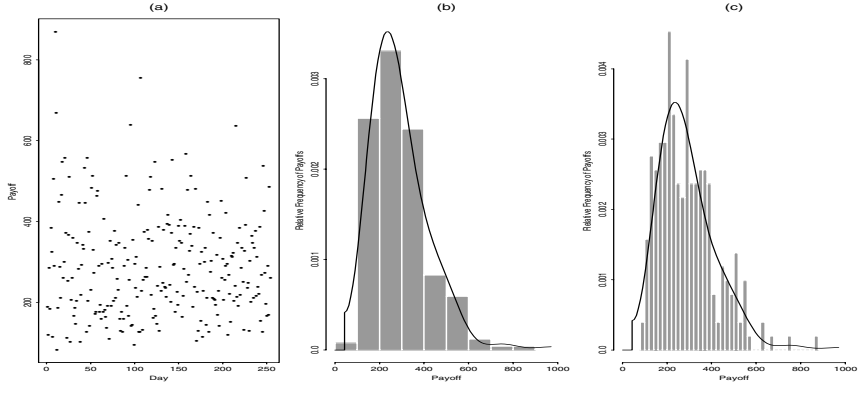

FIGURE 1.3. The time series of the 254 payoffs from May 1975 to March 1976 and two corresponding histograms with different bin widths overlaid by the universal nonparametric estimate (the solid line).

Simulations such as those shown in Figure 1.2, when one knows an underlying density, are very useful for both training and choosing the correct estimates. We shall often use simulated data sets for these purposes.

Now it is time to look at the payoffs. The stream of payoffs is shown in Figure 1.3a. Probably the only interesting information that may be gained from this diagram is that only five times was a payoff larger than \$600, with the largest payoff being \$869.50.

The by now familiar histograms and the nonparametric estimates are shown in two other diagrams. Note that here the default histogram (b) looks reasonable because it shows a frequency distribution with a pronounced mode describing most typical payoffs and a long right tail describing the rarer big payoffs. The more detailed histogram (c) is rather confusing: Are there 3 or even more modes? The nonparametric estimate reasonably well describes both the dynamics of the data and the flat right tail with large but rare payoffs.

Let us look at two more periods of this lottery available in the standard S–PLUS package. Figure 1.4 presents 3 data sets for periods shown in the titles. What we see resembles the pattern of the simulated sets shown in Figure 1.2 and confirms our preliminary conclusion that the drawings are fair.

Figure 1.5 depicts similar descriptive characteristics for the payoffs. Here we see that while the skewed bell shape of the frequency distributions remained the same over the years, a very peculiar change occurred in that the range of the payoffs appeared to be shrinking. Indeed, if during the first year of the lottery both small and large payoffs were recorded, then over the years both the smallest and largest payoffs simply disappeared.

In the beginning of this section we raised questions about the fairness of the drawings and methods to win. We were able to answer the first question, at least heuristically, via analyzing independently simulated data sets. The

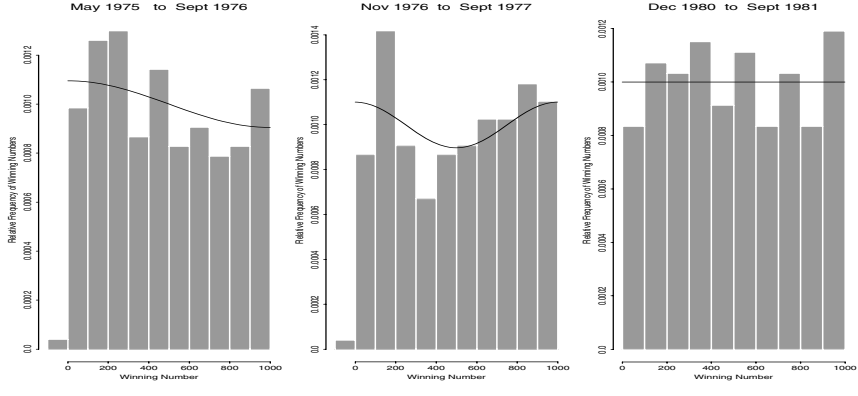

FIGURE 1.4. Default histograms overlaid by the nonparametric density estimates of winning numbers in the New Jersey Pick-It lottery for three time periods.

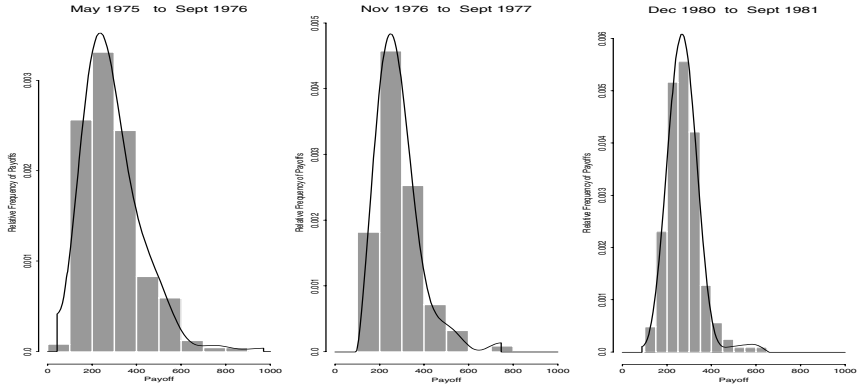

FIGURE 1.5. Default histograms overlaid by the nonparametric density estimates of payoffs in the New Jersey Pick-It lottery for three time periods.

answer to the second question is based on the analysis of the relationship between a picked number and the corresponding payoff (note that if the drawings are fair, then the only chance to win big is to bet on numbers that are out of favor among other players). Such a problem is called a regression, and the regression approach will be considered in the next section. On the other hand, a probability density approach has its own way to give insight into such a problem.

Let us for a moment return to the formula (1.1.1). Assume that we are interested in understanding what is the likelihood that by betting on a number between 120 and 140 the corresponding payoff will be between \$500 and \$550. In other words, we would like to think about likelihood of the pair number–payoff. In this case the set B should be a 2-dimensional rectangle, and the corresponding density becomes a bivariate density $f(x_1, x_2)$. The value of such a density at a point in the x_1x_2-plane tells us about the likelihood of the winning number and the payoff occurring in some vicinity of the point (x_1, x_2). After all, the winning numbers and the payoffs come in pairs for each drawing, so it is natural to look at their joint distribution.

Figure 1.6 shows nonparametric estimates (surfaces) for the three sets of the lottery numbers. These surfaces allow us to make several conclusions. We see that all the surfaces have a saddle-type shape. Of course, the eye is drawn to the crests (the lines along the top of the saddles) and to the summits. The pattern is rather typical, namely, during the first two periods, picking the smaller or larger number led to nice payoffs, while during the third period the largest numbers were more generous to the players. Note that such an outcome is in part due to the observed particular irregularities in the winning numbers drawn and also due to the fact that to win big a picked number should be out of favor among other players. This resembles much more serious (in terms of the amounts of money involved) betting situations, such as sporting events or stock picking, in which to win really

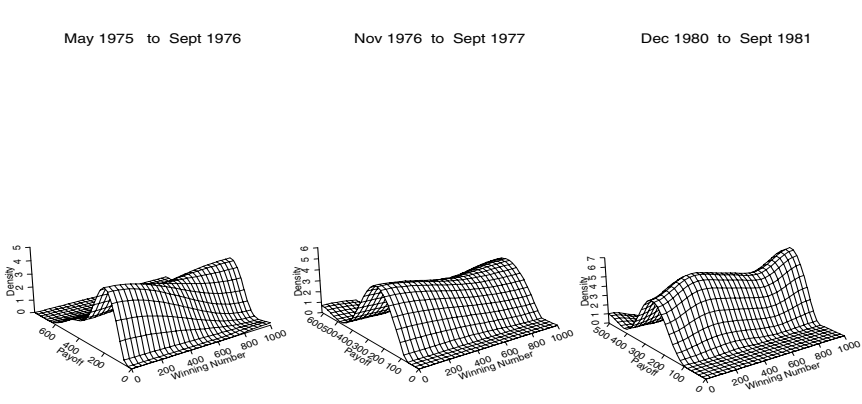

FIGURE 1.6. Nonparametric estimates of bivariate densities of the winning numbers and payoffs in the New Jersey Pick-It lottery for three time periods. The density estimates are multiplied by 10^6.

big, one should not follow the crowd. Figure 1.6 apparently indicates that not all numbers were equally favorable among the players.

Finally, let us explain the basic idea of how the universal nonparametric density estimator works, because this, in turn, explains why Chapter 2 is devoted to the purely mathematical subject of orthogonal series approximation.

Typically, the density $f(x)$ of a continuous random variable X may be well approximated by a series $f_J(x)$,

$$f_J(x) := \sum_{j=0}^{J} \theta_j \varphi_j(x), \tag{1.1.2}$$

as the parameter J becomes large (in this book the notations $:=$ and $=:$ mean "by definition"). Here J is called the *cutoff*, $\{\varphi_j(x), j = 0, 1, \ldots\}$ are some fixed and known functions (typically elements of a classical orthonormal basis), and the coefficients θ_j are calculated by the formula

$$\theta_j = \int_{-\infty}^{\infty} f(x)\varphi_j(x)dx. \tag{1.1.3}$$

Recall that f is the probability density of a random variable X, so by definition of the expectation,

$$\theta_j = \int_{-\infty}^{\infty} f(x)\varphi_j(x)dx =: E\{\varphi_j(X)\}. \tag{1.1.4}$$

Thus, θ_j is the expectation of the random variable $\varphi_j(X)$, and then a natural estimate of θ_j is the sample mean $\hat{\theta}_j := n^{-1}\sum_{l=1}^{n}\varphi_j(X_l)$.

Finally, a series estimate with cutoff J is

$$\hat{f}_J(x) := \sum_{j=0}^{J}\hat{\theta}_j\varphi_j(x). \tag{1.1.5}$$

This is the basic idea of an orthogonal series approach used to construct a universal nonparametric estimate.

1.2 Nonparametric Regression

A classical model of nonparametric regression is defined as follows. We observe n pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and it is supposed that

$$Y_l = f(X_l) + \varepsilon_l, \quad l = 1, 2, \ldots, n, \tag{1.2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are iid realizations of a random variable (error) ε with zero mean and finite variance. The problem is to estimate the *regression function* f. The variables X_l are referred to as the *design points* (*predictors* or *independent variables*), while Y_l are referred to as *responses* (or *dependent variables*). Design points may be either iid realizations of a random variable X or fixed deterministic points; the former model is called *random design regression* and the latter *fixed design regression*.

As with nonparametric density estimation, one of the main aims of nonparametric regression is to highlight an important structure in the data without any assumption about the data. In other words, the nonparametric approach allows the data speak for themselves.

Now we are in a position to answer the question raised in the previous section about a winning strategy for picking the numbers in the New Jersey Lottery. Since the winning numbers and the payoffs come in pairs, to gain information about the structure and a possible relationship between these two variables, it is helpful to construct a *scattergram* for the data. Such a plot, also called a *scatter plot* or *scatter diagram*, exhibits pairs of observations as points in the xy-plane. The hope is that by analyzing such graphs where one variable is plotted versus another, some kind of a pattern or relationship may be discovered.

Scatter plots for the three periods of the New Jersey lottery, where winning numbers are considered as predictors and payoffs as responses, are shown in Figure 1.7. Ignore, for a moment, the solid lines superimposed on

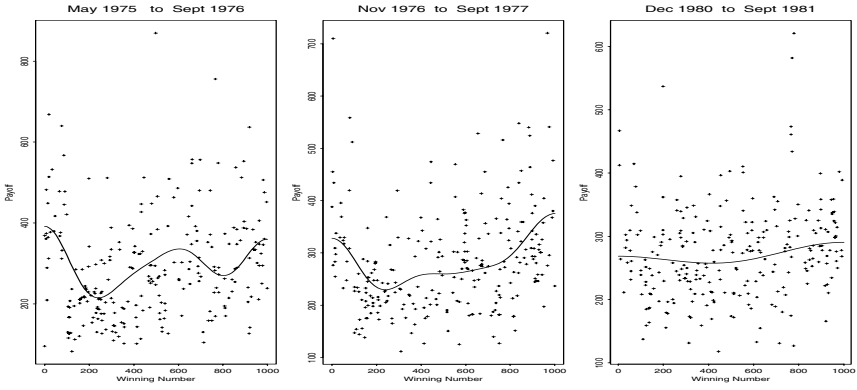

FIGURE 1.7. Nonparametric regression estimates superimposed on the winning number-payoff scatter plots for the 3 periods of the New Jersey Pick-It lottery.

the scatter plots, and try to answer the following questions. Can you see any structure in the data? Does the payoff depend on the winning number? Can you suggest a winning strategy? Perhaps some help is needed to answer all these questions, so let us look at the nonparametric regression estimates (these estimates are based on the same idea discussed at the end of Section 1.1) shown by the solid lines. Now it is simpler to recognize some patterns in the data sets. We see that during all these years, on average, the smallest and largest numbers were most generous to the players. During the onset of the game (the first period) the smallest numbers were overall the best bet, and the largest just a bit behind. Note that this conclusion is not related to the phenomenon, discussed in the previous section, that smaller winning numbers were drawn more often during that period. Here we consider an average payoff given a winning number, and the only fact that matters here is that few players picked small or large numbers during the first year of the lottery. Probably, the players were reluctant to bet on numbers like 003 or 998.

Note that the solid lines may be used to analyze the preferences of players during these periods to choose this or that number. Namely, they are inversely proportional to the likelihood for a number to be chosen by a player. Thus, for instance, the numbers near the minimum of the nonparametric estimate were most popular among the players.

Let us use this approach to see what was the mood among the players during the next years. The second year of the lottery is shown in the middle diagram, and we see that the least favorable numbers were the largest, while the most popular were around 250. Note that while during the second year there was no shift in the players' opinion about what numbers to bet on (of course, on numbers close to 250!), more players decided to bet on smaller numbers and numbers around 600, so now the solid curve has no pronounced mode near 600.

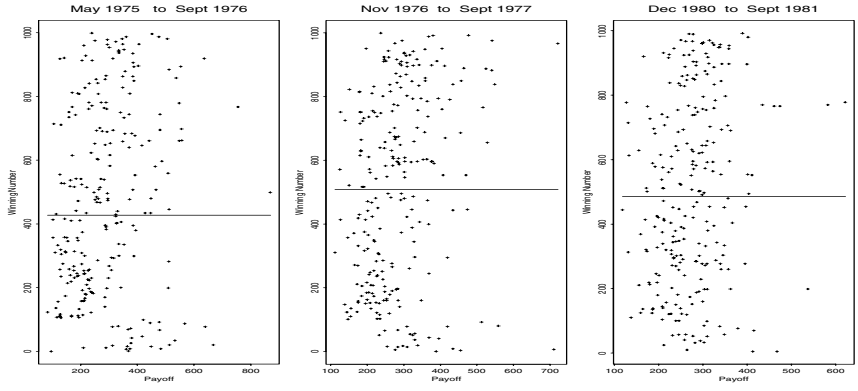

FIGURE 1.8. Nonparametric regression estimates superimposed on the payoff-winning number scatter plot for the New Jersey Pick-It lottery.

Finally, 5 years after the inception of the lottery it looks as if everyone was a bit tired; see the right diagram of Figure 1.7. No longer was there a consensus among the players about what was a lucky number (while still they preferred numbers from 250 to 500). There was again a bit of bias toward larger numbers, but it was absolutely minor in comparison with the second year.

Another curious regression is the inverse one where the payoff is the predictor and the winning number is the response. The point of such a regression is that knowing a payoff, one wants to make a bet on an average winning number. The corresponding scatter plots, overlaid by the nonparametric estimates, are shown in Figure 1.8. As we see, the nonparametric estimates tell us that knowing a payoff does not help us to choose the corresponding winning number. There is additional interesting information that may be gained from these nonparametric estimates. Note that these horizontal estimates show us the average winning numbers. Thus, we see that on average the winning numbers were slightly above 400 during the first period (do you recall the nonparametric estimate in Figure 1.1?), then the average winning number jumped slightly above 500 during the next period (do you recall the nonparametric estimate for this period in Figure 1.4?), and finally it settled down near the expected 500 (again recall the nonparametric estimate shown in Figure 1.4). This remark ends our introductory discussion of the lottery data; see more in Section 3.8.

1.3 Time Series Analysis

Time series analysis is probably the most exciting topic in nonparametric curve estimation. A typical nonparametric problem is the classical decomposition of a realization of a time series into a slowly changing function

known as a "trend component," or simply trend, a periodic function referred to as a "seasonal component," and finally a "random noise component," which in terms of the regression theory should be called the time series of residuals.

The problem of finding a trend may be solved by the methods of nonparametric regression. Estimation of a seasonal component is more involved, since its period is unknown. Here the nonparametrics shines again because the spectral density is the tool to search after the periods. Namely, for a discrete stationary time series $\{X_t, t = 1, 2, \ldots\}$ with zero mean and finite variance, under mild assumptions the spectral density $f(\lambda)$ at the frequency λ is defined as

$$f(\lambda) := (2\pi)^{-1}\theta_0 + \pi^{-1}\sum_{j=1}^{\infty}\theta_j \cos(j\lambda), \quad -\pi < \lambda \le \pi, \tag{1.3.1}$$

where $\theta_j = E\{X_t X_{t+j}\}$ is the covariance at lag j. Then, if the spectral density has a mode at frequency λ^*, then this may indicate a seasonal component with the period

$$T^* = 2\pi/\lambda^*. \tag{1.3.2}$$

Thus, to find the period of a seasonal component one should first estimate the spectral density. It is apparent how to apply the orthogonal series approach to this problem: The basis is given (here it is the cosine basis), and Fourier coefficients are expressed as the mathematical expectation. Thus, if n realizations $X_1, X_2, \ldots, X_n$ are given, then the familiar empirical covariance,

$$\hat{\theta}_j = n^{-1}\sum_{l=1}^{n-j} X_l X_{l+j}, \tag{1.3.3}$$

may be used as an estimator of θ_j.

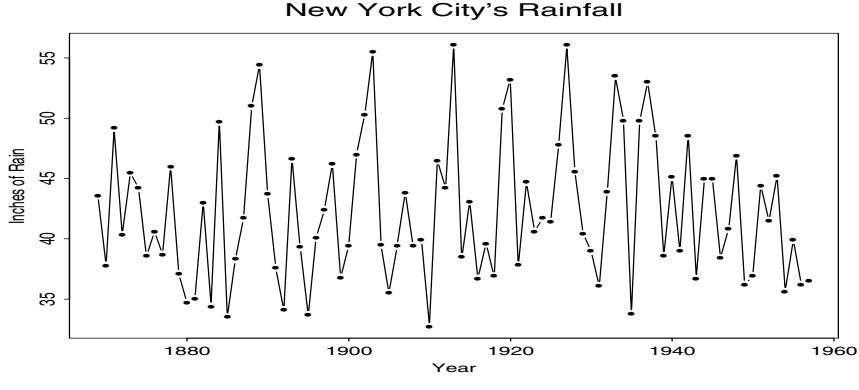

FIGURE 1.9. Rainfall data for New York City.

Let us see how this idea works. As an example, consider the time series of New York City's annual rainfall from 1896 to 1957 (the data file is **rain.nyc1**). In Figure 1.9 the data are shown. The diagram nicely presents the dynamics of the rainfall over the years. Recall that traditional weather-related questions are about patterns: Is there any pronounced pattern in the rainfall? Does the rainfall decrease or increase with passing years? Is the weather more volatile now than it was ten (twenty, sixty, etc.) years ago?

Even using the nice presentation of the data in Figure 1.9, it is not clear how to answer these questions. So below, the answers are exhibited using the approach of Chapter 5.

The first two diagrams in Figure 1.10 repeat the data set using different formats. Diagram 1.10.3 shows us the estimated trend. It tells us that no global change has occurred over those years. We also see that on average New York City had about 42.3 inches of rainfall per year.

Diagram 1.10.4 shows us the data minus the estimated trend (called detrended data). Here the fun begins. Do you see any periodic component here that represents a pattern with a reasonably small period, say between 4 and 12 years? Probably the answer is "no."

This is the place where the nonparametric spectral density estimate may shine. This estimate is shown in diagram 1.10.5, and you can see the pro-

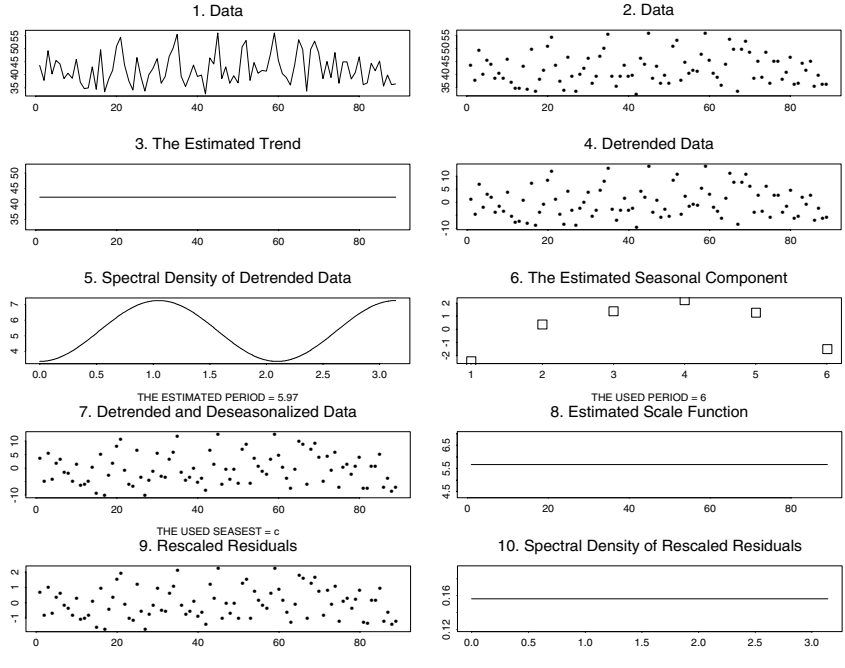

FIGURE 1.10. Nonparametric analysis of rainfall data for New York City.

nounced mode near the frequency 1.05. The period, estimated by formula (1.3.2), is 5.97, and it is shown in the subtitle.

As soon as the period is known, we may estimate the underlying seasonal component. It is shown in diagram 1.10.6. Note that its range is about 5 inches, so for New York City this is about 12 percent of its average rainfall. Of course, this is a minor phenomenon for New York City's rainfall. On the other hand, think about the stock market, where you can recognize a 12 percent seasonal component!

Of course, the data set is relatively small, and this may play a joke on us. However, the message of this example is clear: The nonparametric approach may allow us to find a "needle in a haystack."

The diagrams 7–8 allow us to analyze the residuals. They will be explained in Chapter 5.

One of the most powerful mathematical tools invented quite recently for approximation of spatially inhomogeneous curves and images is wavelets. They are just very special orthonormal elements and may be straightforwardly used in our series approach.

Let us see how the nonparametric wavelet estimate, which will be discussed in Section 4.4, performs for the famous **sunspots** data set, which

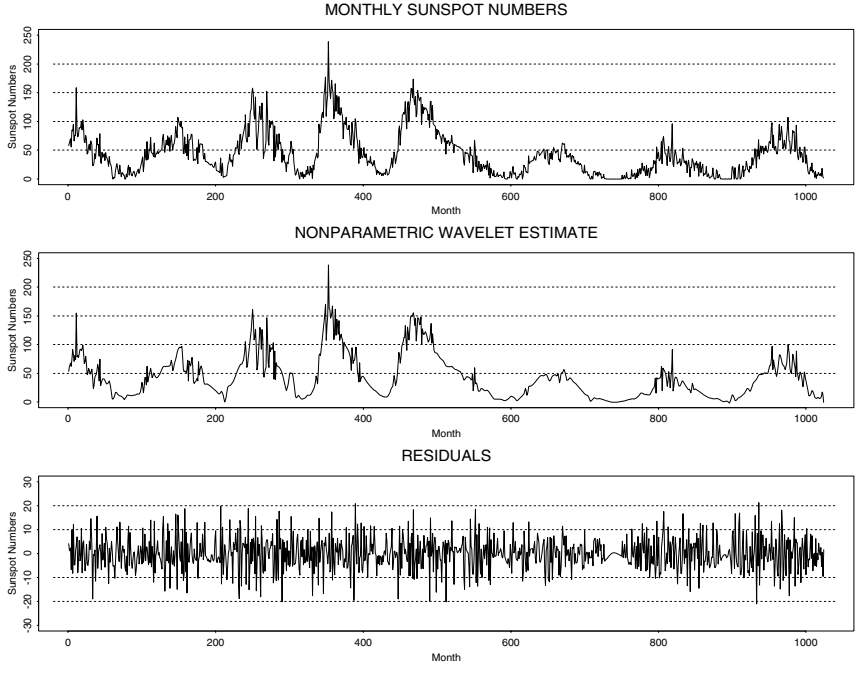

FIGURE 1.11. Wavelet decomposition of monthly sunspot data. The top time series, which starts at January 1749, is the sum of the other two.

contains monthly means of daily numbers of sunspots. The time series starts at January 1749, and here we consider the case of 1024 consecutive months.

The top diagram in Figure 1.11 shows us the complexity of this time series. The wavelet estimate, shown in the middle diagram, does a remarkable job in exhibiting the main structure of the data, and it preserves all but fine details. The residuals are remarkably small; only during extremely volatile periods do they exceed 10 spots. We can think about the sunspots time series as being synthesized by two different instruments, similar to a musical orchestration that is the sum of notes from each instrument. As we discussed earlier, such a decomposition of a time series (a sound) into components (notes) is the main approach of the time series theory. And the nonparametric approach plays an important role in such a decomposition. Note that here you can see a pronounced periodic component (modulated in time) with a period about 130 months. The period is too large for the monthly data frame, but it is perfect for the annual data frame, where the period is about 11 years.

Finally, this chapter is the only one that does not contain a section with exercises. To compensate for such an omission, consider the following problem. Many famous scientists and philosophers have conjectured that the history of our civilization has been greatly affected (and even predetermined) by the sun's activity (this is why the count of sunspots goes back to the eighteenth century). In other words, such events as wars, riots, revolutions, as well as periods of prosperity and peace, are highly correlated with the sunspot numbers. Use your knowledge of history and test this theory.

2

Orthonormal Series and Approximation

The orthonormal series approach is the primary mathematical tool for approximation, data compression, and presentation of curves used in all statistical applications studied in Chapters 3–7. The core topics are given in the first two sections. Section 2.1 considers series approximations via visualization, and Section 2.2 gives a plain introduction in how fast Fourier coefficients can decay. Among special topics, Section 2.3 is devoted to a more formal discussion of the mathematics of series approximation, and it is highly recommended for study or review. Reading other special sections is optional and can be postponed until they are referred to in the following chapters.

2.1 Introduction to Series Approximation

In this section three particular orthonormal systems are introduced and discussed via visualization of their approximations. The first one is the cosine system that will be the main tool in the following chapters. The second one is a polynomial system based on orthonormalization of the powers $\{1, x, x^2, \ldots\}$; this system is an excellent tool for approximating polynomial curves. The third one is a Haar system, which is a good tool for approximation of discontinuous functions; this basis is also of special interest because it is the simplest example of wavelets, which are relative newcomers to the orthogonal series scene.

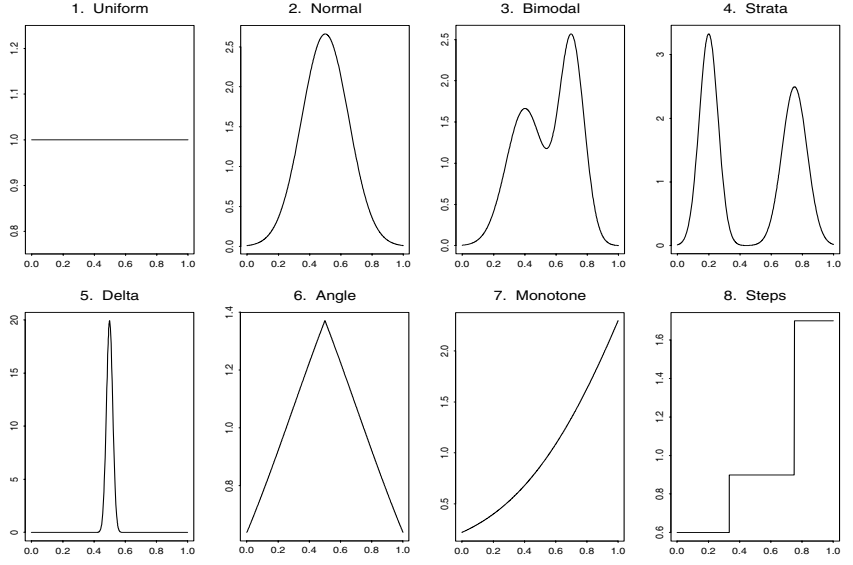

FIGURE 2.1. The corner functions. {This set may be seen on the monitor by calling (after the S–PLUS prompt) > **ch2(f=1)**. A corner function may be substituted by a custom-made one, see explanation in the caption of Figure 2.3.}

For the performance assessment, we choose a set of *corner* (*test*) functions. Corner functions should represent different functions of interest that are expected to occur in practice. In this book eight specific corner functions with some pronounced characteristics are used, and they are expected to be approximated quite well or quite poorly by different systems. The set is shown in Figure 2.1.

To make all statistical simulations as simple as possible, the corner functions are some specific probability densities supported on $[0, 1]$. They are defined via uniform and normal ($d_{\mu,\sigma}(x) := (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$) densities or their mixture.

Below, each of the corner functions is briefly discussed. These functions are arranged in order of the decreasing smoothness of their 1-periodic continuations.

1. *Uniform.* This is a uniform density on $[0, 1]$, that is, $f_1(x) := 1$. The Uniform is the smoothest 1-periodic function in our set, and we shall see that despite its triviality, neither its approximation nor statistical estimation is elementary. Moreover, this function plays a central role in asymptotic theory, and it is an excellent tool for debugging different types of errors.

2. *Normal.* This is a normal density with mean 0.5 and standard deviation 0.15, that is, $f_2(x) := d_{0.5,0.15}(x)$. The normal (bell-shaped) curve is the most widely recognized curve. Recall the rule of three standard deviations, which states that a normal density $d_{\mu,\sigma}(x)$ practically vanishes whenever

$|x - \mu| > 3\sigma$. This rule helps us to understand the curve. It also explains why we do not divide f_2 by its integral over the unit interval, because this integral is very close to 1.

3. *Bimodal.* This is a mixture of two normal densities, $f_3(x) := 0.5d_{0.4,0.12}(x) + 0.5d_{0.7,0.08}(x)$. The curve has two pronounced and closely located modes, which why the curve is included in the set.

4. *Strata.* This is a function supported over two separated subintervals. In the case of a density, this corresponds to two distinct strata in the population. This is what differentiates the Strata from the Bimodal. The curve is obtained by a mixture of two normal densities, namely, $f_4(x) := 0.5d_{0.2,0.06}(x) + 0.5d_{0.7,0.08}(x)$. (Note how the rule of three standard deviations was used to choose the parameters of the normal densities in the mixture.)

5. *Delta.* The underlying idea of the next curve is to have an extremely spatially inhomogeneous curve that vanishes over the entire interval except for an extremely small region at the center ($x = 0.5$) where the function is very large. Such a function resembles many practical situations where a short but abrupt deviation from a normal process occurs. The Delta mimics the theoretical delta function, which has zero width and is integrated to 1. The Delta is defined as a normal density with very small standard deviation, $f_5(x) := d_{0.5,0.02}(x)$.

6. *Angle.* This is a function whose 1-periodic continuation is continuous and extremely smooth except of the points $x = k$ and $x = k + 0.5$, $k = 0, \pm 1, \ldots$, where the derivative changes sign. The Angle is $f_6(x) := (1/0.16095)d_{1,0.7}(x)$ if $0 \le x \le 0.5$ and $f_6(x) := (1/0.16095)d_{0,0.7}(x)$ if $0.5 < x \le 1$.

7. *Monotone.* This function is smooth over the interval, but its 1-periodic continuation has a jump at all integers x. We shall see that this makes approximation of such a function challenging due to boundary effects. This also explains why the Monotone is ranked number 7 among the suggested corner functions. The Monotone is defined by the formula $f_7(x) := d_{2,0.8}(x)/\int_0^1 d_{2,0.8}(u)du$.

8. *Steps.* This is the least smooth function in our set. The function is challenging for smooth series like a trigonometric or polynomial one. Moreover, its approximation is not rosy even for wavelets. The name of the function is clear from the graph. The Steps is defined by $f_8(x) := 0.6$ for $0 \le x < \frac{1}{3}$, $f_8(x) := 0.9$ for $\frac{1}{3} \le x < \frac{3}{4}$ and $f_8(x) := \frac{204}{120}$ for $\frac{3}{4} \le x \le 1$.

Now, let us recall that a *function* $f(x)$ defined on an interval (the *domain*) is a rule that assigns to each point x from the domain exactly one element from the *range* of the function. Three traditional methods to define a function are a table, a formula, and a graph. For instance, we used both formulae and graphs to define the corner functions.

The fourth (unconventional) method of describing a function $f(x)$ is via a series expansion. Suppose that the domain is $[0, 1]$. Then

$$f(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \quad \text{where} \quad \theta_j = \int_0^1 f(x)\varphi_j(x)dx\,. \tag{2.1.1}$$

Here the functions $\varphi_j(x)$ are known, fixed, and referred to as the *orthonormal functions* or *elements* of the *orthonormal system* $\{\varphi_0, \varphi_1, \ldots\}$, and the θ_j are called the *Fourier* coefficients (for a specific system we may use the name of the system in place of "Fourier"; for instance, for a Haar system we may refer to θ_j as Haar coefficients). A system of functions is called *orthonormal* if the integral $\int_0^1 \varphi_s(x)\varphi_j(x)dx = 0$ for $s \neq j$ and $\int_0^1 (\varphi_j(x))^2 dx = 1$ for all j. Examples will be given below.

Note that to describe a function via an infinite orthogonal series expansion (2.1.1) one needs to know the infinite number of Fourier coefficients. No one can store or deal with an infinite number of coefficients. Instead, a *truncated (finite) orthonormal series* (or so-called partial sum)

$$f_J(x) := \sum_{j=0}^{J} \theta_j \varphi_j(x) \tag{2.1.2}$$

is used to approximate f. The integer parameter J is called the *cutoff*.

The advantage of this approach is the possibility of an excellent compression of the data. In statistical applications this also leads to the estimation of a relatively small number of Fourier coefficients. Roughly speaking, the main statistical issue will be how to choose a cutoff J and estimate Fourier coefficients θ_j. Thus, the rest of this section is devoted to the issue of how a choice of J affects visualization of series approximations. This will give us a necessary understanding and experience in choosing reasonable cutoffs.

Below, several orthonormal systems are introduced and then analyzed via the visualization of partial sums.

Cosine orthonormal system on $[0, 1]$**.** The elements are

$$\varphi_0(x) := 1 \quad \text{and} \quad \varphi_j(x) := \sqrt{2}\,\cos(\pi j x) \quad \text{for} \quad j = 1, 2, \ldots. \tag{2.1.3}$$

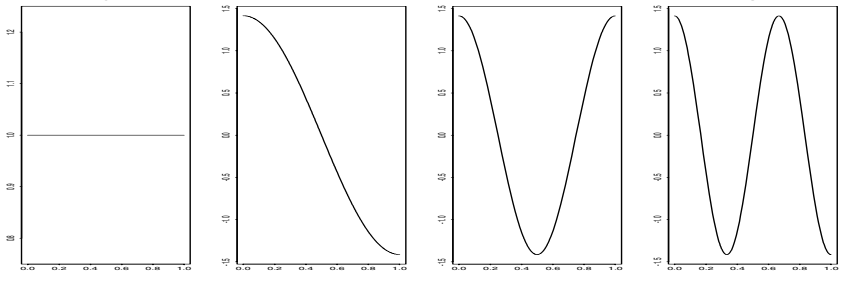

FIGURE 2.2. The first four elements of the cosine system. {Recall that any 4 (or fewer) elements may be visualized using the argument *set.j*.} [*set.j* = *c*(0,1,2,3)]

The first four elements are shown in Figure 2.2. It is not easy to believe that such elements may be good building blocks for approximating different functions, but surprisingly, they do a good job in approximation of smooth functions.

To visualize partial sums, several particular cutoffs, namely $J = 3$, $J = 5$, and $J = 10$, are chosen. Then Fourier coefficients are calculated by (2.1.1), and the partial sums (2.1.2) are shown in Figure 2.3. (Note that here and in what follows an underlying corner function is always shown by the solid line. As a result, all other curves are "hidden" behind a solid line whenever they coincide.)

Consider the partial sums shown. The Uniform is clearly described by the single Fourier coefficient $\theta_0 = 1$, all other θ_j being equal to zero because $\int_0^1 \varphi_j(x)dx = 0$ whenever $j > 0$ (recall that the antiderivative, see the definition below at (2.1.4), of $\cos(\pi j x)$ is $(1/\pi j)\sin(\pi j x)$; thus $\int_0^1 \sqrt{2}\cos(\pi j x)dx = \sqrt{2}(\pi j)^{-1}[\sin(\pi j 1) - \sin(\pi j 0)] = 0$ for any positive integer j). Thus, there is no surprise that the Uniform is perfectly fitted by the cosine system—after all, the Uniform corner function is the first element of this system.

Approximation of the Normal is a great success story for the cosine system. Even the approximation based on the cutoff $J = 3$, where only 4 Fourier coefficients are used, gives us a fair visualization of the underlying function, and the cutoff $J = 5$ gives us an almost perfect fit. Just think about a possible compression of the data in a familiar table for a normal density into only several Fourier coefficients.

Now let us consider the approximations of the Bimodal and the Strata. Note that here partial sums with small cutoffs "hide" the modes. This is especially true for the Bimodal, whose modes are less pronounced and separated. In other words, approximations with small cutoffs oversmooth an underlying curve. Overall, about ten Fourier coefficients are necessary to get a fair approximation. On the other hand, even the cutoff $J = 5$ gives us a correct impression about a possibility of two modes for the Bimodal and clearly indicates two strata for the Strata. The cutoff $J = 10$ gives us a perfect visualization except for the extra mode between the two strata. This is how the cosine system approximates a constant part of a function. We shall see the same behavior in other examples as well.

The approximations of the Delta allow us to summarize the previous observations. The partial sum with $J = 3$ oversmooths the Delta. Approximations with larger cutoffs do a better job in the visualization of the peak, but the valley is approximated by confusing oscillations ("wiggles"). This corner function allows us to gain necessary experience in "reading" cosine approximations. Note that the wiggles are "suspiciously" symmetric about $x = 0.5$, which is the point of the pronounced mode. This will always be the case for approximating a function like the Delta. This is how a trigonometric approximation "tells" us about a spatially inhomogeneous underlying

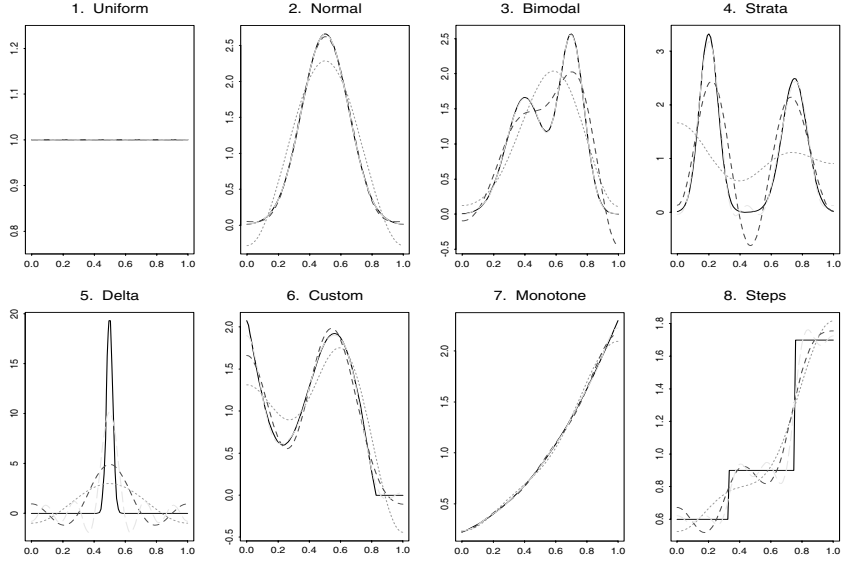

FIGURE 2.3. Approximation of corner functions (solid lines) by cosine series: dotted, short-dashed, and long-dashed lines correspond to cutoffs $J = 3$, $J = 5$, and $J = 10$, respectively. The 6th function is custom-made. {The optional argument *CFUN* allows one to substitute a corner function by a custom-made corner function. For instance, the choice $CFUN = list(3, "2 * x - 3 * \cos(x)")$ implies that the third corner function (the Bimodal) is substituted by the positive part of $2x-3\cos(x)$ divided by its integral over [0,1], i.e., the third corner function will be $(2x - 3\cos(x))_+ / \int_0^1 (2u - 3\cos(u))_+ du$. Any valid S–PLUS formula in x (use only the lower case x) may be used to define a custom-made corner function. This option is available for all Figures where corner functions are used. Only for this figure to visualize approximations of the Angle set *CFUN=list(6,NA)*. The choice of cutoffs is controlled by the argument *set.J*. The smaller number of approximations may be used to make curves more recognizable. On the other hand, even 4 curves are well recognizable on a color monitor. Try > **ch2(f=0)** to test colors. [*set.J = c(3,5,10), CFUN* $= list(6, "2 - 2 * x - \sin(8 * x)")$]

function. Note that here even the cutoff $J = 10$ is not enough for a good representation of the Delta. Clearly the cosine system is not very good for approximation of this particular corner function. On the other hand, if it is known that an underlying function is nonnegative, then a projection onto the class of nonnegative functions creates a dramatically better visualization. This will be discussed in detail in Section 3.1.

The approximations of the custom-made function are fairly good even for $J = 3$ (of course, the representation of the tails needs more Fourier coefficients). Let us use this particular example to discuss the approximation of a function near the boundary points. As we see, the partial sums are flattened out near the edges. This is because derivatives of any partial

sum (2.1.2) are zeros at the boundary points (derivatives of $\cos(\pi jx)$ are equal to $-\pi j \sin(\pi jx)$ and therefore they are zero for $x = 0$ and $x = 1$). In other words, the visualization of a cosine partial sum always reveals small flat plateaus near the edges (you could notice them in all previous approximations as well). Increasing the cutoff helps to decrease the length of the plateaus and improve the visualization. This is the *boundary* effect, and we shall discuss in Section 2.6 how to overcome it.

A similar situation occurs for the Monotone. Here the only reason to increase the cutoff is to diminish the boundary effect.

The approximations of the Steps are not aesthetically appealing, to say the least. On the other hand, it is the purpose of this corner function to "explain" to us how the cosine partial sums approximate a piecewise constant function. In particular, let us look at the long-dashed line, which exhibits overshoots of the steps in the underlying function. This is the famous Gibbs phenomenon, which has to do with how poorly a trigonometric series converges in the vicinity of a jump. The natural conjecture would be that the overshoots vanish as $J \to \infty$, but surprisingly, this does not take place (actually, that overshoots are proportional to a jump).

Note that while cosine approximations are not perfect for some corner functions, understanding how these partial sums perform may help us to "read" messages of these approximations and guess about underlying functions. Overall, for the given set of corner functions, the cosine system does an impressive job in both representing the functions and the data compression.

Polynomial orthonormal system on $[0, 1]$**.** This is probably the most familiar system of functions $\{\varphi_j(x) = \sum_{l=0}^{j} a_{jl}x^l,\ j = 0, 1, 2, \ldots\}$. Here j is called the *degree* of the polynomial φ_j, and the coefficients $\{a_{jl}\}$ are chosen in such a way that the polynomials are orthonormal.

The underlying idea of this system is as follows. It is absolutely natural to approximate a function by a linear combination of the power functions $1,\ x,\ x^2, \ldots$ (this resembles the idea of a polynomial regression). Unfortunately, the power functions are not orthonormal. Indeed, recall that the antiderivative $G(x)$ of x^k is equal to $x^{k+1}/(k+1)$, so $\int_0^1 x^k dx = G(1) - G(0) = (k+1)^{-1}$; see (2.1.4) below. On the other hand, the power functions may be used as building blocks for creating a polynomial orthonormal system using the *Gram–Schmidt orthonormalization* procedure discussed in detail in Section 2.3.

The Gram–Schmidt procedure is very simple and performs as follows. The first function is normalized and becomes the null element of the polynomial basis, namely, $\varphi_0(x) := 1/[\int_0^1 1^2 dx]^{1/2} = 1$. The first element $\varphi_1(x)$ is calculated using x and $\varphi_0(x)$ by the formula

$$\varphi_1(x) := \frac{x - (\int_0^1 \varphi_0(u)udu)\,\varphi_0(x)}{[\int_0^1 (v - (\int_0^1 \varphi_0(u)udu)\,\varphi_0(v))^2 dv]^{1/2}}\ .$$

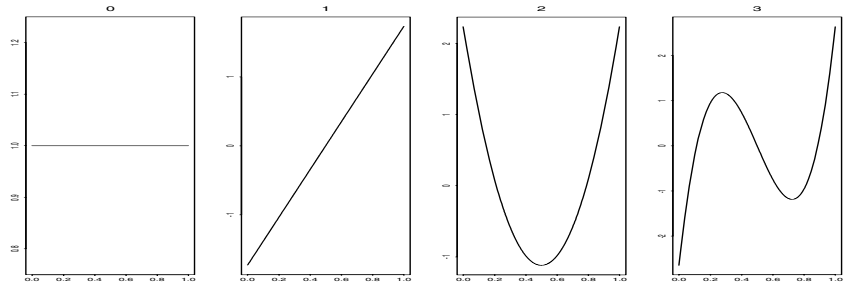

FIGURE 2.4. The first four elements of the polynomial system. {Recall that the information in the square brackets tells us that any 4 (or fewer) elements of this system may be visualized using the argument *set.j*.} [*set.j*=*c*(0,1,2,3)]

A straightforward calculation shows that $\varphi_1(x) = \sqrt{3}(2x-1)$. Then any other element is defined by recursion. For instance, to find the element $\varphi_j(x)$, all previous elements are to be calculated, and then φ_j is defined via the previous elements and x^j by

$$\varphi_j(x) := \frac{x^j - \sum_{l=0}^{j-1}(\int_0^1 u^j\varphi_l(u)du)\varphi_l(x)}{[\int_0^1(v^j - \sum_{l=0}^{j-1}(\int_0^1 u^j\varphi_l(u)du)\varphi_l(v))^2dv]^{1/2}} \ .$$

The first four elements of the polynomial orthonormal system are shown in Figure 2.4.

Note that the idea of the direct approximation of f by a power series $\sum_{j=0}^J b_{Jj}x^j$ is so natural and so appealing that it is worthwhile to explain why in place of a power series the orthonormal series is recommended. The only (but absolutely crucial) reason is the simplicity in calculating polynomial coefficients θ_j. Indeed, we can always write $f_J(x) = \sum_{j=0}^J \theta_j\varphi_j(x) = \sum_{j=0}^J b_{Jj}x^j$. The power series clearly looks simpler and more natural. On the other hand, its coefficients b_{Jj} should be calculated for every J, and there is no simple formula for doing this. Actually, probably the best way to find b_{Jj} is first to calculate θ_j (note that they do not depend on the cutoff J!) and then use them for calculating b_{Jj}.

Prior to the discussion of the partial sums of the polynomial system, it is worthwhile to explain how Fourier coefficients θ_j can be calculated for a particular underlying f. The *fundamental theorem of calculus* states that for a function $g(x)$ continuous on $[0,1]$,

$$\int_0^1 g(x)dx = G(1) - G(0), \tag{2.1.4}$$

where $G(x)$ is an *antiderivative* of g, that is, $dG(x)/dx = g(x)$, $x \in [0,1]$. Thus, if an antiderivative for $f(x)\varphi_j(x)$ is known, then a calculation of the Fourier coefficient $\theta_j = \int_0^1 f(x)\varphi_j(x)dx$ is elementary. Unfortunately, in many cases antiderivatives are unknown, and this natural approach can-

not be used. Also, we should always keep in mind statistical applications where an underlying function is unknown and, typically, only its noisy observations at some particular points are given.

Thus, instead of (2.1.4), a numerical integration based on values of a function at some points may be of a special interest. As an example, consider the widely used trapezoid rule for numerical integration. Let $h = 1/N$ and $x_k = kh$ for $k = 0, 1, \ldots, N$. Assume that the second derivative $\phi^{(2)}$ of the function ϕ is continuous. Then it is possible to show that for some x^* in $[0, 1]$ the following formula holds:

$$\int_0^1 \phi(x)dx = [(h/2)(\phi(x_0) + 2\phi(x_1) + 2\phi(x_2) + \cdots + 2\phi(x_{N-1}) + \phi(x_N))]$$

$$- (1/(12N^2))\phi^{(2)}(x^*). \tag{2.1.5}$$

The first term on the right side gives us the *trapezoid* rule, and the second is called the *discretization* (or *numerical*) error. Note that the discretization error decreases proportionally to $1/N^2$.

As you see, to implement the trapezoid formula it is sufficient to know values of f at $N + 1$ equidistant points. Also, the formula is simple. Of course, a numerical error will be presented. On the other hand, for our purposes of understanding how partial sums perform, these errors can be considered as a positive phenomenon. Indeed, in all statistical applications Fourier coefficients are estimated with some stochastic errors. Here we do not have them, but the numerical errors can simulate for us the effect of stochastic ones. As a result, we shall gain experience in dealing with partial sums whose Fourier coefficients are contaminated by errors.

The trapezoid rule has been used to calculate the polynomial coefficients (with $N = 300$) for the corner functions; these polynomial approximations are shown in Figure 2.5. The eye is drawn to the partial sums for the Uniform. The Uniform serves as an excellent test for debugging all possible errors because we know for sure that all partial sums are to be identical to the underlying function (indeed, the Uniform should be perfectly matched by $\varphi_0(x) = 1$). But what we see is rather puzzling because only the partial sum with the cutoff $J = 3$ gives us a fair representation of the curve. Moreover, the approximations perform inversely to our expectations and previous experience, where larger cutoffs meant better approximation. The reason is the numerical errors, and we see how they affect the partial sums. Note that a larger cutoff implies a larger number of calculated coefficients and therefore a larger cumulative error. This is clearly seen in Figure 2.5.1. Thus, in the presence of errors, an optimal cutoff is not necessarily the largest, and a choice of an optimal cutoff is based on a compromise between a fair approximation and cumulative errors due to incorrectly calculated polynomial coefficients. We shall see that this is also the main issue for all statistical settings where stochastic errors are inevitable.

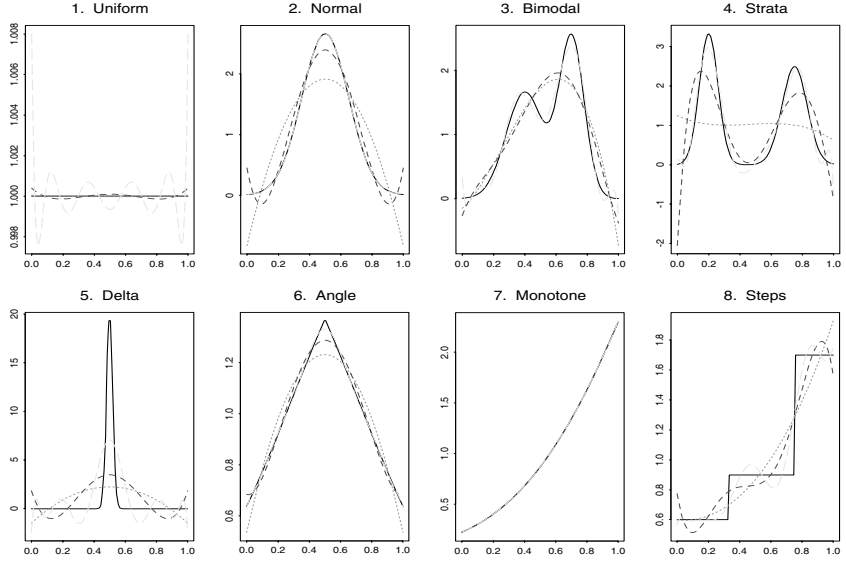

FIGURE 2.5. Approximation of the corner functions (solid lines) by polynomial series: Dotted, short-dashed, and long-dashed lines correspond to cutoffs $J = 3$, $J = 5$, and $J = 10$, respectively. [*set.J=c(3,5,10)*]

Among other approximations shown in Figure 2.5, it is worthwhile to mention the exceptionally good approximation of the Monotone. Here even $J = 3$ gives us a perfect approximation. Also, we see that the polynomial basis has its own boundary effects, and they can be pronounced.

Haar orthonormal system on $[0, 1]$**.** This system is of special interest because it is a good tool to approximate piecewise constant functions and it is the simplest example of wavelets. It is easier to draw elements of the Haar system than to define them by formulae; in Figure 2.6 the first four elements are shown.

The wavelet literature refers to the function $F(x)$ as the *scaling* or wavelet *father* function and to $M(x)$ as the *wavelet function* or wavelet *mother* function. Note that the mother function is integrated to zero, while the father function is integrated to one. The name *mother* is motivated by the fact that all other elements are generated by the mother function. For instance, the next two elements shown in Figure 2.6 are $\sqrt{2}M(2x)$ and $\sqrt{2}M(2x-1)$. Already you have seen the two essential operations for creating the elements: *translation* and *dilation*. Translation is the step from $M(2x)$ to $M(2x-1)$, while dilation is from $M(x)$ to $M(2x)$. Thus, starting from a single mother function, the graphs are shifted (translated) and compressed (dilated). The next *resolution level (scale)* contains functions $2M(2^2x)$, $2M(2^2x-1)$, $2M(2^2x-2)$, and $2M(2^2x-3)$. Note that each of these four functions is supported on an interval of length $\frac{1}{4}$. This procedure

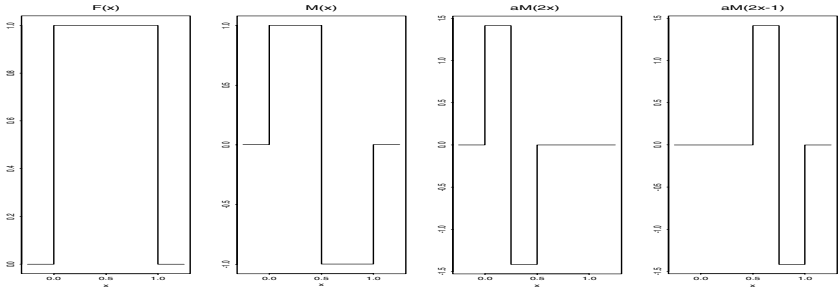

FIGURE 2.6. The first four functions of Haar system. In the title $a = 2^{1/2}$.

may be continued, and in the end we get a Haar system with elements $F(x)$ and $M_{jk}(x) := 2^{j/2}(2^j x - k)$, $j = 0, 1, \ldots$ and $k = 0, \ldots, 2^j - 1$. Here j denotes the resolution level, and k denotes the *shift*.

Elements of a Haar system are *localized*; for instance, $M_{jk}(x)$ is supported on $[2^{-j}k, 2^{-j}(k+1)]$. This is what makes them so special. In short, one can expect that Haar elements will be good building blocks for approximation of nonsmooth functions.

It is customary to write the Haar partial sum as

$$f_J(x) := \theta_0 F(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} \theta_{jk} M_{jk}(x). \tag{2.1.6}$$

Here J is the maximum *multiresolution level* (or the number of *multiresolution components* or *scales*), and thus 2^{J+1} Haar coefficients are used. In other words, a Haar partial sum may be based on 2, 4, 8, 16, etc. terms. Typically, only a small portion of Haar coefficients is significant, and all others are negligibly small. This implies good data compression.

The Haar system is so simple that we can even guess the Haar coefficients. For instance, let $f(x)$ be equal to 1 on the interval $[0, \frac{1}{4}]$ and vanish beyond the interval. Try to guess how to approximate the function by the elements shown in Figure 2.6. (This is a nice puzzle, and the answer is $f(x) = 0.25F(x) + 0.25M(x) + 0.5M(2x)$; of course, we can always solve such a puzzle using the formula (2.1.1).)

Figure 2.7 shows how a Haar system approximates the corner functions: the dotted line is based on 16 Haar coefficients ($J = 3$), and the short-dashed line on 64 Haar coefficients ($J = 5$). We see that the trapezoid rule of numerical integration gives relatively large numerical errors. Here we again do nothing to improve the numerical method of integration (later we shall use the toolkit S+WAVELETS for accurate calculation of these coefficients).

A promising case is the approximation of the Delta function. The localized Delta is almost perfectly (apart of its magnitude and smoothness) represented by the Haar partial sum with $J = 5$ due to the localized prop-

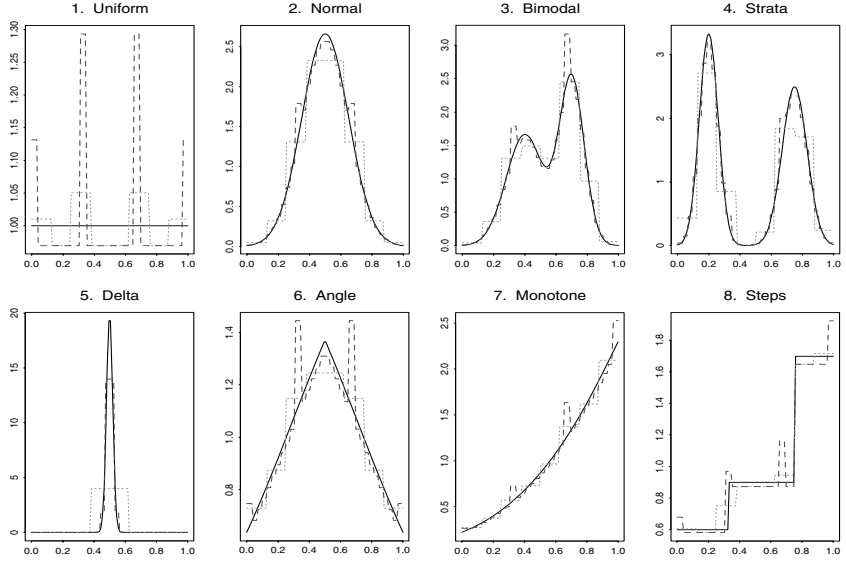

FIGURE 2.7. Approximation of the corner functions (solid lines) by the Haar basis: Dotted and short-dashed lines correspond to the number of multiresolution scales $J = 3$ and $J = 5$, respectively. [*set.J=c(3,5)*]

erty of Haar elements. The Strata is also nicely approximated, and again this is due to the localized nature of Haar elements.

With all other corner functions the situation is not too rosy, and the main issue is not even the nonsmooth approximations but the large number of Haar coefficients necessary to get a fair approximation.

An interesting example is the case of the Steps. By all means, it should be the exhibition case for the Haar system. But we see that while the second jump is perfectly shown (let us ignore the numerical errors), this is not the case for the first jump. The issue is that the second jump is perfectly positioned at the point $x = \frac{3}{4}$, while the first jump is positioned at the point $x = \frac{1}{3}$, which cannot be matched by any dyadic Haar element. Thus, a Haar approximation is forced to use a sequence of elements to approximate the first jump. The important conclusion from the Steps is that even a piecewise constant function cannot be perfectly fitted by the Haar system whenever it has a jump at a point different from 2^{-l}, $l = 0, 1, \ldots$.

To analyze Haar coefficients, the S+WAVELETS module of S–PLUS has two built-in functions: *dwt* and *mra*. The former computes the discrete wavelet transform and allows us to visualize Haar coefficients at different resolution levels. The latter computes the multiresolution analysis and allows us to visualize a set of multiresolution approximations.

Figure 2.8 illustrates the analysis of the Normal and the Steps corner functions based on $64 = 2^6$ equidistant values of the corner functions.

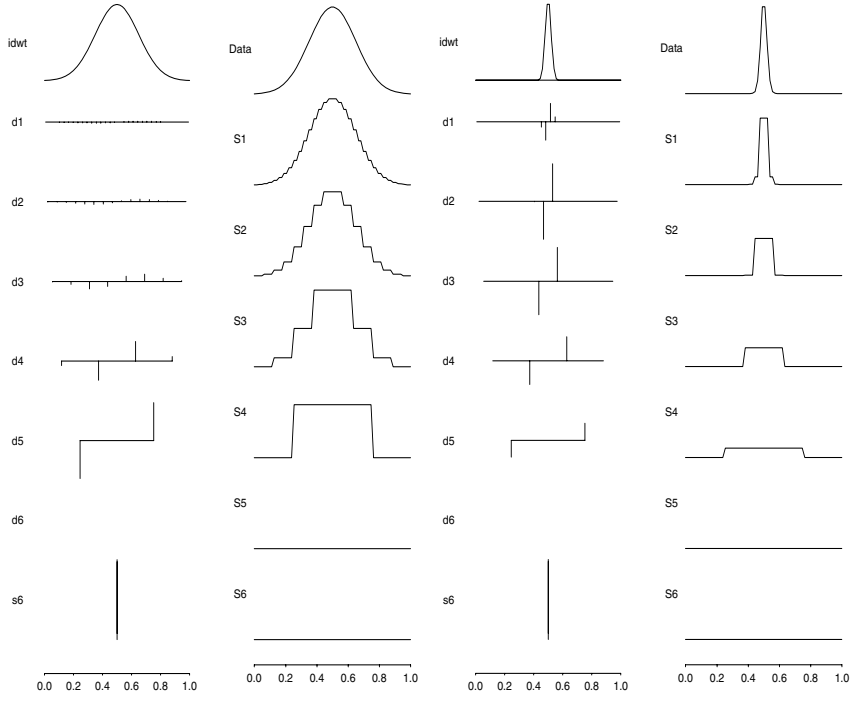

FIGURE 2.8. Haar coefficients and multiresolution approximations of the Normal and the Delta functions by the Haar system. Approximations are based on 2^L equidistant values of the functions; the default is $L = 6$. {The set of approximated corner functions is controlled by the argument *set.cf*. Recall that before using any figure with wavelets, the S+WAVELETS module should be loaded using the command > **module(wavelets)** at the S–PLUS prompt.} [*set.cf =c(2,5), L=6*]

These plots are standard in S+WAVELETS, so let us explain how to read them. The first column of plots shows locations and magnitudes of Haar coefficients for the Normal. The top graph (idwt) shows the underlying curve; you can see that this is indeed the Normal. The bottom of the first column shows a magnitude of the Haar coefficient for the father function (see row s6). The Haar coefficient θ_{00} for the mother function should be shown in row d6, but the Normal function is symmetric about 0.5, so this coefficient is zero and thus not shown. In row d5 we see both approximate locations of elements and magnitudes of corresponding Haar coefficients for $M_{1,0}$ and $M_{1,1}$, etc.

The second column of plots illustrates multiresolution approximations. Row S6 shows the approximation by the father function. This approximation is often referred to as the *low-frequency* approximation. Because

$\theta_{00} = 0$, the approximation $f_0(x)$, shown in row S5, is the same as the low-frequency approximation S6. The Haar partial sum with $J = 1$ is shown in row S4, with $J = 2$ in S3, with $J = 3$ in S2, and with $J = 4$ in S1. Finally, the approximated function is shown in the top row called Data. Note that the approximation with $J = 5$, which includes the finest elements with the Haar coefficients shown in d1, is not exhibited.

Similarly, the third and fourth columns show Haar coefficients and partial sums for the Delta.

These diagrams allow us to count the number of Haar coefficients needed for a "good" approximation of the curves. Let us begin with the Normal. Probably, the approximation S1 (which corresponds to $J = 4$) may be considered as a good one, and then $2^{J+1} = 32$ Haar coefficients should be calculated. However, we see that only 24 of them are significant (to get the number 24, just calculate the total number of coefficients shown in rows s6 and d6–d2). Thus, for the Normal curve the Haar system compresses the data essentially worse than the cosine or polynomial bases (just look again at Figure 2.3, where only 6 Fourier coefficients give us an almost perfect approximation and 4 Fourier coefficients give us a good visualization). Also, Haar approximation S3, based on 7 Haar coefficients, is a caricature of the Normal.

The outcome is quite the opposite for the Delta. Here just 9 Haar coefficients give us a fair visualization S1.

What is the conclusion? We see that there is no magical orthonormal system. Roughly speaking, smooth functions are better approximated by smooth elements, and thus cosine or polynomial systems can be recommended; nonsmooth functions may be better approximated by Haar or other wavelet systems. On the other hand, knowledge of how a particular system approximates a function allows us to recognize a pattern and then, if necessary, change the system. This is the reason why it is worthwhile to know both approximation properties of a particular orthonormal system and different orthonormal systems oriented on approximation of a specific type of function.

2.2 How Fast Fourier Coefficients May Decrease

The previous section introduced us to the world of orthonormal series approximations via visualization of partial sums for 8 corner functions. Another possible approach is a theoretical one that allows us to analyze simultaneously large classes of functions f that are square integrable on $[0, 1]$, i.e., when $\int_0^1 f^2(x)dx < \infty$. This approach is based on the famous Parseval identity. For the cosine or polynomial orthonormal systems this

identity is written as

$$\int_0^1 (f(x) - f_J(x))^2 dx = \sum_{j>J} \theta_j^2, \tag{2.2.1}$$

where f_J is the partial sum (2.1.2), and for the Haar system as

$$\int_0^1 (f(x) - f_J(x))^2 dx = \sum_{j>J} \sum_{k=0}^{2^j-1} \theta_{jk}^2, \tag{2.2.2}$$

where here f_J is the partial sum (2.1.6).

Thus, the faster Fourier coefficients decrease, the smaller cutoff J is needed to get a good global approximation of f by a partial sum $f_J(x)$ in terms of the integrated squared error (ISE). Note that in nonparametric statistics the ISE is customarily called the integrated squared bias (ISB).

The aim of this section is to explain the main characteristics of a function f that influence the rate at which its Fourier coefficients decrease.

First, let us begin with the cosine system. We would like to understand what determines the rate at which Fourier coefficients $\theta_j = \int_0^1 \sqrt{2}\cos(\pi j x) f(x)dx$ of an integrable function f decrease as $j \to \infty$.

To analyze θ_j, let us recall the technique of integration by parts. If $u(x)$ and $v(x)$ are both differentiable functions, then the following equality, called *integration by parts*, holds:

$$\int_0^1 u(x)dv(x) = [u(1)v(1) - u(0)v(0)] - \int_0^1 v(x)du(x). \tag{2.2.3}$$

Here $du(x) := u^{(1)}(x)dx$ is the differential of $u(x)$, and $u^{(k)}(x)$ denotes the kth derivative of $u(x)$.

Assume that $f(x)$ is differentiable. Using integration by parts and the relations

$$d\cos(\pi j x) = -\pi j \sin(\pi j x)dx, \qquad d\sin(\pi j x) = \pi j \cos(\pi j x)dx, \tag{2.2.4}$$

we may find θ_j for $j \geq 1$,

$$\begin{aligned}\theta_j &= \sqrt{2}\int_0^1 \cos(\pi j x) f(x)dx = \sqrt{2}(\pi j)^{-1}\int_0^1 f(x)d\sin(\pi j x) \\ &= \frac{\sqrt{2}}{(\pi j)}[f(1)\sin(\pi j) - f(0)\sin(0)] - \frac{\sqrt{2}}{(\pi j)}\int_0^1 \sin(\pi j x) f^{(1)}(x)dx.\end{aligned}$$

Recall that $\sin(\pi j) = 0$ for all integers j, so we obtain

$$\theta_j = -\sqrt{2}(\pi j)^{-1}\int_0^1 \sin(\pi j x) f^{(1)}(x)dx. \tag{2.2.5}$$

Note that $|\int_0^1 \sin(\pi j x) f^{(1)}(x)dx| \leq \int_0^1 |f^{(1)}(x)|dx$, and thus we may conclude the following. *If a function $f(x)$ is differentiable, then for the cosine*

system,

$$|\theta_j| \le \sqrt{2}(\pi j)^{-1} \int_0^1 |f^{(1)}(x)|dx, \quad j \ge 1. \tag{2.2.6}$$

We established the first rule (regardless of a particular f) about the rate at which the Fourier coefficients decrease. Namely, if f is differentiable and $\int_0^1 |f^{(1)}(x)|dx < \infty$, then $|\theta_j|$ decrease with rate at least j^{-1}.

Let us continue the calculation. Assume that f is twice differentiable. Then using the method of integration by parts on the right-hand side of (2.2.5), we get

$$\theta_j = -\frac{\sqrt{2}}{\pi j} \int_0^1 \sin(\pi j x) f^{(1)}(x)dx = \frac{\sqrt{2}}{(\pi j)^2} \int_0^1 f^{(1)}(x) d\cos(\pi j x)$$

$$= \frac{\sqrt{2}}{(\pi j)^2}[f^{(1)}(1)\cos(\pi j) - f^{(1)}(0)\cos(0)] - \frac{\sqrt{2}}{(\pi j)^2} \int_0^1 \cos(\pi j x) f^{(2)}(x)dx. \tag{2.2.7}$$

We conclude that *if $f(x)$ is twice differentiable then for some finite constant c,*

$$|\theta_j| \le cj^{-2} \int_0^1 |f^{(2)}(x)|dx, \quad j \ge 1 \ . \tag{2.2.8}$$

Thus, the Fourier coefficients θ_j of smooth (twice differentiable) functions decrease with rate not slower than j^{-2}.

So far, boundary conditions (i.e., values of $f(x)$ near boundaries of the unit interval $[0, 1]$) have not affected the rate. The situation changes if f is smoother, for instance, it has three derivatives. In this case integration by parts can be used again. However, now the decrease of θ_j may be defined by boundary conditions, namely by the term $[f^{(1)}(1)\cos(\pi j) - f^{(1)}(0)\cos(0)]$ on the right-hand side of (2.2.7). Note that $\cos(\pi j) = (-1)^j$, so all these terms are equal to zero only if $f^{(1)}(1) = f^{(1)}(0) = 0$. *This is the boundary condition that allows θ_j to decrease faster than j^{-2}.* Otherwise, *if the boundary condition does not hold, then θ_j cannot decrease faster than j^{-2} regardless of how smooth the underlying function f is.*

Now we know two main factors that define the decay of Fourier coefficients of the cosine system and therefore the performance of an orthonormal approximation: smoothness and boundary conditions.

A customary rule of thumb, used by many statisticians, is that an underlying function is twice differentiable. As we have seen, for twice differentiable functions the cosine system yields the optimal decrease of θ_j regardless of the boundary conditions. Thus, the cosine system may be a good tool in statistical applications.

The case of the polynomial basis is discussed in Section 2.6.

Now let us consider a similar problem for the Haar basis. Using the specific shape of the mother function $M(x)$ (see Figure 2.6) we write,

$$|\theta_{jk}| = \int_{k2^{-j}}^{(k+1)2^{-j}} f(x)M_{jk}(x)dx$$
$$\le 2^{-j}(\max|M_{jk}(x)|)(\max f(x) - \min f(x))/2,$$

where both the maximum and the minimum are taken over $x \in [k2^{-j}, (k+1)2^{-j}]$. Because $\max|M_{jk}(x)| = 2^{j/2}$, we get

$$\sum_{k=0}^{2^j-1} |\theta_{jk}| \le 2^{-1-j/2} \sup \sum_{k=1}^{2^{j+1}} |f(t_k) - f(t_{k-1})|, \tag{2.2.9}$$

where the supremum (see definition of the supremum below line (A.45) in Appendix A) is taken over all possible partitions of the unit interval $0 \le t_0 < t_1 < \cdots < t_{2^{j+1}} \le 1$.

The quantity $\mathrm{TV}(f) := \lim_{m\to\infty} \sup \sum_{k=1}^{m} |f(t_k) - f(t_{k-1})|$, where the supremum is taken over all possible partitions $0 \le t_0 < t_1 < \cdots < t_m \le 1$ of the unit interval, is called the *total variation* of the function f on $[0, 1]$. Note that the total variation of a monotone function is equal to $|f(1)-f(0)|$.

Thus, we get from (2.2.9) that

$$\sum_{k=0}^{2^j-1} |\theta_{jk}| \le 2^{-1-j/2}\,\mathrm{TV}(f). \tag{2.2.10}$$

This inequality shows how the sum of absolute values of Haar coefficients at a resolution scale j decreases as j increases. Such behavior is typical for wavelet coefficients (see more in Section 2.5).

Absolutely similarly we establish that

$$\sum_{k=0}^{2^j-1} |\theta_{jk}|^2 \le 2^{-2-j}\,(\mathrm{QV}(f))^2, \tag{2.2.11}$$

where

$$\mathrm{QV}(f) := \lim_{m\to\infty} \sup \left(\sum_{k=1}^{m} |f(t_k) - f(t_{k-1})|^2 \right)^{1/2} \tag{2.2.12}$$

is called the *quadratic variation* of f on $[0, 1]$. Here again the supremum is taken over all possible partitions $0 \le t_0 < t_1 < \cdots < t_m \le 1$ of the unit interval.

These are the fundamentals that we need to know about the series approximation. The topic of how the decay of Fourier coefficients depends on various properties of an underlying function and, conversely, what Fourier coefficients may tell us about an underlying function, is a well-developed branch of mathematics. We shall discuss more formally other interesting mathematical results and approaches in the following sections.

2.3 Special Topic: Geometry of Square Integrable Functions

It would be beneficial to know that square integrable functions, which are the primary target in nonparametric curve estimation, may be viewed like points or vectors in a finite-dimensional Euclidean space, only with their own notion of perpendicular coordinates, distance, angle, Pythagorean theorem, etc.

Denote by $L_2 = L_2([0,1])$ the space of all square integrable functions with domain $[0,1]$. In other words, L_2 is the set of all functions f such that $\|f\| := (\int_0^1 |f(x)|^2 dx)^{1/2} < \infty$. Note that bounded functions belong to L_2 because if $|f(x)| \le c < \infty$, then $\int_0^1 |f(x)|^2 dx \le c^2 \int_0^1 dx = c^2 < \infty$.

Below, the geometry of L_2 is discussed via a sequence of steps that make the similarity between L_2 and k-dimensional Euclidean space $\mathcal{E}_k$ of points $\mathbf{v} = (v_1, \ldots, v_k)$ apparent. We shall also consider vectors in $\mathcal{E}_k$ that are directed line segments like ones shown in Figure 2.9. In what follows we shall denote by $\vec{v}$ the vector from the origin to the point $\mathbf{v}$. Figure 2.9 reminds us the main rule of finding the difference (and respectively the sum) of two vectors.

• L_2 **is a linear space.** If $\vec{v}$ and $\vec{u}$ are two vectors in $\mathcal{E}_k$, then $a\vec{v} + b\vec{u} \in \mathcal{E}_k$ for any real numbers a and b. A space with this property is called *linear* because any linear combination of its elements is again an element of this space. Let us verify that L_2 is linear. Using the *Cauchy inequality* $2|abf(x)g(x)| \le a^2 f^2(x) + b^2 g^2(x)$, which is a corollary of the elementary $|af(x) - bg(x)|^2 \ge 0$, implies

$$\|af + bg\|^2 = \int_0^1 (af(x) + bg(x))^2 dx \le 2a^2\|f\|^2 + 2b^2\|g\|^2 \ . \qquad (2.3.1)$$

Thus, any linear combination of two square integrable functions is again a square integrable function. In short, $af + bg \in L_2$ whenever $f, g \in L_2$ and a and b are real numbers.

• **Distance between two square integrable functions.** If $\mathbf{v}$ and $\mathbf{u}$ are two points in $\mathcal{E}_k$ then the Euclidean distance (length, norm) between them is $[\sum_{j=1}^k (v_j - u_j)^2]^{1/2}$. Note that this definition is based on the Pythagorean theorem and the orthonormality of the basic vectors of the Cartesian coordinate system. In particular, the length (norm) of a vector $\vec{v}$, which is the distance between the origin and the point $\mathbf{v}$, is $[\sum_{j=1}^k v_j^2]^{1/2}$. Also, the length of the difference $\vec{v} - \vec{u}$ of two vectors corresponding to points $\mathbf{v}$ and $\mathbf{u}$ is $[\sum_{j=1}^k (v_j - u_j)^2]^{1/2}$, which is exactly the distance between those points.

For functions we may define the *distance* between two square integrable functions f and g as $\|f - g\|$. In particular, this definition implies that the *norm* of f is $\|f\|$. Below, we shall see that this definition preserves all properties of classical Euclidean geometry.

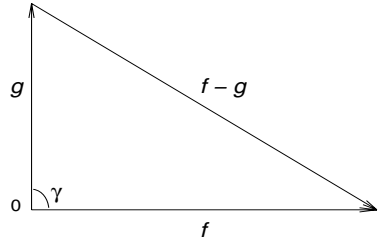

FIGURE 2.9. Illustration to the discussion of the orthogonality between two square integrable functions. The origin is denoted by 0.

• **The orthogonality (perpendicularity) of square integrable functions.** The crown jewel of Euclidean geometry is the Pythagorean theorem. Recall that this famous theorem is about a right triangle whose two sides are perpendicular, i.e., the angle between them is 90° (see Figure 2.9, where the angle γ is 90°). Recall that the side opposed to the right angle is called the hypotenuse, and the other sides are called legs. The Pythagorean theorem states that the sum of the squares of the legs of a right triangle is equal to the square of the hypotenuse. Moreover, this is a property only of right triangles. In other words, to check that two sides are perpendicular it suffices to check that the sum of the squares of these sides is equal to the square of the other side.

Let us use this Pythagorean rule for introducing the notion of orthogonal (or one may say perpendicular) square integrable functions. Figure 2.9 illustrates the underlying idea. Let f and g be two square integrable functions, which may be thought as either points in L_2 or the corresponding vectors. As we have defined earlier, their lengths (norms) in L_2 are $\|f\|$ and $\|g\|$. The Pythagorean rule together with Figure 2.9 implies that if these two functions are orthogonal (perpendicular), then the equality $\|f\|^2 + \|g\|^2 = \|f-g\|^2$ must hold. Let us check when this happens:

$$\|f-g\|^2 = \int_0^1 (f(x)-g(x))^2 dx = \|f\|^2 + \|g\|^2 - 2\int_0^1 f(x)g(x)dx\,. \quad (2.3.2)$$

Thus, we may say that two square integrable functions f and g are *orthogonal* (perpendicular) in L_2 if their *inner product* $\langle f, g\rangle := \int_0^1 f(x)g(x)dx$ is zero. Moreover, the angle γ between two functions in L_2 may be defined via the relation

$$\cos(\gamma) := \langle f, g\rangle / [\|f\|\,\|g\|]. \quad (2.3.3)$$

The definition (2.3.3) fits the geometry of Euclidean space $\mathcal{E}_k$, where the inner product, also referred to as the dot product, is defined as $\langle \vec{v}, \vec{u}\,\rangle = \sum_{j=1}^k v_j u_j$. Let us check that the absolute value of the right side of (2.3.3)

is at most 1. This follows at once from the *Cauchy–Schwarz inequality*

$$\langle f, g\rangle \le \|f\|\,\|g\|, \tag{2.3.4}$$

where equality holds iff $f = ag$ for some real number a. Let us prove this assertion. First, note that if $\|f\|\,\|g\| = 0$, then the assertion clearly holds. Thus, consider the case $\|f\|\,\|g\| > 0$. As in (2.3.1), for $t_1(x) := f(x)/\|f\|$ and $t_2(x) := g(x)/\|g\|$ we may write

$$0 \le \|t_1 - t_2\|^2 = \|t_1\|^2 + \|t_2\|^2 - 2\langle t_1, t_2\rangle.$$

This together with $\|t_1\| = \|t_2\| = 1$ implies (2.3.4), with equality if and only if $\|t_1 - t_2\| = 0$, which is equivalent to $f = ag$.

Finally, to finish our "triangles" business, recall that Euclidean geometry tells us that a side of a triangle is not longer than the sum of the other two sides. Such a property is called the *triangle inequality*. This property also holds in L_2. Indeed, (2.3.2) together with the Cauchy–Schwarz inequality implies

$$\|f \pm g\| \le \|f\| + \|g\| \ . \tag{2.3.5}$$

• **Coordinate system in L_2.** What makes a Euclidean space so transparent and intuitively clear? Why is the procedure of depicting a point in this space so simple? The answer is obvious: the familiar Cartesian (rectangular) coordinates make this space so convenient. Thus, let us briefly recall this coordinate system and then try to introduce its analogue for L_2.

Cartesian coordinates in $\mathcal{E}_k$ are defined by k perpendicular basic unit vectors $\{\vec{b}_1, \dots, \vec{b}_k\}$. By definition, the basic vectors proceed from the origin to the points whose Cartesian coordinates are $\{(1, 0, \dots, 0), \dots, (0, \dots, 0, 1)\}$. Then a vector $\vec{v}$ may be written as $\vec{v} = \sum_{j=1}^{k} v_j \vec{b}_j$, and it is easy to check that $v_j = \langle \vec{v}, \vec{b}_j\rangle$.

Thus, "translation" of the Cartesian system of coordinates into the "language" of the space of square integrable functions is straightforward. Let $\{\varphi_1, \varphi_2, \dots\}$ be a system of square integrable functions that are pairwise orthogonal and have unit norms, that is, $\langle \varphi_j, \varphi_l\rangle = 0$ if $j \ne l$ and $\|\varphi_j\| = 1$. This system is called *orthonormal*. Also, let us assume that this system *spans* L_2, that is, for any $f \in L_2$ and $\varepsilon > 0$ there exist a positive integer n and numbers $c_1, \dots, c_n$ such that $\|f - \sum_{j=1}^{n} c_j \varphi_j\| \le \varepsilon$. If such a system of functions exists, then it is called an *orthonormal basis*, or simply *basis*.

Then the elements of a basis may be declared as the basic unit vectors in L_2. Indeed, they are orthonormal, and it is possible to show that as is the case for a finite-dimensional Euclidean space,

$$\Big\|f - \sum_{j=1}^{n} \theta_j \varphi_j\Big\| = \min_{\{c_j\}} \Big\|f - \sum_{j=1}^{n} c_j \varphi_j\Big\|, \quad \text{where} \ \ \theta_j = \langle f, \varphi_j\rangle. \tag{2.3.6}$$

In other words, the best representation of a function by a linear combination of n basic unit vectors is one in which the coefficients are Fourier coeffi-

cients. Thus Fourier coefficients play the role of coordinates of a function in the space L_2 where the coordinate system is created by the orthonormal elements of the basis.

Let us prove (2.3.6). Write

$$\begin{aligned}\Big\|f-\sum_{j=1}^{n}c_j\varphi_j\Big\|^2 &= \Big\|f-\sum_{j=1}^{n}\theta_j\varphi_j+\sum_{j=1}^{n}(\theta_j-c_j)\varphi_j\Big\|^2\\ &= \Big\|f-\sum_{j=1}^{n}\theta_j\varphi_j\Big\|^2+\Big\|\sum_{j=1}^{n}(\theta_j-c_j)\varphi_j\Big\|^2\\ &\quad+2\Big\langle f-\sum_{j=1}^{n}\theta_j\varphi_j,\sum_{j=1}^{n}(\theta_j-c_j)\varphi_j\Big\rangle\,.\end{aligned}$$

The orthonormality of the elements $\{\varphi_j\}$ together with the definition of the Fourier coefficients θ_j implies that the inner product term is zero. This yields (2.3.6).

• **Gram–Schmidt orthonormalization.** How can one construct a basis in L_2? To answer this question, let us assume that a countable system $\{\psi_1,\psi_2,\dots\}$ of square integrable functions spans L_2 (in other words, this system is *dense* in L_2). Note that we can always discard an element ψ_n of this system if it is a linear combination of the previous elements $\psi_1,\dots,\psi_{n-1}$, that is, if $\psi_n(x)=\sum_{l=1}^{n-1}c_l\psi_l(x)$. Thus, let us assume that this system contains only linearly independent elements.

Then a basis may be constructed using the *Gram–Schmidt orthonormalization procedure*. The first element φ_1 is defined by

$$\varphi_1(x):=\psi_1(x)/\|\psi_1\|. \tag{2.3.7}$$

Then all the following elements are defined by the recursion

$$\varphi_j(x):=\frac{\psi_j(x)-\sum_{l=1}^{j-1}\langle\psi_j,\varphi_l\rangle\varphi_l(x)}{\big\|\psi_j(x)-\sum_{l=1}^{j-1}\langle\psi_j,\varphi_l\rangle\varphi_l(x)\big\|}\,. \tag{2.3.8}$$

• **The projection theorem.** The notion of a projection of a point onto a set of points is well known for Euclidean spaces. The *projection* of a point onto a set is defined as the point of this set that is nearest to the point. If there is more than one such point, then all these points are called projections. For instance, in a plane, the projection of a point onto a straight line is the foot of the perpendicular from the point to the line. In this case the (orthogonal) projection is unique. In three-dimensional space, the projection of a point onto a plane is also unique: This is the foot of the perpendicular from the point to the plane. Of course, there are plenty of examples where a projection is not unique. For instance, a projection of the center of a circle onto the circle is not unique, because all points of the circle are equidistant from its center. Note that the difference between the line–plane case and the circle case is that a line and a plane are linear

subspaces of two-dimensional Euclidean space, while the circle is not. Also, the projection may not exist. For instance, consider points on the real line and let us try to project the point 2 onto the interval $[0, 1)$. There is no nearest point, because the point 1, which could be a natural projection, does not belong to the interval. On the other hand, the projection onto $[0, 1]$ is well-defined, and it is the point 1. Note that the first interval is half open, while the second is closed, and this is what makes the difference.

Keeping these examples in mind, we would like to formulate the result about a unique projection in L_2 of a function f onto a linear subspace. Let us say that a linear subspace $\mathcal{L}$ of L_2 is a *closed* subspace if $\mathcal{L}$ contains all of its limits points, that is, if $f_n \in \mathcal{L}$ and $\|g - f_n\| \to 0$ then $g \in \mathcal{L}$. The following result states that the projection of a function f onto a closed linear subspace is always unique, and moreover, the geometry of this projection is absolutely similar to the examples for Euclidean spaces.

The projection theorem. *Let $\mathcal{L}$ be a closed linear subspace of L_2. Then for each $f \in L_2$ there exists a unique element $f^* \in \mathcal{L}$ (the projection of f onto $\mathcal{L}$) such that*

$$\|f - f^*\| = \inf_{g \in \mathcal{L}} \|f - g\|. \tag{2.3.9}$$

Moreover, f^* is the projection iff the difference $f - f^*$ is orthogonal to all elements of $\mathcal{L}$. (Thus, the projection f^* is unique, and it may be called the orthogonal projection of f onto $\mathcal{L}$.)

The proof of this theorem may be found, for instance, in the textbook by Debnath and Mikusinski (1990).

• **Hilbert space.** The previous step finished our discussion of the geometry of L_2. On the other hand, we are so close to understanding the notion of a Hilbert space that it is irresistible to discuss it here because both Euclidean spaces and L_2 are particular examples of a Hilbert space.

Let $\mathcal{H}$ be a linear space with an inner product. Denote by x, y, and z any 3 elements of $\mathcal{H}$ and by a any real number. An inner product $\langle x, y\rangle$ should satisfy the following properties: $\langle x, y\rangle = \langle y, x\rangle$; $\langle x + y, z\rangle = \langle x, y\rangle + \langle y, z\rangle$; $\langle ax, y\rangle = a\langle y, x\rangle$; $\langle x, x\rangle \geq 0$, with equality iff $x = 0$. The distance in $\mathcal{H}$ between two elements is declared to be $\|x - y\| := \langle x - y, x - y\rangle^{1/2}$. Then the space $\mathcal{H}$ is called a *Hilbert* space if for any sequence of elements $x_n \in \mathcal{H}$ the fact $\|x_n - x_m\| \to 0$ as $n, m \to \infty$ implies that x_n converges to some $x \in \mathcal{H}$ (in other words, any Cauchy sequence converges to an element of the Hilbert space, and this property is called *completeness*). One more definition is due: A Hilbert space is called *separable* if there exists a countable system of elements that approximates any other element from the space.

It is possible to show that L_2, with the inner product defined via the Lebesgue integral, is a separable Hilbert space. A proof of this result may be found in the textbook by Debnath and Mikusinski (1990), and a sketch of a proof will be given in the next section. We do not discuss here the

Lebesgue integral but note that for all practically interesting functions discussed in the book it is equal to the Riemann integral.

• **Two useful relations.** Let $\{\varphi_j\}$ be an orthonormal basis in L_2 and let $\theta_j = \langle f, \varphi_j \rangle = \int_0^1 f(x)\varphi_j(x)dx$ be the jth Fourier coefficient of $f \in L_2$. Then the following relations hold: The *Bessel inequality*

$$\sum_{j=1}^{n} \theta_j^2 \le \|f\|^2 , \ \ n = 1, 2, \dots , \tag{2.3.10}$$

and the *Parseval identity*

$$\|f\|^2 = \sum_{j=1}^{\infty} \theta_j^2. \tag{2.3.11}$$

The Bessel inequality is implied by the line

$$0 \le \Big\| f - \sum_{j=1}^{n} \theta_j \varphi_j \Big\|^2 = \|f\|^2 + \Big\| \sum_{j=1}^{n} \theta_j \varphi_j \Big\|^2 - 2\Big\langle f, \sum_{j=1}^{n} \theta_j \varphi_j \Big\rangle = \|f\|^2 - \sum_{j=1}^{n} \theta_j^2.$$

The fact that $\{\varphi_j\}$ is a basis in L_2 means that $\|f - \sum_{j=1}^{n} \theta_j \varphi_j\| \to 0$ as $n \to \infty$. This together with the last line yields the Parseval identity.

2.4 Special Topic: Classical Trigonometric Series

The classical orthonormal trigonometric Fourier system is defined by

$$\begin{aligned} &\varphi_0(x) := 1, \quad \varphi_{2j-1}(x) := \sqrt{2}\sin(2\pi j x), \\ &\varphi_{2j}(x) := \sqrt{2}\cos(2\pi j x), \quad j = 1, 2, \dots . \end{aligned} \tag{2.4.1}$$

Our first object is to discuss how the *partial trigonometric (Fourier) sums*

$$S_J(x) := \sum_{j=0}^{2J} \theta_j \varphi_j(x) \tag{2.4.2}$$

approximate an underlying integrable function f and why this system is a basis in L_2. In this section $\theta_j := \int_0^1 f(x)\varphi_j(x)dx$ denote the Fourier coefficients, which are well-defined for integrable f.

Fourier sums for the corner functions are shown in Figure 2.10. Be aware that here the Fourier sum S_J is based on $1+2J$ Fourier coefficients. Because the trigonometric elements (2.4.1) are 1-periodic, the Fourier sums are 1-periodic as well. This definitely shows up in the approximation of functions like the Monotone and the Steps (Section 2.6 explains how to improve approximations of aperiodic functions). Also, the approximations of the Steps again exhibit the Gibbs phenomenon of overshooting. Interestingly, as $J \to \infty$, the overshoot approaches approximately 9% of a jump. (Historical

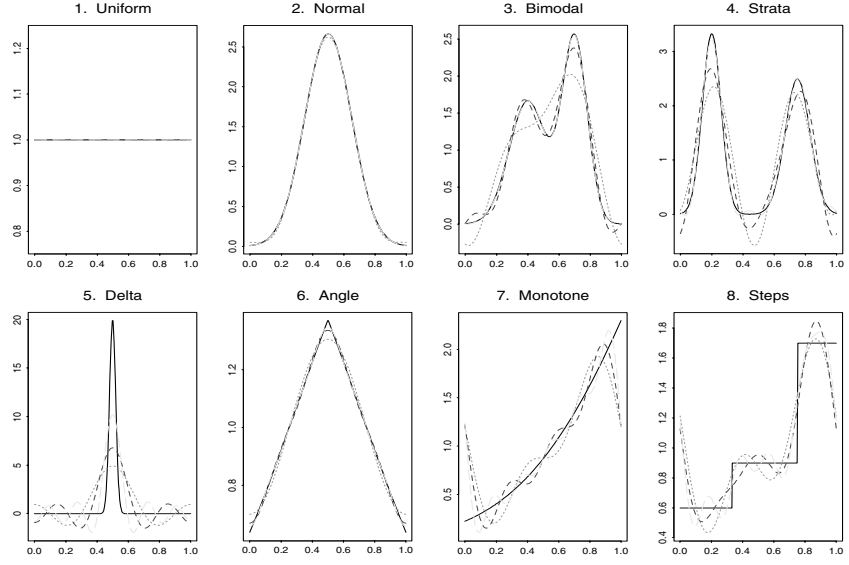

FIGURE 2.10. Approximation of the corner functions (solid lines) by Fourier sums: Dotted, short-dashed, and long-dashed lines correspond to $J = 2$, $J = 3$, and $J = 5$, respectively. [*set.J =c(2,3,5)*]

notes and discussion of this interesting phenomenon may be found in Dym and McKean 1972, Section 1.6.) In general, the Gibbs phenomenon occurs in the vicinity of any jump of a piecewise smooth function, and at this point the Fourier sums converge to the average value of the function. This is clearly exhibited in Figure 2.10.7. (It is worthwhile to know that even some wavelet expansions, discussed in the following section, suffer from overshooting. Thus, in one way or another nonsmooth functions present a challenge for any orthonormal system.)

Let us now focus on the theory of convergence of Fourier sums $S_J(x)$ to $f(x)$ at a given point $x \in [0, 1]$ as $J \to \infty$. Such convergence is called *pointwise*. Substituting the expressions for Fourier coefficients θ_j into the right-hand side of (2.4.2) yields

$$S_J(x) = 2 \int_0^1 f(t) \Big[\frac{1}{2} + \sum_{k=1}^{J} (\cos(2\pi kx) \cos(2\pi kt) + \sin(2\pi kx) \sin(2\pi kt)) \Big] dt,$$

and because $\cos(\alpha - \beta) = \cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta)$, we get

$$S_J(x) = 2 \int_0^1 f(t) \Big[\frac{1}{2} + \sum_{k=1}^{J} \cos(2\pi k(t - x)) \Big] dt. \tag{2.4.3}$$

The expression inside the square brackets may be simplified using the following trigonometric formula and notation,

$$\frac{1}{2}+\sum_{k=1}^{J}\cos(2\pi ku)=\frac{1}{2}\frac{\sin(\pi(2J+1)u)}{\sin(\pi u)}=:\frac{1}{2}D_J(u)\ , \tag{2.4.4}$$

with the understanding that $D_J(0) := 2J+1$. The function $D_J(u)$ is called the *Dirichlet kernel*, and it plays a central role in the study of pointwise convergence.

Note that graphs of $D_J(u-.5)$ resemble the approximations in Figure 2.10.5. Namely, as $J \to \infty$, the peak tends to infinity and the symmetric oscillations to either side of the peak become increasingly rapid, and while they do not die away, on the average they cancel each other. In short, as $J \to 0$, the Dirichlet kernels approximate the theoretical delta function.

From now on let us assume that $f(x)$ is 1-periodic, that is, $f(x+1) = f(x)$ for all x (in this case the unit interval may be thought as a unit circular circumference with identified endpoints 0 and 1). Then the function $f(t)D_J(t-x)$ is also 1-periodic in t. The substitution $z = t - x$ gives

$$S_J(x)=\int_0^1 f(x+z)D_J(z)dz. \tag{2.4.5}$$

Recall that the theoretical delta function is integrated to 1, and from (2.4.5) it is easy to see by choosing $f(x) \equiv 1$ that the Dirichlet kernel has the same property,

$$\int_0^1 D_J(z)dz=1. \tag{2.4.6}$$

Thus we may write

$$S_J(x)-f(x)=\int_0^1 [f(x+z)-f(x)]D_J(z)dz. \tag{2.4.7}$$

An important conclusion from (2.4.7) is that a pointwise approximation should crucially depend on the local smoothness of an approximated function f in the vicinity of x.

The expression (2.4.7) is the key to all the main properties of the Fourier sum. For instance, assume that f is a twice differentiable function, and set $g(x,z) := (f(x-z)-f(x))/\sin(\pi z)$. Under the assumption, the partial derivative $\partial g(x,z)/\partial z$ exists, and let us additionally assume that this derivative is bounded. Then integration by parts implies

$$S_J(x)-f(x)=\int_0^1 g(x,z)\sin(\pi(2J+1)z)dz$$

$$=(\pi(2J+1))^{-1}\Big[g(x,1)+g(x,0)+\int_0^1(\partial g(x,z)/\partial z)\cos(\pi(2J+1)z)dz\Big].$$

Thus, under our assumption (recall that C is a generic positive constant)

$$\max_x |S_J(x) - f(x)| < CJ^{-1}. \tag{2.4.8}$$

This result allows us to make the following two conclusions. First, if an approximated function is sufficiently smooth, then it may be uniformly approximated by Fourier sums. Second, because twice differentiable functions may approximate any square integrable function in the L_2-norm, this together with (2.4.8) implies that the trigonometric system is a basis in L_2.

More properties of Fourier sums may be found in Exercises 2.4.2–4.

For pointwise convergence it might be a good idea to *smooth* (*shrink*) the Fourier coefficients, that is, to multiply them by some real numbers between 0 and 1. Smoothing (shrinkage) is also a key idea of adaptive nonparametric series estimation (as well as many other statistical approaches). Let us consider two famous smoothing procedures.

The *Fejér (Cesáro) sum* is the average of Fourier sums,

$$\sigma_J(x) := [S_0(x) + S_1(x) + \cdots + S_{J-1}(x)]/J. \tag{2.4.9}$$

It is easy to see that σ_J is a smoothed partial sum S_{J-1}. Indeed,

$$\sigma_J(x) = \theta_0\varphi_0(x) + \sum_{j=1}^{J-1}(1 - j/J)[\theta_{2j-1}\varphi_{2j-1}(x) + \theta_{2j}\varphi_{2j}(x)].$$

A remarkable property of the Fejér sum is that if f is nonnegative, then σ_J is also nonnegative (thus the Fejér sum is a bona fide approximation for probability densities). To check this we use (2.4.3)–(2.4.4) and write

$$\sigma_J(x) = J^{-1}\int_0^1 \Big[\sum_{k=0}^{J-1} \sin(\pi(2k+1)z)/\sin(\pi z)\Big] f(x+z)dz.$$

This equality together with $\sum_{k=0}^{J-1}\sin(\pi(2k+1)z) = \sin^2(\pi Jz)/\sin(\pi z)$ implies

$$\sigma_J(x) = J^{-1}\int_0^1 [\sin(\pi Jz)/\sin(\pi z)]^2 f(x+z)dz. \tag{2.4.10}$$

This expression yields the nonnegativity of the Fejér sum whenever f is nonnegative. The function $\Phi_J(z) := J^{-1}[\sin(\pi Jz)/\sin(\pi z)]^2$ is called the *Fejér kernel*. Thus $\sigma_J(x) = \int_0^1 \Phi_J(z)f(x+z)dz$.

Another useful property of the Fejér sum is that if f is continuous and 1-periodic, then $\sigma_n(x) \to f(x)$ uniformly over all $x \in [0,1]$. Also, the Fejér sum does not "overshoot" (does not suffer from the Gibbs phenomenon). The proof may be found in Dym and McKean (1972, Theorem 1.4.3).

The performance of Fejér sums is shown in Figure 2.11 (note that the number of elements of the trigonometric basis used is the same as in Figure 2.10). Due to the smoothing, the Fejér approximations are worse for visualizing modes than the similar trigonometric approximations (just compare

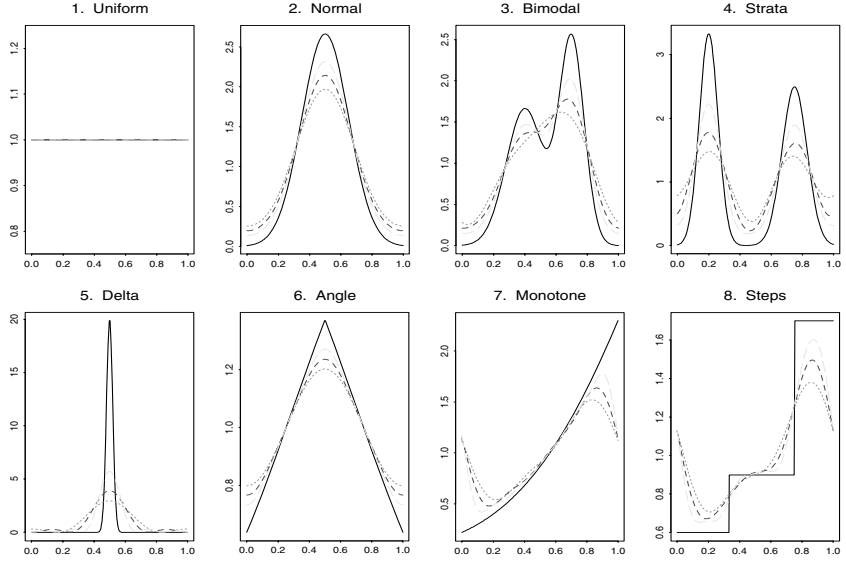

FIGURE 2.11. Approximation of the corner functions (solid lines) by Fejér (Cesáro) sums: dotted, short-dashed, and long-dashed lines correspond to $J = 3$, $J = 4$, and $J = 6$, respectively. [*set.J = c(3,4,6)*]

the approximations for the Bimodal, the Strata, and the Angle). On the other hand, the approximations of the Delta are nonnegative, the first step in the Steps is shown much better, and there are no overshoots. In short, we see exactly what has been predicted by the theory.

One more interesting property of Fejér sums is that $|\sigma_J(x)| \le \max_x |f(x)|$; see Exercise 2.4.7 and Figure 2.11. Fourier sums have no such property. On the other hand, if f is a trigonometric polynomial of degree n, then Fourier sums S_J are equal to f for $J \ge n$, while Fejér sums have no such property. The next smoothing procedure has both these properties.

The *de la Vallée Poussin* sum, which is a trigonometric polynomial of degree $2J - 1$, is given by

$$\begin{aligned} V_J(x) &:= (S_J + S_{J+1} + \ldots + S_{2J-1})/J \qquad (2.4.11)\\ &= \theta_0\varphi_0(x) + \sum_{j=1}^{J-1}[\theta_{2j-1}\varphi_{2j-1}(x) + \theta_{2j}\varphi_{2j}(x)] \\ &\quad + \sum_{j=J}^{2J-1}(2 - j/J)[\theta_{2j-1}\varphi_{2j-1}(x) + \theta_{2j}\varphi_{2j}(x)]. \end{aligned}$$

It is clear that $V_J(x) = f(x)$ if f is a trigonometric polynomial of degree $n \le J$. Also, direct calculations show that $V_J(x) = 2\sigma_{2J}(x) - \sigma_J(x)$, which implies $|V_J(x)| \le 3 \max_x |f(x)|$.

Another remarkable property of this sum is that under very mild conditions it converges uniformly to f, and it is within a factor 4 of the best sup-norm approximation by trigonometric polynomials of degree J. Also, for the case of smooth functions there is a simple formula for the pointwise approximation error.

To describe these properties mathematically, define the (inhomogeneous) *Lipschitz* function space (class) $Lip_{0,\alpha,L}$, $0 < \alpha \le 1$, $0 < L < \infty$, of 1-periodic functions:

$$Lip_{0,\alpha,L} := \{f : \sup_x |f(x)| < \infty, \quad \sup_{x,h} |f(x+h) - f(x)||h^{-\alpha}| \le L < \infty \}. \tag{2.4.12}$$

Here α is the order and L is the constant of the Lipschitz space.

Also, we define a Lipschitz space $Lip_{r,\alpha,L}$ of r-fold differentiable and 1-periodic (including the derivatives) functions:

$$Lip_{r,\alpha,L} := \{f : \sup_x |f(x)| < \infty, \quad f^{(r)} \in Lip_{0,\alpha,L} \}. \tag{2.4.13}$$

Here $f^{(r)}$ denotes the rth derivative of f. (A Lipschitz space of order $\alpha < 1$ is often referred to as a *Hölder* space. We shall use this notion in the next section which is devoted to wavelets, because wavelet coefficients characterize Hölder functions but not Lipschitz functions of order $\alpha = 1$.)

Proposition 2.4.1. Let us restrict our attention to 1-periodic functions f. Then for any trigonometric polynomial T_J of degree J, i.e., $T_J(x) = \sum_{j=0}^{2J} c_j \varphi_j(x)$, the *de la Vallée Poussin inequality* holds:

$$\sup_{f \in L_p} (\|V_J - f\|_p / \|T_J - f\|_p) \le 4, \qquad 1 \le p \le \infty. \tag{2.4.14}$$

Here L_p-norms are defined as $\|g\|_p := (\int_0^1 |g(x)|^p dx)^{1/p}$, $1 \le p < \infty$, $\|g\|_\infty := \sup_{x \in [0,1]} |f(x)|$ is the sup-norm of g, and $L_p := \{g : \|g\|_p < \infty\}$ is the L_p space of functions with finite L_p-norm.

Proposition 2.4.2. For any integer $r \ge 0$, $0 < \alpha \le 1$, and a finite L there exists a constant c such that

$$\sup_{f \in Lip_{r,\alpha,L}} \sup_{x \in [0,1]} |V_J(x) - f(x)| \le c J^{-\beta} \ , \quad \beta := r + \alpha \, . \tag{2.4.15}$$

The proofs of these propositions may be found in Temlyakov (1993, p. 68) and DeVore and Lorentz (1993, p. 205), respectively.

Due to these properties, de la Vallée Poussin sums are the primary tool in pointwise estimation of functions. Figure 2.12 exhibits these sums.

Only for the case $p = 2$ do Fourier sums "enjoy" the nice properties formulated in these propositions. This is not a big surprise, because Fourier coefficients and Fourier sums are specifically designed to perform well for square-integrable functions. On the other hand, it is amazing that a simple smoothing (shrinkage) of Fourier coefficients allows one to attain the best

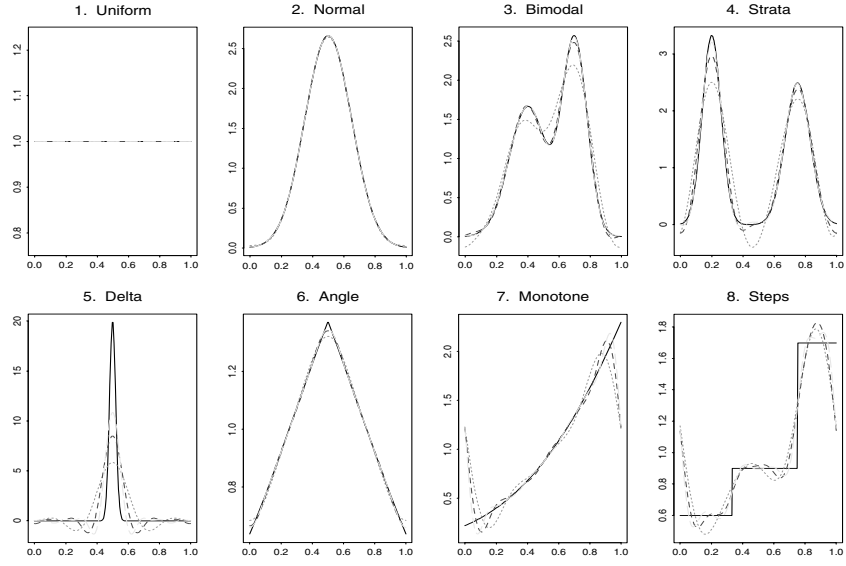

FIGURE 2.12. Approximation of the corner functions (solid lines) by de la Vallée Poussin sums: Dotted, short-dashed, and long-dashed lines correspond to $J = 2$, $J = 3$, and $J = 4$, respectively. [*set.J = c(2,3,4)*]

possible convergence within a reasonable factor in any L_p-norm. This fact tells us that it is worthwhile to use Fourier coefficients as building blocks in approximation of functions.

Let us return to Fourier sums and approximations in the L_2-norm. By the Parseval identity (2.3.11),

$$\|f - S_J\|^2 = \sum_{j>2J} \theta_j^2. \tag{2.4.16}$$

Recall that the approximation theory refers to the left-hand side of (2.4.16) as the integrated squared error, but we shall use here the statistical notion of the integrated squared bias (ISB),

$$\mathrm{ISB}_J(f) := \|f - S_J\|^2 = \sum_{j>2J} \theta_j^2. \tag{2.4.17}$$

According to (2.3.5), S_J is the optimal trigonometric polynomial of degree J for approximation of a square integrable function f under the L_2-norm. Thus, all known results about optimal approximation of functions from specific function classes may be applied to ISB. The next proposition states that ISB_J converges similarly to (2.4.15). It is discussed in DeVore and Lorentz (1993, p. 205), and its direct proof may be found in Bary (1964, Section 2.3).

Proposition 2.4.3. For any integer $r \ge 0$, real $\alpha \in (0, 1]$ and finite L there exists a finite constant c such that

$$\sup_{f \in Lip_{r,\alpha,L}} \mathrm{ISB}_J(f) \le cJ^{-2\beta}, \qquad \beta := r + \alpha. \tag{2.4.18}$$

There are two important function spaces defined via Fourier coefficients. The *Sobolev function space (ellipsoid)* $W_{\beta,Q}$, $0 \le \beta, Q < \infty$, is

$$W_{\beta,Q} := \Big\{ f : \ \theta_0^2 + \sum_{j=1}^{\infty} (1 + (2\pi j)^{2\beta})[\theta_{2j-1}^2 + \theta_{2j}^2] \le Q \Big\}. \tag{2.4.19}$$

Clearly, if $\beta = 0$, then according to the Parseval identity, $W_{\beta,Q}$ is the space of functions whose L_2-norm is at most $Q^{1/2}$. If f is r-fold differentiable and 1-periodic (including the derivatives), then the inequality $\|f + f^{(r)}\|^2 \le Q$ together with the Parseval identity implies $f \in W_{r,Q}$. Recall that $f^{(r)}$ denotes the rth derivative. Exercise 2.4.6 shows that a Sobolev space is larger than just a set of functions whose rth derivatives are square integrable; on the other hand, this set of functions is the main reason why Sobolev functions are considered in statistical applications. (Rules of integration and differentiation of Fourier sums may be found in Bary 1964, Sections 1.23.8–9, 1.24.)

Another important example of a function space that may be defined via Fourier coefficients is the space of *analytic* functions,

$$A_{\gamma,Q} := \{ f \ : \ |\theta_0| \le Q, \ |\theta_{2j-l}| \le Q e^{-\gamma j}, \ l = 0, 1, \ \ j = 1, 2, \ldots \}. \tag{2.4.20}$$

Analytic functions are 1-periodic and infinitely differentiable (i.e., they are extremely smooth), and the parameters (γ, Q) define a region in the xy-plane where a complex-valued function $f(x + iy)$ may be expanded into a convergent power series.

Note that a function belongs to one of the above-defined spaces if and only if the absolute values of the Fourier coefficients satisfy some restrictions. In other words, the signs of Fourier coefficients play no role in the characterization of these function spaces. In this case the basis used is called *unconditional*.

Now let us consider two different bases closely related to the classical trigonometric one.

Half-range trigonometric systems on $[0, 1]$**.** A shortcoming of the classical trigonometric basis is that any partial sum is periodic. The following half-range trigonometric (cosine) basis is popular among statisticians because it allows one to approximate aperiodic functions very nicely. This is also the reason why we introduced it in Section 2.1.

The underlying idea is that a function $f(x)$, $0 \le x \le 1$, is considered as an even 2-periodic function on the interval $[-1, 1]$; that is, $f(-x) = f(x)$. Then the classical trigonometric basis is used, and because $\int_{-1}^{1} f(x) \sin(\pi j x) dx = 0$ for any integrable even function, the only nonzero Fourier coefficients correspond to cosine functions.

Thus, we get the half-range trigonometric (cosine) basis on $[0,1]$ defined by $\{1, \sqrt{2}\cos(\pi x), \sqrt{2}\cos(2\pi x), \dots\}$. To see that the elements are orthonormal it suffices to recall that $\cos(j\pi x)$ is an even function.

For the case of functions vanishing at the boundary points, that is, when $f(0) = f(1) = 0$, the half-range sine basis is defined as $\{\sqrt{2}\sin(\pi x), \sqrt{2}\sin(2\pi x), \dots\}$. To see that the elements are orthonormal, recall that $\sin(\pi j x)$ is an odd function.

Complex trigonometric basis on $[0, 2\pi]$. For the case of the interval of support $[0, 2\pi]$, the classical trigonometric orthonormal system in $L_2([0, 2\pi])$ and its Fourier coefficients are defined similarly to (2.4.1)–(2.4.2), only here $\varphi_0(x) = (2\pi)^{-1/2}$, $\varphi_{2j-1}(x) = \pi^{-1/2}\sin(jx)$, $\varphi_{2j}(x) = \pi^{-1/2}\cos(jx)$, and $\theta_j = \int_0^{2\pi} f(x)\varphi_j(x)dx$.

Then, the famous Euler's formulae

$$\cos(jx) = [e^{ijx} + e^{-ijx}]/2, \quad \sin(jx) = [e^{ijx} - e^{-ijx}]/2i, \tag{2.4.21}$$

where $i^2 := -1$ is the complex unit, imply the expansion

$$f(x) = \sum_{k=-\infty}^{\infty} c_k (2\pi)^{-1/2} e^{-ikx} \tag{2.4.22}$$

of a function $f(x)$ supported on $[0, 2\pi]$. Here

$$c_0 = \theta_0,\ c_k = [\theta_{2k} + i\theta_{2k-1}]/\sqrt{2},\ c_{-k} = [\theta_{2k} - i\theta_{2k-1}]/\sqrt{2},\ k > 0. \tag{2.4.23}$$

This gives us the *complex trigonometric* system $\{e^{isx}, s = 0, \pm 1, \pm 2, \dots\}$. For complex functions the inner product is defined by $\langle f, g\rangle := \int_0^{2\pi} f(x)\bar{g}(x)dx$, where $\bar{g}$ is the complex conjugate of g (i.e., $\overline{a + ib} = a - ib$). For example, $c_k = \langle f, e^{-ikx}\rangle = \int_0^{2\pi} f(x)e^{ikx}dx$ because $\overline{e^{ikx}} = e^{-ikx}$, and $\langle f, g\rangle = \overline{\langle g, f\rangle}$. While similar to the sine–cosine basis, the complex basis is more convenient in some statistical applications like regression with measurement errors in predictors or density estimation with indirectly observed data.

2.5 Special Topic: Wavelets

In Section 2.1 the Haar basis was introduced, which is the simplest example of a wavelet basis. We have seen that the Haar basis has an excellent localization property. On the other hand, because its elements are not smooth, stepwise Haar approximations of smooth functions may be confusing. Thus, if a smooth approximation is desired, then smooth father and mother wavelets should be used.

Smooth wavelets are relative newcomers to the orthogonal approximation scene. Their name itself was coined in the mid 1980s, and in the 1990s interest in them among the statistical community has grown at an explosive rate.

There is both bad and good news about smooth wavelets. The bad news is that there are no simple mathematical formulae to describe them. Figure 2.13 depicts four smooth mother functions. The functions have a continuous, wiggly, localized appearance that motivates the label *wavelets*. Just by looking at these graphs it is clear why there are no nice formulae for these wiggly functions.

The good news is that there are software packages that allow one to employ wavelets and calculate wavelet coefficients very rapidly and accurately (the last is a very delicate mathematical problem by itself, but fortunately, nice solutions have been found). Here we use the module S+WAVELETS mentioned in Section 2.1 Also, we shall see that these wiggly functions are good building blocks for approximation of a wide variety of functions.

Four types of smooth wavelets are supported by S+WAVELETS. The first is the familiar Haar. The second is the *Daublets*. These wavelets are continuous, have bounded support, and they are identified by "dj" where j is an even integer between 4 and 20. The mother wavelets "d4" and "d12" are shown in the first two plots in Figure 2.13. The number j of a wavelet indicates its width and smoothness. Wavelets with larger indices are typically wider and smoother.

The third type of supported wavelets is *Symmlets*, which are also continuous, have bounded support, and are more symmetric than the Daublets. *Symmlet* 8 ("s8"), which is one of the most popular among statisticians, is shown in the third diagram in Figure 2.13. Here again, the larger the index of the Symmlet, the wider and smoother the mother function.

The last type is *Coiflets*. These wavelets have an additional property of vanishing moments. Coiflet "c12" is shown in Figure 2.13.

It is necessary to know that the toolkit S+WAVELETS was created for the analysis of time series (it is assumed that observations are $f(1), f(2), \ldots, f(n)$), so the following multiresolution expansion is very special. Under the assumption that n is divisible by 2^{j_0}, the wavelet partial

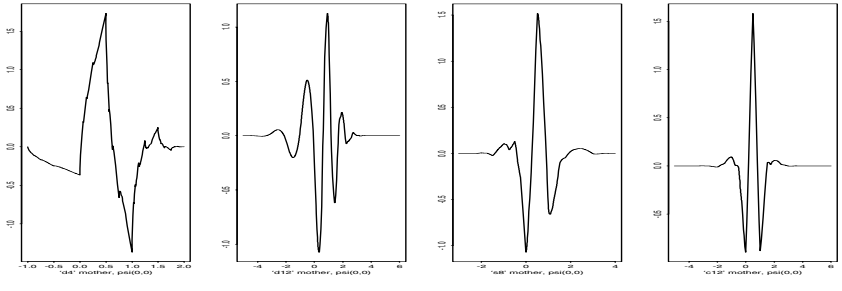

FIGURE 2.13. Four different mother wavelets "d4," "d12," "s8," and "c12." {Recall that before any figure with wavelets is used, the S+WAVELETS module should be loaded with the command > **module(wavelets)**.} [*set.wav=c("d4", "d12", "s8", "c12")*]

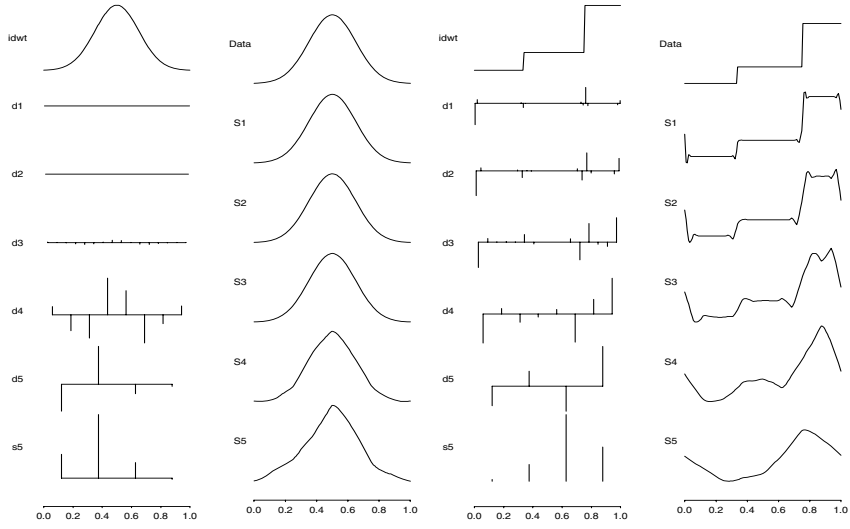

FIGURE 2.14. Wavelet coefficients and default multiresolution approximations of the Normal and the Steps corner functions. Functions are given at $n = 128$ equidistant points, and the wavelet used is Symmlet 8. {The choice of a wavelet and two corner functions is controlled by the arguments *wav* and *set.cf*, respectively.} [*n=128, set.cf= c(2,8), wav= "s8"*]

sum for time series is defined by

$$f_{j_0}(x) := \sum_{k=1}^{n/2^{j_0}} s_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=1}^{j_0}\sum_{k=1}^{n/2^j} d_{j,k}\psi_{j,k}(x). \qquad (2.5.1)$$

Here j_0 is the number of multiresolution components (or scales), $\psi_{j,k}(x) := 2^{-j/2}\psi(2^{-j}x-k)$, $\phi_{j,k}(x) := 2^{-j/2}\phi(2^{-j}x-k)$, $\psi(x)$ is the wavelet function (mother wavelet), and $\phi(x)$ is the scaling function (father wavelet); $s_{j_0,k}$ and $d_{j,k}$ are wavelet coefficients. The notation is the same as in the toolkit.

As we explained in Section 2.1, S+WAVELETS allows us to visualize wavelet coefficients and multiresolution approximations. As in Figure 2.8, let us consider approximation of the Normal and the Steps by the Symmlet 8 ("s8") shown in Figure 2.14. Here the particular case $n = 128$ and $j_0 = 5$ is exhibited.

We see that approximation of the Normal is significantly improved in comparison to the Haar approximation shown in Figure 2.8. Here even approximation by the four father functions (see the second diagram in the bottom row, S5, which is called the low-frequency part) gives us a fair visualization. However, only the partial sum S3 gives us an approximation that resembles the underlying function. Recall that in Figure 2.3 a good

approximation had been obtained by using only 5 Fourier coefficients; here we need at least 16 wavelet coefficients.

The outcome changes for the Steps function. Here the approximations are better than by trigonometric bases but much worse than by the Haar basis. Also, the Gibbs phenomenon (the overshoot of jumps), familiar from trigonometric approximations, is pronouncedly represented.

Now let us consider a wavelet expansion of a function $f(x)$, $-\infty < x < \infty$. Let ϕ' be a father wavelet and let ψ' be a mother wavelet. Denote by $\theta_{j,k} := \int_{-\infty}^{\infty} f(x)\psi'_{j,k}(x)dx$ the wavelet coefficient that corresponds to $\psi'_{j,k} := 2^{j/2}\psi(2^j x - k)$ and by $\kappa_{j,k} := \int_{-\infty}^{\infty} f(x)\phi'_{j,k}(x)dx$ the coefficient that corresponds to $\phi'_{j,k}(x) := 2^{j/2}\phi(2^j x - k)$. Then, for any integer j_1, the *wavelet multiresolution expansion* of a square integrable function f is

$$f(x) = \sum_{k=-\infty}^{\infty} \kappa_{j_1,k}\phi'_{j_1,k}(x) + \sum_{j=j_1}^{\infty}\sum_{k=-\infty}^{\infty} \theta_{j,k}\psi'_{j,k}(x). \tag{2.5.2}$$

Note that if a function f vanishes beyond a bounded interval and the wavelets also vanish beyond a bounded interval, then the number of nonzero wavelet coefficients at the jth resolution level is at most $C2^j$.

Let us consider two function spaces that can be characterized by absolute values of wavelet coefficients (when a wavelet basis is an unconditional basis). The first one is the Hölder space $H_{r,\alpha}$ with $0 < \alpha < 1$, which is defined as the space (2.4.13), only here the assumption about 1-periodicity is dropped. Then the following characterization result holds (the proof may be found in Meyer 1992, Section 6.4): There exist wavelets such that

$$f \in H_{r,\alpha} \;\Leftrightarrow\; |\theta_{j,k}| < c_1 2^{-j(r+\alpha+1/2)},\;\; |\kappa_{j_0,k}| < c_2 \;\;, \tag{2.5.3}$$

where c_1 and c_2 are some constants. No characterization of Lipschitz spaces with $\alpha = 1$ exists. In this case a larger Zygmund function space may be considered; see Meyer (1992, Section 6.4).

The second function space is a *Besov* space B^{σ}_{pqQ}, $1 \le p, q \le \infty$, $0 < \sigma, Q < \infty$, which includes both smooth and discontinuous functions like Hölder functions and functions of bounded total variation. The definition of this space may be found in Meyer (1992, Section 2.9), and it is skipped here. Instead, its characterization via wavelet coefficients is presented (the mathematics of this characterization and assumptions may be found in Meyer 1992, Section 6.10),

$$B^{\sigma}_{pqQ} := \Biggl\{ f : \;\Bigl[\sum_{k=-\infty}^{\infty} |\kappa_{j_1,k}|^p\Bigr]^{1/p} + \Biggl(\sum_{j=j_1}^{\infty}\Bigl[2^{j(\sigma+1/2-1/p)}\Bigl[\sum_{k=-\infty}^{\infty}|\theta_{j,k}|^p\Bigr]^{1/p}\Bigr]^q\Biggr)^{1/q} < Q\Biggr\}. \tag{2.5.4}$$

For instance, a Hölder space $H_{r,\alpha}$ corresponds to $B^{\beta}_{\infty\infty Q}$, $\beta := r + \alpha$, and the space of functions of bounded total variation is a superset of B^1_{11Q} and a subset of $B^1_{1\infty Q'}$. These two examples shed light on the meaning of the parameters σ, p, q, and Q.

2.6 Special Topic: More Orthonormal Systems

This section reviews several orthonormal systems that may be useful for approximation of functions from particular spaces.

• **Polynomials on a bounded interval.** For classical polynomials the customarily studied bounded interval is $[-1, 1]$. An orthonormal basis for the space $L_2([-1, 1])$, with the inner product $\langle f, g\rangle := \int_{-1}^{1} f(x)g(x)dx$, is generated by applying the Gram–Schmidt orthonormalization procedure (2.3.7)–(2.3.8) to the powers $\{1, x, x^2, \ldots\}$. Also, the jth element $G_j(x)$ of this basis may be calculated via the formula

$$G_j(x) = \frac{1}{j!2^j}\sqrt{(2j+1)/2}\frac{d^j}{dx^j}(x^2-1)^j. \tag{2.6.1}$$

It is worthwhile to note that the well-known Legendre polynomials $P_j(x) = \sqrt{2/(2j+1)}G_j(x)$, which are built-in functions in many software packages, are orthogonal but not orthonormal. To compute Legendre polynomials, the recurrence formula

$$P_n(x) = n^{-1}[(2n-1)xP_{n-1}(x) - (n-1)P_{n-2}(x)], \tag{2.6.2}$$

together with the facts that $P_0(x) = 1$ and $P_1(x) = x$, is especially useful.

The following assertion, whose proof may be found in DeVore and Lorentz (1993, Section 7.6), shows how the integrated squared bias of the polynomial approximation decreases.

Proposition 2.6.1. Let f be r-fold differentiable and

$$|f^{(r)}(t) - f^{(r)}(s)| \le Q|t-s|^{\alpha}, \text{ where } t, s \in [-1, 1], \ 0 < \alpha \le 1. \tag{2.6.3}$$

Then there exists a constant c such that the polynomial partial sums $S^*_J(x) := \sum_{j=0}^{J}\langle f, Q_j\rangle Q_j(x)$ satisfy the relation

$$\int_{-1}^{1}(f(x) - \tilde{S}^*_J(x))^2dx \le cJ^{-2(r+\alpha)}. \tag{2.6.4}$$

•**Polynomials on** $[0, \infty)$. Sometimes it is convenient to approximate an underlying function on a half-line $[0, \infty)$. In this case the idea is to modify the inner product in such a way that the integration over the half-line is well-defined. The customary approach is to consider the inner product $\langle f, g\rangle := \int_0^{\infty} f(x)g(x)e^{-x^2}dx$ and then apply the Gram–Schmidt orthonormalization procedure to the polynomials $\{1, x, x^2, \ldots\}$. This defines the *Laguerre* basis.

•**Polynomials on** $(-\infty, \infty)$**.** The approach is the same and the inner product is defined by $\langle f, g\rangle := \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2}dx$. Then the Gram-Schmidt procedure is applied to the polynomials $\{1, x, x^2, \ldots\}$, and the system obtained is called the *Hermite* basis.

•**A set of discrete points**. So far, we have discussed the case of functions from L_2 that are defined on intervals. For many practical problems a function is defined only at a set of discrete points, and there is no interest in values of this function at other points. Let there be m such points $\{x_1, x_2, \ldots, x_m\}$. Then the inner product may be defined as

$$\langle f, g\rangle := \sum_{k=1}^{m} p_k f(x_k)g(x_k), \tag{2.6.5}$$

where p_k are some positive "weights." If these weights are summable to 1, then the inner product is just $E\{f(X)g(X)\}$, where X is a discrete random variable with probability mass function p_k, and $E\{\cdot\}$ denotes the expectation.

Thus, for any system $\{\psi_1(x), \ldots, \psi_J(x)\}$, a problem of best approximation, which is the analogue of the L_2-approach, becomes the familiar problem of finding the coefficients $\{c_j\}$ minimizing $\sum_{k=1}^{J} p_k[f(x_k) - \sum_{j=1}^{m} c_j\psi_j(x_k)]^2$.

It was shown by Chebyshev (the famous probabilistic inequality (A.26) also bears his name) that orthonormalization of the polynomials $\{1, x, x^2, \ldots\}$ gives a complete orthonormal system (basis) in this setting. More details may be found in Kolmogorov and Fomin (1957, Section 7.3.8). Similarly, orthonormalization of trigonometric functions also leads to a basis at discrete points. Note that for the case of identical weights and equidistant points on $[0, 1]$ the trigonometric system is the orthonormal basis.

•**Enriched bases**. So far, we have discussed only classical bases. In some cases it is worthwhile to enrich a classical basis by elements like linear, quadratic, or step functions, which allow one to approximate a set of targeted functions.

As an example, let us begin with the case of the trigonometric sine–cosine system. We have seen in Section 2.4 that its approximations of aperiodic functions were terrible.

The issue is that the elements of the basis are periodic, and the derivatives of the elements are also periodic. Thus, to fit aperiodic functions, this basis should be enriched by aperiodic elements, for instance, by the linear function x and the quadratic function x^2.

Since both the linear and quadratic functions are not orthonormal to the elements of the trigonometric system, the Gram–Schmidt procedure should be used.

Approximations of the corner functions by the trigonometric system enriched by the linear function are shown in Figure 2.15. The partial sums for the Monotone and the Steps look much better. The only remaining pattern

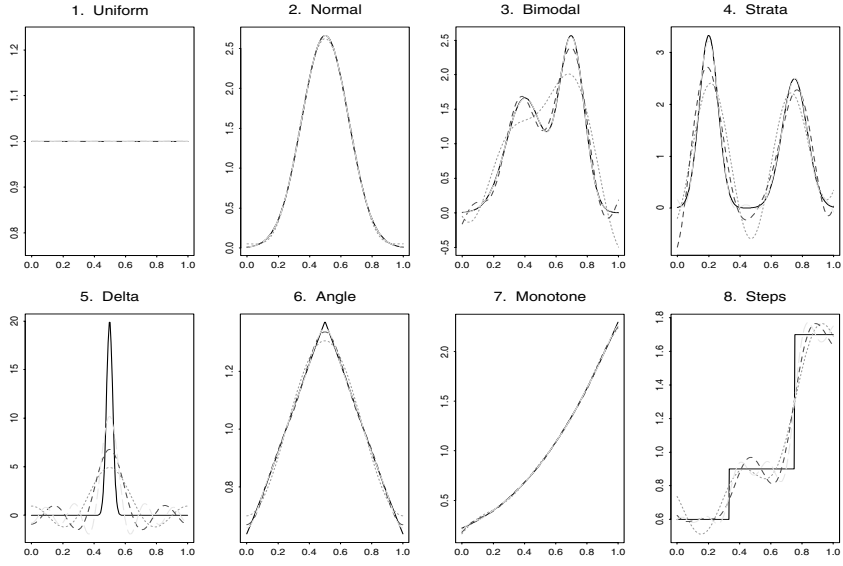

FIGURE 2.15. Approximation of the corner functions (solid lines) by Fourier sums enriched by the linear function: dotted, short-dashed, and long-dashed lines correspond to the cutoffs $J = 2$, $J = 3$, and $J = 5$, respectively. [*set.J =c(2,3,5)*]

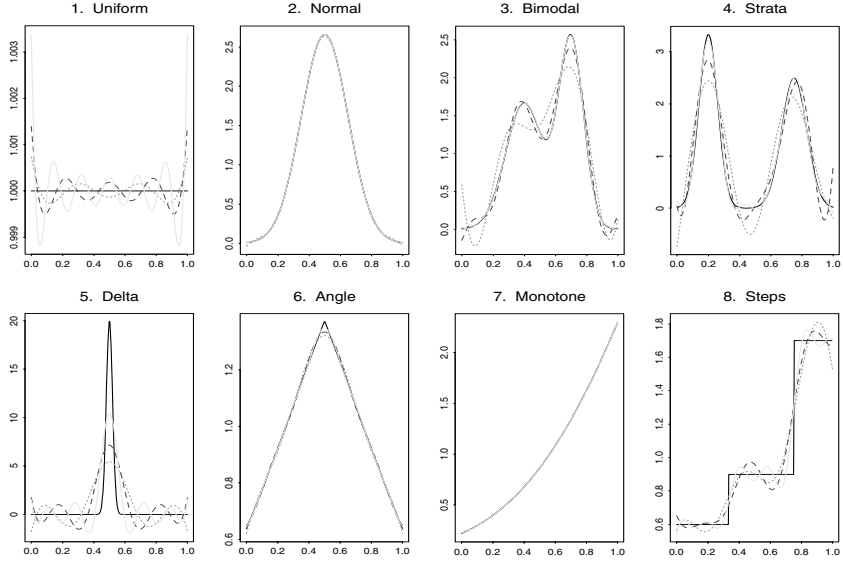

FIGURE 2.16. Approximation of the corner functions (solid lines) by Fourier sums enriched by the linear and quadratic polynomial functions: Dotted, short-dashed, and long-dashed lines correspond to cutoffs $J = 2$, $J = 3$, and $J = 5$, respectively. [*set.J=c(2,3,5)*]

that catches eye is that near the edges, the Monotone (and the Angle) are not represented very well. This is because the derivative of a partial sum is still periodic.

Thus, let us additionally enrich this new basis by the quadratic function. The corresponding approximations are shown in Figure 2.16. As we see, now the Angle and the Monotone are represented near the boundaries much better. On the other hand, some other corner functions are fitted worse (especially for the smallest cutoff) near the edges. So, for the case of small cutoffs there are pros and cons in this enrichment.

Now let us consider the more challenging problem of enriching the cosine basis by a function that mimics a step function at a point a. The aim is to get a perfect fit for the first jump in the Steps.

Set $\phi(x,a)=1$ if $0\le x\le a$ and $\phi(x,a)=0$ if $a<x\le 1$; that is, $\phi(x,a)$ is a step function with unit jump at the point a. We add the step function to the set of the first $1+J$ cosine functions $\{\varphi_0=1,\varphi_j=\sqrt{2}\cos(\pi jx),\ j=1,\dots,J\}$ and then use the Gram–Schmidt orthogonalization procedure to get the $(2+J)$th element of the enriched system,

$$\phi(x,a,J):=\frac{\phi(x,a)-a-\sum_{j=1}^{J}2^{1/2}(\pi j)^{-1}\sin(\pi ja)\varphi_j(x)}{\Big[\int_0^1(\phi(u,x_0)-x_0-\sum_{j=1}^{J}2^{1/2}(\pi j)^{-1}\sin(\pi ja)\varphi_j(u))^2du\Big]^{1/2}}.$$

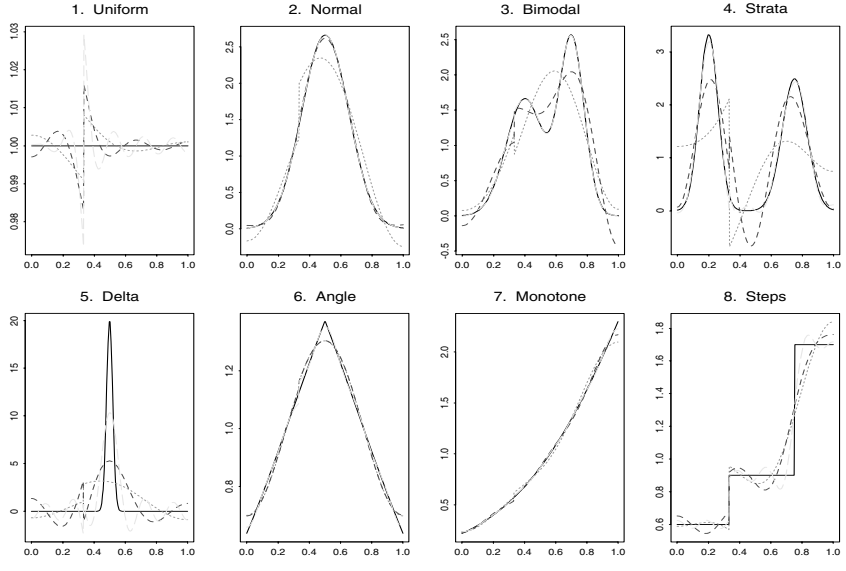

FIGURE 2.17. Approximation of the corner functions (solid lines) by the cosine basis enriched with the step function $\phi(x,a)$ with the default value $a=\frac{1}{3}$: Dotted, short-dashed, and long-dashed lines correspond to $J=3$, $J=5$, and $J=10$, respectively. [*a*=1/3, *set.J=c(3,5,10)*]

Denote $\theta_j = \int_0^1 f(u)\varphi_j(u)du$ and $\kappa(a, J) = \int_0^1 f(u)\phi(u, a, J)du$. Then, a partial sum

$$S_J(x, a) := \sum_{j=0}^{J} \theta_j \varphi_j(x) + \kappa(a, J)\phi(x, a, J) \tag{2.6.6}$$

is used to approximate f. Partial sums are shown in Figure 2.17.

Let us begin the discussion of these approximations with the Uniform. The numerical errors are apparently large because the trapezoid rule is used for nonsmooth functions. Also note how "aggressively" the step function tries to find its place for the smallest cutoffs. The same pattern is clearly seen in the Bimodal and the Strata diagrams. This is because there is plenty of room for the step function when only several cosine functions are used to approximate a spatially inhomogeneous function. On the other hand, the enriched basis does a superb job in visualizing the first jump in the Steps.

2.7 Exercises

2.1.1 Suggest corner functions with 3 and 4 modes using mixtures of normal densities.

2.1.2 Repeat Figure 2.3 with different cutoffs. Answer the following questions: (a) What are the minimal cutoffs (if any) for each corner function that imply a reasonable fit? (b) How does the cosine system approximate a constant part of a function? (c) Indicate graphs that exhibit the Gibbs phenomenon.

2.1.3 Verify that for the polynomial system, $\varphi_1(x) = \sqrt{3}(2x - 1)$. Also, calculate $\varphi_2(x)$.

2.1.4 Find an antiderivative for: (a) $3x - 5x^2$; (b) $5\cos(2x) - 3\sin(5x)$.

2.1.5 Verify (2.1.5).

2.1.6 Repeat Figure 2.5 with different cutoffs. What are the minimal cutoffs for each corner function (if any) that give a reasonable fit? Compare these cutoffs with those obtained for the cosine system in Exercise 2.1.2.

2.1.7 Repeat Figure 2.8 for different corner functions, and discuss the outcomes.

2.1.8 Explain the multiresolution approximation of the Delta function shown in Figure 2.8.

2.2.1 Let $f_J(x) = \sum_{j=0}^{J} \theta_j \varphi_j(x)$. Find $\int_0^1 (f_{J+L}(x) - f_J(x))^2 dx$.

2.2.2 For f_J from the previous exercise, check that $\int_0^1 f_J^2(x)dx = \sum_{j=0}^{J} \theta_j^2$.

2.2.3 Verify (2.2.3).

2.2.4 Assume that the boundary condition $f^{(1)}(0) = f^{(1)}(1)$ holds, and f has either 3 or 4 derivatives. How fast do the Fourier coefficients (for the cosine basis) decrease? Hint: Continue (2.2.7) using integration by parts.

2.2.5 Find how fast the Fourier coefficients (for the cosine system) of the functions x, x^2, x^3, x^4 decrease.

2.2.6 Verify (2.2.9).
2.2.7 Verify (2.2.11).
2.2.8 Find the total and quadratic variations of $\cos(j\pi x)$ on $[0,1]$.
2.3.1 Prove that if f_1 and f_2 are square integrable, then $f_1 f_2 \in L_1$.
2.3.2 Show that both the inner product $\langle f, g\rangle$ in L_2 space and the dot product $\langle \vec{v}, \vec{u}\rangle$ in $\mathcal{E}_k$ satisfy the properties of an inner product formulated in the subsection **Hilbert space**.
2.3.3 Check the orthogonality (in L_2) between the following two functions: (a) 1 and $\sin(ax)$; (b) $\sin(ax)$ and $\cos(bx)$; (c) $\cos(ax)$ and $\cos(bx)$; (d) 1 and $x + a$; (e) x and $ax + bx^2$.
2.3.4 We say that a sequence of functions f_n converges to f in the L_2-norm if and only if $\|f_n - f\| \to 0$ as $n \to \infty$. Let sequences f_n and g_n converge in the L_2-norm to f and g, respectively. Prove that (a) the sum of two sequences $f_n + g_n$ converges to the sum of their limits $f + g$; (b) if a_n is a sequence of real numbers converging to a, then the sequence of functions $a_n f$ converges to af; (c) the following convergence holds,

$$\langle f_n, g_n\rangle \to \langle f, g\rangle. \tag{2.7.1}$$

2.3.5 Let n functions $\{f_j, j = 1, 2, \ldots, n\}$ be orthogonal in L_2. Prove that these functions are also linearly independent, that is, the identity $\sum_{j=1}^{n} a_j f_j(x) \equiv 0$, $x \in [0,1]$ implies, $a_1 = a_2 = \cdots = a_n = 0$.
2.3.6 Establish that if a series $\sum_{j=0}^{n} \theta_j \varphi_j(x)$ converges to a function f in L_2 as $n \to \infty$ and $\{\varphi_j\}$ is an orthonormal system in L_2, then

$$\theta_j = \langle f, \varphi_j\rangle = \int_0^1 f(x)\varphi_j(x)dx. \tag{2.7.2}$$

2.3.7 Let $f_1, f_2, \ldots, f_k$ be pairwise orthogonal, i.e., $\langle f_l, f_j\rangle = 0$ whenever $l \neq j$. Verify that $\|\sum_{l=1}^{k} f_l\|^2 = \sum_{l=1}^{k} \|f_l\|^2$.
2.3.8 Check that for any orthonormal system $\{f_1, f_2, \ldots\}$ the equality $\|f_l - f_j\| = \sqrt{2}$ holds for $l \neq j$.
2.3.9 Using the Gram–Schmidt procedure, orthogonalize the set of functions $\{1, x, x^2\}$. As a result, the first three elements of the Legendre polynomial basis on $[0,1]$ are obtained.
2.3.10 Using the Gram–Schmidt procedure, orthogonalize the following set of trigonometric functions enriched by the power functions $\{1, \sin(2\pi x), \ldots, \sin(2\pi N x), \cos(2\pi x), \ldots, \cos(2\pi N x), x, x^2\}$. As a result, you get a so-called trigonometric-polynomial system.
2.3.11 Show that the element (2.3.8) is orthogonal to the elements φ_s, $s = 1, \ldots, j-1$.
2.3.12 Verify (2.3.6) using the projection theorem.
2.3.13 Find the orthogonal projection (in L_2) of a function $f \in L_2$ onto a subspace of all linear combinations of the functions $\{1, \cos(\pi x), \cos(\pi 2x)\}$.
2.4.1 Repeat Figure 2.10 with different cutoffs, and then for every corner function find a minimal cutoff that gives a fair fit.

2.4.2 Let $f \in L_1$ (i.e., $\int_0^1 |f(x)|dx < \infty$) and let for some $\delta > 0$ the *Díni condition* $\int_{-\delta}^{\delta} |(f(x+t) - f(x))/t|dt < \infty$ hold. Then $S_J(x) \to f(x)$ as $J \to \infty$. Hint: See Theorem 8.1.1 in Kolmogorov and Fomin (1957).
2.4.3 Suppose that f is bounded, has only simple discontinuities, and at every point has left and right derivatives. Then $S_J(x)$ converges to $\lim_{\delta \to 0}[f(x+\delta) + f(x-\delta)]/2$. Hint: See Remark 8.1.1 in Kolmogorov and Fomin (1957).
2.4.4 Let f be bounded and differentiable and let its derivative be square integrable. Then $S_J(x)$ converges to $f(x)$ uniformly over all $x \in [0, 1]$. Hint: See Theorem 8.1.2 in Kolmogorov and Fomin (1957).
2.4.5 Check (2.4.10).
2.4.6 The function $f(x) = |x - 0.5|$ is not differentiable. Nevertheless, show that it belongs to a Sobolev class $W_{1,Q}$ with some $Q < \infty$.
2.4.7 Show that Fejér sums satisfy $|\sigma_J(x)| \le \max_x |f(x)|$. Hint: Begin with considering $f(x) = 1$ and then using (2.4.10) show that the Fejér kernel is integrated to 1.
2.4.8 Use Figure 2.11 to find cutoffs for Fejér sums that give the same visualization of modes as the Fourier sums in Figure 2.10.
2.4.9 Use Figure 2.12 to find cutoffs for de la Vallée Poussin sums that give the same visualization of modes as the Fourier sums in Figure 2.10.
2.5.1 Repeat Figure 2.14 with two other wavelets having the parameter j different from 8. Discuss how this parameter affects the data compression property of wavelet approximations.
2.5.2 Repeat Figure 2.14 for two other corner functions. Discuss how smoothness of an underlying function affects the data compression.
2.5.3 Repeat Figure 2.14 with different n, different corner functions, and different wavelets. Find the best wavelets for the corner functions.
2.6.1 Explain how to calculate the polynomial basis for $L_2([0, 1])$.
2.6.2 Find G_1, G_2, and G_3 using (2.6.1).
2.6.3 For the subsection "Polynomials on $(-\infty, \infty)$" find the first four elements of the Hermite basis. Hint: Recall that $\pi^{-1/2}e^{-x^2}$ is the normal $N(0, .5)$ density.
2.6.4 Prove that a trigonometric basis is orthonormal on a set of equidistant points.
2.6.5 Use Figure 2.17 to analyze how the cutoff J affects the visualization of pseudo-jumps in the smooth corner functions.
2.6.6 Try different parameters a in Figure 2.17. Explain the results.

2.8 Notes

The basic idea of Fourier series is that "any" periodic function may be expressed as a sum of sines and cosines. This idea was known to the Babylonians, who used it for the prediction of celestial events. The history of the

subject in more recent times begins with d'Alembert, who in the eighteenth century studied the vibrations of a violin string. Fourier's contributions began in 1807 with his studies of the problem of heat flow. He made a serious attempt to prove that any function may be expanded into a trigonometric sum. A satisfactory proof was found later by Dirichlet. These and other historical remarks may be found in Dym and McKean (1972). Also, Section 1.1 of that book gives an excellent explanation of the Lebesgue integral, which should be used by readers with advanced mathematical background. The textbook by Debnath and Mikusinskii (1990) is devoted to Hilbert spaces. The books by Bary (1964) and Kolmogorov and Fomin (1957) give a relatively simple discussion (with rigorous proofs) of Fourier series.

The simplest functions of the variable x are the algebraic (ordinary) polynomials $P_n = c_0 + c_1 x + \cdots + c_n x^n$ of degree n. Thus, there is no surprise that they became the first and powerful tool for approximation of other functions. Moreover, a theorem about this approximation, discovered in the nineteenth century by Weierstrass, became the cornerstone of modern approximation theory. For a continuous function $f(x)$, $x \in [0, 1]$, it asserts that *there exists a sequence of ordinary polynomials* $P_n(x)$ *that converge uniformly to* $f(x)$ *on* $[0, 1]$. There are many good books on approximation theory (but do not expect them to be simple). Butzer and Nessel (1971) is the classical reference. Among recent ones, DeVore and Lorentz (1993), Temlyakov (1993), and Lorentz, Golitschek and Makovoz (1996) may be recommended as solid mathematical references.

The mathematical theory of wavelets was developed in the 1980s and it progressively appeared to be useful in approximation of spatially inhomogeneous functions. There are several relatively simply written books by statisticians for statisticians about wavelets, namely, the books by Ogden (1997) and Vidacovic (1999), as well as the more mathematically involved book by Härdle et al. (1998). The book by Mallat (1998) discusses the application to signal processing.

The book by Walter (1994) gives a rather balanced approach to the discussion of all orthogonal systems, and it is written on a level accessible to graduate students with good mathematical background.

3

Density Estimation for Small Samples

This chapter is devoted to the data-driven orthogonal series estimation of a univariate density for the case of small sample sizes. An estimator is defined and discussed in Section 3.1. This estimator is called universal because it will also be used for other statistical models including nonparametric regression and spectral density estimation. Section 3.2 studies risks of this estimator via lower bounds (oracle inequalities). These bounds allow us to say how far the suggested estimate is from a "golden standard." Section 3.3 explores different data-driven estimators, which will be used in some special cases. The remaining sections are devoted to special cases where the applicability of the universal estimator to a broad spectrum of statistical settings is explored. Finally, the practical seminar is devoted to a discussion of how to use the universal estimator for the analysis and presentation of real data sets.

3.1 Universal Orthogonal Series Estimator

In this section the classical (and simplest) model of probability density estimation is considered, where n independent and identically distributed observations $X_1, X_2, \dots, X_n$ of a random variable X are given. It is supposed that X is distributed according to an unknown probability density $f(x)$, and the problem is to estimate $f(x)$ over the interval $[0, 1]$ based only on the data. Remark 3.1.2 at the end of this section explains how to estimate f over an arbitrary interval. In some cases it is also of interest to

estimate f over its support, which may be unknown. Remark 3.2.3 discusses this issue, and this discussion is continued in Section 3.9.

The recommended data-driven estimator is defined below at (3.1.14), so now step by step we consider the motivation behind this estimator, and then we will study this estimator via Monte Carlo simulations.

The underlying idea of an orthogonal series estimator is as follows. As we know from Chapter 2, under mild assumptions a function $f(x)$, $x \in [0,1]$, may be approximated by a partial sum (truncated orthogonal series),

$$f_J(x) := \sum_{j=0}^{J} \theta_j \varphi_j(x),\ 0 \le x \le 1, \quad \text{where } \theta_j = \int_0^1 \varphi_j(x) f(x) dx. \quad (3.1.1)$$

Here $\{\varphi_j\}$ is an orthonormal basis; in this chapter the cosine basis $\{\varphi_0(x) = 1, \varphi_j(x) = \sqrt{2}\cos(\pi j x), j = 1, 2, \ldots\}$, discussed in Section 2.1, is used. Recall that J is called the cutoff, and θ_j is called the jth Fourier coefficient of f corresponding to the jth element φ_j of the basis used.

Note that $\theta_0 = P(0 \le X \le 1)$; thus $\theta_0 \le 1$ with equality iff f vanishes beyond the unit interval.

We also discussed in Section 2.4 that in many cases it was worthwhile to smooth (shrink toward the origin) the Fourier coefficients (to multiply them by constants that took on values between 0 and 1), so consider a smoothed partial sum

$$f_J(x, \{w_j\}) := \sum_{j=0}^{J} w_j \theta_j \varphi_j(x), \quad 0 \le x \le 1, \quad \text{where } 0 \le w_j \le 1. \quad (3.1.2)$$

Thus, the statistical problem of estimation of the density f is converted into that of finding estimates for (i) Fourier coefficients $\{\theta_j\}$; (ii) the cutoff J; (iii) the smoothing coefficients (weights) $\{w_j\}$. It is worthwhile to note that all of the known series estimators differ only by methods of estimating the weights and the cutoff, because the choice of an estimate for Fourier coefficients is straightforward. Indeed, according to (3.1.1) and the fact that f is the probability density, the Fourier coefficient θ_j may be written as

$$\theta_j = E\{I_{\{X \in [0,1]\}} \varphi_j(X)\}. \quad (3.1.3)$$

Recall that $E\{\cdot\}$ denotes the expectation (theoretical mean) and $I_{\{A\}}$ is the indicator of an event A, that is, the indicator is equal to 1 if A occurs and is 0 otherwise. Then the natural estimate of the theoretical mean is the sample mean,

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} I_{\{X_l \in [0,1]\}} \varphi_j(X_l). \quad (3.1.4)$$

This is the estimate used in all known nonparametric density estimators. Note that we use diacritics (e.g., "hat," "tilde," or "bar") above a param-

eter or a function to indicate that this is an estimate (statistic) of the corresponding parameter or function.

The next step is to choose a cutoff J. The choice crucially depends on what kind of goodness of fit one wishes to get from an estimator $\tilde{f}_J(x)$. In this chapter we restrict our attention to a global risk called the *mean integrated squared error* (the abbreviation is MISE) and defined by

$$\text{MISE}(\tilde{f}_J, f) := E\left\{\int_0^1 (\tilde{f}_J(x) - f(x))^2 dx\right\}. \tag{3.1.5}$$

Define $\tilde{f}_J(x) := \sum_{j=0}^J \hat{\theta}_j \varphi_j(x)$. (Note that during the step of choosing J we set $w_j \equiv 1$. The reason is based on the simplicity and the numerical analysis presented in the next sections. Also note that using weights $w_j < 1$ may increase the cutoff.) Then Parseval's identity (recall (2.3.11)) implies

$$\text{MISE}(\tilde{f}_J, f) = \sum_{j=0}^J E\{(\hat{\theta}_j - \theta_j)^2\} + \sum_{j>J} \theta_j^2. \tag{3.1.6}$$

This equality gives us the key idea on how to choose the cutoff J. Because the optimal cutoff minimizes MISE, it minimizes the sum of the two terms in the right-hand side of (3.1.6). Let us consider these two terms.

The first term is the variance of $\tilde{f}_J$, and it is the sum of $J+1$ variances $\text{Var}(\hat{\theta}_j) = E\{(\hat{\theta}_j - \theta_j)^2\}$ of the sample mean estimates $\hat{\theta}_j$. A straightforward calculation, based on the elementary trigonometric equality

$$\cos^2(\alpha) = [1 + \cos(2\alpha)]/2, \tag{3.1.7}$$

shows that for $j > 0$

$$E\{(\hat{\theta}_j - \theta_j)^2\} = \text{Var}(\hat{\theta}_j) = \theta_0 n^{-1} + [\theta_{2j} 2^{-1/2} - \theta_j^2] n^{-1} = d_j n^{-1}, \tag{3.1.8}$$

where $d_j := \theta_0 + \theta_{2j} 2^{-1/2} - \theta_j^2$. As we know from Chapter 2, Fourier coefficients θ_j decay rather rapidly as j increases. Thus, we choose $\hat{d} = \hat{\theta}_0$ as an estimate for all d_j, $j = 0, 1, \ldots$.

The second term in (3.1.6) is the integrated squared bias $\text{ISB}_J(f) = \sum_{j>J} \theta_j^2$. It is impossible to estimate this sum directly because it contains infinitely many terms. Instead, let us note that by Parseval's identity

$$\text{ISB}_J(f) = \int_0^1 f^2(x) dx - \sum_{j=0}^J \theta_j^2 \tag{3.1.9}$$

and that the term $\int_0^1 f^2(x) dx$ is a constant. Thus, the problem of finding a cutoff J that minimizes (3.1.6) is equivalent to finding a cutoff that minimizes $\sum_{j=0}^J (\hat{d} n^{-1} - \theta_j^2)$. Here we used the above-suggested estimate $\hat{d} n^{-1}$ for the variances.

Thus, we need to decide how to estimate θ_j^2. Because $\hat{\theta}_j$ is an unbiased estimate of θ_j (i.e., $E\{\hat{\theta}_j\} = \theta_j$), a natural estimate could be $\hat{\theta}_j^2$. However,

this estimate is biased because $E\{\hat\theta_j^2\} = \theta_j^2 + \mathrm{Var}(\hat\theta_j)$, see (A.5) in Appendix A. Because $\hat d$ is the estimate for $n\mathrm{Var}(\hat\theta_j)$, we choose $\hat\theta_j^2 - \hat d n^{-1}$ as an estimate for θ_j^2.

Remark 3.1.1. An attractive alternative unbiased estimate of θ_j^2 is the U-statistic, $\tilde\theta_j^2 = (2/(n(n-1)) \sum_{1\le l<m\le n} \varphi_j(X_l)\varphi_j(X_m)$. Sometimes this estimate may be very convenient, because it does not depend on estimation of d_j. Also note that $\hat\theta_j^2 - \tilde\theta_j^2$ is an unbiased estimate of $\mathrm{Var}(\hat\theta_j)$.

Now we are in a position to combine these ideas and suggest the following estimate of the cutoff

$$\hat J := \mathrm{argmin}_{0\le J\le J_n} \sum_{j=0}^{J} (2\hat d n^{-1} - \hat\theta_j^2). \tag{3.1.10}$$

Recall that the function $\mathrm{argmin}_{0\le s\le S}\{a_s\}$ returns the value s^* that is the index of the smallest element among $\{a_0, a_1, \dots, a_S\}$. In (3.1.10) the search for the optimal cutoff is restricted from above by some reasonable upper bound J_n. Based on the results of the following two sections, we set $J_n = \lfloor c_{J0} + c_{J1}\ln(n)\rfloor$, where $\lfloor x\rfloor$ is the rounded-down x and c_{J0} and c_{J1} are parameters (coefficients) whose default values are 4 and 0.5, respectively. (Exercises 3.1.14–16 are devoted to finding optimal coefficients for each particular setting; see also Section 3.9.)

Finally, we need to choose the smoothing coefficients w_j. It has been established in Example A.25, see (A.38) in Appendix A, that the best smoothing coefficients minimizing MISE are

$$w_j^* = \frac{\theta_j^2}{\theta_j^2 + E\{(\hat\theta_j - \theta_j)^2\}}. \tag{3.1.11}$$

We have discussed above how to estimate the components of this ratio. Note that θ_j^2 is nonnegative and that we do not want to shrink $\hat\theta_0$ because if $[0,1]$ is the support of f, then $\hat\theta_0 = 1$ implies a series estimate that is correctly integrated to unity. Thus we set

$$\hat w_0 := 1 \quad \text{and} \quad \hat w_j := (1 - \hat d/n\hat\theta_j^2)_+, \quad j > 0\,. \tag{3.1.12}$$

Here $(x)_+ := \max(0, x)$ denotes the positive part of x.

This gives us the estimator

$$\bar f(x) := \sum_{j=0}^{\hat J} \hat w_j \hat\theta_j \varphi_j(x). \tag{3.1.13}$$

The estimator (3.1.13) is a classical smoothed series estimator, which will be discussed in detail in the following two sections. For practical purposes, we would like to make two "improvements" on this estimator. The first one is based on the idea of obtaining a good estimator for spatially inhomogeneous densities like the Delta; see Figure 2.1.5. As we discussed in detail

in Section 2.1, such a density requires a relatively large number of Fourier coefficients for a fair visualization. Thus we add to the estimate (3.1.13) high-frequency terms, which are shrunk by a hard threshold procedure, and get the estimate

$$\tilde{f}(x) := \sum_{j=0}^{\hat{J}} \hat{w}_j \hat{\theta}_j \varphi_j(x) + \sum_{j=\hat{J}+1}^{c_{JM} J_n} I_{\{\hat{\theta}_j^2 > c_T \hat{d} \ln(n)/n\}} \hat{\theta}_j \varphi_j(x). \tag{3.1.14}$$

Here c_{JM} and c_T are again parameters (coefficients) that define the maximal number of elements included into the estimate and the coefficient in the hard threshold procedure. The default values are 6 and 4, respectively. Note that a high-frequency element is included only if the corresponding Fourier coefficient is extremely large and thus cannot hurt estimation of smooth functions like the Normal.

The necessity of the second improvement is obvious. An estimate $\tilde{f}$ is integrated to $\hat{\theta}_0$, which is at most 1, but may take on negative values. Fortunately, it is a simple problem to find a projection of $\tilde{f}$ in $L_2([0,1])$ onto a class of nonnegative functions integrated to $\hat{\theta}_0$. The projection is

$$\hat{f}(x) := (\tilde{f}(x) - c)_+, \quad \text{plus removing small bumps,} \tag{3.1.15}$$

where the constant c is such that $\int_0^1 \hat{f}(x)dx = \hat{\theta}_0$. Then, as is highlighted in (3.1.15), small bumps are removed. This procedure is explained below.

The procedure (3.1.15) is illustrated in Figure 3.1. The solid line in the diagram (a) shows a hypothetical density estimate integrated to $\hat{\theta}_0$. Note that $\hat{\theta}_0$ is the proportion of observations that have fallen into the unit interval. Because the estimate takes on negative values, the area under the curve and above the x-axis is more than $\hat{\theta}_0$. Then a positive constant c exists such that the shaded area under the estimate and above the dashed horizontal line (at the level c) equals $\hat{\theta}_0$. The nonnegative projection (defined at (3.1.15)) is shown in the diagram (b), and it is the nonnegative part of the estimate where the new x-axis is translated to the point c. Note

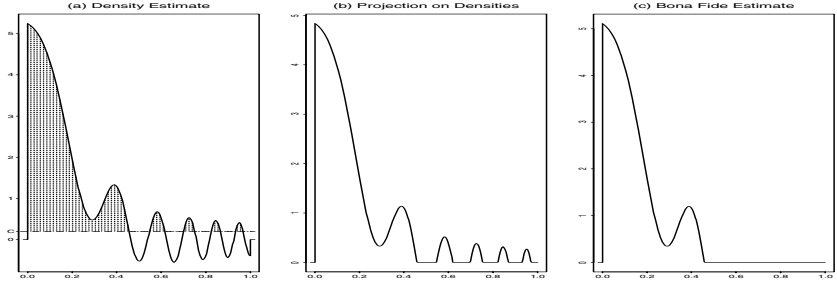

FIGURE 3.1. (a) A density estimate integrated to $\hat{\theta}_0$. (b) Nonnegative projection of the estimate on a class of densities. (c) The projection with small bumps removed.

that the projection is a bona fide density, that is, it is integrated to $\hat{\theta}_0$ and nonnegative. While the projection is an improvement of the original density, the next step, illustrated in the diagram (c), is based on the following idea. We know from Figure 2.3.5 that a cosine series approximates the horizontal tails of the Delta by waves that are similar to those shown in the diagram (a). As a result, in the projection (b) we see several bumps, that have nothing to do with the underlying density but are just the waves "leftovers." Thus, the last step is to remove bumps that we believe are due solely to the oscillatory approximation of the flat parts of an underlying density. The procedure is as follows. Let f_a and f_b be the estimates shown in diagrams (a) and (b). We know that f_b is an improvement of the original estimate, and the decrease in MISE is $\delta := \int_0^1 (f_a(x) - f_b(x))^2 dx$. The value of δ gives us an idea of which bump is significant and which is not. Namely, let $[t_1, t_2]$ be the domain of a bump. If $\int_{t_1}^{t_2} f_b^2(x)dx < c_B\delta$, then this bump may be removed. In all the following figures the coefficient c_B is equal to 2. Finally, since bump-removing decreases the integral of the estimate, the final step is to divide the estimate by this integral and then multiply by $\hat{\theta}_0$. The obtained estimate is again bona fide.

The *universal* estimator is constructed. We call it universal because it will be used with the same values of the coefficients for all the settings and models. On the other hand, these values are not necessarily optimal for a particular setting and model; exercises are devoted to the choice of optimal values.

Several questions immediately arise. First, how does this estimator perform for small sample sizes? Second, is it possible to suggest a better estimator?

The rest of this section is devoted to answering the first question, and the next two sections give an answer to the second question.

To evaluate the performance of the estimator, we use Monte Carlo simulations where data sets are generated according to each of the corner densities shown in Figure 2.1. For these densities $\hat{\theta}_0 = 1$ because they are supported on $[0, 1]$. Then estimates for sample sizes 50, 100, and 200 are shown in Figure 3.2. This figure exhibits results of 8 times 3 (that is, 24) independent Monte Carlo simulations. Note that particular estimates depend on simulated data sets, so they may be better or worse. A particular outcome, shown in Figure 3.2, is chosen primarily with the objective of the discussion of possible estimates.

Let us begin the discussion of the exhibited estimates with a general remark. The main beneficiary of the added high-frequency terms in (3.1.14) and the procedure (3.1.15) of obtaining a bona fide estimate is the Delta. Note that the estimates in Figure 3.2.5 are dramatically better than all the approximations (based on the underlying density (!)) discussed in Chapter 2. Moreover, for $n = 200$ even the magnitude of the peak is shown almost correctly. For all other estimates, as will be clear from the following two

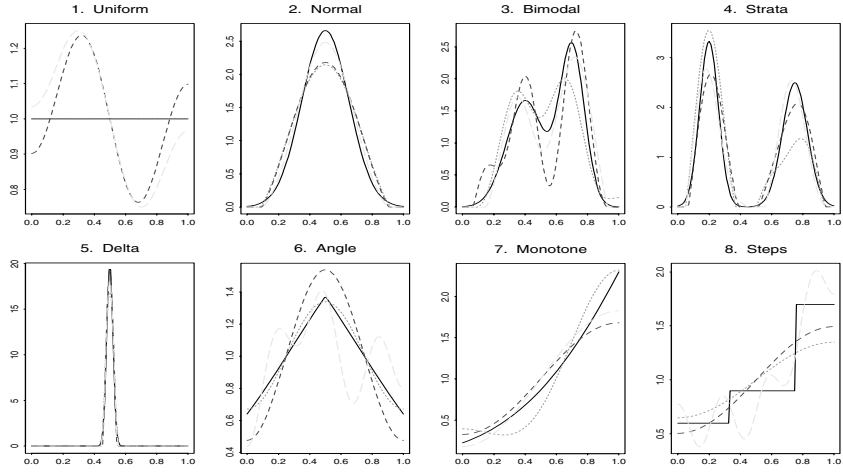

FIGURE 3.2. Data-driven cosine series (universal) estimates: Dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$. The underlying corner densities are shown by solid lines. Note that a solid line may "hide" other lines, and this implies a perfect estimation. For instance, in the diagram Uniform the dotted line coincides with the solid one, and thus it is invisible. {Recall that this figure may be repeated (with other simulated data sets) by calling (after the S–PLUS prompt) the S-function > **ch3(f=2)**. Also, see the caption of Figure 2.3 about a custom-made density. All the arguments, shown below in square brackets, may be changed. Let us review these arguments. The argument *set.n* allows one to choose 3 (or fewer) different sample sizes. Set *sam=T* to see samples for the first sample size. The arguments *cJ0* and *cJ1* control the parameters c_{J0} and c_{J1}, which are used to calculate J_n in (3.1.10). Note that S–PLUS does not recognize subscripts, so we use *cJ*0 instead of c_{J0}, etc. The argument *cJM* is used in (3.1.14) as the factor for the highest possible frequency, and cT is used in (3.1.14) as the coefficient in the hard thresholding. Furthermore, cB is the coefficient in the procedure of removing small bumps. Also recall that below in the square brackets the default values for these arguments are given. Thus, after the call > **ch3(f=2)** the estimates will be calculated with these values of the coefficients. If one would like to change them, for instance to use a different threshold level, say $c_T = 6$, make the call > **ch3(f=2, cT=6)**.} [*set.n=c(50,100,200), sam=F,cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

sections, the estimates (3.1.13) and (3.1.14) typically coincide (there are no large high-frequency terms). Estimation of some densities, and the Strata is a particular example, greatly benefits from the procedure (3.1.15) of bona fide estimation. Indeed, we clearly see two strata in Figure 3.2.4, while some oscillations were always presented in approximations discussed in Section 2.1. Exercise 3.1.15 explains how to visualize the impact of removing bumps.

Now, let us look at other densities. First of all, the examples of the Uniform and the Angle show that a twofold increase in the sample size

does not necessarily improve an estimate. Moreover, it may be dramatically worse. Just look at the diagram for the Angle, where the best estimate is the dotted line corresponding to $n = 50$.

This does not happen often, but the fact that larger sample sizes may lead to worse estimates is counterintuitive.

To understand this phenomenon, let us consider a simple example. Suppose an urn contains 5 chips and we know that 3 of them have one color (the "main" color) and that the other two chips have another color. We know that the colors are red and green but do not know which one is the main color. We draw a chip from the urn and then want to make a decision about the main color. The natural bet is that the color of the drawn chip is the main color (after all, the chances are $\frac{3}{5}$ that the answer is correct). Now let us continue the game and draw two more chips. Clearly, a decision based on three drawn chips should only be better. However, there is a possible practical caveat. Assume that the main color is green and the first chip drawn is also green. Then the conclusion is correct. On the other hand, if two next chips are red (and this happens with probability $\frac{1}{6}$), the conclusion will be wrong despite the increased "sample size."

The estimates for the Normal are good; all the estimates for the Bimodal clearly show the modes; the estimates for the Strata exhibit two strata. For the Monotone the outcomes are not perfect, but you get the correct impression about the underlying density. For the Steps, it will be explained in Section 3.7 how to improve the long-dashed line by taking a projection on a class of monotone densities.

Let us return one more time to the Uniform and the Angle diagrams. Is it possible to avoid the oscillations of the particular estimates? One of the possibilities to do this is to decrease J_n by choosing smaller values for *cJ0* and *cJ1*. The caveat is that in this case the estimation of a density like the Strata or the Bimodal may be worse, because these densities require larger cutoffs. Thus, an optimal choice of these two coefficients is a tradeoff between a better estimation of functions like the Uniform and the Angle and a better estimation of functions like the Bimodal and the Strata. Section 3.2 will shed theoretical light on this choice.

Another possibility to explore the universal estimator is to analyze Figure 3.3. Here the histograms help us to shed light on the underlying data sets (the histogram estimate was discussed in Section 1.1, and a review may be found in Section 8.1). Note that each figure is based on new Monte Carlo simulations; thus the underlying data sets in Figure 3.3 are different from those in Figure 3.2.

For the Uniform we see that the estimate does remarkably well for this complicated data set, whereas the histogram is oscillatory. The estimate for the Normal is skewed, and both tails are shown incorrectly. But can the estimator be blamed for this? The answer is "no." Recall that the density estimate describes the underlying data set. Thus, the fact that the estimate is zero to the right of 0.85 is consistent with the data set, and the same may

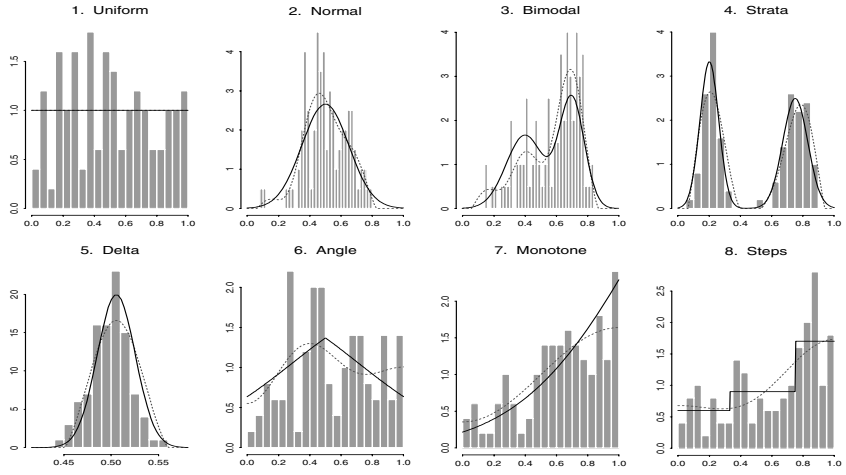

FIGURE 3.3. Performance of the universal estimator for simulated data sets of sample size $n = 100$. Histograms are overlaid by the underlying densities (solid lines) and estimates (dotted lines). [*n =100, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

be said about the left tail. Also, we see that the mode of the estimate is skewed to the left, and it is higher than the mode of the underlying density. But again, this is consistent with the data at hand.

A similar situation is exhibited in the Bimodal diagram. Here the left tail of the estimate does not vanish appropriately, and the histogram explains why. On the other hand, the two modes are well exhibited, and their locations are shown correctly. Note that here the nonparametric estimate does help us to see the shape of the Bimodal, which is not obvious when one tries to analyze the histogram. The estimate for the Strata absolutely correctly shows the two strata, but the magnitude of the left one is shown too conservatively. Here the histogram does a better job in showing the difference between the strata. On the other hand, the location of the main mode is shown more correctly by the nonparametric estimate. For the Delta the universal estimate does a remarkable job in showing us both the location of the peak and the symmetric shape of the Delta, while the histogram is clearly skewed.

Now let us look at the Angle diagram. The estimate is bad; no doubt about this. But can we improve it having this particular data set at hand? The answer is "no." The mode of the dotted line is shifted to the left, but this is what we see in the data. Also, the right tail goes up instead of down, but again, there are no data that could "convince" the estimate to go down (just for comparison, look at the left tail, where the data clearly indicate that the underlying density decreases). The Monotone is another example that explains why density estimation is the art of smoothing. Here both

the left and right tails of the estimate incorrectly represent the underlying density. On the other hand, for this data set the only other option for an estimate to show tails correctly is via waves (indeed, look at the tails of the histogram). Among these two options, the universal estimate chose the better one because at least the monotonic nature of the underlying density was exhibited correctly. On the other hand, the histogram explains why oscillatory nonparametric estimates for densities like the Uniform, the Angle, and the Monotone may appear for particular simulations.

Finally, consider the case of Steps. Here it is worthwhile to put yourself in the shoes of the data-driven estimator. Look at this particular data set, compare it with the underlying density, and then try to answer the following two questions. Is there anything in this data set that indicates the underlying Step density? Can you suggest a procedure of smoothing this particular histogram that will "beat" the universal estimate? If your answers are "yes," then try to apply your procedure to the other histograms. If your algorithm regularly outperforms the universal estimate, then it is better.

Two conclusions may be drawn from the analysis of these particular estimates. First, the visualization of a particular estimate is useful and sheds light on the estimator. Second, a conclusion may not be robust if it is based only on the analysis of just several simulations. The reason is that for any estimator one can find a data set where an estimate is perfect or, conversely, very bad. Thus, every experiment (every figure) should be repeated many times (the rule of thumb, based on the author's experience, is that the results of at least 20 simulations should be analyzed before making a conclusion). Also, there exist more rigorous methods for assessing the quality of estimation. We shall discuss them in the following two sections.

Remark 3.1.2. In the case where the density is estimated over a given interval $[a, b]$ (or data are given only for this interval), to convert the problem to the case of the $[0, 1]$ interval one first should rescale the data and compute $Y_l := (X_l - a)/(b - a)$. The rescaled observations are distributed according to a density $f^Y(y)$, which is then estimated over the unit interval $[0, 1]$. Let $\tilde{f}^Y(y)$, $y \in [0, 1]$, be the obtained estimate of the density $f^Y(y)$. Then the corresponding estimate of $f^X(x)$ over the interval $[a, b]$ is defined by $\tilde{f}^X(x) := (b - a)^{-1}\tilde{f}^Y((x - a)/(b - a))$, $x \in [a, b]$. Figure 3.4 shows how this approach works for the interval $[0.4, 0.9]$. The particular outcome for the Normal is the worst one. We see that the dotted line (the estimate) is oscillatory, and this is very confusing. On the other hand, these oscillations perfectly fit the data set over the interval $[0.4, 0.9]$. The decision of the estimator is not so clear for the case of the Uniform, but the message of the estimate is that there are more observations near 0.9 than near 0.4, and this is what we may agree with. Also, keep in mind that for this setting the estimator has at hand only those of the 100 observations that belong to the interval $[0.4, 0.9]$. This decreases the size of the data sets and does not allow the estimator to make a conclusion about the smoothness of an underlying

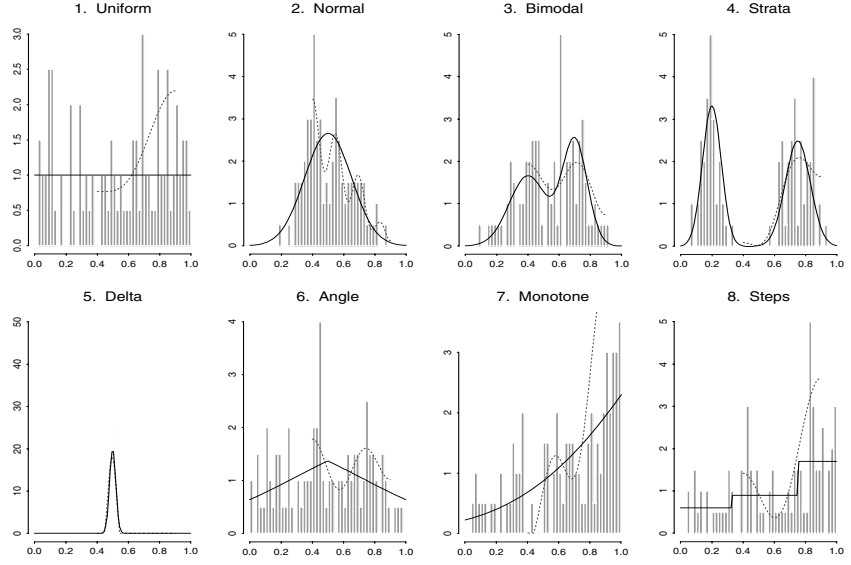

FIGURE 3.4. The universal estimates over a subinterval $[a, b] = [0.4, 0.9]$. Histograms (except for the Delta) and underlying densities (solid lines) are shown over the support $[0, 1]$, and the estimates (dotted lines) are shown over $[a, b]$. The sample size is $n = 100$. [*n =100, a=.4, b=.9, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

density based on other observations. On the other hand, this figure sheds light on the idea of a "local" estimation when one estimates density at a point x based only on the observations nearest to that point. Apparently, this may be a good idea for estimation of spatially inhomogeneous densities.

Remark 3.1.3. Consider the setting where one would like to estimate an underlying density over its finite support $[a, b]$, which is unknown. In other words, both the density and its support are of interest. (Recall that by the support we mean a minimal interval beyond which the density vanishes.) In this case the only sure fact about the support is that, according to Exercise A.12 in Appendix A, $P(a < X < X_{(1)}) = P(X_{(n)} < X < b) = 1/(n + 1)$, where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are the ordered observations. Thus, $a =: X_{(1)} - d_1$ and $b =: X_{(n)} + d_2$, where both $d_1 > 0$ and $d_2 > 0$ should be estimated. Let us use the following approach. If an underlying density is flat near the boundaries of its support, then for a sufficiently small positive integer s we have $(X_{(1+s)} - X_{(1)})/s \approx X_{(1)} - a = d1$, and similarly $(X_{(n)} - X_{(n-s)})/s \approx b - X_{(n)} = d2$. The default value of s is 1. Thus, we set

$$\hat{d}_1 := (X_{(1+s)} - X_{(1)})/s, \quad \hat{d}_2 := (X_{(n)} - X_{(n-s)})/s. \tag{3.1.16}$$

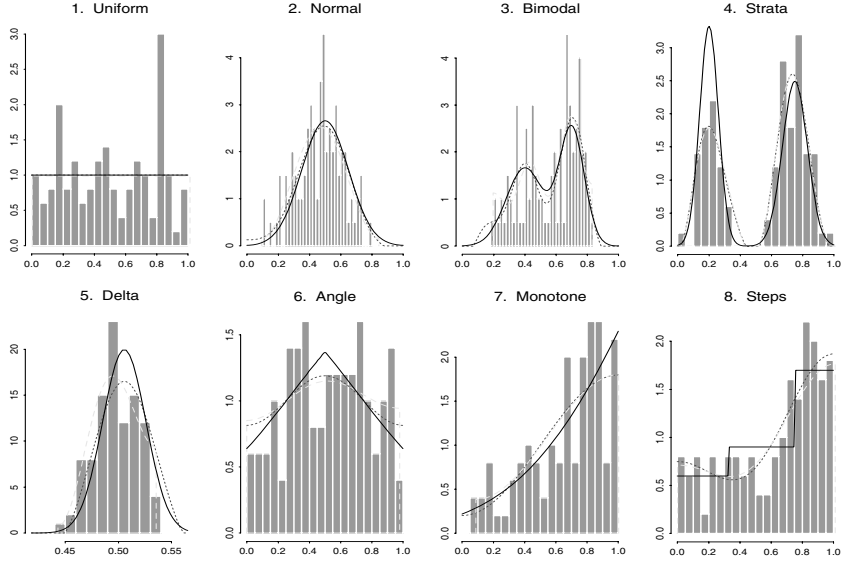

FIGURE 3.5. The universal estimates based on the support $[0, 1]$ and a support estimated according to (3.1.16). These estimates are shown by dotted and dashed lines, respectively. Underlying histograms and densities (solid lines) are shown as well. [*n =100, s=1, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

More precise estimation of d_1 and d_2 requires estimation of both the density and its derivatives near $X_{(1)}$ and $X_{(n)}$, and this is a complicated problem for the case of small sample sizes.

Figure 3.5 illustrates how this idea works. The histograms for simulated data sets of size $n = 100$ are overlaid by estimates that use the support $[0, 1]$ (the dotted lines) and by estimates that use the estimated support $[X_{(1)} - \hat{d}_1, X_{(n)} + \hat{d}_2]$ (the dashed lines). We see that for densities that are flat near the edges the recommended method performs well. All the other estimates are apparently affected by the fact that the support is unknown. And this takes its toll for the cases of densities with light tails. Just look at the Normal, the Bimodal, and especially the Delta. There is no way for small samples to indicate that the support is $[0, 1]$, and as a result, these estimates are discontinuous over $[0, 1]$.

What if it is known that a density is continuous over the real line? In other words, let an underlying density vanish (be equal to zero) at boundary points of the support. Then an estimated support may be defined as a minimal interval where the data-driven estimate (3.1.15) vanishes at its boundary points.

Figure 3.6 illustrates this approach, and it also allows us to understand how this or that hypothetical support (the interval $[a, b]$) affects the universal estimate. Here the left diagram (a) shows a histogram for simulated

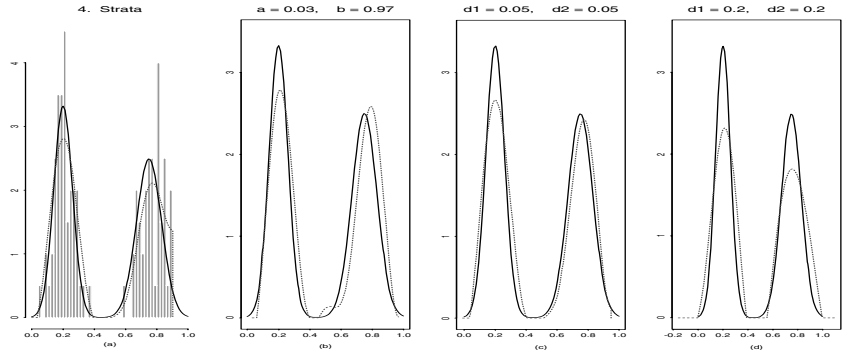

FIGURE 3.6. The effect of a chosen interval (support) $[a, b] = [X_{(1)} - d_1, X_{(n)} + d_2]$ on the universal estimate (the dotted line). The case of the sample size $n = 100$ is considered. The solid line shows the underlying density. The diagram (a) shows the histogram overlaid by the universal estimate over an interval calculated by formula (3.1.16). For the same data set the diagram (b) shows the estimate that is based on a minimal support $[a, b]$, shown in the title of the diagram, such that the estimate vanishes at the boundary points. Two other diagrams (c) and (d) exhibit outcomes for particular d_1 and d_2 shown in the title. {The diagrams (c) and (d) are plotted interactively after entering (at the prompt "1:") $d1$ and $d2$ from the keyboard (use a space between these numbers and then press ENTER (RETURN)). The argument *corden* allows one to change the underlying density.} [*n =100, corden = 4, s=1, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

data. This histogram is overlaid by the underlying density (the solid line) and the estimate (the dotted line) based on a support calculated with the help of (3.1.16). We see that the estimate does a good job apart from the discontinuity in the right tail. The diagram (b) again depicts the underlying density and exhibits the estimate with the minimal support where the estimate vanishes at its boundary points. This support is larger, and we also see the pronounced flat appendix to the left tail of the right stratum, which is also seen in the histogram. The other two diagrams show estimates for some particular values of d_1 and d_2. In general, the larger a hypothetical support, the smoother the corresponding estimate.

Remark 3.1.4. We shall see that for different settings, which include regression, filtering, and spectral density estimation, the parameter $d := \lim_{j,n\to\infty} nE\{(\hat{\theta}_j - \theta_j)^2\}$, where $\hat{\theta}_j$ is an appropriate sample mean estimate of θ_j, defines the factor in changing a sample size that makes estimation of an underlying curve comparable, in terms of the same precision of estimation, with the density estimation model over the known support $[0, 1]$ when $\hat{d} = d = 1$. In other words, the problem of density estimation may be considered as a basic one for analyzing other models. Thus we shall refer to the coefficient d as the *coefficient of difficulty*. We shall see that it is a valuable tool, which allows us to judge the complexity of a problem based on experience with the density estimation.

3.2 Lower Bounds (Oracle Inequalities)

The aim of this section is to develop lower bounds for MISE (mean integrated squared error) that could be used to answer the question of Section 3.1 on how far this or that estimate is from a "golden standard." Also, we shall discuss the ideas of how to suggest a "good" data-driven estimator. In this section it is assumed that the support of an estimated function is $[0, 1]$.

Asymptotic (when $n \to \infty$) theory obtains lower bounds theoretically, basically using limit theorems discussed in Appendix A. An estimate is said to be optimal if its MISE is close to a lower bound. This is also a prudent method to rank estimators. For the case of small sample sizes we cannot use limit theorems. Thus we employ the idea of oracle inequalities for a family of corner functions.

The idea is as follows. We choose an ideal oracle (pseudo-estimator) that is based on both data and an underlying density. Then for a particular underlying density an exact risk (i.e., MISE) of the oracle is calculated, and it serves as the lower bound ("golden standard").

To explain the idea more precisely, consider the setting of Section 3.1 with $[0, 1]$ being the support, and recall that then an orthonormal series estimator may be written as

$$\hat{f}(x, \{w_j\}) := 1 + \sum_{j=1}^{\infty} w_j \hat{\theta}_j \varphi_j(x). \tag{3.2.1}$$

Then, the only difference between estimators is in the method of choosing the weights w_j. Assume that these weights may depend on an underlying density f. Then we refer to this estimator as an *oracle* (*guru* or *supervisor*) because the oracle knows both the data and the underlying density. On the other hand, an oracle may use the underlying density f only for choosing the weights, and oracles differ by how they choose the weights.

Below we define linear, raw truncated, smoothed truncated, and hard-threshold oracles.

(i) *Linear (optimal smoothing) oracle.* This is a pseudo-estimator with weights w_j^* defined in (3.1.11). As we know from Example A.25 in Appendix A, this estimator has the minimal MISE over all possible estimates (3.2.1). It is worthwhile to repeat here the explanation why this is the case. Using Parseval's identity we write

$$\begin{aligned} \text{MISE}(\hat{f}, f) &:= E\left\{\int_0^1 (\hat{f}(x, \{w_j\}) - f(x))^2 dx\right\} = \sum_{j=1}^{\infty} E\{(w_j \hat{\theta}_j - \theta_j)^2\} \\ &= \sum_{j=1}^{\infty} (w_j^2 E\{\hat{\theta}_j^2\} - 2 w_j E\{\hat{\theta}_j\}\theta_j + \theta_j^2). \end{aligned} \tag{3.2.2}$$

Note that the sample mean is an unbiased estimate of the mean. Thus here we have $E\{\hat{\theta}_j\} = \theta_j$, and therefore $E\{\hat{\theta}_j^2\} = \mathrm{Var}(\hat{\theta}_j) + \theta_j^2$ due to (A.5). Substituting these relations into (3.2.2) we get

$$\begin{aligned}\mathrm{MISE}(\hat{f}, f) &= \sum_{j=1}^{\infty}[w_j^2(\theta_j^2 + \mathrm{Var}(\hat{\theta}_j)) - 2w_j\theta_j^2 + \theta_j^2] \\ &= \sum_{j=1}^{\infty}(\theta_j^2 + \mathrm{Var}(\hat{\theta}_j))[w_j - \theta_j^2/(\theta_j^2 + \mathrm{Var}(\hat{\theta}_j))]^2 \\ &\quad + \sum_{j=1}^{\infty}\theta_j^2\mathrm{Var}(\hat{\theta}_j)/(\theta_j^2 + \mathrm{Var}(\hat{\theta}_j)). \qquad (3.2.3)\end{aligned}$$

Thus, the MISE is minimized by the smoothing (shrinkage) coefficients $w_j^* = \theta_j^2/(\theta_j^2 + Var(\hat{\theta}_j))$ that are optimal pseudo-coefficients (they depend on f). The corresponding estimator is called a *linear oracle*. We have also obtained the *exact* expression for the MISE of the linear oracle,

$$\mathrm{OMISEL} = \sum_{j=1}^{\infty} w_j^*\mathrm{Var}(\hat{\theta}_j). \qquad (3.2.4)$$

The abbreviation OMISEL stands for the oracle MISE of linear estimate. Also, recall that the formula to calculate $\mathrm{Var}(\hat{\theta}_j)$ is given in (3.1.8).

The OMISEL is a natural candidate for the lower bound (*oracle inequality*) for the MISE of any estimator. If the MISE of an estimator is close to the OMISEL, then we may conclude that the estimator is *efficient* (at least for a given underlying density).

On the other hand, it is not clear how to mimic the linear oracle by a data-driven estimate because this oracle uses (and thus knows) absolutely all Fourier coefficients. Thus, let us introduce several less "informed" oracles that know only some Fourier coefficients and therefore may be mimicked by a data-driven estimator.

(ii) *Raw truncated oracle.* This is the estimator (3.2.1) with $w_j = 1$ for $j \le J^*$ and $w_j = 0$ otherwise (the name is clear from the procedure where only the cutoff J^* may depend on f), that is,

$$\hat{f}_T(x) = 1 + \sum_{j=1}^{J^*}\hat{\theta}_j\varphi_j(x).$$

This oracle is very simple, and the only parameter chosen by the oracle is the optimal cutoff $J^* := \mathrm{argmin}\{R^*(k),\ k = 0, 1, \ldots\}$, where $R^*(k)$ is the MISE of the raw truncated oracle, that is, according to (3.1.6), $R^*(k) = \sum_{j=1}^{k}\mathrm{Var}(\hat{\theta}_j) + \sum_{j>k}\theta_j^2$.

The MISE of the raw truncated oracle is equal to $R^*(J^*)$ and we call it the OMISET (here the letter T stands for "truncated," and we shall often

refer to this estimator as truncated). Thus,

$$\text{OMISET} = \sum_{j=1}^{J^*} \text{Var}(\hat{\theta}_j) + \sum_{j>J^*} \theta_j^2. \tag{3.2.5}$$

To simplify mimicking this oracle, as in Section 3.1 we use n^{-1} as the estimate of $\text{Var}(\hat{\theta}_j)$. In general, this may imply a cutoff J different from the optimal J^*. However, direct calculations of OMISET for all the corner densities and sample sizes from 25 to 1000 have revealed that $J^* = J$. Thus, from now on the cutoff J is used in place of J^*.

The raw truncated oracle is appealing due to its simplicity. At the same time, this oracle ignores the possibility of improving the performance by shrinking (multiplication by weights) the chosen J Fourier coefficients. The next oracle does exactly this.

(iii) *Smoothed truncated oracle.* This oracle performs like the raw truncated one, only in addition, it shrinks the first J Fourier coefficients via multiplication by optimal smoothing coefficients w_j^*. The exact MISE of this oracle is simply calculated,

$$\text{OMISES} = \sum_{j=1}^{J} w_j^* \text{Var}(\hat{\theta}_j) + \sum_{j>J} \theta_j^2. \tag{3.2.6}$$

Here the letter S in the abbreviation OMISES stands for "smoothed truncated," and we shall refer to this oracle as smoothed.

We shall see that overall a data-driven estimator that mimics this oracle is the best one.

The smoothed oracle shrinks the estimated coefficients, the alternative being to "keep" or "kill" them. This is the approach of the next oracle.

(iv) *Hard-threshold truncated oracle.* This oracle again considers the first J Fourier coefficients as the raw truncated oracle does, but then it keeps or kills them according to the rule $w_j := I_{\{\theta_j^2 > 2\ln(n)/n\}}$ (recall that $I_{\{A\}}$ is the indicator of an event A). The truncation is motivated by the asymptotic theory discussed in Section 7.4. The abbreviation for the MISE of this oracle is OMISEH. Here the letter H stands for "hard-threshold," and the oracle is referred to as hard-threshold. Then

$$\text{OMISEH} = \sum_{j=1}^{J} [\text{Var}(\hat{\theta}_j) I_{\{w_j=1\}} + \theta_j^2 I_{\{w_j=0\}}] + \sum_{j>J} \theta_j^2. \tag{3.2.7}$$

The belief that this oracle should perform well for small sample sizes is based on the conjecture that a majority of squared Fourier coefficients has the "large-small" property, that is, they are either large or small in comparison with n^{-1}.

Now all the four oracles have been introduced, and we are in a position to study them. A special remark is to be made about computing the MISE for the Uniform density. For this density OMISEL = OMISET = OMISEH

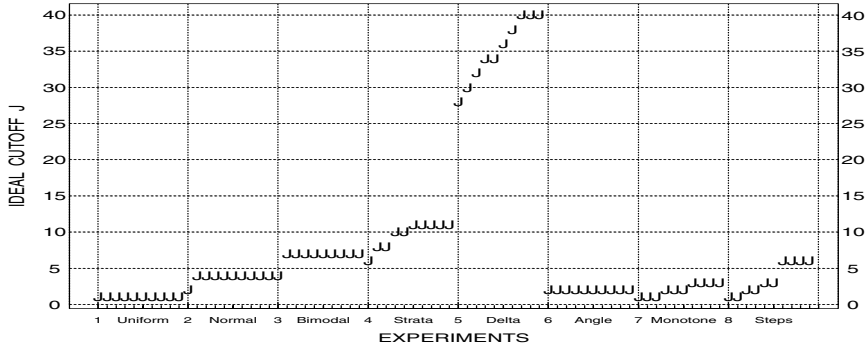

FIGURE 3.7. Ideal cutoff J.

= OMISES = 0 and $J = J^* = 0$. This is not a "fair" lower bound, especially if we would like to study ratios of risks. Thus only for the Uniform density do we "ask" all the oracles to set $J = J^* = 1$ and OMISEL = OMISET = OMISEH = OMISES = n^{-1}.

Let us analyze MISE of the oracles for our set of 8 corner densities shown in Figure 2.1 and for a set of 10 sample sizes (25, 50, 75, 100, 150, 200, 400, 600, 800, and 1000). Overall, we have 80 experiments (examples), and each experiment is devoted to a particular density and sample size. To visualize all the experiments simultaneously in one figure (see Figure 3.7 as an example) we refer to the horizontal axis as "experiments," where coordinate $i.0$ corresponds to density $\#i$ with a sample size 25; similarly, coordinates $i.1, i.2, \ldots, i.9$ correspond to density $\#i$ with sample sizes of 50, 75, 100, 150, 200, 400, 600, 800, and 1000, respectively.

First let us consider the optimal cutoffs J shown in Figure 3.7 (it is worthwhile to repeat that for all the experiments $J = J^*$). When you look at the ideal cutoffs, it is striking that except for the Delta the cutoffs are surprisingly small. In no way do they try to match the sample size. Thus, we see that an orthogonal series estimator is a good tool for data compression and developing simple formulae for underlying densities (of course, here this conclusion is based only on the analysis of the corner densities, but it is also supported by the asymptotic theory). The Delta is a very special case, and it is worthwhile to explain why. Due to Parseval's identity, the sum of the squared Fourier coefficients is equal to the integral of the squared Delta density, which is relatively large (for a theoretical delta function the integral is equal to infinity). This implies that there are many large (in comparison with $n^{-1/2}$) Fourier coefficients that are to be estimated even for the smallest sample sizes. This example is worthwhile to keep in mind because it shows that the integral of a squared function may have a significant effect on the optimal cutoff. Note also the relatively large ideal cutoffs for the Strata explained by the same reason.

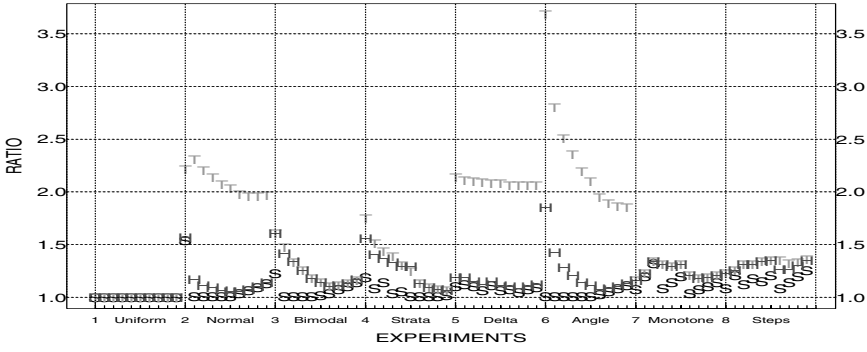

FIGURE 3.8. Oracle ratios OMISES/OMISEL, OMISET/OMISEL, and OMISEH/OMISEL shown by the letters S, T, and H, respectively.

The other interesting conclusion is that the ideal cutoff as a function in n is not necessarily a strictly increasing function. The Normal and the Bimodal are particular examples that show this clearly.

Now let us consider MISE of the oracles. Recall that OMISEL may be used as a lower bound or the benchmark for the other oracles. Moreover, we shall see in Chapter 7 that asymptotically it is the best among all possible estimators over a wide variety of function classes. Thus we refer to the inequalities MISE $\geq$ OMISEL and MISE/OMISEL ≥ 1 as the *lower bound* and the *oracle inequality*, respectively, and to the ratio MISE/OMISEL as the *oracle ratio*.

Let us consider the oracle ratios for the "less informed" oracles. The oracle ratios are shown in Figure 3.8. The ratios confirm the previous theoretical conclusion that the smoothed oracle performs better than the other two. Secondly, for all the experiments OMISES is close to OMISEL, and this supports our choice of the cutoff J that minimizes the approximation $n^{-1}J + \sum_{j>J} \theta_j^2$ of MISE of the raw truncated oracle, our simplified estimation of $\mathrm{Var}(\hat{\theta}_j)$ by n^{-1}, and the approach to choose J by minimizing MISE of the raw truncated oracle instead of the smoothed one.

The truncated oracle does not perform as well as the smoothed one. This is primarily due to the phenomenon discussed above of small and large Fourier coefficients. For instance, all odd Fourier coefficients of the Normal density are zero, and this explains why this oracle does not perform well for the Normal density. A similar situation occurs for the Delta and the Angle. However, the truncated oracle is much simpler than the others, and this is a plus for a data-driven estimator that mimics it, as we shall see later.

For some experiments OMISET is close to OMISES (see the Bimodal and the Strata). As a result, a simpler data-driven estimator, mimicking the raw truncated oracle, may perform better than a more complicated

smoothed adaptive estimator. This remark is important because the choice of a data-driven estimator should not be made solely on the merits of an underlying oracle. The simplicity of mimicking is another important factor, and the truncated oracle is the simplest to mimic.

The analysis of the ratio OMISEH/OMISEL and the fact that OMISEH is typically significantly smaller than OMISET for the smallest sample sizes also confirms our conjecture about the "large–small" property of squared Fourier coefficients in comparison with n^{-1}. As we have discussed, for densities like the Normal and the Angle we clearly have "large" and "small" Fourier coefficients, while for the Monotone all coefficients are "large." The latter is due to the the fact that $|\theta_j|$ decreases at the rate j^{-2}; recall (2.2.8) and the discussion below that line.

3.3 Data-Driven Estimators

In this section we study data-driven (adaptive) estimators that mimic the above-introduced oracles (except for the linear oracle) and bear their names. The linear oracle serves as an "unmanageable" benchmark and gives us the lower bound.

Let us formally define the estimators. Recall that $[0, 1]$ is the support, so $\hat{d} = \hat{\theta}_0 = 1$.

An adaptive (raw) truncated estimator is

$$\hat{f}_t(x) := 1 + \sum_{j=1}^{\hat{J}} \hat{\theta}_j \varphi_j(x), \tag{3.3.1}$$

where $\hat{J}$ is defined in (3.1.10).

A smoothed estimator mimics the smoothed truncated pseudo-estimator and is defined in (3.1.13).

A hard-threshold estimator mimics the hard-threshold truncated oracle and is defined by

$$\hat{f}_h(x) := 1 + \sum_{j=1}^{\hat{J}} I_{\{\hat{\theta}_j^2 > 2\ln(n)/n\}} \hat{\theta}_j \varphi_j(x). \tag{3.3.2}$$

Finally, to make all these estimators bona fide, the procedure (3.1.15) is used.

To analyze the estimators, a Monte Carlo study is used. For each experiment (that is, for a density and a sample size), we make a sufficiently large number m of repeated simulations (here $m = 5000$) and then calculate the *average integrated squared error* (AISE) defined as $m^{-1} \sum_{l=1}^{m} \int_0^1 (\hat{f}_l(x) - f(x))^2 dx$, where $\hat{f}_l$ is an estimate based on data generated by the lth Monte Carlo simulation. We shall use the natural notation

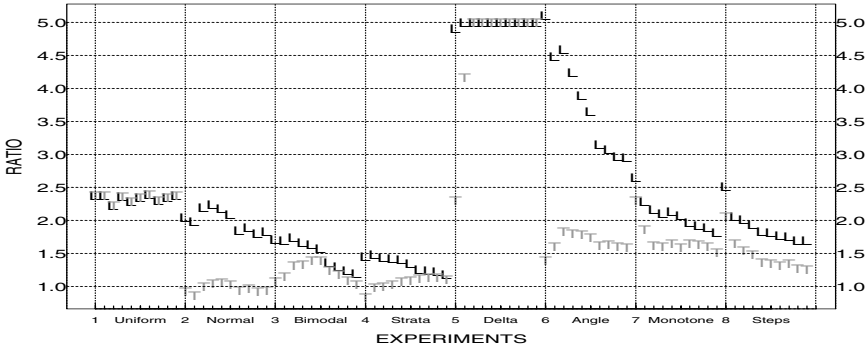

FIGURE 3.9. The ratios AISES/OMISEL and AISES/OMISET shown by the letters L and T, respectively. For the Delta the ratios are truncated at 5.0.

AISES, AISET, and AISEH for the average risks of smoothed, truncated, and hard-threshold estimators.

Let us begin the analysis with the oracle inequalities for the smoothed estimator. The ratios AISES/OMISEL and AISES/OMISET are shown in Figure 3.9. The ratios for the Delta are truncated at level 5 because they are very large (go to 90) for reasons discussed previously (recall that because of this we are adding high-frequency terms in (3.1.14)). Overall, it is fair to say that this adaptive estimator performs well in comparison with the oracles. Note that for a majority of the experiments the ratio is less than 2, and this is an absolutely amazing result because the OMISEL is the lower bound (oracle inequality) for the MISE. Moreover, we see that for some experiments the data-driven estimate outperforms the truncated oracle.

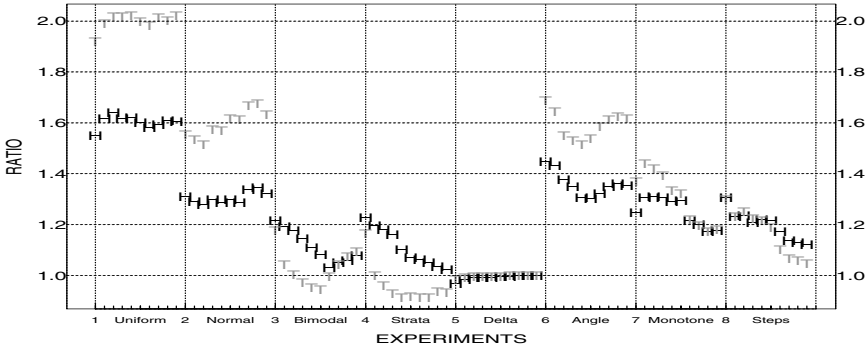

FIGURE 3.10. Ratios AISET/AISES and AISEH/AISES shown by the letters T and H.

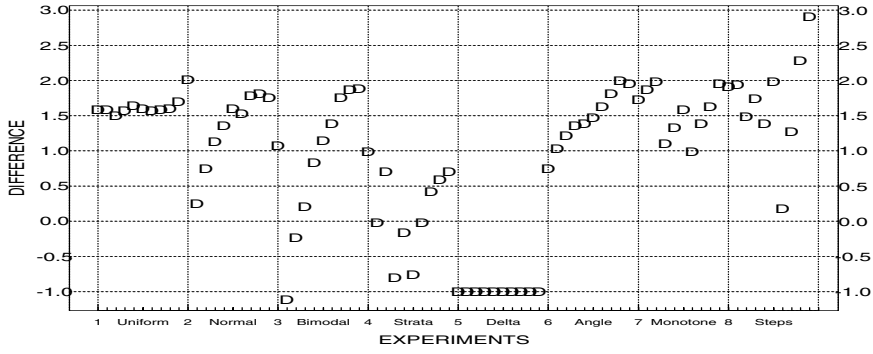

FIGURE 3.11. Differences $\bar{J} - J$.

Now let us consider how the other estimators perform. In Figure 3.10 the ratios AISET/AISES and AISEH/AISES are shown.

We see that for some experiments with the Bimodal and the Strata densities the truncated estimator outperforms the smoothed one, but overall the smoothed estimator is clearly the "winner."

Finally, let us consider the important part of our estimators—mimicking J by the estimate (3.1.10). The differences between the average (over the same $m = 5000$ Monte Carlo simulations) value $\bar{J}$ of the estimates $\hat{J}$ and the ideal J are shown in Figure 3.11. For the case of the Delta the differences are very large; thus we just set them to -1; this is the reason why the thresholded terms are added in (3.1.14). Overall, we see that the estimate does a good job in finding optimal cutoffs.

3.4 Case Study: Survival Analysis

Survival analysis is a class of statistical methods that were originally applied to the study of deaths, thus explaining the name. However, for now this is an important part of statistics that is useful for studying different events including equipment failures, stock market crashes, job terminations, births, arrests, and retirements. It is worthwhile to note that in different fields scientists have given their own names to the topic; for example, reliability analysis (engineering), duration analysis (economics), event history analysis (sociology), and transition analysis (economics). The different names do not imply different mathematical methods but emphasize different applied aspects.

In this section, as an example, we consider the problem of density estimation for the case of right-censored observations.

Let X denote a lifetime random variable whose observations may be *right-censored*. An observation X is right-censored if all you know about X

is that it is greater than some given value. An example that motivated the name is as follows. Suppose that X is a person's age at death (in years), and you know only that it is larger than 61, in which case the time is censored at 61 (this may occur if the person at age 62 moved to another part of the country or to another country and can no longer be traced). Another example is the lifetime of a light bulb that may accidently break before it burns out. But in no way is the notion of censoring restricted to event times. For instance, if you know only that the invoice price of a car is greater than \$20,000, then the price is right-censored at \$20,000.

Clearly, censoring changes the problem, and we cannot use the previous algorithms directly. Nevertheless, the underlying idea of series estimation again may be used.

Let us put right censoring into a statistical framework. Assume that X is a random variable with density $f^X(x)$, $x \in (-\infty, \infty)$, and that there exist n iid realizations $X_1, \ldots, X_n$ of X.

The problem is to estimate f^X over the interval $[0, 1]$ when the realizations are not available to us directly, but instead, the data (Y_l, δ_l), $l = 1, 2, \ldots, n$, are given, where $Y_l = \min(X_l, T_l)$, $\delta_l = I_{\{X_l \le T_l\}}$, and T_l are iid random variables that "censor" the random variable of interest X. Such data are called *censored on the right* (or right-censored), and all the examples discussed above fall into this framework. Recall that $I_{\{A\}}$ is the indicator of an event A, that is, it is equal to 1 if the event occurs and 0 otherwise. Thus, the data at hand consist of either realizations of X that are not larger than corresponding realizations of T or realizations of T otherwise.

Probably, the first "natural" idea of how to estimate f is to use only the uncensored realizations of X. Let us see what happens in this case. Assume that T is uniformly distributed on $[0, 1.5]$ and then use the idea together with the data-driven density estimate (3.1.15). The results are shown in Figure 3.12. We see that the estimates are skewed to the left (with respect to the underlying corner densities), and this is a natural outcome. Indeed, by using only uncensored observations, the proportion of smaller observations is increased, and this implies the skewness. This is probably most clearly seen for the Uniform, the Strata, the Angle, and the Steps.

Thus, the "naive" approach does not work. To understand how to find a series estimate, we again begin with writing a partial sum,

$$f_J^X(x) = \sum_{j=0}^{J} \theta_j \varphi_j(x)\,, \; 0 \le x \le 1, \quad \text{where } \theta_j = \int_0^1 f^X(x)\varphi_j(x)dx\,. \tag{3.4.1}$$

Recall that the main idea of a series estimate is to write down the Fourier coefficient θ_j as the expectation of a function $\psi_j(Y, \delta)$ of the observed pair of random variables (Y, δ), that is, to find $\psi_j(Y, \delta)$ such that $\theta_j = E\{\psi_j(Y, \delta)\}$. Then the sample mean estimate $\hat\theta_j = n^{-1} \sum_{l=1}^n \psi_j(Y_l, \delta_l)$ may be used.

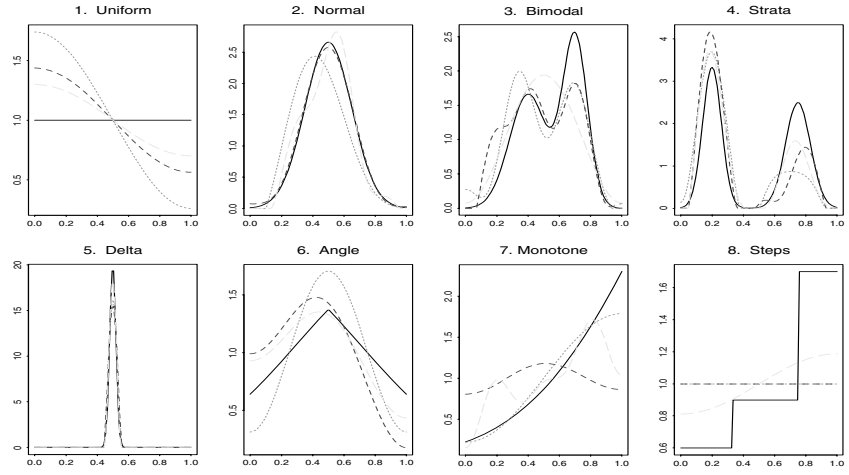

FIGURE 3.12. The universal estimates based only on noncensored realizations: Dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$. Data are censored on the right by the uniform random variable $U(0, a)$. The solid lines show the underlying corner densities. [*set.n =c(50,100,200), a=1.5, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

To choose a function ψ_j, we should find the distribution of the observed pair (Y, δ). Recall that δ takes on values 0 or 1. Write

$$P(Y \le y,\ \delta = 1) = P(\min(X, T) \le y,\ X \le T) = P(X \le y,\ X \le T)$$
$$= \int_{-\infty}^{y} f^X(x)(1 - F^T(x))dx. \tag{3.4.2}$$

Here $F^T(y) := P(T \le y)$ is the cumulative distribution function (cdf) of T. Recall that the function $G^T(y) := 1 - F^T(y)$ is called the *survivor function* corresponding to the cdf F^T. Then the relations

$$\theta_j = \int_0^1 f^X(y)\varphi_j(y)dy = \int_{-\infty}^{\infty} I_{\{0\le y\le 1\}} f^X(y)G^T(y)[\varphi_j(y)/G^T(y)]dy$$
$$= E\{\delta I_{\{0\le Y\le 1\}}[\varphi_j(Y)/G^T(Y)]\}$$

show that one can choose $\psi_j(Y, \delta) := \delta I_{\{0\le Y\le 1\}}\varphi_j(Y)/G^T(y)$. This implies the following sample mean estimate:

$$\hat\theta_j := n^{-1}\sum_{l=1}^{n} \delta_l I_{\{0\le Y_l\le 1\}}\varphi_j(Y_l)/G^T(Y_l). \tag{3.4.3}$$

Of course, it is assumed that $G^T(Y_l)$ is not equal to 0.

As an exercise, let us check that (3.4.3) is indeed an unbiased estimate of θ_j. Write

$$E\{\hat\theta_j\} = E\{\delta I_{\{0\le Y\le 1\}}\varphi_j(Y)/G^T(Y)\}$$

$$= E\{(I_{\{\delta=0\}} + I_{\{\delta=1\}})\delta I_{\{0\le Y\le 1\}}\varphi_j(Y)/G^T(Y)\}$$
$$= E\{I_{\{\delta=1\}}I_{\{0\le Y\le 1\}}\varphi_j(Y)/G^T(Y)\} = \theta_j,$$

where (3.4.2) was used to get the last equality.

To use the data-driven estimate (3.1.14), we need to suggest an estimate $\hat{d}$ of $n\mathrm{Var}(\hat{\theta}_j)$ (recall that for the uncensored case the estimate $\hat{d} = \hat{\theta}_0$ is used). This is not a complicated issue. Write

$$n\mathrm{Var}(\hat{\theta}_j) = E\big\{\big(\delta I_{\{0\le Y\le 1\}}\varphi_j(Y)/G^T(Y)\big)^2\big\} - \theta_j^2$$
$$= \int_0^1 \varphi_j^2(y)(f^X(y)/G^T(y))dy - \theta_j^2.$$

Then using (3.1.7) we get

$$n\mathrm{Var}(\hat{\theta}_j) = \int_0^1 (f^X(y)/G^T(y))dy + \int_0^1 \varphi_{2j}(y)(f^X(y)/G^T(y))dy - \theta_j^2.$$

Note that $\int_0^1 \varphi_{2j}(y)(f^X(y)/G^T(y))dy$ is the $(2j)$th Fourier coefficient of $f^X(x)/G^T(y)$, and, as we know from Section 2.2, Fourier coefficients practically vanish for large j. Thus, to estimate the variance we need to estimate $\int_0^1 (f^X(y)/G^T(y))dy = E\{\delta I_{\{0\le Y\le 1\}}(G^T(Y))^{-2}\}$. Again, to estimate the expectation a sample mean estimate can be recommended,

$$\hat{d} := n^{-1}\sum_{l=1}^{n} \delta_l I_{\{0\le Y_l\le 1\}}(G^T(Y_l))^{-2}. \tag{3.4.4}$$

Thus, if the survivor function $G^T(y)$ is given, then the data-driven estimate (3.1.14) can be used straightforwardly with $\hat{\theta}_j$ and $\hat{d}$ defined at (3.4.3) and (3.4.4), respectively. What do we do if the survivor function is unknown? Fortunately, it is a well-known problem to estimate a survivor function for censored data. Here the only point that should be clarified is that T is left censored by X. Then one of the widely used estimates is the *product-limit (Kaplan–Meier)* estimate,

$$\tilde{G}^T(x) := 1, \ \ x < Y_{(1)}; \quad \tilde{G}^T(x) := 0, \ \ x > Y_{(n)};$$
$$\tilde{G}^T(x) := \prod_{i=1}^{l-1}[(n-i)/(n-i+1)]^{1-\delta_{(i)}}, \quad Y_{(l-1)} < x \le Y_{(l)}, \tag{3.4.5}$$

where $(Y_{(l)}, \delta_{(l)})$ are ordered Y_l's with their corresponding δ_l's, $l = 1, \dots, n$. The survivor function $G^T(y)$ is assumed to be positive at the point $y = 1$ (recall that the function is used in the denominator of the ratio (3.4.3)). Thus the product-limit estimate is also truncated from below,

$$\hat{G}^T(x) := \max(\tilde{G}^T(x), 1/\ln(n)). \tag{3.4.6}$$

Now the estimate (3.1.14) may be used straightforwardly with (3.4.3) and (3.4.4) as estimates of θ_j and d, and $\hat{G}^T$ used in place of G^T. Finally, the bona fide procedure (3.1.15) is implemented.

The estimator obtained is completely data-driven, and Figure 3.13 shows how it performs. Overall, the estimates are fairly good, but definitely the censoring takes its toll and makes the estimation worse in comparison with the case of direct data. Is it possible somehow to quantify the effect of the censorship and understand what characteristics of X and T affect the estimation? In other words, if someone is experienced in density estimation based on direct data, what is the needed increase in the number of observations to get a comparable quality of estimation?

This important question is not simple even for asymptotic theory. The issue is that traditional asymptotic theory studies rates of the MISE convergence, and these rates are not affected by the censoring. Thus constants of the MISE convergence should be studied. Section 7.7 shows that for an m-fold differentiable function, the convergence of MISE is proportional to $(d_c/n)^{2m/(2m+1)}$, where $d_c := \int_0^1 (f^X(y)/G^T(y))dy$; see (7.7.1). Note that $d_c \geq \int_0^1 f^X(x)dx$, with equality only if no censoring occurs. We refer to d_c as the coefficient of difficulty due to censoring (CDC) and to $r_c := d_c/d = \int_0^1 (f^X(x)/G(x))dx / \int_0^1 f^X(x)dx$ as the *relative* CDC (RCDC). The meaning of RCDC is as follows: For large sample sizes, $r_c n$ censored observations give us about the same precision of estimation as n direct observations of X.

Can this simple asymptotic rule be used for the case of small samples? To answer this question, let us consider some particular cases. For X distributed according to the corner densities (ordered as shown in Figure 3.13)

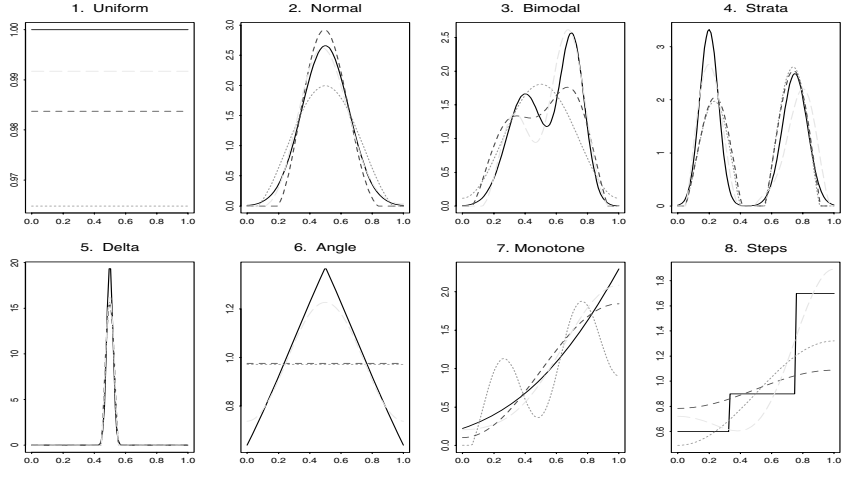

FIGURE 3.13. The universal estimates for right-censored data. Data are censored by a uniform $U(0, a)$ random variable, $a = 1.5$. The dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$. The solid lines show the underlying densities. [*set.n =c(50,100,200), a=1.5, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

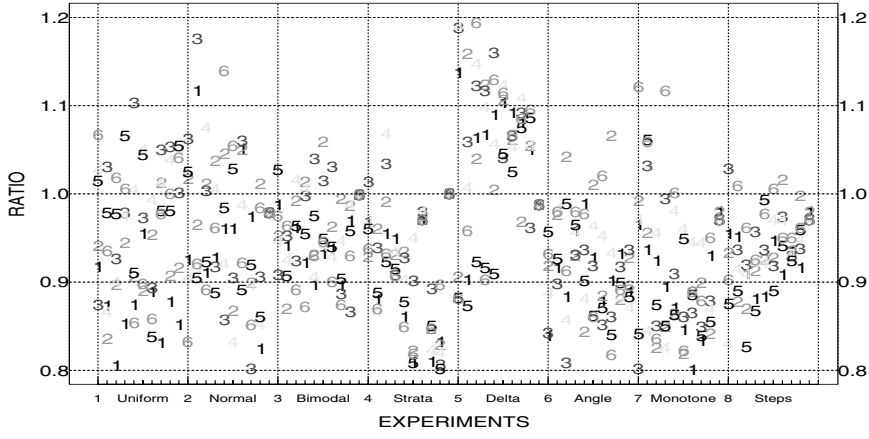

FIGURE 3.14. Ratios AISEC/AISED, where for censored data the sample size of directly observed data sets is multiplied by the RCDC. Points 1, 2, 3 correspond to Uniform $U(0, a)$ censoring with $a = 1.2$, 1.5, and 2. Points 4, 5, 6 correspond to Exponential E(λ) censoring with $\lambda = 1$, 1.5, and 2.

and T being $U(0, 1.5)$, the RCDC are equal to 1.7, 1.5, 1.6, 1.6, 1.5, 1.6, 2, and 1.9. Thus, for instance, for the Monotone we need twice as many censored observations to match the case of uncensored data. One more example. Let T have Exponential E(2) distribution, i.e., $G^T(y) = e^{-\lambda y}$, $y \geq 0$, and the rate λ is equal to 2. Then the RCDC are 3.2, 2.8, 3.2, 3, 2.7, 3.1, 4.3, and 3.9. Thus, the exponential censoring makes the problem extremely complicated, and one needs dramatically larger sample sizes to get estimates comparable to estimates based on direct data.

To verify that RCDC is a reasonable coefficient for finding a necessary increase in the size or small samples, the following experiment was conducted. Six censoring distributions were considered: 3 Uniform $U(0, a)$ with a equal to 1.2, 1.5, and 2; and 3 Exponential E(λ) with λ equal to 1, 1.5, and 2. For all these experiments, i.e., the corner functions, the censored distributions, and the sample sizes 25, 50, 75, 100, 150, 200, 400, 600, 800, and 1000 (see the discussion about an experiment in Section 3.2), 500 Monte Carlo simulations were made with sample sizes multiplied by the corresponding RCDC. Then calculated average integrated squared errors (for the censored observations) AISEC were compared with the previously obtained (in Section 3.2) AISED for the case of direct observations.

These ratios are shown in Figure 3.14. The result clearly supports the possibility to use the RCDC as a measure of difficulty due to censorship.

Remark 3.4.1. (Left Censoring) A symmetric problem to right censoring is *left censoring*, where the variable of interest is censored on the left. For instance, when a physician asks a patient about the onset of a particular

disease, a typical answer is that it occurred prior to some specific date. In this case the variable of interest is left-censored. To "translate" this setting into right censoring, choose a value A that is not less than all available left-censored observations and then consider a new data set that is A minus the given left-censored observations. The new data set is right-censored. Find the universal estimate for this right-censored data, and this estimate is the mirror image of the desired estimate for the left-censored data.

3.5 Case Study: Data Contaminated by Measurement Errors

There are many situations where a random variable is not observed directly. The case of censored data discussed above is a particular example. In this section another case is discussed where realizations of X are measured with some nonnegligible errors (the data are error contaminated). Note that errors occur not necessarily due to an imperfect measurement tool; in many cases, like a study of the behavior of insects, only indirect measurements are possible. Another classical example is a score on an IQ test that should measure the IQ of a person. Quantities that cannot be directly measured are sometimes called *latent*.

In this section a rather simple mathematical model of independent additive errors is considered, where realizations of $Y = X + \varepsilon$ are given and X is a random variable of interest.

To make this section even more interesting and enrich the "toolbox" of models, we shall consider the practically important case of *directional* (*angular*, *circular*) data where data are measured in the form of angles. Such data may be found almost everywhere throughout science. Typical examples include departure direction of birds and animals from points of release, wind and ocean current directions, times of accidents occurrence, and energy demand over a period of 24 hours. Thus, it is worthwhile to be familiar with these random variables. At the end of this section it will be explained how to solve the problem for data on the real line.

It is customary to measure directions in radians with the range $[0, 2\pi)$ radians. In this case the mathematical procedure of translation of any value onto this interval by *modulo* 2π (the shorthand notation is $[\mathrm{mod}\, 2\pi]$) is useful. As an example, $3\pi[\mathrm{mod}\, 2\pi] = 3\pi - 2\pi = \pi$, and $-2.1\pi[\mathrm{mod}\, 2\pi] = -2.1\pi + 4\pi = 1.9\pi$. In words, you add or subtract $j2\pi$ (where j is an integer) to get a result in the range $[0, 2\pi)$.

The statistical setting is as follows. The data are n independent and identically distributed realizations (so-called directions) Y_l, $l = 1, 2, \ldots, n$, of a circular random variable Y that is defined by $Y := (X + \varepsilon)[\mathrm{mod}\, 2\pi]$ (or $Y := (X[\mathrm{mod}\, 2\pi] + \varepsilon[\mathrm{mod}\, 2\pi])[\mathrm{mod}\, 2\pi]$), where the random variable ε is independent of X. The variable ε is referred to as the measurement error.

The problem is to estimate the probability density $f^X(x)$, $0 \le x < 2\pi$, of the random variable $X[\mathrm{mod}\, 2\pi]$.

Before explaining the solution, several comments about circular random variables should be made. Many examples of circular probability densities are obtained by *wrapping* a probability density defined on the line around the circumference of a circle of unit radius (or similarly one may say that a continuous random variable on the line is wrapped around the circumference). In this case, if Z is a continuous random variable on the line and X is the corresponding wrapped random variable, then

$$X = Z[\mathrm{mod}\, 2\pi]\,, \quad f^X(x) = \sum_{k=-\infty}^{\infty} f^Z(x + 2\pi k). \tag{3.5.1}$$

While the notion of a wrapped density is intuitively clear, the formulae are not simple. For instance, a wrapped normal $N(\mu, \sigma^2)$ random variable has the circular density (obtained after some nontrivial simplifications)

$$f^X(x) = (2\pi)^{-1}\Big(1 + 2\sum_{k=1}^{\infty} e^{-k^2\sigma^2/2}\cos(k(x-\mu))\Big).$$

Fortunately, these complications with wrapped densities are not crucial for an orthogonal series estimation whenever a correct basis is chosen. For this setting a complex trigonometric basis, discussed in Section 2.4, see (2.4.22)–(2.4.23), is the most convenient. Indeed, a partial Fourier sum may be written as

$$f_J^X(x) := (2\pi)^{-1} \sum_{j=-J}^{J} h^X(j) e^{-ijx}, \tag{3.5.2}$$

where

$$h^X(j) := \int_0^{2\pi} f^X(x) e^{ijx} dx = E\{e^{ijX}\} \tag{3.5.3}$$

is the *characteristic function* of the random variable X. Here i is the imaginary unit, that is, $i^2 = -1$.

Note that $|h^X(j)| \le 1$, and the characteristic function is real whenever the random variable is symmetric about 0.

For the case of a wrapped distribution we get, according to (3.5.1),

$$h^X(j) = \int_0^{2\pi} \sum_{j=-\infty}^{\infty} f^Z(x + 2\pi k) e^{ijx} dx = h^{Z[\mathrm{mod}\, 2\pi]}(j) = h^Z(j). \tag{3.5.4}$$

While formula (3.5.1) for a wrapped density is not very convenient, formula (3.5.4) for the corresponding characteristic function is simple. Moreover, if we let X and ε be independent, then

$$h^{X+\varepsilon}(j) = E\{e^{ij(X+\varepsilon)}\} = h^X(j) h^\varepsilon(j). \tag{3.5.5}$$

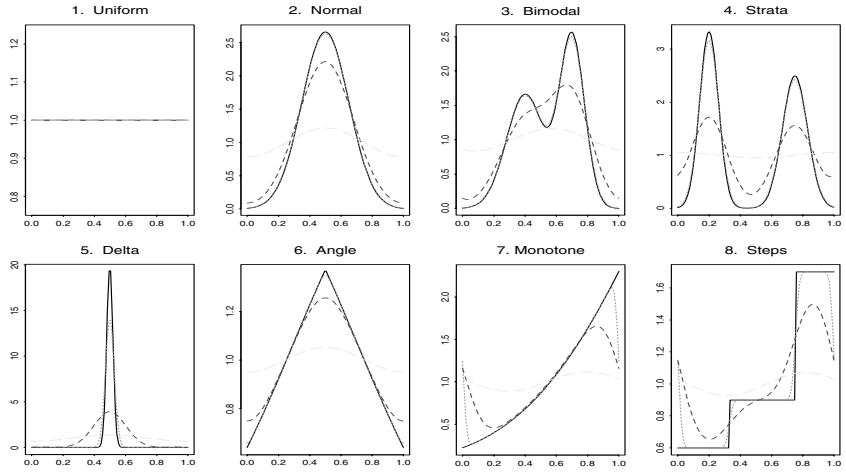

FIGURE 3.15. Densities of $(X + \varepsilon)$[mod 1], where X is distributed according to the corner densities (shown by solid lines) and ε is $N(0, \sigma^2)$. The dotted, short-dashed, and long-dashed lines correspond to σ equal to 0.02, 0.1, and 0.3, respectively. {σ is the Greek letter "sigma," so the corresponding argument is named *sigma*.} [set.sigma=c(.02,.1,.3)]

Thus if the sum $Y = X + \varepsilon$ is observed, then $h^X(j) = h^Y(j)/h^\varepsilon(j)$. It is assumed that the characteristic function $h^\varepsilon(j)$ of ε does not vanish, and recall that the last formula is valid for both circular and linear random variables. This together with (3.5.2) explains the underlying idea of an estimate. Namely, the characteristic function of the observed random variable Y is estimated, and then it is divided by the characteristic function of the measurement error ε. This gives us an estimate of $h^X(j)$, which may be used in (3.5.2).

Equation (3.5.5) also explains why the problem is said to be *ill-posed*. The reason is that if $h^\varepsilon(j)$ is small, then even a large change in $h^X(j)$ leads to a relatively small change in $h^Y(j)$. Because only the $h^Y(j)$ may be estimated, this makes the problem ill-posed.

To "visualize" the ill-posedness, Figure 3.15 shows densities of $Y = (X + \varepsilon)$[mod1] for X distributed according to the corner densities (shown by solid lines) and ε being normal $N(0, \sigma^2)$ with σ equal to 0.02, 0.1, and 0.3. Below, we refer to the density of Y as a convolved density; the reason for this adjective is explained in Appendix A; see (A.18). The normal error with the smallest standard deviation, 0.02 (see the dotted lines), makes a difference only in the visualization of the convolved Monotone and the convolved Steps because they become periodic. Also, the sharp angle in the Angle becomes a smooth one, the Delta resembles a typical normal density, and the sharp steps in the convolved density Steps disappear. In short, such a small measurement error makes a circular random variable continuous,

and while it is easy to recognize the original densities, the problem of testing for the presence of jumps or sharp angles becomes extremely complicated.

As we see, the situation dramatically changes for the case of normal errors with larger standard deviations. Here the recognition of the underlying densities becomes a "puzzle."

It is of a special interest to look at the case of Figure 3.15.1. We see that the convolved densities are always uniform. This is because the characteristic function of the Uniform density is $h^U(0) = 1$ and $h^U(j) = 0$ for any integer $j \neq 0$. As a result, if a measurement error is uniformly distributed, then no recovery of a convolved density is possible. Such an extreme setting is called *irregular*. Note that we avoid such a setting due to the assumption that $h^\varepsilon(j)$ does not vanish at integers j.

Now we are in a position to explain the problem of estimation (recovery) of the density of X based on observations of the sum $Y = (X + \varepsilon)[\mathrm{mod}\, 2\pi]$. The observations of Y allow us to estimate the characteristic function $h^Y(j)$ by the *empirical characteristic function*

$$\hat{h}^Y(j) := n^{-1} \sum_{l=1}^{n} e^{ijY_l}. \tag{3.5.6}$$

Then the natural estimate of $h^X(j)$ is $\hat{h}^X(j) := \hat{h}^Y(j)/h^\varepsilon(j)$. Consider the mean squared error of estimation of $h^X(j)$ by $\hat{h}^X(j)$,

$$\begin{aligned} E\{|\hat{h}^X(j) - h^X(j)|^2\} &= E\{|\hat{h}^Y(j) - h^Y(j)|^2\}/|h^\varepsilon(j)|^2 \\ &= n^{-1}(1 - |h^Y(j)|^2)/|h^\varepsilon(j)|^2. \end{aligned} \tag{3.5.7}$$

As an example, if ε is normal $N(0, \sigma^2)$, then $h^\varepsilon(j) = e^{-j^2\sigma^2/2}$, that is, the decrease in $h^\varepsilon(j)$ is extremely fast. The asymptotic theory shows that this implies an extremely slow logarithmic convergence of MISE (after all, this is an ill-posed problem). A normal error is the worst-case scenario (for instance, Cauchy error is better and Gamma error is dramatically better), but it is also the most typical measurement error.

After all these "scary" asymptotic scenarios, it is necessary to explain why we have a chance to get a reasonable estimation (in comparison with no-error case) for small sample sizes. According to Figure 3.7, typical cutoffs are not large. Thus, if $h^\varepsilon(j)$ is not too small for the first values of j (and typically this is the case), then a reasonable recovery, which is comparable with the case of direct observations, is possible. In short, for small sample sizes we see only the onset of ill-posedness. Of course, for moderate and large samples no fair competition between the cases of direct and indirect data is possible. Also, there is no way to nicely restore a density like the Delta, where too many high-frequency components should be estimated for a decent recovery.

Finally, prior to writing down a recommended data-driven estimate, consider the practically important case of an unknown characteristic function $h^\varepsilon(j)$. In this case measurement errors are to be studied via this or that

sampling procedure, and there is no way to "bypass" this. As an example, let us assume that m realizations ε'_l, $l = 1, 2, \ldots, m$, of the measurement error ε are given. (Another possible sampling procedure is discussed in Exercise 3.5.9.) Then the empirical characteristic function

$$\hat{h}^{\varepsilon}(j) := m^{-1} \sum_{l=1}^{m} e^{ij\varepsilon'_l} \tag{3.5.8}$$

may be used in place of an unknown characteristic function h^{ε}. The rule of thumb is that $m = n$ is sufficient, and in many practically interesting settings m may even be less than n.

Let us now define a data-driven estimate. Set b_n to be the integer part of $[c_b \ln(\ln(n+20))]^{-1}$, and J'_n to be the rounded-up $d_0 + d_1[\ln(n+20)]^{1/(d_2 b_n)}$ with the default parameters $c_b = 8$, $d_0 = 2$, $d_1 = 0.5$, and $d_2 = 10$. The recommended data-driven estimator is

$$\begin{aligned} \tilde{f}_n(x) := (2\pi)^{-1} \sum_{j=-J'_n}^{J'_n} (1 - |\hat{h}^Y(j)|^{-2} n^{-1})_+ \\ \times (\hat{h}^Y(j)/h^{\epsilon}(j)) I_{\{|h^{\epsilon}(j)| > c_H n^{-1/2+b_n}\}} e^{-ijx}. \end{aligned} \tag{3.5.9}$$

Recall that $(z)_+ := \max(z, 0)$ is the positive part of z, and the default value of c_H is 1. To make the estimate bona fide, the nonnegative projection $\hat{f}_n(x) = (\tilde{f}_n(x) - c)_+$ is used, where the nonnegative constant c is such that $\hat{f}_n(x)$ is integrated over $[0, 2\pi]$ to unity. Finally, the procedure of removing small bumps is used.

Particular estimates for normal $N(0, \sigma^2)$, $\sigma = 0.1$, measurement error and sample sizes $n = 50$ and $n = 1000$ are shown in Figure 3.16, where we again use the corner densities to generate X, which is then wrapped around the circumference of a circle of unit length. Thus, all the underlying densities are 1-periodic. Please, do not read the caption. Can you guess which line corresponds to which sample size? This would be a trivial question for the case of direct data, but not so here because MISE has only a logarithmic rate of convergence. Probably, the cases of the Uniform, the Angle, the Monotone, and the Steps would vote for the dotted line being the estimate based on 1000 observations, while the Bimodal would vote for the dashed line being the estimate based on 1000 observations. Curiously, for the Normal, the Strata, and the Delta the twentyfold increase in the sample size has no feasible effect on the estimates. This is what makes this problem so specific and complicated. On the other hand, we see that for $n = 50$ (the dotted lines) the estimates are not bad in comparison with the estimates based on direct data and shown in Figures 3.2–3.

Of course, we have seen the worst-case scenario of a normal measurement error. A majority of other errors imply much better recovery because their characteristic functions decrease much more slowly.

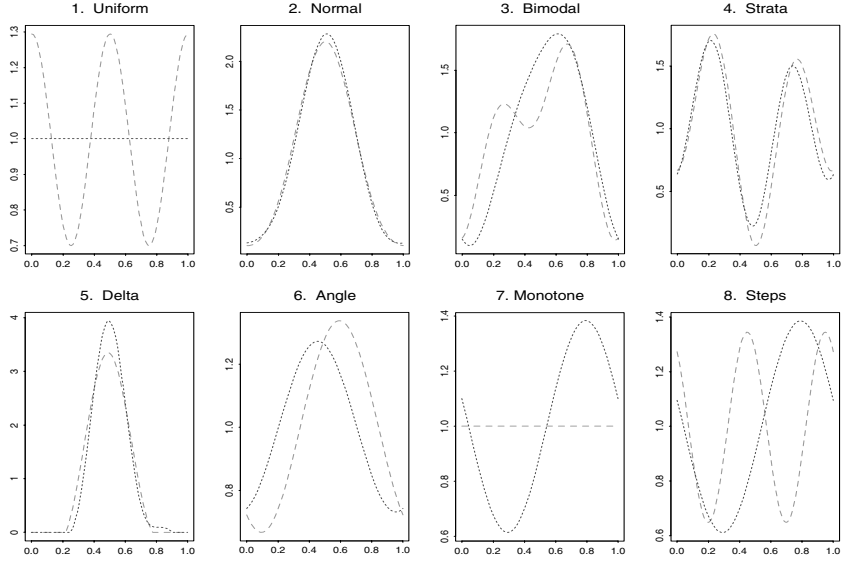

FIGURE 3.16. Estimates of the underlying corner densities (which are not shown) for the case of a normal $N(0, \sigma^2)$, $\sigma = 0.1$, measurement error and sample sizes 50 (dotted lines) and 1000 (dashed lines). The data are circular with the range [0,1]. {Subscripts cannot be used in S-functions, so c_b is denoted by *cb*, d_0 by *d*0, etc.} [*set.n=c(50,1000), sigma=.1, cb=8, d0=2, d1=.5, d2=10, cH=1, cB=2*]

Finally, let us consider the case where X is not circular and is supported on an interval $[0, T]$. Write,

$$f(x) = T^{-1} + (2/T) \sum_{j=1}^{\infty} \text{Re}\{h^X(j\pi/T)\} \cos(j\pi x/T), \tag{3.5.10}$$

where $\text{Re}\{z\}$ is the real part of a complex number z. Then the direct analogue of the estimate (3.5.9) is the estimate

$$\tilde{f}_n(x, h^\varepsilon) := T^{-1} + (2/T) \sum_{j=1}^{2J_n'} (1 - |\hat{h}^Y(j)|^{-2}/n) \text{Re}\{\hat{h}^Y(j\pi/T)/h^\varepsilon(j\pi/T)\}$$

$$\times I_{\{|h^\varepsilon(j\pi/T)| > c_H n^{-1/2+b_n}\}} \cos(j\pi x/T). \tag{3.5.11}$$

The estimator becomes simpler when errors are symmetric about zero. In this case $h^\varepsilon(j)$ is real, and we may write

$$\tilde{f}_n(x, h^\varepsilon) = T^{-1} + (2/T) \sum_{j=1}^{2J_n'} (1 - \hat{\theta}_j^{-2} n^{-1})_+$$

$$\times [\hat{\theta}_j / h^\varepsilon(j\pi/T)] I_{\{|h^\varepsilon(j\pi/T)| > c_H n^{-1/2+b_n})\}} \cos(j\pi x/T), \tag{3.5.12}$$

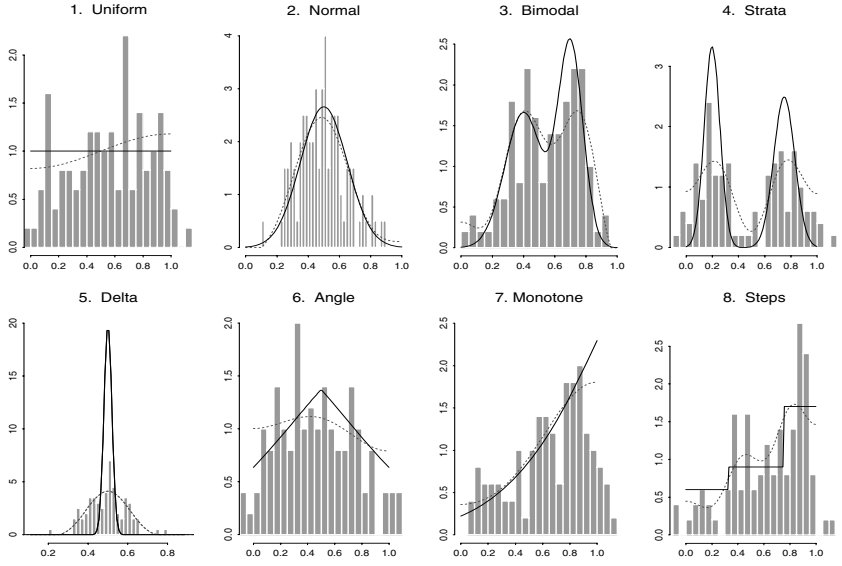

FIGURE 3.17. Histograms of $n = 100$ realizations of $Y := (X + \varepsilon)[\mathrm{mod}\, 1]$ overlaid by underlying densities (solid lines) and estimates (3.5.12) of the density f^X shown by dotted lines. The measurement error ε is normal $N(0, \sigma^2)$. [*n=100, sigma=.1, cb=8, d0=2, d1=.5, d2=10, cH=1, cB=2*]

where

$$\hat{\theta}_j = n^{-1} \sum_{l=1}^{n} \cos(j\pi Y_l/T)\,. \tag{3.5.13}$$

As usual, the bona fide procedure (3.1.15) is the final step.

Figure 3.17 is a convenient tool to understand both the problem and the suggested solution. Here $T = 1$, and let us look, as an example, at the Delta diagram. Does the histogram exhibit the underlying Delta density? Is it symmetric? Do you see that the range of the data is about 0.6, that is, much larger than the tiny support of the Delta? All these questions show the complexity of this setting and allow us to conclude that the estimator performs reasonably well under these circumstances.

3.6 Case Study: Length-Biased Data

This case is another example of indirect observations where an observed random variable Y is supported on $[0, 1]$ and has the probability density

$$f^Y(y) := g(y) f^X(y)/\mu, \tag{3.6.1}$$

where $g(y)$ is a given positive function and f^X is a probability density of interest that is also supported on $[0, 1]$. Thus,

$$\mu = \int_0^1 g(x) f^X(x) dx, \tag{3.6.2}$$

and note that because f^X is unknown, the parameter μ is also unknown.

It is clear that the problem is indirect because one has observations of Y and would like to estimate the density of an unobserved X. But why are the data called length-biased? The simplest way to understand this is to consider an example of the setting.

Suppose that a researcher would like to know the distribution of the ratio of alcohol in the blood of liquor-intoxicated drivers traveling along a particular highway. The data are available from routine police reports on arrested drivers charged with driving under the influence of alcohol (a routine report means that there are no special police operations to reveal all intoxicated drivers). Because a drunker driver has a larger chance of attracting the attention of the police, it is clear that the data are length-biased toward higher ratios of alcohol in the blood. Thus, the researcher should make an appropriate adjustment in a method of estimation of an underlying density of the ratio of alcohol in the blood of all intoxicated drivers.

There are many other similar examples in different sciences where a likelihood for an observation to appear in a sample depends on its value. In many cases a linear $g(x)$ is recommended, but in general the function $g(x)$ should be studied via additional experiments.

Probably the first idea of how to solve the problem is to estimate f^Y and then divide it by g. This is a good idea, but it does not lead to an optimal estimation according to the asymptotic theory. Also, for small sample sizes some problems may arise for a set of points x where $g(x)$ is relatively small.

Thus, let us use our standard series approach and try to estimate the Fourier coefficient $\theta_j = \int_0^1 \varphi_j(x) f^X(x) dx$ via the expectation $E\{\psi(Y)\}$ of some function $\psi(Y)$. Assume for a moment that μ is given. Then the straightforward choice $\psi(y) := \mu\varphi_j(y)/g(y)$ leads to the sample mean estimate based on n iid realizations $Y_1, \ldots, Y_n$ of Y,

$$\hat{\theta}_j := \mu n^{-1} \sum_{l=1}^{n} \varphi_j(Y_l)/g(Y_l). \tag{3.6.3}$$

Let us check that (3.6.3) is an unbiased estimate of θ_j. Write

$$E\{\hat{\theta}_j\} = \mu E\left\{ \frac{\varphi_j(Y)}{g(Y)} \right\} = \mu \int_0^1 \frac{f^Y(y)\varphi_j(y)}{g(y)} dy = \int_0^1 \varphi_j(y) f^X(y) dy = \theta_j.$$

The parameter μ used in (3.6.3) is, of course, unknown. A recommended estimate is (we use the notation $g^{-k}(x) := (1/g(x))^k$)

$$\hat{\mu} := \frac{1}{n^{-1}\sum_{l=1}^{n} g^{-1}(Y_l)}. \tag{3.6.4}$$

The idea of this estimate is that $1/\hat{\mu}$ is an unbiased estimate of $1/\mu$. Indeed,

$$E\{1/\hat{\mu}\} = E\{g^{-1}(Y)\} = \mu^{-1}\int_0^1 g(y)f^X(y)g^{-1}(y)dy = 1/\mu.$$

Finally, to use the data-driven estimate (3.1.14), we need to find an estimate $\hat{d}$ for $nE\{(\hat{\theta}_j - \theta_j)^2\}$. Again, assuming that μ is given, we write for $\hat{\theta}_j$ defined at (3.6.3),

$$nE\{(\hat{\theta}_j - \theta_j)^2\} = E\{(\mu\varphi_j(Y)/g(Y))^2\} - \theta_j^2.$$

Then, using (3.1.7) we get

$$E\{(\mu\varphi_j(Y)/g(Y))^2\} = \mu^2 E\{g^{-2}(Y)\} + \mu^2 2^{-1/2}\int_0^1 f^Y(y)g^{-2}(y)\varphi_{2j}(y)dy.$$

Note that the second term is the $(2j)$th Fourier coefficient of the function $2^{-1/2}\mu^2 f^Y(y)\, g^{-2}(y)$. As we know from Section 2.2, under mild conditions these coefficients vanish for large j. Thus, we may define the coefficient of difficulty due to length-biased data:

$$d := \mu\int_0^1 f^X(y)g^{-1}(y)dy = \mu^2 E\{g^{-2}(Y)\}. \tag{3.6.5}$$

Then a natural estimate of the coefficient of difficulty is

$$\hat{d} := \hat{\mu}^2 n^{-1}\sum_{l=1}^{n} g^{-2}(Y_l). \tag{3.6.6}$$

The suggested estimator is based on the function $g(x)$, which may be unknown. It is impossible to estimate both f and g based only on observations of Y. Thus, an additional experiment should be done. For instance, if f^X is given, then observations of Y could be used to estimate $g(x)$ by the above-recommended estimator. Indeed, $g(x)$ can always be thought of as a density (otherwise divide it by its integral), and then the problem is symmetric with respect to g and f^X.

Estimates for the case of $g(x) = 0.1 + 0.9x$ and Monte Carlo simulated data with the corner functions being the underlying densities are shown in Figure 3.18. As we see, for this particular case the estimates, except of several "bad guys," are relatively good. This indicates that the coefficient of difficulty should be moderate. To tackle this coefficient note that d, defined in (3.6.5), may be written as

$$d = E\{g(X)\}E\{g^{-1}(X)\}. \tag{3.6.7}$$

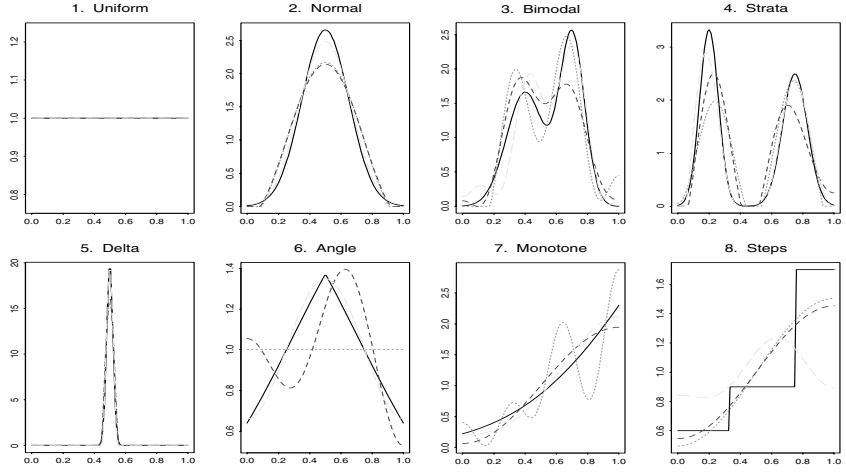

FIGURE 3.18. Estimates for length-biased data with $g(x) = a + bx$, $a = 0.1$ and $b = 0.9$. The dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. The solid lines show the underlying corner densities. [*set.n=c(50,100,200), a=.1, b=.9, cJ0=4, cJ1 =.5, cJM=6, cT=4, cB=2*]

Using Cauchy–Schwarz inequality (A.9) with $Z_1 = \sqrt{g(X)}$ and $Z_2 = 1/Z_1$ we get that $d \geq 1$, with equality for the case of direct data. For the given g and the corner densities, the coefficients of difficulty, rounded to the first digit after the decimal point, are 1.4, 1.1, 1.1, 1.3, 1.0, 1.3, 1.2, 1.4. Recall that the support is $[0, 1]$, so as in Section 3.4 the coefficient of difficulty is equal to the relative coefficient of difficulty. Thus, the coefficient of difficulty shows the increase in a sample size that allows us to get about the same precision of estimation as for the case of direct data. The numbers presented support our preliminary conclusion that this setting is not much more complicated than the case of direct observations. On the other hand, the cases of the Uniform and the Steps may sometimes present a surprise. Such a "surprise" is clearly seen in the Steps diagram.

Finally, it is worthwhile to comment on how the Monte Carlo simulations of Y with the probability density (3.6.1) have been made. Here the two-steps *acceptance-rejection* method has been used (see Exercise 3.6.5).

Step 1. Simulate X according to f^X and independently simulate a uniform $U(0, 1)$ random variable U.

Step 2. Find a (preferably minimal) constant $c \geq 1$ such that $g(x)/\mu \leq c$ for all x. If $U \leq g(X)/c\mu$, then set $Y := X$, otherwise return to Step 1.

3.7 Case Study: Incorporating Special Features

It is quite common that some qualitative characteristics of a curve are known a priori or are being sought. For instance, it may be known a priori that an underlying density is monotone; in this case an estimate should be monotone as well. Another example is the case where it is important to know the shape of a curve near boundaries; since the cosine estimates are always flat near boundaries, we should address such an issue. In all these cases it is desirable to have an estimator that incorporates these features.

Let us begin with the case of estimation of monotone densities, to be specific, nondecreasing ones. Clearly, for the case of monotone densities the series estimate (3.1.15) may be not bona fide, see the examples in Figures 3.2.1, 3.2.7, and 3.2.8.

There are two possible approaches to solve this problem: Either suggest a new special procedure for estimation of monotone densities, or use any estimator and then find its projection onto a class of monotone curves. The asymptotic theory (Section 7.7) tells us that the second approach leads to optimal estimation for large sample sizes. Thus, if we believe that an estimator is good, then we may use it for estimation of any density, and then, if it is given that an underlying density is monotone, use a monotonic projection. Such an approach is convenient and robust because the original estimate may be visualized together with its monotonic projection.

A *monotonic projection* is extremely simple. Let $\hat{f}_1, \dots, \hat{f}_m$ be the values of a nonmonotone estimate at points $x_1 < x_2 < \cdots < x_m$. Then:

(i) Start with the pair $(\hat{f}_1, \hat{f}_2)$ and find the first pair $(\hat{f}_j, \hat{f}_{j+1})$ such that $\hat{f}_j > \hat{f}_{j+1}$, i.e., the first pair from the left where monotonicity fails.

(ii) Replace both $\hat{f}_j$ and $\hat{f}_{j+1}$ by their average value, i.e., by $(\hat{f}_j + \hat{f}_{j+1})/2$.

(iii) Beginning with the pair $(\hat{f}_{j-1}, \hat{f}_j)$, check that after the modification (according to step (ii)) of the original estimate, all pairs to the left, that is, the pairs $(\hat{f}_{j-1-s}, \hat{f}_{j-s})$, $s = 0, 1, \dots, j-2$, satisfy the monotonicity requirement $\hat{f}_{j-1-s} \le \hat{f}_{j-s}$. If for some s^* monotonicity fails, then replace all the elements in the triplet $(\hat{f}_{j-1-s^*}, \hat{f}_{j-s^*}, \hat{f}_{j-s^*+1})$ by their average value $(\hat{f}_{j-1-s^*} + \hat{f}_{j-s^*} + \hat{f}_{j-s^*+1})/3$.

(iv) If the modified estimate is monotone, then stop. This is the monotonic projection. If not, then return to step (i).

Let us see how this procedure works using Figure 3.19. The top row shows data-driven series estimates (3.1.15) for simulated data, and the bottom row shows the corresponding monotonic projections. First of all, we see that if an estimate is already monotone, then the monotonic projection simply reproduces it. If an estimate is not monotone, then the monotonic projection replaces it by a "ladder-like" function. The short-dashed lines in the Monotone and Steps diagrams clearly exhibit how the monotonic projection works. The case of the long-dashed lines in the Steps diagrams exhibits a "miracle" of the monotonic projection when an "ugly" oscillatory

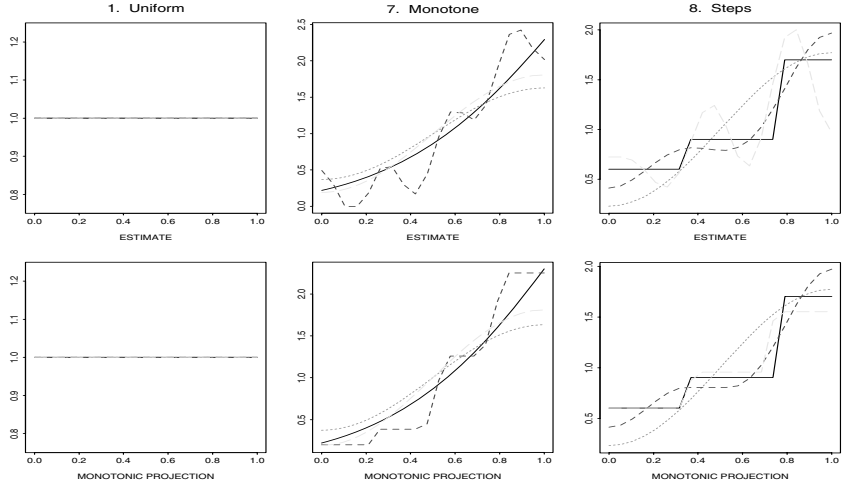

FIGURE 3.19. Data-driven estimates and their monotonic projections. The solid lines are the underlying densities. The dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$. The number of knots used is reduced to make the calculation of the projection faster. This implies rougher lines and slopes in the stepwise curves. [*set.n=c(50,100,200), cJ0=4, cJ1 =.5, cJM=6, cT=4, cB=2*]

estimate in the top diagram becomes an almost perfect estimate of the Steps shown in the right bottom diagram.

Now let us explain how to improve the performance of a cosine series estimate near edges. Recall that the main issue is that this estimate always flattens out near edges; more precisely, its derivative is always zero at the boundary points. The problem is to get a correct visualization of an underlying density like the Monotone and at the same time preserve a good estimation for functions like the Normal where no improvement of the cosine estimate is required. In short, any addition that targets a particular corner function (here the Monotone) should not hurt estimation of others.

The key idea of solving this boundary problem was explained in detail in Section 2.6, and it was to enrich the cosine basis by polynomial terms x and x^2 that should take care of derivatives near edges. (Note that if the issue is, say, the second derivative near edges, then the cosine basis should be enriched by x, x^2, and x^3.)

Technically, a cosine-polynomial data-driven estimate is constructed similarly to the cosine data-driven estimate (3.1.13), so it is worthwhile to explain it using the steps that led us to (3.1.13). First, instead of a partial sum (3.1.1) based on the elements $(1, \varphi_1, \ldots, \varphi_J)$ we use a partial sum based on the elements $(1, \varphi_1', \ldots, \varphi_J')$, where these elements are obtained by applying the Gram–Schmidt orthonormalization procedure (2.3.8) to

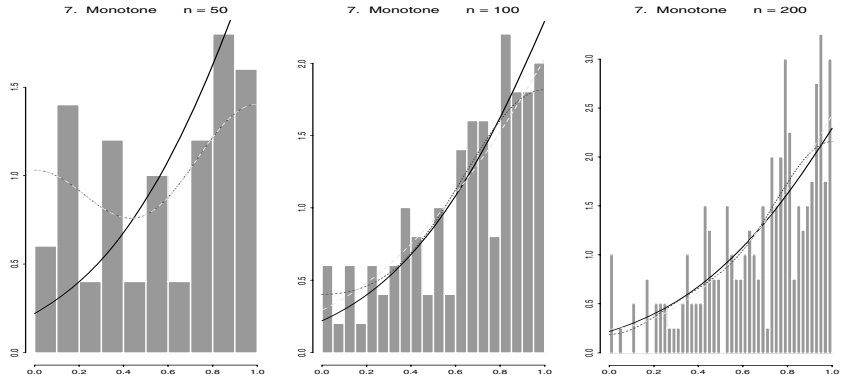

FIGURE 3.20. Histograms overlaid by the underlying density (solid line), cosine estimates (dotted line), and cosine-polynomial estimates (dashed line). The sample sizes are shown in the titles. [*set.n=c(50,100,200), corden=7, cTP=4, cJ0=4, cJ1 = .5, cJM=6, cT=4, cB=2*]

$(1, \varphi_1(x), \ldots, \varphi_{J-2}(x), x, x^2)$. Note that $\varphi'_j = \varphi_j$ for $j \le J-2$ and only φ'_j for $j = J-1, J$ need to be calculated (Exercise 3.7.2).

Then, using φ'_j in place of φ_j, we obtain the estimate $\bar{f}$ defined at (3.1.13). In short, this is the optimal smoothed estimate that always uses the two polynomial terms. Then, the nonnegative projection is used.

Now we make the last step, which is to use or not to use the cosine-polynomial estimate. The issue is that we would like to use it only if the Fourier coefficients for the polynomial terms are statistically significant. To do this, we use the same approach as in the decision to include or not to include high-frequency terms in (3.1.14). Namely, if $\hat{\theta}_j^2 < c_{TP} \ln(n)/n$, $j = J-1, J$, then the cosine estimate (3.1.15) is used; otherwise, the cosine-polynomial estimate is used. The default value for c_{TP} is 4. As you see, the approach is extremely cautious toward including the polynomial terms.

Figure 3.20 illustrates the performance of the estimator obtained. Here we show the underlying Monotone density, estimates obtained by the estimator of Section 3.1 based on the cosine basis, the estimates based on the cosine-polynomial basis, and the histograms.

Let us discuss the particular outcomes shown in Figure 3.20. First of all, consider the case of 50 observations. Here the estimates coincide, i.e., no polynomial terms are used. This is well justified because nothing in the underlying histogram indicates that there is a need to improve the tails of the cosine estimate. Note that while this particular estimate is very poor, it correctly describes the data set at hand, which reveals no monotonicity in the underlying density. For the case of 100 observations the cosine-polynomial estimate significantly improves the visually aesthetic appeal. It also correctly shows the dynamic of the Monotone near edges. The outcome for the case of 200 observations is even more impressive.

3.8 Special Topic: Goodness-of-Fit Tests

This is, finally, the section where we will discuss the question raised in Section 1.1 about fairness of the drawings in the New Jersey Pick-It lottery. Recall that in that section we tried to answer this question via analyzing several Monte Carlo simulations. This approach is not scientific but very convincing, so it is worthwhile to have it in the "toolkit." (A review of the last part of Appendix A devoted to parametric hypothesis testing is recommended; below we use notions and notations introduced there.)

The problem may be stated as a hypothesis testing where the null hypothesis H_0: the underlying distribution for winning numbers is uniform on [000, 999] (i.e., the lottery is fair) is tested against the alternative hypothesis H_a: the underlying distribution for winning numbers is not uniform (the lottery is not fair). A corresponding test is called a *goodness-of-fit* test. We shall consider several goodness-of-fit tests (be prepared that they may give sharply different conclusions, similar to the opinions of expert witnesses called by prosecutors and defense lawyers in a trial). Also, we shall see that there is no loss in generality in considering a uniform distribution as the null hypothesis.

a. Tests based implicitly on the empirical cumulative distribution function. These are "oldy but still goody" classical goodness-of-fit tests when one might be interested in testing H_0: $F = F_0$ versus H_a : $F \neq F_0$. Here F is an underlying cdf (cumulative distribution function) of n iid observations $X_1, \ldots, X_n$ generated according to this cdf. In what follows we shall always assume that F is continuous (this is, of course, the case where the probability density exists).

It has been explained in Appendix A that empirical cdf $\bar{F}_n$, defined at (A.32), is a good estimate for an underlying cdf. Exercise 3.8.2 summarizes some basic properties of the empirical cdf. Then, consider a distance (not necessarily metric) $D(\bar{F}_n, F_0)$ between the empirical cdf and the cdf F_0 of the null hypothesis. We may expect that the distance will be small under the null hypothesis and large under the alternative hypothesis. Thus, the rejection region, based on this distance, should be

$$R := \{(X_1, \ldots, X_n) : \ D(\bar{F}_n, F_0) > c_\alpha \delta_n \}, \tag{3.8.1}$$

where the decaying sequence δ_n and the constant (for a fixed α) c_α are such that the probability of the rejection region given the null hypothesis (i.e., the first-type error) is equal to the level of significance α.

The main issue is to choose a distance that should satisfy the following two requirements: (i) c_α and δ_n are easily computed (at least asymptotically); (ii) the test is consistent, that is, given an alternative distribution, its power (the probability of the rejection region given this alternative hypothesis) tends to 1. Recall that the power is 1 minus the second-type error; thus the consistency implies that the second-type error asymptotically vanishes.

Now let us consider the main particular test discussed in this subsection.

a1. Kolmogorov test. Consider the Kolmogorov–Smirnov distance introduced in Appendix A (see the paragraph above the line A.33),

$$D_K(\bar{F}_n, F) := \sup_{x\in(-\infty,\infty)} |\bar{F}_n(x) - F(x)|. \tag{3.8.2}$$

The goodness of fit test, based on this distance, is called the *Kolmogorov test*. The following proposition,

$$\lim_{n\to\infty} P(D_K(\bar{F}_n, F_0) > cn^{-1/2} | F = F_0) = K(c) := 2\sum_{l=1}^{\infty}(-1)^{l+1}e^{-2l^2c^2}, \tag{3.8.3}$$

proved by Kolmogorov, allows us (by a rule of thumb for $n > 30$) to set $\delta_n = n^{-1/2}$ and choose c_α as the solution to the equation $K(c) = \alpha$.

The Kolmogorov test may be easily inverted to get a corresponding confidence band for an underlying cdf with the confidence coefficient $1-\alpha$,

$$\{F: \ D_K(\bar{F}_n, F) \le c_\alpha n^{-1/2}\}. \tag{3.8.4}$$

Let us apply this method for the **lottery.number** data set discussed in Section 1.1. Because there are 254 observations, we apparently may use the asymptotic formula (3.8.3) for computing the p-value. Recall that p-value is the smallest level of significance for which the null hypothesis would be rejected given the observed data.

Figure 3.21 illustrates both the Kolmogorov test and confidence band for this data set. The solid line is the empirical cdf, which is a step function, and the dashed line is the uniform cdf (the null hypothesis). The Kolmogorov test looks after the largest distance between these lines, and the location of the largest distance $\hat{D}_K$ is highlighted by the dashed-dotted vertical line. For this particular case, $\hat{D}_K = 0.08$ (it is not shown in this figure). Two dotted lines, parallel to the empirical cdf, show the Kolmogorov confidence band with the confidence coefficient $1-\alpha$.

For solving the hypothesis testing problem, Figure 3.21 reports the p-value for the Kolmogorov test. This p-value is denoted by p-valK and shown in the title. Its calculation is based on the formula

$$\text{p-valK} := K(\hat{D}_K n^{1/2}), \tag{3.8.5}$$

where $\hat{D}_K$ is the observed Kolmogorov–Smirnov distance (statistic). The calculated p-value is 0.11. Thus, if the level of significance $\alpha \ge 0.11$, then the Kolmogorov test rejects the null hypothesis with the judgment that the lottery is unfair, and conversely for the case $\alpha < 0.11$.

a2. Several related tests. There are many alternatives to the Kolmogorov test where different distances (not necessarily metrics) are used. Here we mention several of them just to give a flavor of this approach.

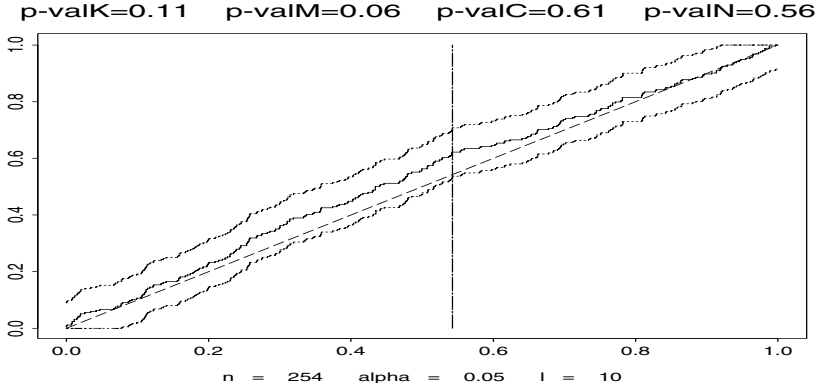

FIGURE 3.21. The Kolmogorov goodness-of-fit test and the corresponding confidence interval. The data set is **lottery. number** rescaled onto $[0, 1]$. Recall that this is the set of winning numbers for the New Jersey Pick-It lottery from May 1975 to March 1976 discussed in Section 1.1. The solid line shows the empirical cdf; the dashed line shows the cdf of the Uniform distribution, which is the null hypothesis. The dashed-dotted vertical line shows the location of the largest distance $\hat{D}_K$ (the value of the Kolmogorov–Smirnov distance) between these two functions (between the solid and dashed lines). The title shows the calculated p-value of the Kolmogorov test, denoted by p-valK. The two dotted lines show the Kolmogorov confidence band with the confidence coefficient $1 - \alpha$, where the value $\alpha = 0.05$ is shown in the subtitle. This figure also shows in the title p-values for the Moran, chi-squared, and nonparametric series tests denoted by p-valM, p-valC, and p-valN, respectively. {Recall that α is the Greek letter *alpha*, so we denote the argument α by *alpha*. Any *alpha* from the set $\{.01, .02, .05, .1, .15, .2, .25\}$ may be chosen. The choice of a data set is controlled by the argument *DATA*. The sample size n of a data set *DATA* is shown in the subtitle. This figure allows one to test any data set available in the S–PLUS environment. Remark 3.8.1 explains how to use this Figure for the case of an arbitrary (not necessarily Uniform) null distribution. The number l of bins for the chi-squared test is given in the subtitle, and it is controlled by the argument l. The arguments *m, cJ0*, and *cJ1* control the parameters of the nonparametric test.} [*DATA = lottery.number, alpha = .05, l=10, m=100, cJ0=4, cJ1=.5*]

Smirnov test. This is probably the closest one to the Kolmogorov test. It is based on the *one-sided* Kolmogorov–Smirnov distance,

$$D^+(\bar{F}_n, F) := \sup_{x \in (-\infty, \infty)} [\bar{F}_n(x) - F(x)] \,. \tag{3.8.6}$$

It was established by Smirnov that

$$\lim_{n \to \infty} P(D^+(\bar{F}_n, F_0) > cn^{-1/2} | F = F_0) = e^{-2c^2}, \tag{3.8.7}$$

which makes the calculation of rejection regions and p-values elementary.

ω^2 (von Mises–Smirnov) test. This is the test based on the normed integrated squared error of the empirical cdf,

$$D_{\omega^2}(\bar{F}_n, F) := n \int_{-\infty}^{\infty} (\bar{F}_n(x) - F(x))^2 dx. \tag{3.8.8}$$

There exists a closed formula that allows one to calculate c_α, but it is too complicated to present here.

Before considering several tests based on different ideas, let us recall the following classical result of probability theory.

Let X be distributed according to a continuous cdf $F(x)$. Define a new random variable $Y := F(X)$. It is clear that this new random variable is supported on the unit interval $[0, 1]$, and it is also not difficult to show that Y is uniformly distributed on $[0, 1]$. Indeed, denote by $F^{[-1]}(y)$ (read "F-inverse") the inverse of F. Note that the inverse is unique because $F(x)$ is continuous in x, and write for any $0 \le y \le 1$,

$$P(Y \le y) = P(F(X) \le y) = P(X \le F^{[-1]}(y)) = y. \tag{3.8.9}$$

This implies the uniform distribution of $F(X)$.

Let us formulate this assertion as a remark.

Remark 3.8.1. Let X be a random variable with a continuous cdf $F(x)$. Then the random variable $F(X)$ is uniformly distributed on $[0, 1]$. Thus, by considering statistics $F_0(X_1), \dots, F_0(X_n)$ in place of the original observations $X_1, \dots, X_n$, one may convert testing the null hypothesis H_0: $F = F_0$ into testing the null hypothesis H_0: F is uniform on $[0, 1]$. In particular, this method allows us to use Figure 3.21 for the case of any continuous distribution F_0.

Now we are in a position to consider several more nonparametric tests.

b. Moran test. This is a very simple and intuitively clear goodness-of-fit test. According to Remark 3.8.1, it is sufficient to consider the case where the null hypothesis states that the underlying distribution is uniform on $[0, 1]$. Then the Greenwood–Moran test statistic is defined as the sum of squared spacings,

$$\hat{M}_n := \sum_{l=0}^{n} (X_{(l+1)} - X_{(l)})^2. \tag{3.8.10}$$

Here $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are ordered observations, $X_{(0)} := 0$ and $X_{(n+1)} := 1$.

The underlying idea of this test statistic is as follows. It is not difficult to show (Exercise 3.8.4) that the sum $\sum_{l=1}^{n} y_l^2$ takes on its minimum, given $\sum_{l=1}^{n} y_l = 1$, at $y_1 = y_2 = \cdots = y_n = 1/n$. Thus, the Greenwood–Moran test statistic is minimal for equidistant observations. This implies that the rejection region should correspond to large values of this test statistic, which indicate irregular spacings.

There is a very simple asymptotic rule for finding the α-level test due to the formula

$$\lim_{n\to\infty} P(n^{1/2}(n\hat{M}_n/2-1)\geq c)=1-\Phi(c), \tag{3.8.11}$$

where $\Phi(x)$ is the cdf of a standard normal random variable. This implies the following rejection region (use it if the sample size is at least 30):

$$R_M:=\{(X_1,\dots,X_n):\ n^{1/2}(n\hat{M}_n/2-1)\geq z_\alpha\}, \tag{3.8.12}$$

where z_α is defined as the solution to the equation $\Phi(z)=1-\alpha$.

Let us use the Moran test for the lottery data. The p-value, denoted by p-valM and calculated by the formula p-valM:$=1-\Phi(n^{1/2}(n\hat{M}_n/2-1))$, is shown in the title of Figure 3.21. It is 0.06, and thus, according to the Moran test, the null hypothesis about fairness of the lottery is accepted only for levels of significance that are less than 0.06.

We are now in a position to consider a goodness-of-fit test that is a clear favorite among data analysts. It is called the chi-squared test because the limit distribution of the test statistic is chi-squared.

c. Chi-squared test. There are many practical problems where hypotheses should be tested based on grouped data. Recall the example in Section 1.1 where testing the fairness of the lottery was based on visualizing the histogram in Figure 1.1(b), which is an example of a grouped data set, with the experts' conclusion "...The histogram looks fairly flat—no need to inform a grand jury...." Here we would like to discuss how one can make such a conclusion based solely on the analysis of a histogram, and we begin with this lottery histogram.

Assume that the null hypothesis H_0 is that the underlying distribution for the winning numbers is uniform on $[000, 999]$. Then, ignoring the left bin in the default histogram in Figure 1.1(b), which corresponds to only one winning number 000, we get ten bins (cells). Thus, theoretically, under the null hypothesis a winning number belongs to each bin with the same probability $\frac{1}{10}$. Let us denote, just for generality, by p_k the probability that a winning number belongs to the kth bin, and by l the number of bins (for our particular case $p_k=\frac{1}{10}$ and $l=10$ because we ignore the left bin with only one number 000). Then, on average, the kth bin should contain np_k numbers (Exercise 3.8.5).

Denote by X_k the observed number of the winning numbers from the kth bin. Then the chi-squared test statistic is

$$\hat{\chi}^2:=\sum_{k=1}^{l}(X_k-np_k)^2/(np_k). \tag{3.8.13}$$

(χ is the Greek letter "chi," so χ^2 should be read as "chi squared").

This test statistic is absolutely natural because large deviations of observed X_k from the expected np_k express lack of fit of the observed data to the null hypothesis. Note that because $X_1+X_2+\cdots+X_l=n$, there

are only $l-1$ independent observations, or in other words, $l-1$ degrees of freedom.

The rejection region of the chi-squared test is defined as

$$R_{\chi^2} := \{(X_1, \ldots, X_l) : \ \hat{\chi}^2 > c_{l-1,\alpha}\}. \tag{3.8.14}$$

Here $c_{l-1,\alpha}$ is the quantity similar to the familiar z_α, only here, instead of a standard normal random variable, a chi-squared random variable χ^2_{l-1} with $l-1$ degrees of freedom is used, and $c_{l-1,\alpha}$ is defined as the solution to the equation

$$P(\chi^2_{l-1} \geq c_{l-1,\alpha}) = \alpha. \tag{3.8.15}$$

There are both special tables and software that allow one to find these values. Note that $\chi^2_l := \sum_{j=1}^{l-1} \xi_j^2$, where ξ_j are iid standard normal. This indicates that the formulated test should be used when all np_k are relatively large; the rule of thumb is $np_k \geq 8$ whenever $\alpha \geq 0.01$. Also, if the number of bins is more than 30, then the distribution of $(1/\sqrt{2l})(\chi^2_l - l)$ may be approximated by a standard normal distribution (see also Exercise 3.8.6). Exercise 3.8.7 discusses a particular rule of thumb for choosing m.

For the Lottery data the p-value for the chi-squared test, denoted by p-valC, is shown in the title of Figure 3.21. This value is 0.61, so the chi-squared test based on the default histogram positively supports the conclusion of the experts who just visualized that histogram.

d. Nonparametric series test. Here we consider a test that is motivated by the following idea of Neyman (1937). Consider a density supported on $[0, 1]$. Parseval's identity implies that the integral of the squared density is equal to 1 if and only if it is uniform. Otherwise, this integral is larger than 1. Thus, the integral $\int_0^1 f^2(x)dx$ may check the null hypothesis that the underlying distribution is uniform. Then the natural test statistic is (here J_n is the same as in (3.1.10))

$$\hat{T}_n = \sum_{j=1}^{J_n} \hat{\theta}_j^2. \tag{3.8.16}$$

Because any data set may be rescaled onto $[0, 1]$, this approach is general. Of course, for small sample sizes it may be a problem to calculate the distribution of this test statistic. Here we bypass this step by using Monte Carlo simulations. This is a "trick" that is worthwhile to discuss on its own merits. The idea is as follows. For practical applications we need to know the p-value. To get it, let us simulate m samples of length n according to the uniform distribution (which is the distribution under the null hypothesis) and then count the number Y of samples whose test statistics (3.8.16) are larger than the observed $\hat{T}_n$. Clearly, Y is distributed according to the Binomial distribution $B(m, p)$, where p is the estimated p-value. Then $\bar{p} = Y/m$ is the natural unbiased estimate of p, and recall that $E\{(\bar{p}-p)^2\} =$

$p(1-p)/m$. The last formula allows one to choose m. Note that this test may be used for very small sample sizes, and its simplicity is appealing.

In Figure 3.21 the p-value, based on $m = 100$ Monte Carlo simulations, is shown as p-valN and it is 0.56. This p-value has a very simple meaning: Among 100 samples simulated according to the uniform distribution, 56 had the value of the test statistic larger than one observed for the lottery numbers.

Thus, as in the outcome of the simulations shown in Figure 1.2, the nonparametric series test strongly supports the fairness of the lottery.

As we have seen from the analysis of the lottery data, conclusions of different tests may differ dramatically: They range from the prudent belief of the chi-squared and nonparametric tests in the fairness of the lottery to a rather suspicious opinion of the Kolmogorov and Moran tests. Thus, it is necessary to be trained in using these tests. Figure 3.22 is a tool to gain such experience. It is based on intensive Monte Carlo study of these tests. Namely, for an underlying corner density and a sample size, 100 simulations are analyzed by these four tests, and as in Figure 3.21, p-values are calculated.

Results of 24 particular experiments are presented by boxplots. A *boxplot* is a way to look at the overall shape of a data set (here 100 p-values). The central box shows the data between "hinges" which are approximately the first and third quartiles of the p-values. Thus, about 50% of the data are

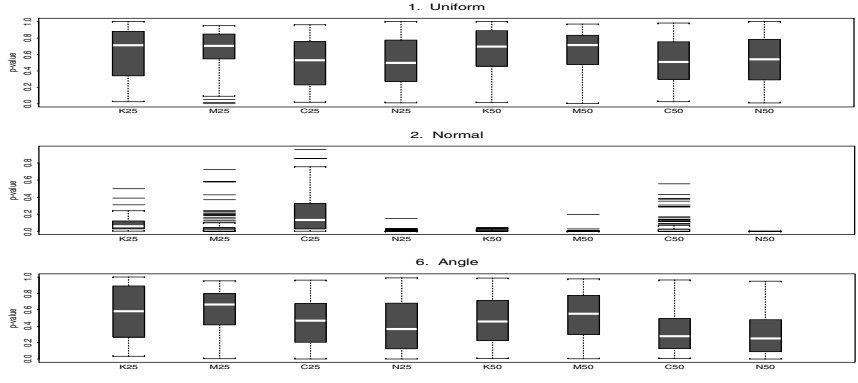

FIGURE 3.22. The boxplots of 100 p-values based on repeated Monte Carlo simulations according to the indicated underlying corner densities and sample sizes. For instance, the boxplot K25 in the top diagram shows the overall shape of 100 p-values obtained by the Kolmogorov test for 100 independent samples of size 25 from the Uniform corner density. {The arguments *set.nn*, *set.l*, *set.cden*, and *reps* allow one to change sample sizes, the corresponding numbers of bins for the chi-squared test, corner densities, and the number of repeated simulations.} [*set.nn=c(25,50), set.l=c(3,5), set.cden=c(1,2,6), reps=100, alpha = .05, m=100, cJ0=4, cJ1=.5*]

located within this box and its height is equal to the interquartile range. The horizontal line in the interior of the box is located at the median of the data, it shows the center of the distribution for the p-values. The *whiskers* (the dotted lines extending from the top and bottom of the box) extend to the extreme values of the data or a distance 1.5 times the interquartile range from the median, whichever is less. Very extreme points are shown by themselves.

The experiment shows that for small sample sizes and the Uniform underlying density the observed p-values may range from very small to almost 1. Here some preference may be given to the Moran test. On the other hand, this test is not the best for the two other densities where the nonparametric test is the best.

3.9 Special Topic: Basis Selection

We discussed the case of two different bases (cosine and cosine–polynomial) in Section 3.7. Because the cosine–polynomial basis is a cosine one with several extra polynomial elements, the selection has been made via analyzing the magnitude of the Fourier coefficients of the extra polynomial terms.

What can be done if two different bases are considered? Let us consider an approach that is motivated by the ideas of Sections 3.1–3.

Suppose that g_j, $j = 0, 1, 2, \ldots$, is a basis, $\tilde{f}$ is a corresponding data-driven estimate (3.1.14), and $\theta_j = \int f(x)g_j(x)dx$ are the Fourier coefficients of an underlying density f. For the sake of simplicity, let us assume that in (3.1.14) the weights $\hat{w}_j$ are either 0 or 1 (the more general case is left as Exercise 3.9.1). Then this estimate tries to match the oracle

$$\tilde{f}^*(x) := \sum_{j=0}^{c_{JM}J_n} w_j\hat{\theta}_j g_j(x), \tag{3.9.1}$$

where $w_j = 1$ if $\theta_j^2 > E\{(\hat{\theta}_j - \theta_j)^2\}$ and $w_j = 0$ otherwise. The MISE of this oracle is

$$\begin{aligned}\text{MISE}(\tilde{f}^*, f) &= \sum_{j=0}^{c_{JM}J_n} E\{(w_j\hat{\theta}_j - \theta_j)^2\} \\ &= \sum_{j=0}^{c_{JM}J_n} w_j E\{(\hat{\theta}_j - \theta_j)^2\} + \Big[\sum_{j=0}^{c_{JM}J_n}(1-w_j)\theta_j^2 + \sum_{j>c_{JM}J_n}\theta_j^2\Big]. \end{aligned}\tag{3.9.2}$$

The term in square brackets is the integrated squared bias. Using Parseval's identity it may be written as

$$\sum_{j=0}^{c_{JM}J_n}(1-w_j)\theta_j^2 + \sum_{j>c_{JM}J_n}\theta_j^2 = \int_0^1 f^2(x)dx - \sum_{j=0}^{c_{JM}J_n} w_j\theta_j^2. \tag{3.9.3}$$

This implies

$$\text{MISE}(\tilde{f}^*, f) = \sum_{j=0}^{c_{JM}J_n} w_j[E\{(\hat{\theta}_j - \theta_j)^2\} - \theta_j^2] + \int_0^1 f^2(x)dx. \qquad (3.9.4)$$

Note that $\int_0^1 f^2(x)dx$ is a constant; thus a basis minimizes the MISE if and only if it minimizes the risk

$$R(\{g_j\}) := \sum_{j=0}^{c_{JM}J_n} w_j[E\{(\hat{\theta}_j - \theta_j)^2 - \theta_j^2]. \qquad (3.9.5)$$

It was discussed in Section 3.1 how to estimate $E\{(\hat{\theta}_j - \theta_j)^2\}$ and θ_j^2. Thus, we may use a plug-in estimate $\tilde{R}(\{g_j\})$ of the risk $R(\{g_j\})$. If several bases are considered, then the smaller estimated risk indicates the better basis. This selection method may be referred to as a method of *empirical risk minimization*.

Let us check this approach for the two half-range trigonometric bases discussed in Section 2.4. Recall that the first one is the cosine one, which has been used in all the sections, and the second one is the sine one, i.e., $\psi_j(x) = 2^{1/2}\sin(\pi jx)$, $j = 1, 2, \ldots$.

The sine basis may be used only for densities that vanish at the boundary points, so let us restrict our attention to the corner densities Normal, Bivariate, Strata, and Delta. Note that while both these bases are trigonometric, they have different specifics. The cosine basis will always give us an estimate integrated to unity because 1 is its first element. The sine basis does not necessarily imply an estimate integrated to unity, but the endpoints of the estimate will be correctly equal to zero.

For both these bases n^{-1} may be used as the estimate of $E\{(\hat{\theta}_j - \theta_j)^2\}$ and thus $\hat{\theta}_j^2 - n^{-1}$ as a natural estimate of $\hat{\theta}_j^2$ (Exercise 3.9.4).

The result of a particular experiment is shown in Figure 3.23. Here the dotted and dashed lines depict the cosine and sine estimates. The choice based on empirical risk minimization is shown in the title of each diagram. Also, in brackets the "correct" choice is shown based on the minimal integrated squared error, i.e., on $\text{ISE} := \int_0^1 (\hat{f}(x) - f(x))^2 dx$.

This figure is interesting from several points of view. Firstly, we may analyze how the universal estimate performs for the two bases. Overall, we see no dramatic differences (repeated simulations show that sometimes estimates have different shapes, but this does not occur often). On the other hand, the differences are clear in the estimates of the Bivariate. Here the interesting situation is that while the estimates are poor, there is nothing in the data that may indicate the correct magnitudes of modes of the underlying Bivariate density. Another curios case is the Strata. Look at the right tails of the estimates. The sine estimate (the dashed line) correctly vanishes because it cannot perform differently, while the cosine estimate (the dotted line) shows the right tail incorrectly. On the other hand, the

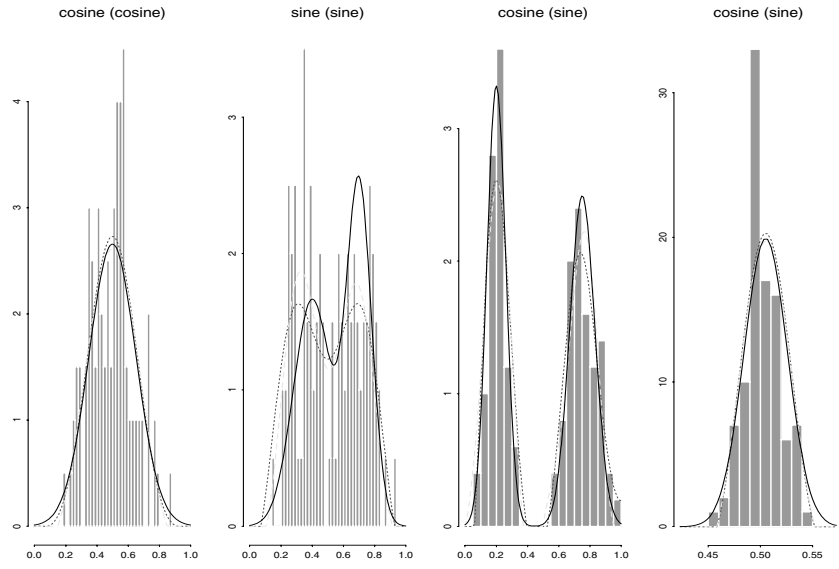

FIGURE 3.23. Basis selection based on 4 sets of $n = 100$ observations depicted by histograms. The compared bases are the half-range cosine and sine ones. The underlying densities (Normal, Bivariate, Strata, and Delta) are shown by solid lines; the cosine estimates are shown by dotted lines, and the sine estimates by dashed lines. The choice based on a smaller empirical risk is shown in the title of each diagram, and the choice based on a smaller integrated squared error (the benchmark) is shown in brackets. [*n=100, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

"incorrect" right tail of the cosine estimate is supported by the particular histogram, while the "correct" right tail of the sine estimate is solely due to the nature of the sine basis.

Secondly, we can understand how the visualization of an underlying density by an estimate is correlated with its ISE. Here the basis shown in brackets is the one that should be chosen if one believes that ISE is the right criterion. For the Normal this choice looks right because the dotted line gives a better visualization. The Bivariate case is a difficult one because none of the estimates are good, but it allows us to understand the meaning of the smaller ISE because here the dashed line has a smaller ISE. For the Strata the dashed line has a smaller ISE, and this is a reasonable outcome. For the Delta, the dashed line has a smaller ISE, and this again agrees with the better visualization of the Delta by the dashed line.

Finally, let us discuss the particular recommendations of the empirical risk minimization method. In 2 cases out of 4 the recommendations were wrong. For the Strata, the wrong recommendation is to use the dotted line. But is it possible by looking at the data, to make a different recommendation, i.e., to recommend the dashed line? It is not an easy task. The Delta

diagram is another interesting example. Here again the conclusion of the empirical risk minimization method is wrong, and the recommended dotted line is worse near the peak than the rejected sine estimate (dashed line). But look at the pronounced peak of the histogram. This peak explains the decision made by the empirical risk minimization procedure because the dotted line fits the data better. On the other hand, both these estimates give us a similar visualization and almost perfect magnitudes of the Delta.

3.10 Practical Seminar

The objective of this seminar is to use the universal estimator (3.1.15) for the analysis of real (not simulated) data sets and to explore the effect of different parameters of this estimator on obtained estimates. Also, our attention will be devoted to estimation of a density over its support. As we know from Section 3.1, this is a rather complicated problem, since the support is typically unknown (of course, there are settings like the lottery winning numbers where the support is known).

As an example, let us consider New York City's rainfall in inches for every year from 1869 to 1957. The rainfall (as observations over the years) was shown in Figure 1.9.

For the rainfall data, Figure 3.24(a) shows the default S–PLUS histogram (recall the discussion in Section 1.1 that this histogram assumes a normal underlying density) and the universal nonparametric estimate calculated over the range $[32.7, 58.7]$ of the observed rainfall. Recall that the caption of Figure 3.2 contains a cumulative review of all the arguments.

The nonparametric estimate shown in the diagram (a) is rather shocking. As in the histogram, it is skewed, and the largest mode is positioned similarly to the "mode" of the histogram, but the second mode and the right tail of the universal estimate are eye-catching. It looks as if something is wrong with the universal estimate.

However, before criticizing the universal estimate, let us try to understand why we so trust in the histogram estimate and do not believe in the universal estimate. Probably, the reason is that a histogram is the most widely used and known tool to present data, and a particular histogram is created by the respected statistical software. However, the issue is that a histogram is just one of many nonparametric tools to visualize the data. Thus, in Figure 3.24(a) we see two different estimates, and neither has a "birthright" superiority.

Let us zoom in on the data by increasing the number of bins from 6 to 40. The corresponding histogram, overlaid by the same estimate, is shown in Figure 3.24(b); it explains the "strange" shape of the universal estimate. We see that the two modes are justified, and moreover, the right tail of the estimate corresponds to the zoomed-in data.

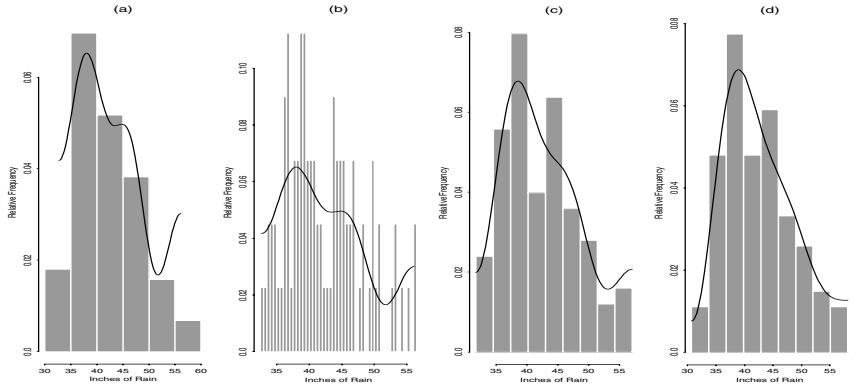

FIGURE 3.24. Effect of a support on histogram and universal estimate. The underlying data set is the New York City's rainfall data. (a) The default histogram overlaid by the universal estimate estimated over the interval $[a, b] := [32.7, 56.1]$, which is the range of the observed rainfall. (b) The histogram with 40 bins overlaid by the same universal estimate as in (a). (c) Both the histogram with 9 bins and the universal estimate are over the interval $[31.7, 57.1]$. (d) Both the histogram with 9 bins and the universal estimate are over the interval $[30.7, 58.1]$. {The supports in diagrams (c) and (d) are $[a - del1, b + del1]$ and $[a - del2, b + del2]$, respectively.} [*del1=1, del2=2, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

Thus, the series estimate has alerted us that the data set is not as simple as it is presented by the default histogram, and it is worthwhile to look at the data more closely. And this is one of the *main aims* of nonparametric curve analysis—to give us a first look at the data and alert us to possible deviations from traditionally assumed parametric distributions.

Let us continue our study of the rainfall data and the exploration of the series estimate. Recall that the only parameter of the data-driven series estimate that has been chosen manually is the interval of support. In the diagrams (a)–(b) we made the assumption that the support was the range of the rainfall. From a probabilistic point of view, it is not realistic to expect that the observed minimal and maximal rainfalls coincide with the possible minimal and maximal rainfalls in New York City (the probabilistic part of the issue is discussed in Exercise A.12 of Appendix A). Thus, just to get a feeling for the issue, let us add 1 inch to the left and right sides of the observed range (this is a rather reasonable increase). Then the corresponding universal estimate together with the corresponding histogram based on 9 bins (and that also covers the increased interval) is shown in Figure 3.24(c). Similarly, Figure 3.24(d) shows the histogram (with 9 bins) overlaid by the universal estimate over the interval with 2 inches added to each side of the range. We see that increasing the support interval dramatically affects the series estimate.

The example also shows that the smoothness of an underlying curve plays a crucial role in nonparametric estimation. "To smooth or not to

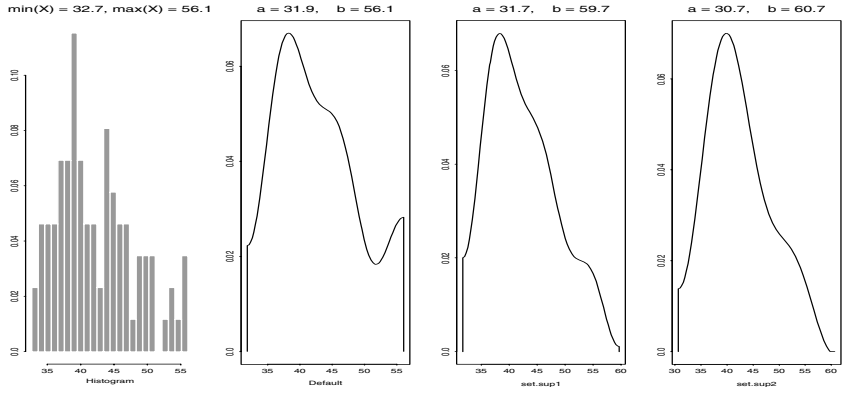

FIGURE 3.25. An analysis of how the interval $[a, b]$, which is the supposed support of an estimated density, affects the universal estimate for the New York City's rainfall data. {A data set is chosen by the argument *DATA*. The left diagram shows a histogram with l bins running from the minimal to maximal observations shown in the title. The next diagram shows the universal estimate with the support $[a, b]$ shown in the title and calculated using (3.1.16). The two right diagrams correspond to the universal estimates with supports controlled by the arguments *set.sup1=c(a,b)* and *set.sup2 = c(a,b).*} [*Data=rain.nyc1, l=25, s=1, set.sup1=c(31.7, 59.7), set.sup2=c(30.7, 60.7), cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

smooth, this is the question" in nonparametric curve analysis. This also explains why professional statisticians working in this area are called the "smoothing community."

Figure 3.25 allows one to choose any S–PLUS compatible data set and explore how a possible support affects the estimate. The left diagram is a histogram with l bins (this number is controlled by the argument l, which allows one to zoom in and out data). The second diagram shows the estimate where (3.1.16) is used to choose the support. Finally, the two right diagrams allow one to look at estimates with manually chosen supports. Note the change in the right tail of the estimate as b increases.

Figure 3.26 allows one to explore how parameters of the universal estimate affect the presentation of a data set at hand. In particular, here the effect of c_{J0} is considered. Also recall that the caption to Figure 3.2 reviews all parameters of the universal estimate.

Figure 3.26 considers the data set **auto.stats**, which is a matrix whose rows are data concerning 72 automobiles and whose columns are different variables. The variables include fuel consumption ("Miles per Gallon"), price ("Price"), dimensions ("Trunk," "Headroom," "Length"), weight ("Weight"), and clearance required to make a U-turn ("Turning Circle"). Here we consider the clearance required to make a U-turn. The left diagram shows a histogram with 25 bins; its title depicts the maximal and minimal observations, and the subtitle shows the sample size. Three other diagrams

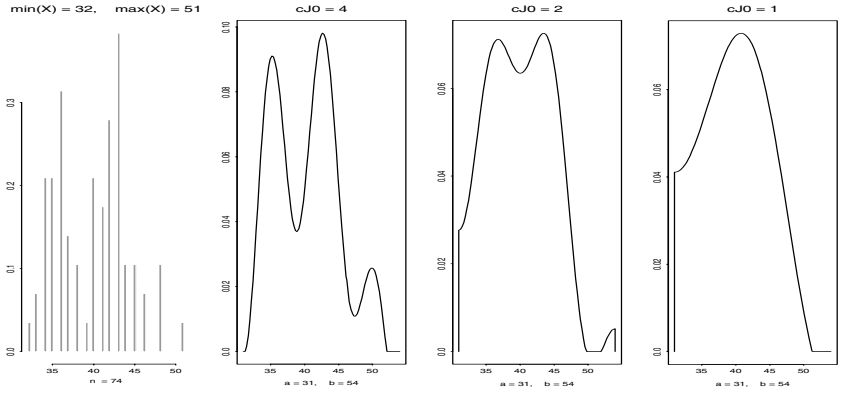

FIGURE 3.26. Effect of a parameter on the universal estimate. Here the parameter c_{J0} is explored that affects the choice of J_n used in the procedure (3.1.10) of calculation of optimal cutoff. The data set is the diameters of turning circles of 74 cars. {A data set is chosen by the argument *DATA*; the default set is *DATA = auto.stats*[, *"Turning Circle"*]. If one wishes to analyze a different variable of this data set, say weights, then set *DATA = auto.stats*[, *"Weight"*]. The left diagram shows a histogram of the *DATA* with l bins running from the minimal to maximal observations shown in the title. The sample size is shown in the subtitle. The next 3 diagrams correspond to 3 different values of the running argument *arg*, which may be *cJ0, cJ1, cJM, cT,* or *cB*. The values of the running argument are controlled by the argument *set.arg*, and they are shown in the corresponding titles. The considered interval (support) is the same for all the 3 diagrams and is controlled by *set.sup=c(a,b)*. If $a = -99.9$, then the default support is calculated using (3.1.16). The interval $[a, b]$ is shown in the subtitles.} [*arg= "cJ0", set.arg=c(4,2,1), Data= auto.stats*[, *"Turning Circle"*], *l=25, s=1, set.sup=c(-99.9, 0), cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

show the universal estimate with 3 different values of c_{J0} shown in their titles. The support $[a, b]$ is shown in the subtitles.

Let us analyze the graphs. The default estimate with $cJ0 = 4$ shows 3 modes and indicates a possibility of 3 clusters. Here the distribution of diameters of turning circles is considered. It is natural to assume that these 3 clusters correspond to economic, full size, and luxury (like the Lincoln) cars. Thus, this estimate, based on only 72 observations, looks rather realistic. The smaller $cJ0$, the smaller the number of Fourier coefficients that may participate in the estimate (recall that this argument is used in (3.1.10) to calculate J_n). Here, because the underlying density clearly has several modes, a decrease in J_n takes its toll in terms of smoothing the data and hiding the clusters. We see that the case $cJ0 = 2$ implies a bimodal density with a peculiar right tail, and the case $cJ0 = 1$ just smoothes everything and hides the structure of this data set.

3.11 Exercises

3.1.1 Let $[0, 1]$ be the support of an estimated density f. Explain why in this case there is no need to estimate θ_0.
3.1.2 First, show that (3.1.4) is an unbiased estimate of θ_j. Second, explain why the estimate $w_j\hat{\theta}_j$ is biased when $w_j \neq 1$ and $\theta_j \neq 0$. Finally, why would one prefer to use a biased estimate $w_j\hat{\theta}_j$ in place of the unbiased $\hat{\theta}_j$?
3.1.3 Verify (3.1.8).
3.1.4 Verify (3.1.9).
3.1.5 Let $\hat{\theta}_j$ be the sample mean estimate (3.1.4). Find $E\{\hat{\theta}_j^2\}$ and $\text{Var}(\hat{\theta}_j^2)$.
3.1.6 Show that the first sum on the right-hand side of (3.1.6) is the variance of $\tilde{f}_J$, while the second one is the integrated squared bias.
3.1.7 Explain the underlying idea of (3.1.10).
3.1.8 Remark 3.1.1 defines the U-statistic $\tilde{\theta}_j^2$. Find $E\{\tilde{\theta}_j\}$, $E\{\tilde{\theta}_j^2\}$, and $\text{Var}\{\tilde{\theta}_j\}$. Also verify that $\hat{\theta}_j^2 - \tilde{\theta}_j^2$ is an unbiased estimate of $\text{Var}(\hat{\theta}_j)$.
3.1.9 Using notations of the previous exercises, suggest an estimate of the optimal weight $w_j^* = \theta_j^2/(\theta_j^2 + \text{Var}(\hat{\theta}_j))$ that is based on $\hat{\theta}_j^2$ and $\tilde{\theta}_j^2$.
3.1.10 Explain how (3.1.12) is obtained.
3.1.11 Write down and comment on all the steps of calculating the universal estimate.
3.1.12 Explain all the parameters (coefficients) of the universal estimate.
3.1.13 Repeat Figure 3.2 approximately 10 times (note that each time new data sets are generated). Print hard copies, and then make your conclusion about the worst, the best, and "typical" estimates for each corner density.
3.1.14 Choose 3 sample sizes. Then, using Figure 3.2, find optimal parameters (arguments) of the universal estimate for every corner density. Also, try to find parameters that are reasonably good for all the corner densities.
3.1.15 Use Figure 3.2 with $cB = 0$. This will exhibit estimates without removing bumps. Discuss the outcomes and then, again using this figure, suggest an optimal cB for the Strata and the Delta.
3.1.16 Use Figure 3.3 to find optimal arguments for (a) the Uniform and the Normal; (b) the Bimodal and the Strata; (c) the Uniform, the Delta, and the Monotone; (d) the Angle and the Strata.
3.1.17 Repeat Figure 3.4 with different intervals of estimation. Then explain how an interval affects the universal estimates.
3.1.18 Use Figure 3.6 to explain how a choice of the support affects estimation of the Delta, the Bimodal, and the Strata.
3.1.19 Use Figure 3.6 and explain how a larger support (say, $[-1, 3]$) affects estimation of the Uniform.
3.1.20 For densities supported on $[0, 1]$, the coefficient of difficulty is 1 (recall the definition given in Remark 3.1.4). What is the value of the coefficient of difficulty for densities supported on an interval $[a, b]$?
3.1.21 Consider a pair (X, Y) of nonnegative random variables uniformly distributed under a nonnegative curve $f(x)$ integrated to unity on $[0, 1]$,

that is, the pair is uniformly distributed on the set $\{(x, y) : \ 0 \le y \le f(x),\ 0 \le x \le 1\}$. What is the marginal density of X?

3.2.1 Can the data-driven estimator of Section 3.1 be written as (3.2.1)?

3.2.2 Suggest an example of the density where $J^* \ne J$ for the hard-threshold oracle.

3.2.3 Check (3.2.6). Hint: Recall (3.2.3)–(3.2.4).

3.2.4 Verify (3.2.7).

3.2.5 Explain for what kind of densities the ideal cutoff J should be relatively large even for the smallest sample sizes.

3.2.6 What can be said about odd Fourier coefficients (for the cosine system) of the Normal corner density?

3.2.7 Explain why the ideal cutoff J does not increase with n running from 50 to 1000 for a particular density shown in Figure 3.7. Hint: Recall how fast Fourier coefficients of analytic functions decrease.

3.2.8 According to Figure 3.8, the linear oracle always performs better than the smoothed one, and the smoothed oracle always performs better than the truncated one. Explain why.

3.3.1 How close is AISE to MISE for $m = 5000$?

3.3.2 Figure 3.9 indicates that for some experiments the data-driven estimate outperforms the truncated oracle. Is this possible?

3.3.3 Consider Figure 3.9. Why are the ratios AISES/OMISEL so large for the Angle?

3.3.4 Figure 3.10 indicates that despite the fact that the smoothed oracle performs better than the truncated one, for some experiments the truncated estimator (which mimics the truncated oracle) performs better than the smoothed estimator (which mimics the smoothed oracle). Explain this phenomenon.

3.3.5 Based on Figure 3.11, can you suggest any improvements in the estimation of the optimal cutoff?

3.3.6 Explain why for the Uniform the difference $\bar{J} - J$ is always positive (see Figure 3.11).

3.4.1 Explain all steps in establishing (3.4.2). Then find $P(Y \le y, \delta = 0)$.

3.4.2 Calculate the variance of the estimate (3.4.3).

3.4.3 Find the expectation and the variance of the estimate (3.4.4).

3.4.4 Consider a particular realization $\{(Y_l, \delta_l), l = 1, 2, \dots, 10\}$ and draw a graph of the product-limit estimate (3.4.5).

3.4.5 Repeat Figure 3.13 with different sets of arguments for the estimate. What is a good set for each corner function and all the corner functions? Is the recommended set robust to the sample size?

3.4.6 Suppose that a reliable estimation of the Normal density requires at least 75 observations. A data analyst may have to deal with a data set that may be right-censored by either uniform $U(0, 1.5)$ or Exponential with the rate $\lambda = 2$ random variable. For these two cases, what minimal sample sizes may be recommended for a reliable estimation?

3.4.7 Verify the idea of dealing with left censoring suggested in Remark 3.4.1. Then write down the estimate.
3.5.1 Let $Y = (X + \varepsilon)[\text{mod}2\pi]$, where ε is uniform on $[0, 2\pi]$. Show that Y is also uniform regardless of the distribution of X.
3.5.2 Let X and ε be random variables with the densities f^X and f^ε. Find the density of the sum $Y = X + \varepsilon$.
3.5.3 Let $h^X(u) := E\{e^{iuX}\}$ be the characteristic function of a random variable X. Show that $h^X(0) = 1$, $|h^X(u)| \le 1$, $h^X(u)$ is a real function if X is symmetric about zero.
3.5.4 Calculate characteristic functions for Uniform, Cauchy, Exponential, and Gamma random variables. Then explain how measurement errors with these distributions will affect the recovery of an underlying density.
3.5.5 Explain the formulae (3.5.4)–(3.5.5).
3.5.6 Repeat Figure 3.15 with different values of σ (it is controlled by *sigma*). What is the value of this argument when the difference between the convolved Normal and the convolved Bimodal practically disappears?
3.5.7 What are the expectation and the variance of the estimate (3.5.6)?
3.5.8 Use Figure 3.16 to find optimal arguments.
3.5.9 Assume that the characteristic function $h^\varepsilon(j)$ of the measurement error is real. Based on twice-repeated observations $Y_{ls} = X_l + \varepsilon_{ls}$, $s = 1, 2$ and $l = 1, 2, \ldots, n$, suggest an estimate of f^X.
3.5.10 Explain why the right tail decreases in the histogram shown in Figure 3.17.7.
3.5.11 Use Figure 3.17, and for every corner function find a minimal σ such that the deconvolution is practically impossible. Then analyze the result.
3.5.12 Use Figure 3.17 to analyze data sets for the cases of σ equal to 0.05, 0.1, and 0.2. Explain why the problem of recovery of an underlying density f^X is called ill-posed. Make hard copies of cases where there is no way to realize an underlying density from the analysis of data.
3.5.13 Use Figure 3.17 to find optimal arguments for the cases $\sigma = 0.05$ and $\sigma = 0.1$.
3.6.1 Suggest an example of length-biased data.
3.6.2 Find the expectation and the variance of the estimate (3.6.3).
3.6.3 Find the expectation and the variance of the estimate (3.6.4).
3.6.4 Show that the coefficient of difficulty (3.6.5) is at least 1, with equality if $g(x) \equiv 1$.
3.6.5 Show that the variable Y, generated by the acceptance–rejection method, has the desired density f^Y. Hint: see Rubinstein (1981, p. 46).
3.6.6 Use Figure 3.18 to find optimal arguments of the estimate. Are they robust to changes in the function g?
3.7.1 Repeat Figure 3.19 with arguments that lead to highly oscillatory estimates for all the three corner densities. Then compare the projections and analyze them.
3.7.2 Find elements $(1, \varphi'_1, \ldots, \varphi'_J)$ of the cosine–polynomial basis.

3.7.3 Using Figure 3.20, find optimal arguments for the Monotone density. Then check how the estimate performs for the other densities.
3.8.1 Let $F(x)$ be the cdf that has the probability density. Show that this cdf is continuous in x.
3.8.2 Consider the empirical cdf $\bar{F}_n(x)$ defined at (A.32) in Appendix A, and let $F(x)$ be an underlying cdf. Show that for each fixed $x \in (-\infty, \infty)$, (a) $\bar{F}_n(x)$ is an unbiased estimate of $F(x)$; (b) $\mathrm{Var}(\bar{F}_n(x)) = F(x)(1 - F(x))/n$; (c) $\bar{F}_n(x)$ is asymptotically normal $N(F(x), F(x)(1 - F(x))/n)$.
3.8.3 What can be said about the Kolmogorov confidence band (3.8.4) in terms of the probability of covering an unknown underlying cdf?
3.8.4 Given $\sum_{l=1}^{n} y_l = 1$, show that $\sum_{l=1}^{n} y_l^2 \ge 1/n$ and the minimum is attained by $y_1 = \cdots = y_n = 1/n$.
3.8.5 Consider n identical trials where a random variable X belongs to a set A with the probability p. Find the expectation and the variance of the number of trials when $X \in A$. Hint: Recall the binomial random variable.
3.8.6 Consider the $\hat{\chi}^2$ test statistic defined at (3.8.13). Prove that

$$E\{\hat{\chi}^2\} = l - 1, \quad \mathrm{Var}(\hat{\chi}^2) = 2(l-1) + n^{-1}\Big[\sum_{k=1}^{l} p_k^{-1} - l^2 - 2l + 2\Big].$$

3.8.7 A typical rule of thumb to choose the number l of bins for the chi-squared test is to choose an l between $4(2n^2/z_\alpha^2)^{1/5}$ and half that value. As a result, for the particular case of $\alpha = 0.05$ the choice $l = \lfloor 2n^{2/5} \rfloor$ is often recommended. Test this rule using Figure 3.22.
3.8.8 Use Figure 3.21 to test two other years of the lottery (the data sets are **lottery2.number** and **lottery3.number**). Draw a conclusion about the fairness of the lottery.
3.8.9 Use Figure 3.22 with different sample sizes. Then analyze the outcomes and rank the tests in terms of the accurate acceptance of the Uniform and accurate rejection of the other corner function.
3.8.10 Using Figure 3.22, try to find optimal parameters for the tests.
3.9.1 Consider the general case of an estimate (3.1.14) with $0 \le \hat{w}_j \le 1$. Suggest analogue of (a) The oracle (3.9.1); (b) The oracle's risk (3.9.5); (c) The empirical risk.
3.9.2 Show that (3.9.1) is the oracle for the estimate (3.1.14) with $\hat{w}_j$ being either 0 or 1.
3.9.3 Check (3.9.2).
3.9.4 Prove that for the sine basis and a density supported on $[0, 1]$, the relation $nE\{(\hat{\theta}_j - \theta_j)^2\} \to 1$ as $n \to \infty$ holds. Hint: Follow along the lines (3.1.7)–(3.1.8) and use $\sin^2(\alpha) = [1 - \cos(2\alpha)]/2$.
3.9.5 How do coefficients of the universal estimate affect the sine and cosine estimates? Hint: Use Figure 3.23.
3.10.1 Repeat Figures 3.24–6 for different variables of the data sets **air** and **auto.stats**.

3.12 Notes

3.1 The first result about optimality of Fourier series estimation of nonparametric densities is due to Chentsov (1962). Chentsov never was satisfied with the fact that this estimate could take on negative values. Thus, later he recommended to estimate $g(x) := \log(f(x))$ by a series estimate $\hat{g}(x)$ and then set $\hat{f}(x) := e^{\hat{g}(x)}$; see Chentsov (1980) and also Efron and Tibshirani (1996). Clearly, the last estimate is nonnegative. Recall that we dealt with this issue by using the projection (3.1.15).

The idea of smoothing Fourier coefficients is due to Watson (1969). A historical overview of smoothing procedures may be found in the book by Tarter and Lock (1993, Sections 4.5–4.6). Also, in Sections 4.2–4 of that book different algorithms of choosing cutoffs are discussed. The reader familiar with the ideas of *Akaike's information criteria*, presented in Akaike (1973), and *penalized* estimation, see Bírge and Massart (1997), may find the similarity between (3.1.10) and these approaches striking.

Series density estimators are discussed in the books by Devroye and Györfi (1985, Chapter 12), Thompson and Tapia (1990, Section 2.4), Tarter and Lock (1993, Chapter 4), and Hart (1997, Section 3.3), where also further references may be found.

Asymptotic justification of using data-driven Fourier series density estimators is given in Efromovich (1985).

3.2–3 The asymptotic minimaxity of the linear oracle over all possible estimators and the possibility to mimic it by a data-driven estimator for smooth densities is established in Efromovich and Pinsker (1982) and Efromovich (1985). Using wavelet bases allows one to establish similar results for spatially inhomogeneous densities as well, see the review in Härdle et al. (1998, Chapter 10) and Efromovich (1999a).

Similar results for a different set of 18 corner densities are presented in Efromovich (1996b).

3.4 The books by Collett (1994) and Venables and Ripley (1977, Chapter 12) discuss the survival analysis and give plenty of examples.

3.5 The statistical analysis of directional data is as old as the analysis of linear data. For instance, the theory of errors was developed by Gauss primarily to analyze directional measurements in astronomy. There are several excellent books about directional data, for instance, Mardia (1972) and Fisher (1993). Practical examples of using the universal estimator are discussed in Efromovich (1997a). Different proposed methods for deconvolution of an underlying density are discussed in the book by Wand and Jones (1995, Section 6.2.4).

3.6 Examples of length-biased data may be found in articles by Zelen (1974) and Morgenthaler and Vardi (1986). A review of different estimators may be found in the book by Wand and Jones (1995, Section 6.2.2).

3.7 A very nice discussion of estimation of monotone densities may be found in the textbook by Devroye (1987, Chapter 8). The book by Barlow et al. (1972) is another useful reference. The book by Härdle (1990, Section 8.1) discusses the same problem for regression models. Robust estimation is discussed in Huber (1981).

Probably the first results about using enriched trigonometric–polynomial bases may be found in the textbook by Krylov (1955). Eubank and Speckman (1990) discuss the use of this basis for regression setting, and Efromovich (1997d) discusses its use for a data-driven density estimation.

3.8 A comprehensive review of classical goodness-of-fit techniques may be found in the textbook edited by D'Agostino and Stephens (1986). The book by Hart (1997) discusses both theoretical and applied aspects of nonparametric model checking. A comprehensive review of asymptotically minimax tests for nonparametric hypotheses may be found in Ingster (1993).

3.9 The article by Marron and Tsybakov (1995) discusses some aspects of visual error criteria for qualitative smoothing. A book-length treatment of visualizing data is given by Cleveland (1993).

3.10 Discussion of practical examples may be found, for instance, in the books by Silverman (1986) and Simonoff (1996).

4

Nonparametric Regression for Small Samples

This chapter is devoted to data-driven orthogonal series estimators for different models of nonparametric regression where a data analyst wishes to know how one variable responds to changes in another variable. The simplest model of additive homoscedastic regression is discussed in Section 4.1. We shall see that estimation of a regression function is similar to the density estimation discussed in Sections 3.1–3.3 (an asymptotic equivalence of these two models is discussed in Section 7.2). There are two important corollaries from this fact. First, estimators and oracles developed and studied for the density model can be directly used for the regression model. Second, we may use the coefficient of difficulty defined in Chapter 3 to assess sample sizes for regression models that give us comparable (with density estimation) precision of estimation.

More complicated and practically interesting heteroscedastic regression models are considered in Section 4.2. In particular, the standard deviation of additive errors may be a function of the predictor; in this case it is referred to as the scale (spread, volatility) function. Estimation this function is an important topic in many applications, and it is discussed in Section 4.3. The case of spatially inhomogeneous regression functions and wavelet series estimators is considered in Section 4.4.

These four sections constitute the core material. All the others are devoted to different special cases; in particular, robust regression is discussed in Section 4.6. Section 4.12, "Practical Seminar," is devoted to employing the universal estimator for the analysis of real data sets.

4.1 Classical Model of Homoscedastic Nonparametric Regression

The aim of regression curve estimation is to find a relationship between variables X and Y that allows one to quantify the impact of X on Y.

The simplest mathematical model is as follows. Let n pairs of observations $(X_1, Y_1), \dots, (X_n, Y_n)$ be given such that

$$Y_l = f(X_l) + \sigma\varepsilon_l\,, \quad l = 1, 2, \dots, n\,. \tag{4.1.1}$$

Here Y_l are called *responses* (or *dependent variables*), and the so-called *predictors* (*independent variables* or *covariates*) $X_1, \dots, X_n$ are either iid realizations of a uniform random variable $U(0, 1)$ or fixed equidistant points $X_l = l/(n + 1)$. The random variables ε_l are iid realizations of a random variable ε with zero mean and unit variance. (The abbreviation iid stands for independent identically distributed. Also, recall that σ and ε are the Greek letters "sigma" and "epsilon," respectively.) The positive constant σ defines the standard deviation (spread or scale) of the additive error $\sigma\varepsilon$. Depending on the model of predictors, the regression is referred to as random- or fixed-design regression. The regression model is called *homoscedastic* if the variance of errors is constant (does not depend on predictors) and predictors are either equidistant or uniformly distributed. Note that for the case of random design, pairs (X_l, Y_l) are iid realizations of a pair of random variables (X, Y), where $Y = f(X) + \sigma\varepsilon$.

The problem is to estimate the regression function $f(x)$, $0 \le x \le 1$, by an estimate $\hat{f}_n$ with minimal mean integrated squared error (the shorthand notation is MISE), which is $E\{\int_0^1 (\hat{f}_n(x) - f(x))^2 dx\}$.

A plot of the pairs (X_l, Y_l) in the xy-plane (so-called *scattergram* or *scatter plot*) is a useful tool to get a first impression about a data set at hand. Consider a Monte Carlo simulation of observations according to (4.1.1) for a fixed-design regression with $n = 50$, $\sigma = 1$, ε being standard normal and the regression functions being the corner functions shown in Figure 2.1. The scattergrams are displayed in Figure 4.1, where dots show the simulated pairs of observations. The scattergrams are overlaid by linear least-squares regression lines calculated by the S–PLUS function **lsfit**.

An appealing nature of the regression problem is that one can easily appreciate its difficulty. To do this, try to draw curves $f(x)$ through the middle of the cloud of dots in the scattergrams that, according to your own understanding of the data give a good fit (describe a relationship between X and Y) according to the model (4.1.1). Or even simpler, because in Figure 4.1 the underlying regression functions f are known, try to recognize them in the cloud of dots.

In Figure 4.1.1 the relationship clearly depends on your imagination. Note that even the linear least-squares regression is confused, despite the fact that this is the best tool of regression analysis for fitting a curve like

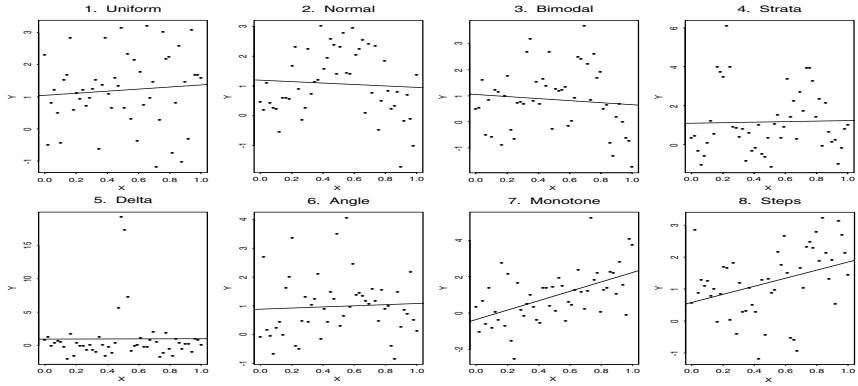

FIGURE 4.1. Simulated, according to (4.1.1), fixed-design scattergrams of size $n = 50$ overlaid by linear regression lines. The coefficient of difficulty is $d = \sigma^2 = 1$, so the precision of estimation of a regression function should be about the precision of estimation of a probability density discussed in Section 3.1. {To repeat this figure with new simulated data sets type at the S–PLUS prompt > **ch4(f=1)**. The sample size n and the standard deviation σ of the additive error are controlled by the arguments n and *sigma*.} [*n=50, sigma=1*]

the Uniform. The underlying Normal function is recognizable in Figure 4.1.2 (especially if you know that the underlying function is the Normal), but note that linear regression is again not a big help. For the case of the Bimodal regression function, only knowledge of the underlying function may help one to see a line through these points that resembles the Bimodal. The Strata is another example where no relationship can be easily recognized. The Delta is the interesting case, where knowledge of the underlying function helps to visualize it. Otherwise, linear regression would be a good choice with the 4 points near $x = 0.5$ announced as "clear" outliers. The Angle is a very complicated case, where it is difficult to recognize the underlying function, which has no sharp features. The Monotone is nicely visualized, and here the linear regression does a good job. The Steps regression function is another complicated case, where it is difficult (if not impossible) to recognize the underlying function. Moreover, it looks reasonable to suggest that the regression function is increasing as $x \to 0$.

Overall, except for the several "lucky" corner functions, manual analysis via visualization is not too helpful, and this explains why a data-driven estimation, where data speak for themselves, is important.

The underlying idea of a series estimator is to approximate f by a partial sum,

$$f_J(x) := \sum_{j=0}^{J} \theta_j \varphi_j(x), \ \ 0 \le x \le 1, \quad \text{where} \ \ \theta_j := \int_0^1 \varphi_j(x) f(x) dx \ \ (4.1.2)$$

are Fourier coefficients and J is a cutoff. The functions φ_j can be elements of any basis $\{\varphi_j,\ j = 0, 1, \ldots\}$ in $L_2([0, 1])$. As in Sections 3.1–3.3, here the cosine basis $\{\varphi_0(x) = 1, \varphi_j(x) = \sqrt{2}\cos(\pi jx),\ j = 1, 2, \ldots\}$ is used.

The problem is to find good estimators for the Fourier coefficients θ_j and an optimal cutoff J that minimizes MISE. A natural estimator for each Fourier coefficient θ_j is

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} Y_l \varphi_j(X_l). \tag{4.1.3}$$

Indeed, for fixed design regression the estimator (4.1.3) is a naive numerical integration formula (Exercise 4.1.2), while for random design this is a sample mean estimator because the assumption $E\{\varepsilon\} = 0$ implies

$$E\{\hat{\theta}_j\} = E\{Y\varphi_j(X)\} = \int_0^1 f(x)\varphi_j(x)dx = \theta_j.$$

Then, according to Section 3.1, a good estimator of an optimal cutoff J is based on a good estimator of $d_j := nE\{(\hat{\theta}_j - \theta_j)^2\}$. Under some mild assumptions on f, a straightforward calculation (Exercise 4.1.3) shows that *for fixed-design regression*,

$$nE\{(\hat{\theta}_j - \theta_j)^2\} = \sigma^2 + r_{nj} =: d + r_{nj}\ , \tag{4.1.4}$$

where r_{nj} vanishes for large j and n. Thus, if σ^2 is known, then the natural choice of the estimate $\hat{d}_j$ for d_j is $\hat{d}_j = \hat{d} := \sigma^2$.

Recall that for the density estimation problem, where a density was estimated over its support $[0, 1]$, a similar estimate was $\hat{d} = 1$. Thus, if $\sigma^2 = 1$, then the fixed-design regression and the density estimation settings are similar in the sense that they have the same coefficient of difficulty $d = 1$. (In Section 7.2 this conclusion will be supported by asymptotic equivalence.)

This similarity is useful for the analysis of both the regression and the density models. Indeed, the advantage for the regression model is that the data-driven estimators, oracles, and oracle inequalities developed in the previous chapter can be used straightforwardly for the regression. For the probability density model the advantage is that the regression model can serve as a simple tool to understand and visualize the coefficient of difficulty, which for the regression is just the variance of errors. Recall that the ratio of coefficients of difficulty (which is called the relative coefficient of difficulty) gives us a rule of thumb on how to change a sample size to get a comparable precision of estimation; see the discussion in Section 3.4. For instance, if $d = \frac{1}{4}$, then one needs a quarter of the observations to get a quality of estimation comparable with the setting where $d = 1$ (for instance, comparable with the density estimation studied in Section 3.1 or the regression with $\sigma = 1$). Thus, to visualize this rule of thumb for $d = \frac{1}{4}$ it suffices to simulate data according to (4.1.1) with $\sigma = 0.5$.

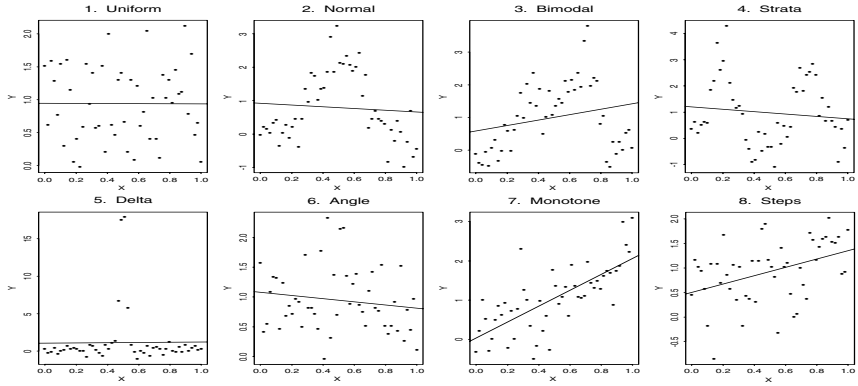

FIGURE 4.2. Simulated, according to (4.1.1), fixed-design scattergrams overlaid by linear regression lines, $n = 50$ and $\sigma = \frac{1}{2}$. The coefficient of difficulty is $d = \sigma^2 = \frac{1}{4}$, so the precision of estimation of a regression function should be about the precision of estimation for the case $\sigma = 1$ and the quadruple sample size, i.e., 200 observations. {This figure was created by the call **ch4(f=1, n=50, sigma=.5)**.}

Such Monte Carlo simulations (with $\sigma = 0.5$) are shown in Figure 4.2. Comparison of Figures 4.1 and 4.2 shows the meaning of the coefficient of difficulty. In Figure 4.2 all the underlying regression functions, with the possible exception of the Uniform, the Angle, and the Steps, are quite recognizable. On the other hand, the scattergrams for the Uniform, the Angle, and the Steps indicate that for the sample size $n = 50$ even the coefficient of difficulty $d = \frac{1}{4}$ is not sufficiently small to recognize these functions "manually." This conclusion will be important for the evaluation of the performance of data-driven estimators. (See also Exercise 4.1.1.)

Now let us return to the search for a good estimator of $nE\{(\hat\theta_j - \theta_j)^2\}$ for the case of random design. According to Exercise 4.1.4,

$$d'_j := nE\{(\hat\theta_j - \theta_j)^2\} = \mathrm{Var}(Y\varphi_j(X)) = \int_0^1 (f^2(x) + \sigma^2)\varphi_j^2(x)dx - \theta_j^2$$
$$= \Big(\int_0^1 f^2(x)dx + \sigma^2\Big) + \Big[2^{-1/2}\int_0^1 f^2(x)\varphi_{2j}(x)dx - \theta_j^2\Big]. \qquad (4.1.5)$$

According to Section 2.2, the term in the square brackets practically vanishes for large j, so (4.1.5) can be approximated by $d' := \int_0^1 f^2(x)dx + \sigma^2$. Comparison with (4.1.4) reveals that the estimator $\hat\theta_j$ performs worse for the random design. This is not the curse of a random design, and in the next section a better estimator will be suggested. On the other hand, the simplicity of $\hat\theta_j$ is so appealing that it makes its consideration worthwhile.

Let us stress that d' is not the coefficient of difficulty for the random-design regression; the next section shows that both these designs have the same coefficient of difficulty d.

Below, we continue the discussion only for the case of the fixed design. The crucial difference between the density and regression settings is the necessity for the regression setting to suggest a good estimator of $d := \sigma^2$. (Recall that $\hat{\theta}_0$ was a good estimator of d for the density model.) Thus, several options to estimate d will be suggested.

First, the following universal estimate may always be used,

$$\hat{d}_{UV} := n(J_{2,n})^{-1} \sum_{j=J_{1,n}+1}^{J_{1,n}+J_{2,n}} \hat{\theta}_j^2, \tag{4.1.6}$$

where $J_{1,n}$ and $J_{2,n}$ are some slowly increasing sequences. For instance, $J_{1,n} = J_n$ and $J_{2,n} = 2J_n$ serve well (the sequence J_n was defined in Section 3.1 and will be reminded below). The underlying idea of (4.1.6) is that $nE\{\hat{\theta}_j^2\} = d + r_{jn}$, where $r_{jn} \to 0$ as j and n increase; see (4.1.5) and Exercise 4.1.5. An attractive modification of (4.1.6) is instead of averaging $\{\hat{\theta}_j^2,\ J_{1,n} < j \le J_{1,n}+J_{2,n}\}$ to consider the squared normed sample median of absolute values $\{|\hat{\theta}_j|,\ J_{1,n} < j \le J_{1,n} + J_{2,n}\}$, that is,

$$\hat{d}_{UM} := n[1.48\ \text{median}(\{|\hat{\theta}_j|,\ J_{1,n} < j \le J_{1,n} + J_{2,n}\})]^2.$$

The underlying idea of the last formula is that for a normal $N(0, \sigma^2)$ random variable ξ the following approximate relation between the variance σ^2 and the median of the of the random variable $|\xi|$ holds, $\sigma^2 \approx [1.48\ \text{median}(|\xi|)]^2$. Because the sample median is essentially more robust to outliers than the sample mean, using $\hat{d}_{UM}$ may be a good idea for robust estimation. We shall use this method in Sections 4.4 and 4.10.

Second, if $f(x)$ is smooth, then

$$Y_{l+1} - Y_l = \sigma(\varepsilon_{l+1} - \varepsilon_l) + [f(X_{l+1}) - f(X_l)] = \sigma(\varepsilon_{l+1} - \varepsilon_l) + o_n(1),$$

where $o_n(1) \to 0$ as $n \to \infty$ (recall that we consider the case where $X_l = l/(n+1)$). Thus, these differences allow one to estimate d because $E\{(\varepsilon_{l+1} - \varepsilon_l)^2\} = 2$. We shall use this approach in Section 5.8.

Another possibility to estimate d, which will be our main approach in the chapter due to its versatility, is to straightforwardly use its definition. The approach is as follows. Let us for a moment assume that an underlying regression function f is known. Then, it is natural to estimate $d = \sigma^2$ by the sample variance estimator

$$\tilde{d}_V := n^{-1} \sum_{l=1}^{n} (Y_l - f(X_l))^2.$$

Since f is unknown, we may plug in a naive truncated estimate,

$$\tilde{f}_J(x) := \sum_{j=0}^{J} \hat{\theta}_j \varphi_j(x). \tag{4.1.7}$$

Such a plug-in estimate with a relatively small J is a good choice for all the corner functions except the Delta. The problem is that the Delta has extremely large Fourier coefficients θ_j even for relatively large j (in other words, the pilot estimate $\tilde{f}$ does not give a good fit of the Delta especially for x near the point 0.5; see the ideal approximations in Figure 2.3.5. Of course, the cutoff J may be increased, but then it leads to overfitted estimates for smooth corner functions (after all, recall that we need to find d for a good choice of a cutoff).

One of the possible approaches to solve this puzzle is as follows. For the case of smooth underlying regression functions even a relatively small J makes all the residuals $Y_l - \tilde{f}_J(X_l)$ approximately the same in the sense that there should be no outliers. (Recall that *outliers* are sample values that cause surprise in relation to the majority of the sample.) The situation is inverse for a function like the Delta, where we shall see several outliers. On the other hand, for large J there are no outliers even for the Delta case. Thus, we can use the notion of outliers to choose an initial cutoff $\tilde{J}$. The only point that is to be clarified is how to determine the outliers.

Here we use the following approach. We assume that if the sample variance of the residuals is smaller than a coefficient r times the squared normed sample median of absolute residuals, then there are no outliers; otherwise, outliers are presented, and $\tilde{J}$ should be increased. In other words, here we use the familiar statistical fact that the sample variance is not resistant to outliers, while the sample median is.

Thus, to find an initial pilot cutoff $\tilde{J}$, we begin with $J = 0$ and calculate two estimates of d: the sample variance estimate

$$\tilde{d}_V(\tilde{f}_J) := n^{-1} \sum_{l=1}^{n} (Y_l - \tilde{f}_J(X_l))^2 \tag{4.1.8}$$

and the squared normed sample median estimate

$$\tilde{d}_M(\tilde{f}_J) := [1.48 \text{ median}(\{|Y_l - \tilde{f}_J(X_l)|, l = 1, \dots, n\})]^2. \tag{4.1.9}$$

If

$$\tilde{d}_V(\tilde{f}_J) < r\tilde{d}_M(\tilde{f}_J), \tag{4.1.10}$$

then we stop and $\tilde{J} = J$; otherwise, we increase J and repeat the previous step. The maximal considered J is $c_{JM}J_n$.

Using this initial cutoff, the initial estimate of d is calculated by (4.1.8). Then we use the estimate (3.1.14) of f with (4.1.3) being the estimate of θ_j. This estimate of f is used as a pilot estimate in (4.1.8) for calculating the estimate $\hat{d}$ of d. Using $\hat{d}$ in (3.1.14) we obtain an estimate $\hat{f}$ of f. Note that recursion of these steps is possible

Now, after the discussion of all these ideas, let us formally write down the universal data-driven estimate. First, we estimate all the Fourier coefficients θ_j, $0 \le j \le c_{JM}J_n$, using (4.1.3). Recall that J_n is the rounded down

$c_{J0} + c_{J1}\ln(n)$ and c_{J0} and c_{J1} are coefficients with the default values 4 and 0.5 (the same as in Section 3.1).

Second, the initial cutoff $\tilde{J}_0$ is estimated as explained above; namely, this is the minimal $\tilde{J}$, $0 \le \tilde{J} \le c_{JM} J_n$, such that (4.1.10) holds. Third, the initial estimate $\tilde{d}$ is calculated by (4.1.8) with J being the initial estimate $\tilde{J}_0$. Fourth, we make the first iteration in using the estimate (3.1.10) to calculate a pilot estimate of J. Namely, a pilot estimate of the cutoff is calculated by the formula

$$\tilde{J} := \operatorname{argmin}_{0 \le J \le J_n} \sum_{j=0}^{J} (2\tilde{d}n^{-1} - \hat{\theta}_j^2). \tag{4.1.11}$$

Then the smoothing weights (coefficients)

$$\tilde{w}_j := (1 - n^{-1}\tilde{d}/\hat{\theta}_j^2)_+ \tag{4.1.12}$$

are calculated (recall that $(x)_+ := \max(0, x)$ denotes the positive part). Then, according to (3.1.14), the pilot estimate of f is calculated as

$$\tilde{f}(x) := \sum_{j=0}^{\tilde{J}} \tilde{w}_j \hat{\theta}_j \varphi_j(x) + \sum_{j=\tilde{J}+1}^{c_{JM} J_n} I_{\{\hat{\theta}_j^2 > c_T \tilde{d} \ln(n)/n\}} \hat{\theta}_j \varphi_j(x). \tag{4.1.13}$$

Here c_T is the same coefficient of thresholding as in (3.1.14) with the default value 4, and similarly, $c_{JM} = 6$.

Finally, we calculate the estimate $\hat{f}$ by repeating the previous steps: (i) the estimate $\hat{d}$ of d is calculated by (4.1.8) with the use of the pilot estimate $\tilde{f}$; (ii) the calculated estimate $\hat{d}$ is used in (4.1.11) to find the optimal cutoff $\hat{J}$; (iii) the estimate $\hat{d}$ is used to calculate the optimal weights $\hat{w}_j$ defined at (4.1.12); (iv) the optimal weights $\hat{w}_j$ and the optimal cutoff $\hat{J}$ are used in (4.1.13) to calculate the *universal* data-driven estimate $\hat{f}$.

Figure 4.3 illustrates the performance of the universal data-driven estimate for the fixed-design regression and the particular case of standard normal errors. Here $d = \sigma^2 = 1$, so the coefficient of difficulty is the same as for the problem of estimation of the corner densities. On the other hand, here the estimator should estimate the coefficient of difficulty, that is, σ^2, and this is a complicated problem by itself. Keeping this in mind, we see that these estimates are relatively good and resemble the density estimates shown in Figures 3.2 and 3.3.

We may conclude from this particular figure that our idea (4.1.10) of comparison between two estimates of d worked out nicely: Both smooth (low-frequency) functions like the Uniform and spatially inhomogeneous (high-frequency) functions like the Delta are well estimated. Let us look more closely at the short-dashed lines, which correspond to the underlying scattergrams. We see that all deviations from the underlying regression functions are justified by the scattergrams. For instance, the estimate for

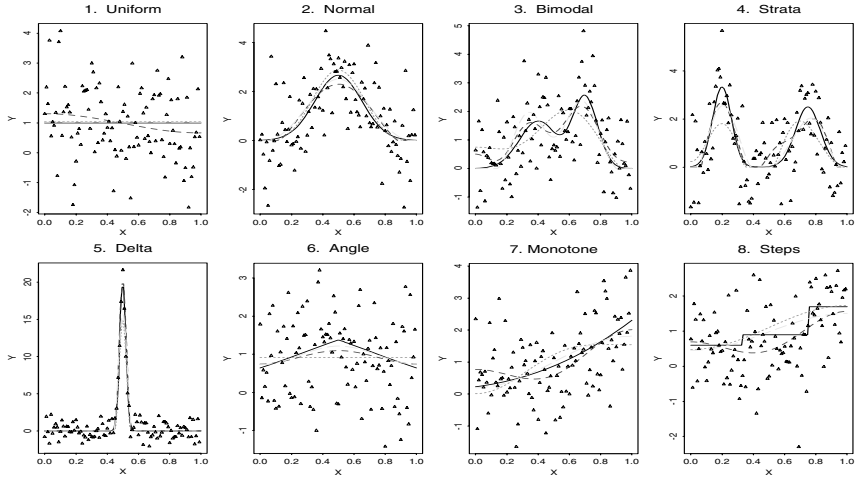

FIGURE 4.3. The universal estimates for fixed-design regression with standard normal errors ($\sigma = 1$): Dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. The underlying regression functions are shown by solid lines. Scatter plots for $n = 100$ are shown by triangles. {The argument *set.n* controls the sample sizes, and n controls the scattergrams; thus it should belong to *set.n*. The argument *sigma* controls σ.} [*set.n=c(50,100,200), n=100, sigma=1, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

the Uniform slopes downward, but we see the same trend in the scattergram as well. The second stratum in the Strata is shown too wide, but the scattergram again supports this particular shape of the universal estimate. The case of the Monotone is another example where the estimate (short-dashed line) has a wrong left tail but this tail does fit the data. In short, it is important to keep in mind that while we know an underlying regression function, the universal estimator does not. Thus, to "judge" an estimate, it is always worthwhile to look at the underlying data.

4.2 Heteroscedastic Nonparametric Regression

In this section we relax two important assumptions of the classical homoscedastic regression setting discussed in the previous section. First, predictors are no longer necessarily equidistant or uniformly distributed. Second, σ is no longer necessarily constant and may be a function of the predictor. Such a setting is called heteroscedastic. More formally, it is assumed that n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfy

$$Y_l = f(X_l) + \sigma(X_l)\varepsilon_l, \tag{4.2.1}$$

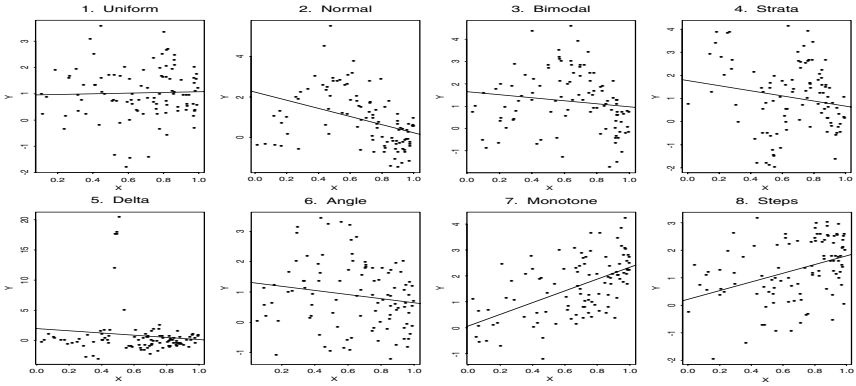

FIGURE 4.4. Scattergrams simulated according to (4.2.1) and corresponding linear regression lines. The design is random with the design density $h(x)$ being the Monotone density, ε_l are standard normal, and the scale function $\sigma(x)$ is the Angle function times a constant σ; the sample size is $n = 100$. {The argument *desden* controls the choice of design density $h(x)$. The scale function $\sigma(x) = \sigma f_j(x)$ is chosen by the argument *sigma*, which controls σ, and by the argument *scalefun*, which controls j.} [*n=100, desden=7, scalefun=6, sigma=1*]

where ε_l are iid realizations of a random variable ε with zero mean and unit variance, and $\sigma(x)$ is a nonnegative function. Set $X_0 = 0$ and $X_{n+1} = 1$. Then fixed-design predictors are defined by the formula

$$\int_{X_l}^{X_{l+1}} h(x)dx = \frac{1}{n+1}, \tag{4.2.2}$$

where $h(x)$ is a probability density supported on $[0, 1]$ and bounded below from zero on this interval; that is, $\int_0^1 h(x)dx = 1$ and $h(x) > C > 0$ for $0 \le x \le 1$. The corresponding case of a random design has predictors $X_1, X_2, \ldots, X_n$ that are iid realizations of a random variable X with the probability density $h(x)$. In both these cases the density $h(x)$ is called the *design* density, and $\sigma(x)$ is called the *scale* (*spread* or *volatility*) function.

To highlight some of the difficulties of a heteroscedastic setting, consider Monte Carlo simulations for the case of a random design with a standard normal ε, $n = 100$, the Monotone corner function being the design density $h(x)$, and the Angle corner function times a constant σ being the scale function $\sigma(x)$. As usual, the corner functions (shown in Figure 2.1) serve as underlying regression functions. Scatter plots are displayed in Figure 4.4. The analysis of these clouds of data sets reveals that a special artistic imagination is necessary to recognize the underlying regression functions. Note that even the Normal, which was so easily recognizable in Figure 4.1.2, is no longer easily recognizable, and the linear regression is extremely confused; see Figure 4.4.2.

Among other interesting features it is worthwhile to note that in Figure 4.2.4 only one observation is available to the left of $X = 0.17$. This is because the design density is the Monotone, i.e., predictors are skewed to the right edge of the unit interval. As a result, we may expect a poor estimation of the left tail of a regression function. Also, in Figure 4.2.8 the Angle scale function makes the responses more spread out (sparse) in the middle of the unit interval. This can easily lead one to the wrong conclusion that the underlying regression curve decreases in the left half of the interval. The Delta is another curious story, where the heterogeneity of errors allows one to see the pronounced wave in the left half of the interval.

Overall, it is clear that a heteroscedastic setting may be extremely complicated for a manual fitting. It is also fair to say that the problem is objectively very difficult, since neither regression function nor design density nor distribution of errors nor scale function are known. And each of these components can make the problem of finding a relationship between X and Y very complicated.

On the other hand, despite the fact that the problem becomes essentially more complicated, we shall see that the slightly modified estimator of Section 4.1 is still a good choice.

First of all, let us consider some possible modifications of the estimate (4.1.3) for θ_j. Define

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} \frac{Y_l \varphi_j(X_l)}{\hat{h}(X_l)}, \tag{4.2.3}$$

where $\hat{h}(x) := \max(\tilde{h}(x), 1/[10 + \ln(n)])$ and $\tilde{h}(x)$ is the estimate (3.1.15) of an underlying design density $h(x)$ based on n observations $X_1, \ldots, X_n$. (For the case of a fixed design an underlying density $h(x)$ is known, so it may be used instead of $\hat{h}(x)$.) Assuming that $\hat{h}(x) \equiv h(x)$, this is again a naive numerical integration for the fixed design and an unbiased estimate for the random design, because

$$E\{\hat{\theta}_j\} = E\{Y\varphi_j(X)/h(X)\} = \int_0^1 f(x)\varphi_j(x)dx = \theta_j. \tag{4.2.4}$$

Then, either the universal estimate (4.1.6) of d is used, or we use the sample variance or the sample median estimates of d only here, as in (4.2.3), based on the normed differences $(Y_l - \tilde{f}_J(X_l))/\hat{h}(X_l)$.

The outlined data-driven estimate is reliable, predictable, robust, and well understood.

However, as was explained in the previous section, for the random design this estimator performed worse than for the fixed design. Below, we explain how to overcome this caveat.

This is not a difficult problem to suggest a unique estimator that is optimal for both random and fixed designs. Surprisingly, the only needed change is in the estimator of the Fourier coefficients. Namely, we begin with

ordering pairs of observations in ascending order according to predictors (note that pairs are always ordered in this way for a fixed-design regression). In other words, we arrange the predictors in ascending order and denote them and the corresponding responses by $(X_{(l)}, Y_{(l)})$, $l = 1, 2, \ldots, n$. Also, define artificial $X_{(l)} := 2X_{(1)} - X_{(2+l)}$ for $l < 1$ and $X_{(l)} := 2X_{(n)} - X_{(2n-l)}$ for $l > n$. Set s to be the rounded-up $s_0 + s_1 \ln(\ln(n+20))$ with the default parameters $s_0 = s_1 = 0.5$. Then, define the estimator of θ_j by

$$\tilde{\theta}_j := (2s)^{-1} \sum_{l=1}^{n} Y_{(l)} \int_{X_{(l-s)}}^{X_{(l+s)}} \varphi_j(x)dx \ . \tag{4.2.5}$$

Note that this estimator is similar to (4.1.3), since it is again a kind of naive numerical integration. Also, for the case of a random design, under very mild assumptions, the difference $X_{(l+s)} - X_{(l-s)}$ is inversely proportional to $nh(X_{(l)})/(2s)$, so the estimator is similar to (4.2.3). Indeed, we can write that "approximately"

$$\begin{aligned}(2s)^{-1} Y_{(l)} \int_{X_{(l-s)}}^{X_{(l+s)}} \varphi_j(x)dx &\approx Y_{(l)}\varphi_j(X_{(l)})(X_{(l+s)} - X_{(l-s)})/(2s) \\ &\approx n^{-1} Y_{(l)}\varphi_j(X_{(l)})/h(X_{(l)}).\end{aligned}$$

The reason why the integral of $\varphi_j(x)$ is taken in (4.2.5) is explained by the fact that $\varphi_j(x)$ for large j is a highly oscillatory function, and therefore the integration gives a more accurate estimate. This integration also does not make the computations more complicated because for $\varphi_0(x) = 1$, $\varphi_j(x) = \sqrt{2}\cos(\pi j x)$ the integrals are easily calculated, namely;

$$\begin{aligned}\hat{D}_{jls} &:= (2s)^{-1} \int_{X_{(l-s)}}^{X_{(l+s)}} \varphi_j(x)dx \qquad (4.2.6)\\ &= (2s)^{-1}(\sqrt{2}/\pi j)[\sin(\pi j X_{(l+s)}) - \sin(\pi j X_{(l-s)})], \quad j > 0\end{aligned}$$

and $\hat{D}_{0ls} = (2s)^{-1}(X_{(l+s)} - X_{(l-s)})$.

Surprisingly, the estimator (4.2.5) implies asymptotically efficient estimation as $n \to \infty$ and outperforms the sample mean estimator (4.2.3) for the case of a random design. This is why we shall use this slightly more complicated estimator.

Then, all the steps defined in the previous section of computing the universal estimator $\hat{f}$ are the same with the only modification in (4.1.8)–(4.1.9), where now in place of the residuals $Y_{(l)} - \tilde{f}_J(X_{(l)})$ the weighted residuals $n(Y_{(l)} - \tilde{f}_J(X_{(l)}))\hat{D}_{0ls}$ are used.

The underlying idea of using these weighted residuals is that they allow us to estimate the coefficient of difficulty,

$$d := \int_0^1 \frac{\sigma^2(x)}{h(x)} dx, \tag{4.2.7}$$

for heteroscedastic regression; see Exercise 4.2.9.

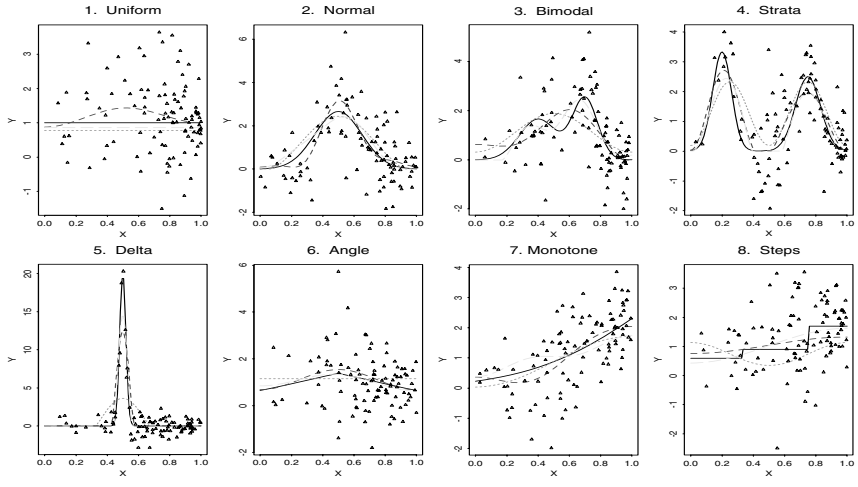

FIGURE 4.5. The universal estimates for heteroscedastic random-design regression with the design density being the Monotone, the scale function being the Angle, and standard normal ε_l. The dotted, short-dashed, and long-dashed lines correspond to sample sizes 50, 100, and 200; scatter plots for $n = 100$ are shown by triangles, and underlying regression functions are shown by solid lines. {Let us review all the arguments given below in square brackets. The arguments *set.n* and n control the sample sizes of the simulated data sets and the scattergrams, respectively. The argument *desden* controls the choice of design density $h(x)$, and the choice of scale function $\sigma(x) = \sigma f_j(x)$ is controlled by the arguments *sigma*=σ and *scalefun*=j. Other arguments control the coefficients of the universal estimate. The arguments *s0* and *s1* control the coefficients s_0 and s_1, which define s used in the estimate (4.2.5) of the Fourier coefficients. Recall that s is the rounded-up $s_0 + s_1 \ln(\ln(n + 20))$. Arguments *cJ0* and *cJ1* control the coefficients c_{J0} and c_{J1}, which define J_n used for choosing the optimal cutoff (4.1.11). Recall that J_n is the rounded-down $c_{J0} + c_{J1} \ln(n)$. The arguments *cJM* and *cT* control the coefficients c_{JM} and c_T, which define the high-frequency part of the estimate (4.1.13). The argument *cB* controls the coefficient c_B in the bump-removing procedure discussed in Section 3.1. Finally, r is used to find a pilot cutoff in the procedure (4.1.10).} [*set.n=c(50,100,200), n=100, desden=7, sigma=1, scalefun=6, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT =4, cB=2, r=2*]

Data-driven estimates $\hat{f}$ for data sets, generated similarly to the data sets shown in Figure 4.4, are plotted in Figure 4.5. The estimates resemble those for homoscedastic regression, but it is fair to say that they are worse, and this is apparent for the smallest sample sizes. The reason is that for this particular case the coefficient of difficulty, calculated according to (4.2.7), is $d = 1.5$. Also, on a top of this complication, in many particular cases only a few realizations are observed near the left edge. For instance, Figure 4.5.4 shows a particular case where only 7 observations (from a hundred!) fall in the region $x \le 0.25$, i.e., a quarter of the domain is covered by only

7 observations from a hundred. This is what may make a particular data set of heteroscedastic regression so complicated for analysis.

Finally, it is worthwhile to discuss the coefficient of difficulty d defined in (4.2.7). The coefficient does not depend on the underlying f; thus it is possible to find an optimal design density that minimizes this coefficient. A simple calculation (see Exercise 4.2.8) shows that the optimal design density $h^*(x)$ and the corresponding minimal coefficient of difficulty d^* are defined by the formulae

$$h^*(x) := \frac{\sigma(x)}{\int_0^1 \sigma(x)dx} \quad \text{and} \quad d^* := \Big(\int_0^1 \sigma(x)dx\Big)^2. \tag{4.2.8}$$

Of course, the optimal design density $h^*(x)$ depends on the scale function $\sigma(x)$, which is typically unknown. This is one of many reasons why the next section is devoted to estimation of the scale function.

4.3 Estimation of Scale Function

Consider the model (4.2.1) with an additional assumption that the random variable ε, which generates the iid $\varepsilon_1, \dots, \varepsilon_n$, has a finite fourth moment, i.e., $E\{\varepsilon^4\} < \infty$. Also recall that $E\{\varepsilon\} = 0$ and $E\{\varepsilon^2\} = 1$. The objective of this section is to estimate the scale function $\sigma(x)$. To avoid possible confusion, let us set $g(x) := \sigma^2(x)$ and recall that $\sigma(x)$ is nonnegative.

We begin the discussion with recalling the idea of estimating the scale parameter σ in the classical parametric location-scale model $Y = \theta + \sigma\varepsilon$ where n iid realizations $Y_1, \dots, Y_n$ of Y are given. The customarily used estimate of the squared scale parameter $g := \sigma^2$ is

$$\hat{g} := n^{-1} \sum_{l=1}^{n} (Y_l - \bar{\theta})^2, \quad \text{where } \bar{\theta} := n^{-1} \sum_{l=1}^{n} Y_l. \tag{4.3.1}$$

This classical parametric estimate has two steps. The first step is to estimate the location parameter θ, and usually the sample mean estimate $\bar{\theta}$ is used. Then, this estimate is subtracted from the observations. Note that if $\bar{\theta} \approx \theta$ (this notation means that $\bar{\theta}$ is approximately equal to θ), then $Y_l - \bar{\theta} \approx \sigma\varepsilon_l$, and thus $\hat{g}$ in (4.3.1) is again the sample mean estimate of g. Indeed, write $Z'_l := (Y_l - \bar{\theta})^2 \approx g + g(\varepsilon_l^2 - 1)$, and because by assumption $E\{\varepsilon^2\} = 1$, the estimate $\hat{g}$ in (4.3.1) is a sample mean estimate.

This idea can be straightforwardly expanded to the nonparametric case. The first step is to estimate f (the location function) by the estimate $\tilde{f}$ suggested in the previous section. Then statistics Z_l are calculated by the formula

$$Z_l := (Y_l - \tilde{f}(X_l))^2. \tag{4.3.2}$$

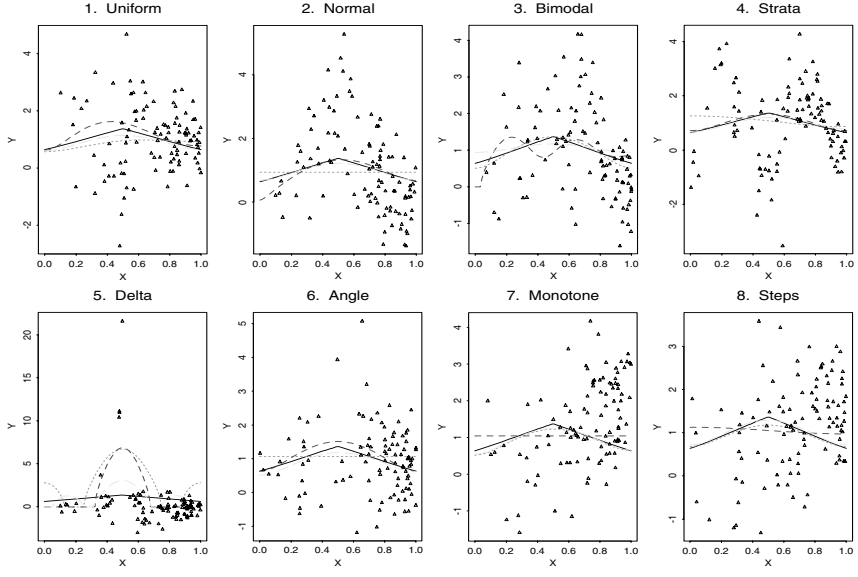

FIGURE 4.6. The universal estimates of the scale function (shown by solid lines), which is the Angle function. The model is a heteroscedastic random design regression (4.2.1) with design density being the Monotone and standard normal error ε. Underlying regression functions are the corner functions. The dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. Scatter plots for $n = 100$ are shown by triangles. {The scale function is $\sigma f_j(x)$ where j =*scalefun.*} [*set.n=c(50,100,200), n=100, desden=7, sigma=1, scalefun=6, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT =4, cB=2, r=2*]

Note that as in the parametric case,

$$Z_l \approx g(X_l) + g(X_l)(\varepsilon_l^2 - 1). \tag{4.3.3}$$

The relation (4.3.3) resembles the heteroscedastic nonparametric regression model discussed in Section 4.2, and thus the universal estimator of that section can be used directly for the pairs (Z_l, X_l) in place of (Y_l, X_l). This gives us the estimate $\hat{g}(x)$.

Finally, the universal estimate of the scale function $\sigma(x)$ is defined by $\hat{\sigma}(x) = \sqrt{(\hat{g}(x))_+}$. Recall that $(x)_+$ denotes the positive part of x.

Let us see how this procedure works for the case, considered in Section 4.2, of the Angle scale function, the Monotone design density, and a standard normal ε. Estimates of the Angle scale function are plotted in Figure 4.6. Here the data sets are simulated similarly to sets shown in Figure 4.5. To assess the complexity of the problem, recall that the issue here is to evaluate the spread of responses around an unknown underlying regression curve as a function in predictor. Keeping this in mind, the universal estimator performs surprisingly well. Even for $n = 50$ (dotted lines) we get some impression of the Angle scale function, and the particular cases of the

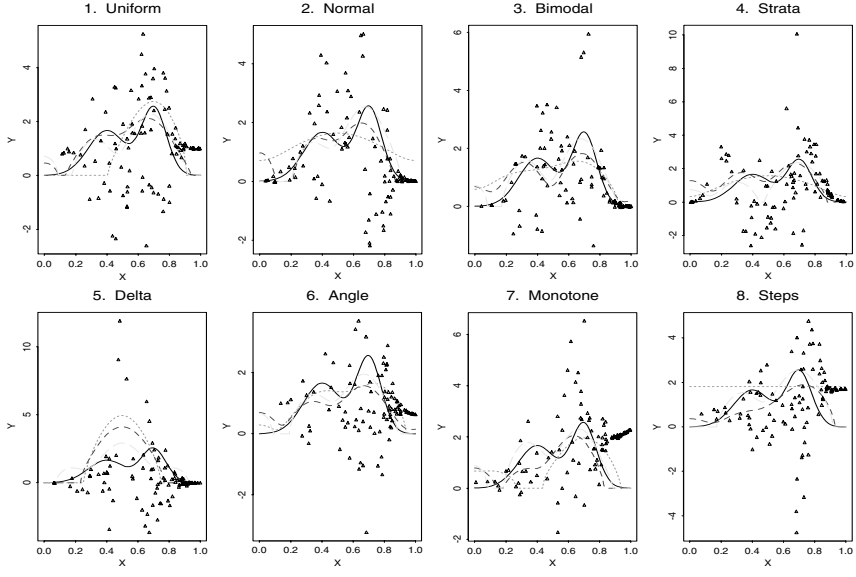

FIGURE 4.7. This figure is similar to Figure 4.6, only here the Bimodal is the scale function. {This figure was created by the call **ch4(f=6, scalefun=3)**.}

Bimodal, the Monotone, and the Steps are good. Also recall that the Angle has been a very complicated corner function for the density and the regression estimation discussed in Sections 3.1 and 4.2. The case of the Delta regression function is a disappointment; however, this is a very special case, where a "poor" estimation of the regression function near $x = 0.5$ is the issue. On the other hand, estimates in Figure 4.6.8 are surprisingly good, where despite typically poor estimation of the underlying regression function, the scale function is recovered well. Here even the short-dashed line is relatively not bad in comparison with estimates shown in Figures 4.3.6 and 4.5.6. The short-dashed line in the Bimodal diagram is a disappointment, but it is a good example of what one may get if the underlying regression function is poorly estimated.

What will be the quality of estimation of a scale function like the Bimodal? Will we be able to see the modes? To answer these questions, we just repeat Figure 4.6 with the Bimodal used in place of the Angle. A particular outcome is shown in Figure 4.7. Apart from the left tails, the estimator performs well, keeping in mind the previously discussed complexity of the problem. Recall that the left tail phenomenon is due solely to the Monotone design density; there is nothing radical that may be done to improve the estimation. Thus, for a heteroscedastic regression it is worthwhile to visualize simultaneously an underlying design density (using the estimator 3.1.15) and a particular scattergram, because this allows us to be alert to the possibility of this phenomenon.

4.4 Wavelet Estimator for Spatially Inhomogeneous Functions

In this section we consider the equidistant regression model of Section 4.1 and explain how the universal estimator of Section 3.1 (and respectively the estimator of Section 4.1) can be used for the case of a wavelet basis. Because wavelet bases are most interesting for the case of spatially inhomogeneous functions, in this section we use a new set of corner functions that represents different types of spatial inhomogeneity. Another interesting aspect of this section is that we discuss how to compare two data-driven estimators using Monte Carlo simulations. In particular, in this section we compare the universal estimator (we shall often refer to it as Universal), which is defined below, with an excellent data-driven wavelet estimator SureShrink supported by the S+WAVELETS toolkit and developed for wavelet bases.

A review of Section 2.1, specifically the part about the Haar basis, and Section 2.5 is recommended.

We begin with several examples that shed light on the problem. Results of two numerical experiments are shown in the two columns of Figure 4.8. Let us begin with the left column. The top diagram is a scatter plot with connected points (time series) based on 1024 equidistant observations of an underlying signal plus normal errors. The signal-to-noise ratio (snr) is 3, and this small ratio makes the setting very challenging because the snr defines σ in (4.1.1) by the formula $\sigma = \mathrm{sdev}(f)/\mathrm{snr}$, where $\mathrm{sdev}(f)$ is the sample standard deviation for $\{f(X_1), \dots, f(X_n)\}$.

Is an underlying signal recognizable? It is clear that there is a sine-like trend, but all other "fine" details are not so obvious. Now, let us look at the diagram below. Here a signal, recovered by Universal, is shown. While the trend of the estimate looks rather satisfactory, the pronounced vertical spike "spoils" the general picture. After all, there are plenty of similar vertical lines in the noisy signal, so it looks as if this spike is simply unfiltered noise. Now let us look at the third diagram, where a signal recovered by SureShrink is depicted. This estimate looks very reasonable: There are no pronounced vertical spikes, and low-frequency oscillations resemble the possible performance of a smart moving average filter. Overall, due to that vertical spike, the estimate depicted by SureShrink looks more reasonable than the universal estimate.

Now let us look at the underlying signal "singcubic" shown in the bottom diagram. It reveals that Universal performs essentially better than SureShrink in terms of both recovering the smooth background and depicting the pronounced vertical spike, which has a width of about 0.005.

Since the underlying signal is known, you can notice that the SureShrink estimate also shows the vertical spike, but it is so drastically shrunk that it is almost invisible among the low-frequency oscillations that surround the spike.

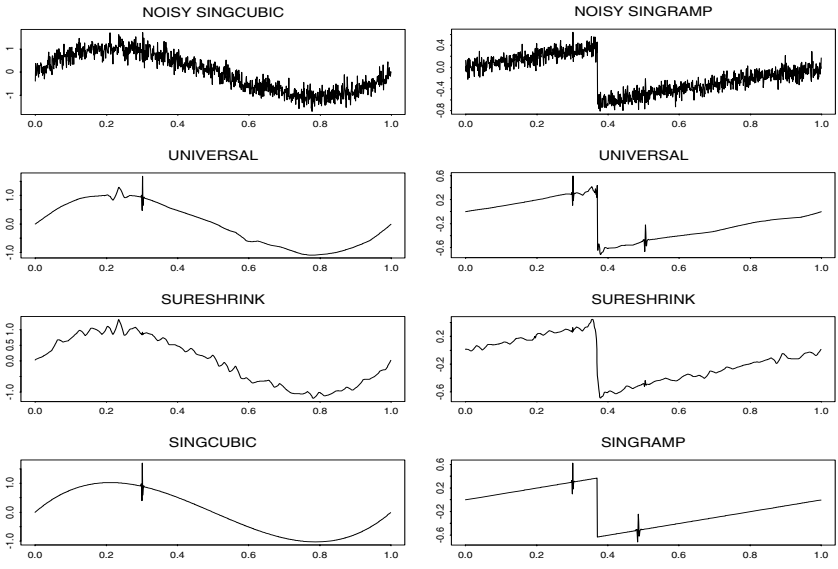

FIGURE 4.8. Recovery of two spatially inhomogeneous signals by universal and SureShrink data-driven wavelet estimators. The signal-to-noise ratio is 3, and $n = 1024$ equidistant observations are made. The wavelet is Symmlet 8. {The signal-to-noise ratio is controlled be the argument *snr*. The choice of a wavelet basis is controlled by the argument *wavelet*. The arguments *j0, cJ, cT*, and *cU* control the coefficients j_0, c_J, c_T, and c_U of the universal estimate (4.4.2). Recall that to reproduce a figure with wavelets, the S+WAVELETS module should be installed by calling > **module(wavelets)**. [*wavelet= ″s8″, snr=3, n=1024, j0=6, cJ=1, cT=4, cU=1*]

On the other hand, the "attitude" of the estimators toward the pronounced noisy bump near the point 0.2 is quite the opposite: Universal smoothes this bump, and SureShrink leaves it untouched. This reveals that these two adaptive estimators use different strategies for low and high frequencies.

Now let us consider the right column of diagrams in Figure 4.8, where a different signal is considered. We begin the discussion with the top diagram showing the data. It is given that the underlying signal has a smooth background, one pronounced jump, and two pronounced vertical spikes; can you recognize such a signal by analyzing this diagram? Now let us see how the estimators have solved this puzzle; to judge the "answers," the underlying signal "singramp" is shown in the bottom diagram. As we see, Universal again outperforms SureShrink in both the quality of smoothing the low-frequency parts of the signal and depicting the two vertical spikes. Note that SureShrink also indicates these spikes, but they are drastically

shrunk and difficult to recognize. Also, we see that Universal competes with SureShrink in the visually aesthetic presentation of the jump discontinuity.

One of the important characteristics of any wavelet estimator is its data compression property, that is, how many nonzero wavelet coefficients are used to reconstruct a signal. To analyze this property, we use the discrete wavelet transform (DWT) function of the S+WAVELETS toolkit, which exhibits wavelet coefficients $d_{j,k}$ and $s_{j,k}$ of an orthogonal wavelet partial sum

$$f_{j_0}(t) := \sum_{k=1}^{n/2^{j_0}} s_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=1}^{j_0}\sum_{k=1}^{n/2^j} d_{j,k}\psi_{j,k}(t). \tag{4.4.1}$$

Here we follow the notation and notions of this toolkit, namely, a given data set is an equidistant regression (a regular time series), j_0 is the number of multiresolution components (or *scales*) used, and the functions $\phi_{j_0,k}$ and $\psi_{j,k}$ are wavelet functions that are generated from the father wavelet (or scaling function) ϕ and the mother wavelet (or wavelet function) ψ through scaling and translation as follows: $\phi_{j,k}(t) := 2^{-j/2}\phi(2^{-j}t - k)$ and $\psi_{j,k}(t) := 2^{-j/2}\psi(2^{-j}t - k)$. To simplify the discussion, only dyadic sample sizes $n = 2^l$ are considered.

Note that the coefficients $d_{1,k}$ correspond to the finest scale and $d_{j_0,k}$ to the coarsest one. On the other hand, it is worthwhile to recall that roughly speaking, the coarsest scales represent the underlying smooth behavior of a signal, while the finest scales are responsible for its high-frequency behavior (fine details). Also, due to the dyadic nature of a wavelet basis, the number of nonzero wavelet coefficients on a scale can be equal to the total number of wavelet coefficients on the coarser scales (in other words, a good data compression property of a wavelet estimator is primarily defined by how it deals with the finest scales).

Now we know what to look for in the DWT. The four columns of Figure 4.9 show us the DWT of the "bumps" signal, noisy signal, universal estimate, and SureShrink estimate. The left column of plots displays the DWT of the signal "bumps." The original signal is plotted in the top row. The wavelet coefficients are plotted in the remaining rows, going downward from the finest scale **d1** to the coarsest scales **d6** and **s6** in the bottom two rows. The wavelet coefficients are plotted as vertical lines extending from zero, they are plotted at approximately the position of the corresponding wavelet function. The wavelet coefficients for mother wavelets are plotted on the same vertical scale, so that the relative importance can be measured by comparing the magnitudes (but the scale is different for the signal and **s1** level). The second column shows the DWT of the signal plus a Gaussian white noise. The third column shows the DWT of the universal estimate, and the right column shows the DWT of the SureShrink estimate.

We see that both estimates nicely capture the bumps, but it is fair to say that the smooth background is better shown by Universal (just compare

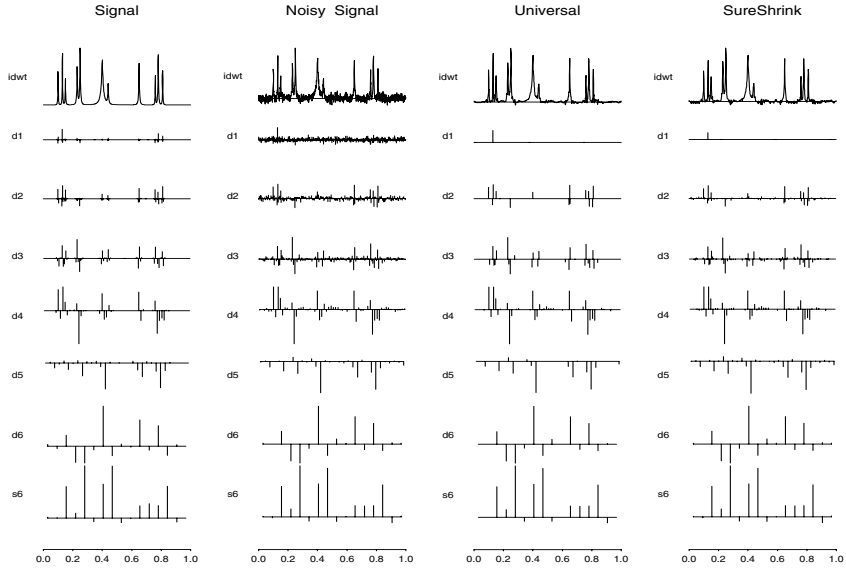

FIGURE 4.9. DWT of the signal "bumps," noisy "bumps," universal estimate, and SureShrink estimate. [*signal= "bumps",wavelet= "s8", snr=3, n=1024, j0=6, cJ=1, cT=4, cU=1*]

the valleys). Now let us look at the corresponding DWT. It is worthwhile to recall that the estimators have at hand the noisy wavelet coefficients shown in the second column, so let us see how they use them. Let us begin with the coarsest level **s6**. As you see, Universal smoothes (shrinks) the noisy coefficients, while SureShrink simply keeps all the noisy coefficients. Such a strategy is also clearly pronounced in the scale **d5**. On the finest scales the estimators "exchange" their strategies of smoothing, which is clearly seen in scale **d1**. Also, on the finest scales Universal is more "picky" in its choice of the coefficients (just look at **d3** or **d2**), but then it leaves them untouched. On the other hand, SureShrink keeps essentially a larger number of shrunk noisy coefficients. As a result, it can be expected that Universal compresses data better than SureShrink, and below, an intensive Monte Carlo study will confirm this conclusion.

Let us summarize our preliminary conclusions. We have seen that Universal and SureShrink employ quite opposite algorithms on how to pick and smooth (shrink) wavelet coefficients, and this leads to rather different approximation and data compression properties that favor the Universal estimator. Of course, we have seen only several examples. Are they repeatable? This is the next issue that we would like to explore.

First, let us define the universal estimator using the notation of the S+WAVELETS toolkit. For an equidistant regression of a sample size n, which is interpreted by S+WAVELETS as a regular time series of length

n, the S+WAVELETS toolkit allows one to calculate estimated wavelet coefficients $\hat{d}_{j,k}$ and $\hat{s}_{j,k}$. Using the same notations as in (4.4.1), define a nonadaptive (since a cutoff J is not specified) universal estimator as

$$\tilde{f}(t, J) \tag{4.4.2}$$

$$:= \sum_{j=1}^{j_0-J} \sum_{k} \hat{d}_{j,(k)} I_{\{|\hat{d}_{j,(k)}|>\hat{\sigma} n^{-1/2} \min(c_T 2^{j_0-J-j}, (2c_U \log(n))^{1/2})\}} \psi_{j,(k)}(t) \tag{4.4.3}$$

$$+ \sum_{j=j_0-J+1}^{j_0} \sum_{k=1}^{n/2^j} (1-\hat{\sigma}^2/n\hat{d}_{j,k}^2)_+ \hat{d}_{j,k}\psi_{j,k}(t) + \sum_{k=1}^{n/2^{j_0}} (1-\hat{\sigma}^2/n\hat{s}_{j_0,k}^2)_+ \hat{s}_{j_0,k}\phi_{j_0,k}(t). \tag{4.4.4}$$

In (4.4.3) the summation over k is from 1 to $c_J n 2^{-2(j_0-J)-1+j}$.

As with SureShrink, the default value of j_0 is 6; the estimate $\hat{\sigma}^2$ is calculated via the finest-scale wavelet coefficients $d_{1,k}$ by the robust median-scale estimator (4.1.9) with $\hat{d}_{1,k}$ used in place of the residuals. Here $\hat{d}_{j,(k)}$ are ordered (descending) empirical wavelet coefficients (i.e., $|\hat{d}_{j,(1)}| \ge |\hat{d}_{j,(2)}| \ge \cdots$), and $\psi_{j,(k)}(t)$ are the corresponding mother wavelets. The default values of the coefficients are $c_J = 1$, $c_T = 4$, and $c_U = 1$. Also recall that $(x)_+$ denotes the positive part of x.

As in (3.1.14) and (4.1.13), we see two terms (4.4.3) and (4.4.4) in the universal estimate that are high- and low-pass filters. The low-pass filter (4.4.4) uses smoothing, while the high-pass filter (4.4.3) uses thresholding. There is a tiny difference between the thresholding procedures used for the cosine basis in (3.1.14) and here in (4.4.3) for a wavelet basis. For the cosine basis the threshold level is the same for all components. Here, because a wavelet basis has a dyadic structure, the threshold level exponentially increases, and the maximal number of nonzero top wavelet coefficients exponentially decreases for finer scales (as the frequency increases). Apart from this modification, (4.4.2) is identical to the universal Fourier estimate.

Finally, a data-driven cutoff $\hat{J}$ is chosen absolutely similarly to (4.1.11) by the procedure of an empirical risk minimization,

$$\hat{J} := \text{argmin}_{1\le J\le j_0} \Big\{ 2\hat{\sigma}^2 N_J - \int_0^1 (\tilde{f}(t, J))^2 dt \Big\}. \tag{4.4.5}$$

Here N_J is the number of nonzero wavelet coefficients used by $\tilde{f}(t, J)$. The universal wavelet estimator $\tilde{f}(t, \hat{J})$ is defined.

Let us briefly recall the algorithm of the estimator SureShrink (a more detailed discussion is given in Section 7.4, where the name is also explained). SureShrink chooses (one at a time) threshold levels for 4 finest-resolution scales **d1**–**d4**, while for the coarsest scales **s6**, **d6**, and **d5** threshold levels are zero by default, i.e., all noisy wavelet coefficients on these coarsest scales are kept unfiltered. The procedure of choosing a data-driven threshold level

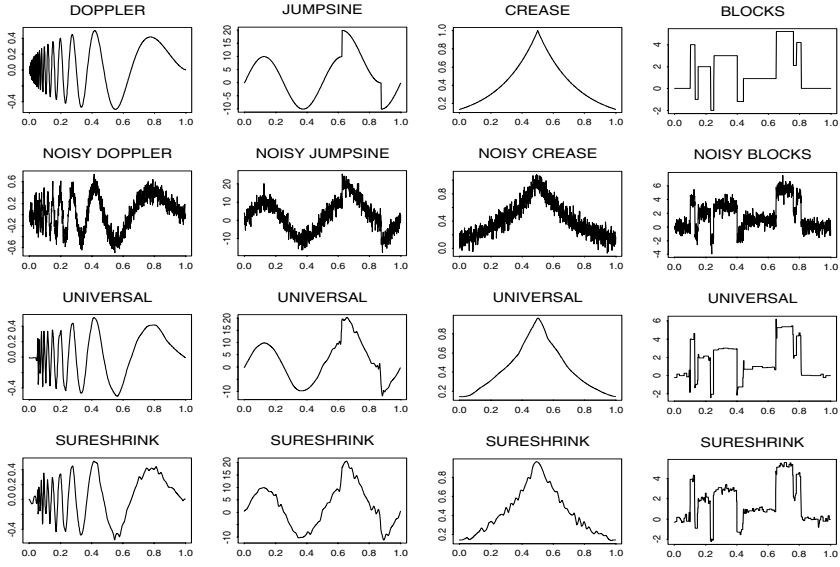

FIGURE 4.10. Four different spatially inhomogeneous signals and the corresponding noisy signals, universal, and SureShrink estimates. [*set.signal=c("doppler", "jumpsine", "crease", "blocks"), set.wavelet=c("s8", "s8", "s8", "haar"), snr=3, n=1024, j0=6, cJ=1, cT=4, cU=1*]

$\hat{t}_j$ for the jth scale is an empirical risk minimization. The shrinkage is soft thresholding, where an estimated wavelet coefficient is $\mathrm{sgn}(\hat{d}_{j,k})(|\hat{d}_{j,k}| - \hat{t}_j \hat{\sigma} n^{-1/2})_+$. The motivation of using the soft thresholding instead of a hard thresholding is that the soft thresholding is continuous in $\hat{d}_{j,k}$, and this implies an easier procedure for computing the empirical risk.

Now we are ready to begin a Monte Carlo comparison of these two data-driven estimators. First, we begin with the analysis of the visual appeal of particular estimates for 4 signals supported by S+WAVELETS and called "doppler," "jumpsine," "crease," and "blocks." These signals represent functions with different types of spatial inhomogeneity and are "classical" corner functions used in the wavelet literature. In the examples the default Symmlet 8 wavelet is used for all signals except "blocks," where the Haar wavelet is used. Recall that these wavelets are customarily used by statisticians.

Signals, noisy signals, universal, and SureShrink estimates are shown in Figure 4.10. For "doppler" both estimates give a fair visualization of the time-varying frequency of the signal, and the universal apparently better restores the smooth background. The signal "jumpsine," with two pronounced jumps (two change-points of the first order), is shown in the second column. Here the universal estimate nicely restores these two jumps (at least not worse than the SureShrink). Note that for other simulations

Universal may "catch" spikes (as it did in Figure 4.8) created by noise because the signal-to-noise ratio is extremely small. Exercise 4.4.4 is devoted to choosing optimal coefficients, which may either improve the ability of Universal to search after spikes or attenuate it.

The signal "crease," which has one pronounced change-point of the second order (the derivative has a jump), is shown in the third column. Here the universal estimate outperforms the SureShrink estimate in all aspects of signal restoration. An interesting situation occurs with the signal "blocks," whose estimation is the trademark of SureShrink. Here the universal estimate looks "cleaner" over the tops of the blocks, but the SureShrink is better near some change-points where the overshootings hurt the universal estimate. Recall our discussion in Section 2.5 that wavelets are not immune against the Gibbs phenomenon (overshooting).

So far, we have discussed only the visual appeal of restored signals. Since the compared estimators have their own pluses and minuses, and for one particular simulation SureShrink looks better and for another the outcome may flip over, we should develop a more consistent method for comparison of the two estimators. One of the possibilities is based on using an intensive Monte Carlo study. The approach is as follows.

Let us, as in Section 3.2, define an experiment as a combination of a sample size n and an underlying signal. Consider sample sizes from the set $\{512, 1024, 2048\}$ and signals from the set {"doppler", "jumpsine", "crease", "blocks", "cubic"}; overall, 15 different experiments. Here the signal "cubic" is a smooth cubic polynomial $f(x) = 32x(x-1)(2x-1)/3$ supported by S+WAVELETS. Then for every experiment we (i) repeat independently 1000 numerical simulations that are similar to the above-discussed; (ii) calculate the sample mean and standard deviation (over these 1000 simulations) of the ratios: the integrated squared error (ISE) of the universal estimate / the integrated squared error of SureShrink; (iii) calculate the sample mean (again over these 1000 simulations) of the ratios: number of nonzero wavelet coefficients used by universal / number of nonzero wavelet coefficients used by SureShrink.

Step (ii) allows us to compare the estimates in terms of integrated squared errors (if the ratio is smaller than 1, then the Universal performs better, and vice versa), while step (iii) allows us to compare the data compression properties of these adaptive estimators (again, if the ratio is smaller than 1, then the Universal compresses data better, and vice versa).

The results of this intensive Monte Carlo study are presented in Table 4.1. They confirm our preliminary conclusion made via visualization of particular estimates in Figures 4.8–4.10. We see that for all the experiments with the exception of the case $n = 512$ and the signal "jumpsine," the approximation property of the universal estimator is better, and the difference in the performances becomes more pronounced for the larger sample sizes.

Moreover, this nice approximation property is combined together with the superior data compression property of the universal estimator.

Table 4.1. *Sample Means (Sample Standard Deviations) of Ratios of ISE and Data Compression: Universal/SureShrink*

Sample Size	Mean Ratio (Standard Deviation) of ISE				
	"doppler"	"jumpsine"	"crease"	"blocks"	"cubic"
512	.95 (.13)	1.06 (.3)	.77 (.6)	.6 (.1)	.68 (.5)
1024	.84 (.09)	.85 (.2)	.5 (.3)	.6 (.1)	.39 (.2)
2048	.67 (.1)	.66 (.09)	.39 (.09)	.57 (.1)	.34 (.1)
	Mean Ratio of Data Compression				
512	.57	.81	.42	.66	.36
1024	.54	.51	.31	.48	.28
2048	.5	.34	.27	.34	.25

This study explains how to compare two data-driven estimators via an intensive Monte Carlo study. Note that here again we used the approach of analyzing the quality of estimation via a set of corner (test) regression functions. In general, this set should be chosen based on prior experience or intuition about possible underlying regression functions.

4.5 Case Study: Binary and Poisson Regressions

In many practical settings the regression model is slightly different from the model (4.2.1). Namely, for a pair of observations (X, Y) with the predictor X and the response Y, the regression function $f(x)$ is defined as the conditional expectation of Y given X, that is,

$$f(x) := E\{Y|X = x\}. \tag{4.5.1}$$

Thus, the regression function f is interpreted as an average value of the response Y given $X = x$. (Recall that the notion of the conditional expectation and its properties are reviewed in Appendix A.) As in Section 4.1, depending on the design of predictors, the regression can be either fixed or random.

This approach has a clear geometric sense, namely, for a given x one searches for means of responses and then connects the means by a regression line.

Note that all the previously considered models could be written as (4.5.1) because the error ε had been supposed to be independent of predictor X, and therefore $E\{Y|X = x\} = f(x) + E\{\varepsilon|X = x\} = f(x)$.

Model (4.5.1) can be rewritten in a form that resembles the previously studied additive models, namely as

$$Y = f(X) + \eta(X, f), \quad \text{where} \quad E\{\eta(X, f)|X = x\} = 0. \tag{4.5.2}$$

The difference between models (4.5.2) and (4.2.1) is that in model (4.5.2) the additive error can be rather complicated and depend on both the regression function and the predictor. Nevertheless, we shall see that the universal estimator $\hat{f}$ of Section 4.2, developed for the case of heteroscedastic regression, can be used again for model (4.5.2) whenever $E\{\eta^2|X=x\} \le C < \infty$.

Let us consider two classical examples of such a setting.

• **Example of Binary Regression.** First, let us recall the notion of a classical Bernoulli random variable. Suppose that a trial, or an experiment, whose outcome can be classified as either a "success" or as a "failure," is performed. Let $Y = 1$ when the outcome is a success and $Y = 0$ when it is a failure. Then the probability mass function of Y is given by $P(Y = 1) = f$ and $P(Y = 0) = 1-f$, where a constant f, $0 \le f \le 1$, is the probability that the trial is a success. Then a direct calculation shows that the expectation of Y is equal to f and the variance to $f(1 - f)$.

Note that the outcome of the trial can be written as $Y = f + \eta$, where $E\{\eta\} = 0$ and $E\{\eta^2\} = f(1 - f)$. Thus, if n independent trials are repeated, then the problem of estimation of the probability of success f can be considered as a nonparametric regression (4.5.1), where $f(x)$ is constant.

Now let us make this Bernoulli setting more complicated. Assume that the probability of a success f is a function of a predictor x, that is, $f = f(x)$. Then, the pair (X, Y) is given such that $Y = 1$ with the probability $f(X)$ of a success and $Y = 0$ with the probability $1 - f(X)$ of a failure. The problem is to estimate the function $f(x)$.

Such a regression problem is an absolutely natural generalization of the Bernoulli setting, and it occurs in a number of fields from biological assay and gambling to the testing of explosives and reliability theory. Several specific examples are as follows. Let Y be equal to 1 if the level of patient's cholesterol is less than a given threshold after 100 days of taking a dosage X of some medicine. Then $f(X)$ is the probability of a successful treatment as a function of the dosage X. Repeated experiments with hitting a target by a missile with different distances X between the point of launching the missile and the target is another example where $f(X)$ is the probability of success that is a function of the distance. One more example is the probability of dropping out of graduate school as a function of GRE score.

For such a problem the response can be written as $Y = f(X) + \eta(X, f)$, where $E\{\eta(X, f)|X = x\} = 0$ and $E\{(\eta(X, f))^2|X = x\} = f(x)(1-f(x) \le \frac{1}{4}$, so the universal estimator of Section 4.2 can be used directly.

Let us check how this universal data-driven estimator performs for the case of the corner functions being the probability of success. Here all the corner functions, shown in Figure 2.1, with the exception of the Uniform are divided by their maximal value, and in place of the Uniform the function $f(x) = \frac{3}{4}$ is used (but here we again refer to this f as the Uniform).

Estimates for Monte Carlo simulated data, where underlying regression functions are the corner functions and predictors are generated according

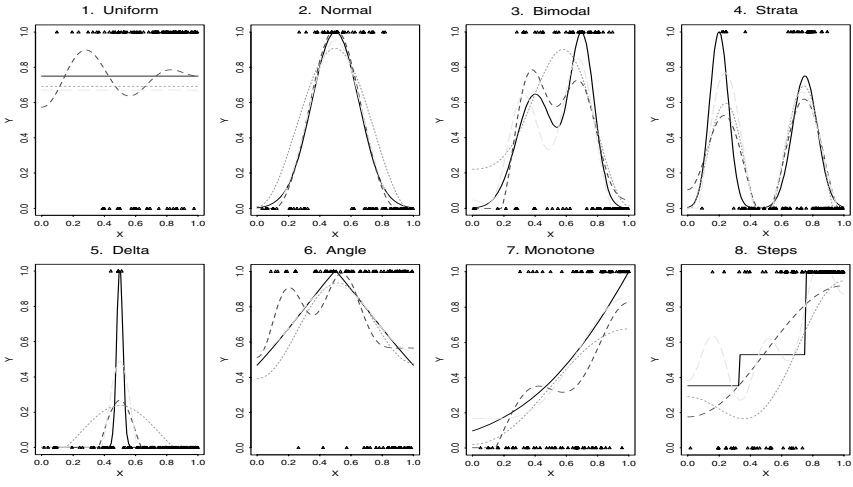

FIGURE 4.11. The estimates for binary random-design regression with design density being the Monotone: Dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$; scatter plots for $n = 100$ are shown by triangles; the underlying regression functions are shown by solid lines. [*set.n=c(50,100,200), desden=7, n=100, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

to the Monotone density, are shown in Figure 4.11. For the case of $n = 100$ the scatter plots are shown by triangles; these plots show the complexity of the binary regression. For instance, for the Normal it is difficult even to realize that the scattergram corresponds to a regression function symmetric about $x = 0.5$, so it is surprising how well the estimator performs.

Several other particular cases are even more interesting. Let us begin with the Strata diagram and recall that we look at the short-dashed line, which corresponds to the scatter plot. Note that here the universal estimator should "create" the left tail of an estimate based just on several observations. More precisely, there are just 2 observations to the left of $x = 0.2$. And despite this obstacle, we can clearly see the shape of the Strata.

An even more impressive outcome is for the Angle. The estimate (short-dashed line) is clearly bad, but is it possible to suggest a better one for this particular data set? First of all, the estimate correctly shows the magnitude of the Angle. Second, the right tail is not really bad, keeping in mind the underlying observations. The left tail is the reason why we refer to this particular estimate as "bad." However, let us look at the data at hand. Any reasonable "handmade" estimate would increase as x changes from .4 to 0. Thus, the fact that the estimate correctly decreases near the left edge and moreover approaches the correct value at the left boundary is amazing. Overall, it is fair to say that the universal estimator performs well under

these very complicated circumstances. Also note that no "local" estimator can deal with such a setting.

Now let us derive the coefficient of difficulty for this setting. According to (4.2.7), the coefficient of difficulty for binary regression is

$$d = \int_0^1 f(x)(1 - f(x))h^{-1}(x)dx. \tag{4.5.3}$$

Note that $f(x)(1-f(x)) \le 0.25$ and this allows one to get an upper bound for the coefficient of difficulty. Also, according to (4.2.8), the optimal design density for binary regression is

$$h^*(x) = \frac{f(x)(1 - f(x)}{\int_0^1 f(u)(1 - f(u))du}. \tag{4.5.4}$$

• **Example of Poisson Regression.** First, let us recall the notion of a Poisson random variable. A random variable Y taking on one of the values $0, 1, 2, \dots$ is said to be a Poisson random variable with parameter f if for some $f > 0$, $P(Y = k) = e^{-f} f^k/k!$. Note that $E\{Y\} = f$ and $\mathrm{Var}(Y) = f$.

Customary examples of random variables that obey the Poisson probability law are as follows: The number of misprints on a page of a book; the number of wrong telephone numbers that are dialed in a day; the number of customers entering a shopping mall on a given day; the number of α-particles discharged in a fixed period of time from some radioactive material; the number of earthquakes occurring during some fixed time span.

It is not difficult to imagine that in all the above-mentioned examples the parameter f can depend on another measured parameter (predictor) X, and this defines the Poisson regression that satisfies (4.5.1) because $E\{Y|X = x\} = f(x)$. Also, $E\{(Y-f(X))^2|X = x\} = f(x)$, so the universal estimator of Section 4.2 can be used again. Note that here the coefficient of difficulty is

$$d = \int_0^1 f(x)h^{-1}(x)dx, \tag{4.5.5}$$

and the optimal design density is

$$h^*(x) = \frac{f(x)}{\int_0^1 f(u)du}. \tag{4.5.6}$$

An example of Monte Carlo simulations and the corresponding universal estimates is shown in Figure 4.12. Here the underlying regression functions and the design of predictors are identical to the binary regression case. Also, the scatter plots are shown by triangles for the case $n = 100$. The scatter plots are rather specific, so some training is necessary to get used to them. But the main rule of thumb is the same: The larger the response for $X = x$, the larger the underlying regression function $f(x)$. Keeping this in mind, we see that while the short-dashed estimate for the Normal is bad, it perfectly describes the underlying scattergram. The same may be

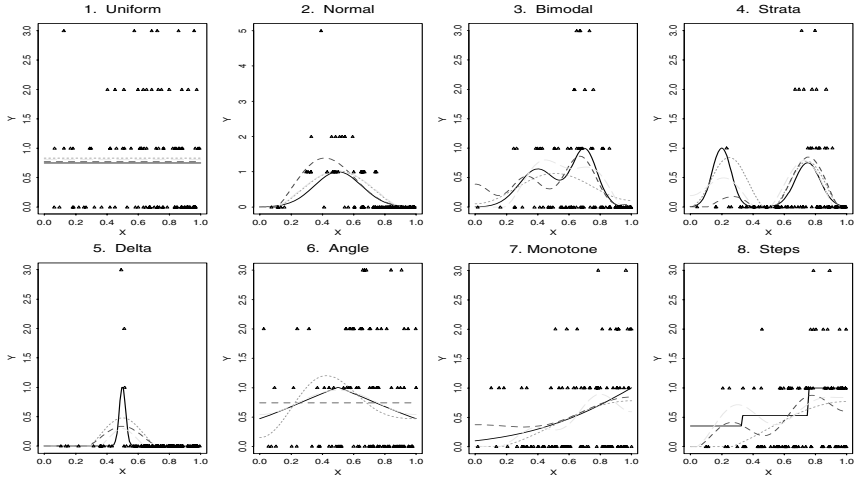

FIGURE 4.12. The estimates for Poisson random-design regression with the Monotone design density: Dotted, short-dashed, and long-dashed lines correspond to $n = 50$, $n = 100$, and $n = 200$; scatter plots for $n = 100$ are shown by triangles; the underlying regression functions are shown by solid lines. [*set.n=c(50,100,200), n=100, desden=7, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

said about the Strata. Here the short-dashed line shows the left stratum absolutely incorrectly. As a result, instead of a pronounced peak we see a small bump. But are there any data to indicate the correct stratum? There are no such data at all. The same may be said about the Monotone diagram. Also, look at the Angle diagram. Can you see any indication in the scatter plot on the underlying Angle regression function? The universal estimator sees none, and this looks like a reasonable answer. On the other hand, here the estimate based on 200 observations is almost perfect apart from the small flat tails.

4.6 Case Study: Quantile and Robust Regression

In many practically interesting cases the relationship between a predictor X and a response Y cannot be interpreted as a mean or a conditional mean, simply because these means do not exist. For instance, consider the model $Y = f(X) + \varepsilon$, where the additive error ε is distributed according to a Cauchy distribution with the density

$$f^{\varepsilon}(x) := \frac{k}{\pi(k^2 + x^2)}, \quad -\infty < x < \infty, \tag{4.6.1}$$

where k is the *scale* parameter.

Figure 4.13 illustrates the complexity of this setting by exhibiting scatter plots for the underlying corner regression functions, Cauchy error with the scale parameter $k = 0.5$, uniformly distributed predictors, and the regression model

$$Y = f(X) + \varepsilon. \tag{4.6.2}$$

The scatter plots are overlaid by least-squares regression lines.

First of all, please look at the scales for the responses. The eye is drawn to Figure 4.13.7, where one response is about -200. Clearly, this is an outlier (and this is the reason why the Cauchy distribution customarily serves as a test for the robustness of an estimator for the presence of outliers), and it may be easily removed. However, this does not solve the problem. Let us, for example, consider the diagram Strata. What are the outliers here? Because the underlying regression function is known, clearly the 3 observations with responses larger than 7 are outliers. On the other hand, the diagram Delta shows us that there are only 3 points that indicate the presence of the pronounced mode of the Delta, and these points should not be removed. In short, while for the case of smooth functions (like those studied by classical linear regression theory) one can suggest smart and relatively simple procedures for identifying outliers, this becomes an extremely difficult issue for the case of spatially inhomogeneous nonparametric curves.

What is so special about Cauchy random variables? The answer is that a Cauchy random variable has "heavy tails" and, as a result, has neither a variance nor an expectation. Moreover, a Cauchy random variable has the peculiar property that the sample mean of its iid realizations has the

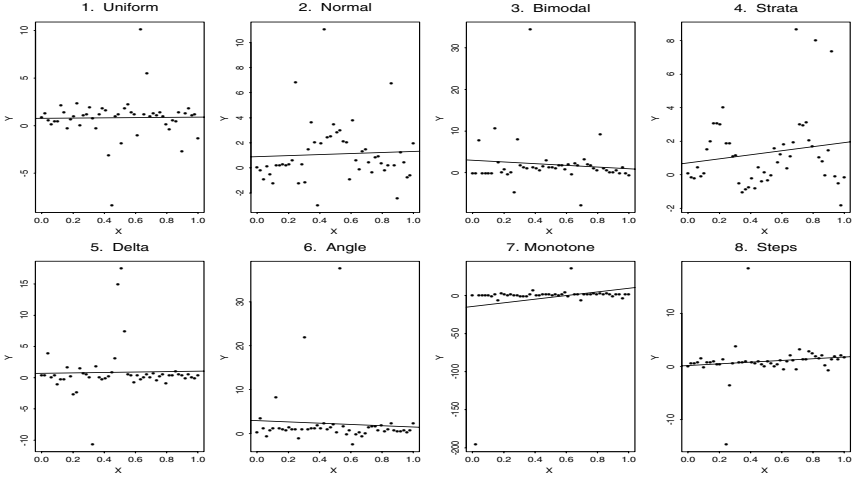

FIGURE 4.13. Simulated scattergrams based on 50 observations and overlaid by linear regression lines. The underlying regression functions are the corner functions; the errors have Cauchy distribution (4.6.1) with $k = 0.5$. [*k=.5, n=50*]

same distribution as the underlying Cauchy random variable. In short, the sample mean cannot serve as an estimator, and all the familiar asymptotic results such as the law of large numbers and the central limit theorem fail.

On the other hand, the probability density (4.6.1) of the Cauchy distribution is symmetric about zero, and therefore its median is equal to zero. Thus, the regression model (4.6.2) has a different and absolutely natural meaning here: The underlying regression function $f(x)$ is a curve such that in a small vicinity of a point x, half of the responses are larger and the other half smaller than $f(x)$. Apparently, this regression function is unaffected by any extreme responses in a set of data. Thus, the median regression may be an attractive alternative to the mean regression when one wants to describe the relationship between predictor and response under the assumption that the errors may have a distribution like Cauchy.

The caveat of such a regression is that the estimation of functions like the Delta will suffer. Also, it should be clear from the beginning that for traditional normal errors a median regression implies worse estimation. To shed light on the issue, consider the problem of estimation of a parameter θ based on n iid observations $Y_l = \theta + \varepsilon_l$, $l = 1, 2, \dots, n$, where ε_l are standard normal. If $\bar\theta_n := n^{-1} \sum_{l=1}^n Y_l$ is the sample mean and $\tilde\theta_n := \text{median}(\{Y_l, l = 1, 2. \dots, n\})$ is the sample median, then the ratio $E\{(\tilde\theta_n - \theta)^2\}/E\{(\bar\theta_n - \theta)^2\}$ is, for instance, 1.47 for $n = 20$ and increases to 1.57 as $n \to \infty$.

Now let us return to the search of an estimator for the median regression. The following explains the underlying idea of the recommended approach. It is known that a sample median for a Cauchy random variable is an unbiased estimate of the median, and it also has a finite second moment whenever the sample size is at least 5; see Exercise 4.6.2.

Thus, our strategy is as follows. First, we seek a moving sample median for neighbor predictors (assuming that an underlying regression function is smooth). Then we use the universal estimate of Section 4.2.

To make the first step, as in Section 4.2, the pairs of observations are arranged in ascending order according to the predictors. Recall that the notation for such pairs is $\{(Y_{(l)}, X_{(l)}), l = 1, 2, \cdots, n\}$, where $X_{(1)} \le \cdots \le X_{(n)}$. Then, for every l such that $m < l < n - m$ the sample median of $\{Y_{(l-m)}, Y_{(l-m+1)}, \dots, Y_{(l+m-1)}, Y_{(l+m)}\}$ is calculated and denoted by $Y'_{(l)}$, where m is the rounded-down $m_0 + m_1 \log(\log(n))$ with the default values $m_0 = 2$ and $m_1 = .3$. For boundary points we set $Y'_{(l)} = Y_{(m+1)}$ for $1 \le l \le m$ and $Y'_{(l)} = Y_{(n-m)}$ for $n - m \le l \le n$.

As a result, the new sequence of pairs (Y', X) of observations is created, and

$$Y' = f(X) + \eta, \tag{4.6.3}$$

where now $E\{\eta\} = 0$ and $E\{\eta^2\} \le C < \infty$. Then the universal estimator of Section 4.2 is used.

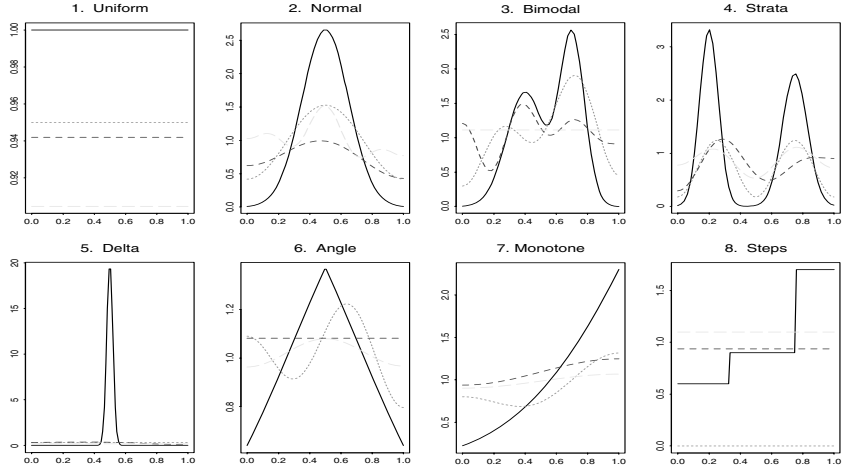

FIGURE 4.14. The estimates for median regression with Cauchy errors (the scale coefficient $k = 0.5$) and uniform design predictors: Dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. The underlying regression functions are shown by solid lines. {Here the arguments *m0* and *m1* control the coefficients m_0 and m_1 used in the formula for calculation of the window-width for the moving sample median.} [*set.n=c(50,100,200), k=.5, m0=2, m1=.3, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

Figure 4.14 depicts the estimates based on data sets simulated by uniformly distributed predictors and Cauchy errors with the scale parameter $k = 0.5$. It is not a surprise that the estimates are worse than for all the previously discussed settings, and the visualization of the Delta is lost. Here simply more observations are needed for a reliable estimation, and we leave this as Exercise 4.6.8.

Despite the poor outcomes for the small sample sizes, which have been predicted, there is no need to be too pessimistic about median regression. Cauchy error is a classical test example that is simply far too extreme for many applications. Some "milder" examples of test errors are given in Exercises 4.6.4–4.6.5.

The sample median is a characteristic of a data set that splits ordered observations into two equal parts. A sample α-quantile is another characteristic of a data set that divides it into two parts such that the αth proportion of observations is less and the $(1-\alpha)$th proportion is larger than this quantile. Classical examples of quantiles are the first ($\alpha = 0.25$) and the third ($\alpha = 0.75$) quartiles, which together with the median ($\alpha = 0.5$) divide a set of data into four equal parts.

There are at least two important reasons to study quantile regression. The first one is that in some cases an underlying regression function is

defined as a conditional αth quantile,

$$P(Y \le f_\alpha(x)|X = x) = \alpha. \tag{4.6.4}$$

(Median regression is a particular example with $\alpha = .5$.) The second reason is that for the case of an additive heteroscedastic regression

$$Y = f(X) + \sigma(X)\varepsilon, \tag{4.6.5}$$

it may be of interest to visualize both f and an αth quantile curve defined in (4.6.4) because together they shed light on the volatility function $\sigma(x)$. Also, in some cases quantile curves have their own merits. A particular example is tracking the price of a particular stock over a period of time where one is typically interested not only in the average price but also in its quantiles over the time period.

The approach of median regression can be straightforwardly extended to the case of quantile regression, only here a sample quantile estimator needs to be used in place of a sample median estimator; the parameter m is defined as the rounded-down $m_0 + m_1 \log(\log(n)) + m_2|\alpha - 0.5|$ with the default $m_2 = 6$.

The performance of the quantile estimator may be analyzed with the help of Figure 4.15, which shows scatter plots ($n = 100$) simulated according to equidistant fixed-design regressions with additive normal errors and

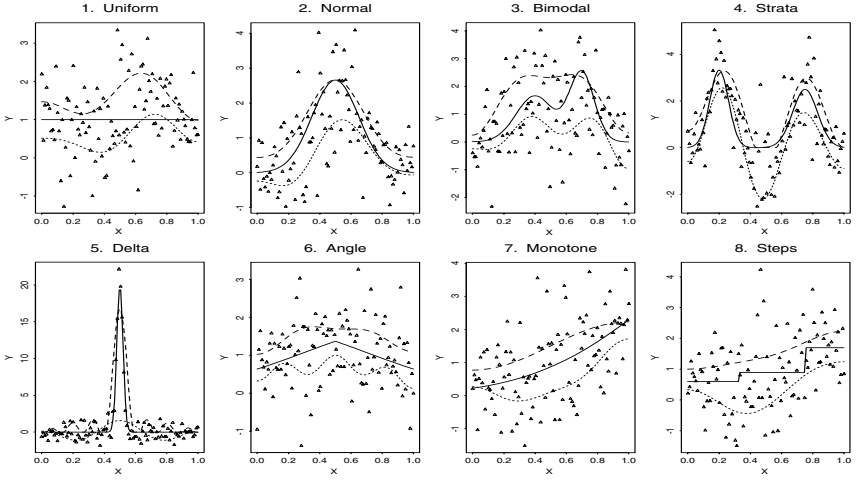

FIGURE 4.15. Interquartile bands for fixed-design heteroscedastic regressions with standard normal error and the scale function being the Angle. Scatter plots for $n = 100$ are shown by triangles. Dotted and dashed lines are the estimates of the first and third quartiles, and together they present the estimated interquartile bands. The underlying regression functions are shown by solid lines. {The scale function is $\sigma f_j(x)$ where $j = scalefun$.} [*n=100, sigma=1, scalefun=6, m0=2, m1=.3,m2=6, s0=.5, s1=.5, cJ0=2, cJ1=.5, cJM=6,cT =4, r=2*]

the scale function being the Angle. The scatter plots are overlaid by the underlying regression functions (solid lines), the first quartile regression estimates (dotted lines), and the third quartile regression estimates (dashed lines). Note that first and third quartile regression lines should be parallel for the case of a constant scale function and not parallel otherwise. More precisely, the width of an interquartile band is proportional to the underlying scale function. We may clearly see this in the estimated bands for the Normal, the Strata, the Monotone, and the Steps.

Finally, let us consider the notion of a *robust* nonparametric regression. The primary concern of this regression is to be less sensitive to outliers, i.e., to responses which cause surprise in relation to the majority of the sample. Median regression discussed earlier and shown in Figure 4.14 is a particular example of this type of regression. Recall that its idea was to replace responses by sample medians $Y'_{(l)}$ of local subsamples $\{Y_{(l-m)}, Y_{(l-m+1)}, \dots, Y_{(l+m-1)}, Y_{(l+m)}\}$, and then use the universal estimator. It was also explained that if noise was normal then using a local sample mean, in place of the local sample median, implied a better estimation. On the other hand, a sample mean is not robust to outliers.

There are several reasonable compromises between the sample mean and sample median estimates, for instance, trimmed means or linear combinations of order statistics (L-estimators). Here we shall explore a local *Huber estimate*

$$Y^h_{(l)} := \operatorname{argmin}_a \sum_{k=l-m}^{l+m} \rho_h((Y_{(k)} - a)/\hat{s}), \tag{4.6.6}$$

where the loss function ρ_h is defined by $\rho_h(x) := x^2/2$ if $|x| \le h$ and $\rho_h(x) := h|x| - h^2/2$ if $|x| > h$, and $\hat{s}$ is the normed median estimate of the scale factor introduced in Section 4.1. Note that as h gets larger, the loss function $\rho_h(x)$ will agree with $x^2/2$ over most of its range, so that $Y^h_{(l)}$ comes closer to the sample mean. As h gets smaller, the absolute loss function implies that $Y^h_{(l)}$ comes closer to the sample median. The value $h = 1.45$ is considered as a default one. Below we shall refer to a universal estimate based on $Y^h_{(l)}$ in place of $Y_{(l)}$ as the Huber estimate.

Figure 4.16 is devoted to exploring three nonparametric estimators: the universal of Section 4.2, median and Huber. As an example, Tukey errors are used, they are defined in Exercise 4.6.5 and the caption of Figure 4.16. Scatter plots in Figure 4.16 show how Tukey errors create outliers, and they also explain why the problem of estimation of a regression function with these errors is so complicated (just look at the Angle scattergram). Universal, median, and Huber estimates are shown by dotted, short-dashed, and long-dashed lines, respectively. The robust estimators perform rather similarly, only for the Delta and the Monotone the particular Huber estimates are better. As about their comparison with the universal, we see that the Bimodal, the Strata, and the Delta functions are better estimated

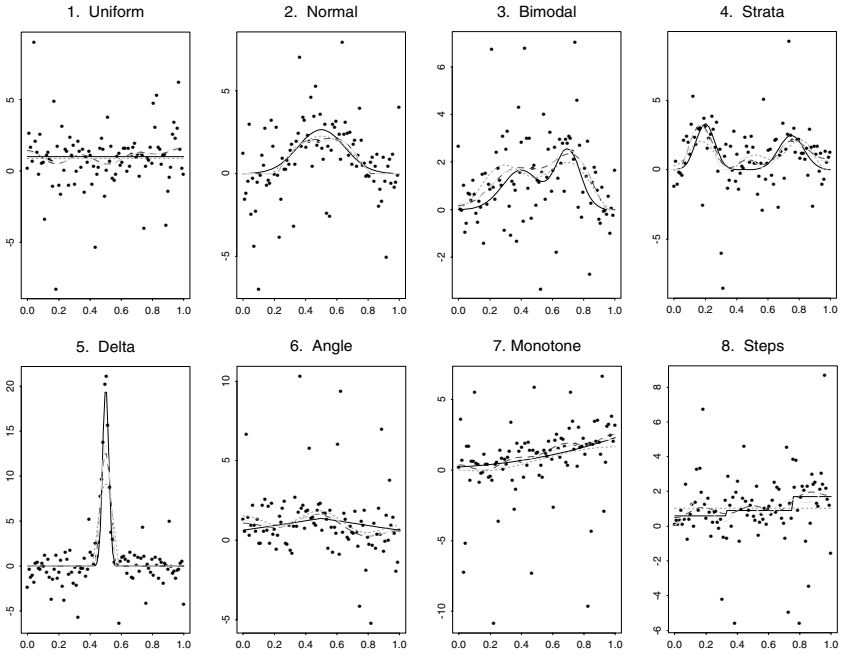

FIGURE 4.16. Performance of the universal, median, and Huber estimators. The corresponding estimates are shown by the dotted, short-dashed, and long-dashed lines. Underlying functions are shown by solid lines. The sample size is $n = 100$, the errors are σ times Tukey random variables with cdf $F(x) = (1 - t_1)\Phi(x) + t_1\Phi(x/t_2)$. Here $\Phi(x)$ is the standard normal cdf, $\sigma = 1$, $t_1 = 0.2$, and $t_2 = 4$. {One may change the parameter h of the Huber estimate and the parameters of errors.} [*n=100, h=1.45, t1=.2, t2=4, sigma=1, m0=2, m1=.3,m2=6, s0=.5, s1=.5, cJ0=2, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

by the robust estimators. On the other hand, the smoother functions, as the Uniform and the Normal, are better estimated by the universal.

Overall, apart of the extreme cases like Cauchy errors, the universal estimator of Section 4.2 performs relatively well and it is robust.

4.7 Case Study: Mixtures Regression

Let us begin with the parametric case where $f(x) := \theta$. Assume that ζ and ξ are two independent random variables with known and different means and some finite variances. Also, let Z be a Bernoulli random variable that takes on the value 1 with probability θ and 0 with probability $1 - \theta$. Then suppose that a new random variable Y is generated by the formula

$$Y := Z\zeta + (1 - Z)\xi. \tag{4.7.1}$$

In words, Y is equal to ζ with probability θ and to ξ with probability $1-\theta$. The random variable Y is called a mixture because the cumulative distribution function $F^Y(c)$ of Y is

$$F^Y(c) = \theta F^\zeta(c) + (1-\theta)F^\xi(c). \tag{4.7.2}$$

The classical parametric problem is to estimate the parameter θ based on iid observations $Y_1, \dots, Y_n$ of Y. Define $E\{\zeta\} =: \mu_\zeta$, $E\{\xi\} =: \mu_\xi$, and assume that $\mu_\zeta \neq \mu_\xi$. Then

$$E\{Y\} = \theta\mu_\zeta + (1-\theta)\mu_\xi,$$

that yields

$$\theta = \frac{E\{Y\} - \mu_\xi}{\mu_\zeta - \mu_\xi}.$$

Thus, a sample mean estimator for the rescaled data $(Y_l - \mu_\xi)/(\mu_\zeta - \mu_\xi)$, $l = 1, 2 \dots, n$, can be used as an estimator of the parameter θ.

One of the classical applications of mixtures models is to describe a *change-point* problem. For example, let ζ correspond to the case where an object functions normally and ξ corresponds to the case where it functions abnormally. Then changing θ from 1 to 0 implies abnormal functioning. Thus, estimation of θ allows one to find a change point.

Now let us consider a nonparametric setting where f is a function of some predictor X (for change point problems X is typically a time). In this case the data at hand are the pairs of observations $(X_l, Y_l), l = 1, 2, \dots, n$, where

$$Y_l := Z_l\zeta_l + (1 - Z_l)\xi_l, \tag{4.7.3}$$

and Z_l are independent Bernoulli random variables that take on values 1 and 0 with probabilities $f(X_l)$ and $1 - f(X_l)$, respectively, and ζ_l and ξ_l are iid realizations of ζ and ξ.

The problem is to estimate the regression function $f(x)$.

As in the parametric case, set $Y' := (Y - \mu_\xi)/(\mu_\zeta - \mu_\xi)$. This implies that

$$E\{Y'|X = x\} = f(x). \tag{4.7.4}$$

Thus, the mixtures regression is a particular case of model (4.5.1), and it can be solved by the method discussed in Section 4.5. Note that $0 \le f(x) \le 1$, so the modified corner functions of that section may be used here.

In Figure 4.17 Monte Carlo simulated scatter plots (for $n = 100$) are shown by triangles (the regression design is fixed equidistant), and the scatter plots are overlaid by estimates. Here the corner functions are modified according to Section 4.5, ζ is $N(muzeta, sdzeta^2)$, and ξ is $N(muxi, sdxi^2)$ with the default values $muzeta = 0$, $muxi = 1$, $sdzeta = 0.7$, and $sdxi = 0.5$.

Note that the equidistant regression is similar to a time series, so the estimates allow us to observe what has been happening over time. Overall,

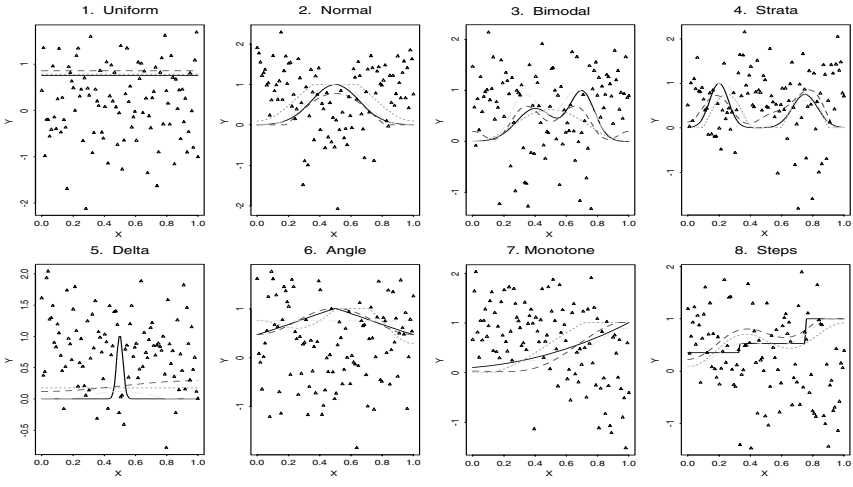

FIGURE 4.17. The estimates for the equidistant mixtures regression: Dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. Scatter plots are shown by triangles for $n = 100$. Underlying regression functions (shown by solid lines) are the same as the functions used for the binary regression. [*set.n=c(50,100,200), n=100, muzeta=0, muxi=1, sdzeta=.7, sdxi=.5, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

except for the Delta case, the estimates are good. For the Delta case the sample sizes are too small to catch the "short-time" change of the underlying distribution. Also note that the scatter plots are so specific that manual analysis via visualization is complicated.

4.8 Case Study: Dependent Errors

Consider again the setting of Section 4.1, where one observes n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that

$$Y_l = f(X_l) + \varepsilon_l, \tag{4.8.1}$$

and the predictors $X_1, \ldots, X_n$ are either iid realizations of a uniform random variable $U(0, 1)$ or fixed equidistant points $X_l = l/(n + 1)$.

The important assumption made in Section 4.1 was that the errors ε_l were independent. What will happen if errors are dependent?

The aim of this section is to show that the dependency among errors crucially affects fixed-design regression but may have a relatively mild effect on random-design regression. In short, a random design is more resistant (robust) to possible deviations from the case of independent errors.

To shed light on the issue, we begin with the case of a constant regression function $f(x) = \theta_0$ and consider the variance of the sample mean estimator

$\hat\theta_0 = n^{-1}\sum_{l=1}^n Y_l$ defined in (4.1.3). In this section it is assumed that the errors are zero-mean and have uniformly bounded variances, that is, $E\{\varepsilon_l^2\} \le C < \infty$ for all l. Also, in general the errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ may have different distributions.

A simple calculation, based on the property (A.10) of the expectation, shows that $E\{\hat\theta_0\} = \theta_0$, so the sample mean estimator is always unbiased regardless of the dependency among the errors.

The situation dramatically changes for the variance of this estimator. Let $\gamma(i,j) := E\{\varepsilon_i\varepsilon_j\}$ denote the autocovariance function of the errors. Recall that if the errors are independent, then $\gamma(i,j) = 0$ for all $i \ne j$ and $\gamma(i,i) = E\{\varepsilon_i^2\} = \mathrm{Var}(\varepsilon_i)$. Then the variance of the sample mean estimator $\hat\theta_0$ is calculated straightforwardly:

$$\begin{aligned}\mathrm{Var}(\hat\theta_0) &= E\{(\hat\theta_0 - E\{\hat\theta_0\})^2\}\\ &= E\Big\{\Big(n^{-1}\sum_{l=1}^n \varepsilon_l\Big)^2\Big\} = n^{-2}\sum_{i,j=1}^n \gamma(i,j). \qquad (4.8.2)\end{aligned}$$

Thus, if the errors are independent, then $\mathrm{Var}(\hat\theta_0) = n^{-1}[n^{-1}\sum_{i=1}^n \gamma(i,i)] < Cn^{-1}$, that is, we get the familiar rate n^{-1} of decreasing the variance as the size of a sample increases. Otherwise, the variance may decrease more slowly.

A practically interesting example is the case of *long-memory* errors of order α where the autocovariance function $\gamma(i,j) = \gamma(i-j)$ is proportional to $|j-i|^\alpha$, $0 < \alpha < 1$ (a Monte Carlo example with such errors will be considered later). Direct calculations show that in this case $\mathrm{Var}(\hat\theta_0)$ is proportional to $n^{-\alpha}$.

Thus dependent errors can make the parametric regression problem extremely complicated. But does this imply a curse for nonparametric regression?

Asymptotic theory shows that there is no way to avoid the curse of dependent errors for fixed-design regression. On the other hand, the outcome is not so glum for random-design regression.

To explain the "miracle" of random design, let us calculate the variance of the estimator $\hat\theta_j$ for some positive j. Write

$$\begin{aligned}E\{(\hat\theta_j - \theta_j)^2\} &= n^{-2}E\Big\{\Big[\sum_{l=1}^n (Y_l\varphi_j(X_l) - \theta_j)\Big]^2\Big\} \qquad (4.8.3)\\ &= n^{-2}E\Big\{\Big[\sum_{l=1}^n ((f(X_l)\varphi_j(X_l) - \theta_j) + \varepsilon_l\varphi_j(X_l))\Big]^2\Big\}\\ &= n^{-2}E\Big\{\sum_{l,t=1}^n \Big[(f(X_l)\varphi_j(X_l) - \theta_j)(f(X_t)\varphi_j(X_t) - \theta_j)\\ &\quad + 2(f(X_l)\varphi_j(X_l) - \theta_j)\varepsilon_t\varphi_j(X_t) + \varepsilon_l\varepsilon_t\varphi_j(X_l)\varphi_j(X_t)\Big]\Big\}.\end{aligned}$$

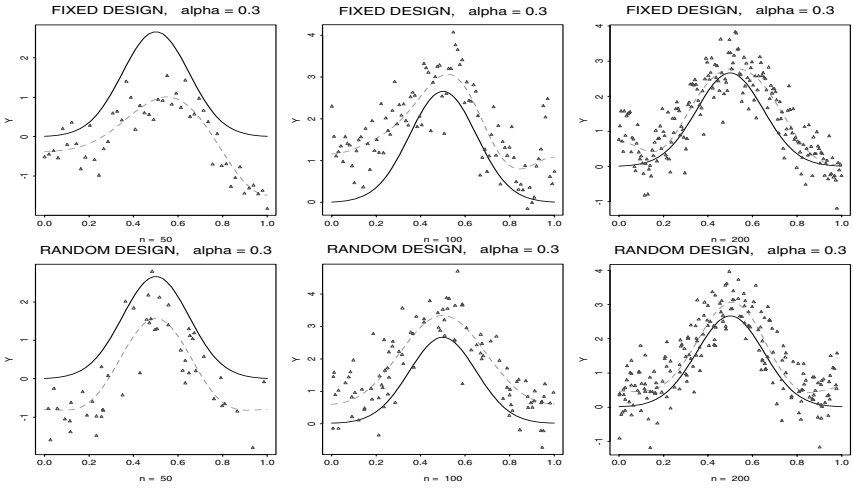

FIGURE 4.18. The case of long-memory errors of order α. The errors are the same for both fixed- and random-design regressions. The solid lines show the underlying Normal regression function, the dashed lines show the estimates, and triangles show scatter plots. {It is a very good idea to repeat this figure with different values of α (controlled by *alpha*) and n to understand the nature of dependent errors and how they affect scatter plots. The choice of an underlying regression function is controlled by the argument *corfun*.} [*set.n=c(50,100,200), corfun=2, alpha=.3, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

Using the relation $E\{f(X)\varphi_j(X)\} = \int_0^1 f(x)\varphi_j(x)dx = \theta_j$, where X is uniform $U(0,1)$, the orthonormality of elements $\{\varphi_j\}$ of the cosine basis, independence the predictors and the errors, and independence of predictors, we get

$$E\{(\hat{\theta}_j-\theta_j)^2\} = n^{-1}\Big[\int_0^1 \big(f(x)\varphi_j(x)-\theta_j\big)^2 dx$$

$$+ n^{-1}\Big(2\int_0^1 f(x)\varphi_j^2(x)dx\sum_{l=1}^n E\{\varepsilon_l\} + \sum_{l=1}^n E\{\varepsilon_l^2\}\Big)\Big], \quad j \ge 1. \quad (4.8.4)$$

Note that (4.8.4) holds even if errors are not zero-mean. Thus, even if $E\{\varepsilon_l\} \ne 0$ and the errors are dependent, (4.8.4) implies that the variance of the sample mean estimator $\hat{\theta}_j$, $j \ge 1$, converges to zero proportionally to n^{-1}.

To use this promising result, let us consider the problem of estimating the *shape* of a regression function f defined as $\psi(x) := f(x) - \int_0^1 f(u)du$. In other words, the shape of a curve is the curve minus its average value, and in a majority of practical problems the shape is the main concern, since it describes the dynamics of a curve. Because $\psi(x) = \sum_{j=1}^\infty \theta_j\varphi_j(x)$, we can use directly the estimator of Section 4.1. This together with (4.8.4) implies

that for the considered random-design regression the dependent errors may have no significant affect on the precision of estimation. On the other hand, the accuracy of estimation of the average value θ_0 is always affected by the dependency.

Now let us explore the merits of this theoretical heuristic for the case of reasonable sample sizes. We use two approaches. First, in Figure 4.18 we show estimates of Section 4.1 for the case of long-memory errors of order α for both fixed equidistant design (the top row) and random uniform design (the bottom row). Because the errors are the same for both these designs, we clearly see how design affects the estimation. First of all, the particular case of $n = 50$ (the left column of diagrams) shows the meaning of the long-memory errors: Here except for only 2 realizations out of 50 all the errors are negative. Note that the errors are zero-mean, but simply many of them are needed to get a negligible sample mean. This is what makes the setting so complicated. As a result, both estimates are shifted down.

On the other hand, there is a remarkable difference in shapes of the estimates. We see that the estimate based on the random design apparently resembles the shape of the Normal, while the estimate based on the fixed design does not. The particular scatter plots also explain why for the fixed design the long-memory dependency is translated into long-memory spatial dependence, whereas the random design spreads the errors along the x-axis and makes them more spatially "independent." The same outcomes occur for the cases of larger sample sizes.

The second recommended approach to explore the merits of the theory is to use intensive Monte Carlo simulations. Here, as an example, we compare precisions of the universal estimator of Section 4.1 for the fixed and random designs via intensive Monte Carlo study.

In the Monte Carlo study the underlying regression function f is the normal density with parameters $(0.3, 0.2^2)$, and the errors are either iid or long-memory Gaussian (defined above) with $\alpha = 0.5$ and $\alpha = 0.3$.

For every experiment with a particular family of errors and a design of predictors, the average integrated squared error (AISE) is calculated based on 1000 repeated simulations, and we write AISEF and AISER for the cases of fixed and random designs. We also consider two different estimands: the function (f) and its shape (ψ).

In Table 4.2 ratios of different AISEs are shown. Column (a) indicates the estimand that is either the function f or its shape ψ. Column (b) indicates what type of errors have been used in the ratio; for instance, the ratio 0.5/iid means that the AISE of the numerator has been calculated for the case of long-memory errors with $\alpha = 0.5$, while the AISE of the denominator has been calculated for the case of iid standard normal errors. Column (c) indicates the design of regression, for instance, F/R means that the AISE of the numerator has been calculated for fixed-design regression and that the AISE of the denominator has been calculated for random-design regression.

In other columns the ratios are shown for indicated sample sizes 50, 100, 200, 400, and 1000.

Table 4.2. *Ratios of AISEs for the case of random and fixed designs, iid and long-memory errors, and the estimand being either the underlying regression function or its shape*

(a)	(b)	(c)	n=50	n=100	n=200	n=400	n=1000
f	iid/iid	F/R	0.50	0.49	0.43	0.55	0.56
f	.5/iid	F/F	3.52	5.15	7.81	9.04	13.54
f	.5/iid	R/R	2.08	2.46	2.88	3.72	4.86
f	.3/iid	F/F	5.30	9.46	15.42	23.04	35.56
f	.3/iid	R/R	2.96	4.24	5.74	9.44	13.51
ψ	iid/iid	F/R	0.49	0.47	0.44	0.49	0.53
ψ	.5/iid	F/F	2.30	3.22	4.46	6.90	9.12
ψ	.5/iid	R/R	1.64	1.62	1.60	1.58	1.52
ψ	.3/iid	F/F	1.28	4.18	6.96	10.78	16.62
ψ	.3/iid	R/R	1.72	1.68	1.68	1.66	1.66

In the first 5 rows f is the estimand. The first line compares the performance of the estimate for the different designs but the same iid errors. We see that the estimate performs better for the fixed design, and this is not a surprise, due to the discussion in Section 4.1.

The second through fifth rows explore the robustness of the estimate for different errors using the same design. For instance, in the second row for the fixed design the average ratios of the AISE for long-memory errors with $\alpha = 0.5$ to the AISE for iid errors are shown. Overall, we see that the random design is essentially more robust.

In the sixth through tenth rows the shape is the estimand. These results clearly support the conclusion of asymptotic theory that the random design together with the universal estimate makes the estimation robust, while this is not the case for the fixed design.

Numerical experiment confirms the above theoretical conclusion about the superiority of the random design over the fixed in terms of robustness for possible deviations from independent errors.

The conclusion of this section is as follows. Whenever errors in responses are dependent, then the estimation of an underlying regression function may become dramatically more difficult. There is a striking difference between random- and fixed-design regressions, namely, while for a random design estimation of the shape of a regression function is almost immune (robust, resistant) to the dependency, for a fixed design there is no cure against the dependency. Thus, if there is a choice between these two designs, the preference should be given to a random design.

4.9 Case Study: Ordered Categorical Data

The aim of this section is to discuss the case where responses are categorical. Examples of such responses are a car that has been driven with speed below 35, between 35 and 55, or above 55 miles per hour; a patient who has no pain, mild pain, moderate pain, severe pain, or acute pain; a person who drinks no beers a day, 1 beer a day, more than 1 but fewer than 5 beers a day, and at least 5 beers a day; the overall rating of a proposal can be poor, fair, good, very good, or excellent.

Note that all these categorical responses have a natural logical ordering and thus are referred to as *ordinal* responses. This is a class of categorical responses that we shall consider. (On the other hand, so-called *nominal* responses have no natural logical ordering; examples are the color of eyes or the place of birth of a respondent to a survey.)

Assume that an ordinal categorical data set is obtained by a grouping (discretization) of responses. This is a clear-cut case for the examples with car speed and number of beers drunk per day (indeed, for instance, a person can drink on average 4.75 beers per day). On the other hand, even for the examples about ratings of proposals and responses about pain, the categorical response can be modeled as a grouping of continuous responses.

Then it is easy to imagine similar nonparametric regression problems such as how a dosage of this or that medicine affects pain, or how the length of a rehabilitation program affects drug-addiction, or how the number of previously published papers affects the rating of a proposal.

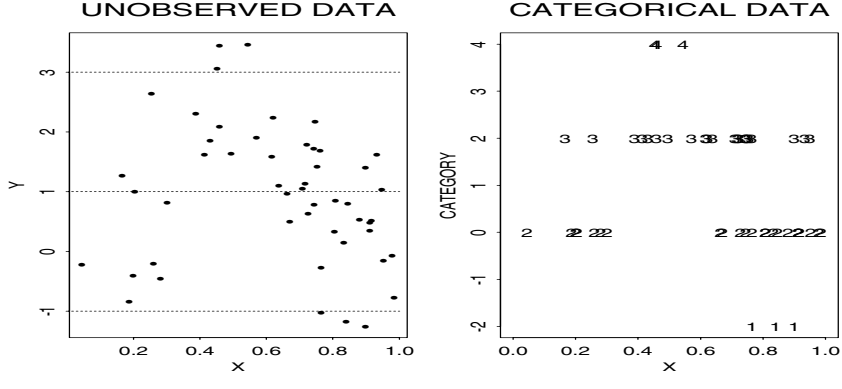

FIGURE 4.19. Simulated example of a categorical random-design regression with the Normal corner function being the underlying regression function, the Monotone being the design density, normal $N(0, \sigma^2)$ independent additive errors, and $n = 50$; the left diagram shows the unobserved data and the bounds for 4 categories; the right diagram shows the corresponding categorical data. [*n=50, corfun=2, desden=7, sigma=1*]

To give an impression about categorical nonparametric regression, let us consider the numerically simulated data shown in Figure 4.19. The left diagram shows an example of simulated classical additive regression, which was discussed in Section 4.1. The scatter plot is overlaid by boundaries for 4 categories: $Y < -1$, $-1 \le Y < 1$, $1 \le Y < 3$, and $3 \le Y$. This data set is not available to a practitioner who gets only the grouped responses (categorical data) shown in the right diagram. Thus, instead of the traditional pairs (X_l, Y_l), where $Y_l = f(X_l) + \varepsilon_l$, there is a set of pairs (X_l, Z_l) where Z_l is the number of a cell (category) for an unobserved Y_l. Note that on top of the traditional complications of classical regression, categorical data give no information on how underlying unobserved responses are spread out over cells. In particular, this can severely influence estimation of such functions as the Delta or the Strata because the information about values of the extreme observations is hidden.

On the other hand, categorical regression is a curious example, where additive errors may help. Indeed, consider a case where a regression function is $f(x) = 0$ and cells are as shown in Figure 4.19. If there are no additive errors, then the categorical data are $(X_l, 2)$, $l = 1, 2, \dots, n$. Thus, the only given information is that all responses are within the cell $[-1, 1)$, and there is no way to estimate the underlying regression function; it may be either $f(x) = 0$ or $f(x) = 0.5 \sin(5x)$ or $f(x) = x - 0.5x^2$. Moreover, even if there are additive errors but their range is not large enough, for instance ε_l are uniform $U(-0.49, 0.49)$, then for all the above-mentioned regression functions the responses are still $Z_l = 2$. See also Exercise 4.9.1.

Let us begin the discussion of a possible estimator with the parametric case $f(x) \equiv \theta$ and the model of grouping data shown in Figure 4.19. Let $\bar{p}$ be the proportion of observations that have categories 3 or 4. Then the probability $P(\theta + \varepsilon \ge 1) =: p$, which is the theoretical proportion of observations in the third and fourth categories, is

$$p = P(\varepsilon \ge 1 - \theta) = 1 - F^{\varepsilon}(1 - \theta). \tag{4.9.1}$$

The foregoing shows that for this example a natural estimate of θ is

$$\bar{\theta} = 1 - Q^{\varepsilon}(1 - \bar{p}), \tag{4.9.2}$$

where $Q^{\varepsilon}(\alpha)$ is the quantile function, that is, $P(\varepsilon \le Q^{\varepsilon}(\alpha)) = \alpha$.

This example shows how to solve the problem of categorical regression because we can convert it into binary regression, discussed in Section 4.5.

Thus, the suggested procedure is as follows.

Step 1. Combine the ordered categories into two groups of "successes" and "failures." Ideally, the boundary in responses that separates these two groups should be such that both successes and failures spread over the domain of predictors. For instance, for the example shown in Figure 4.19 the only reasonable grouping is $\{(1, 2), (3, 4)\}$.

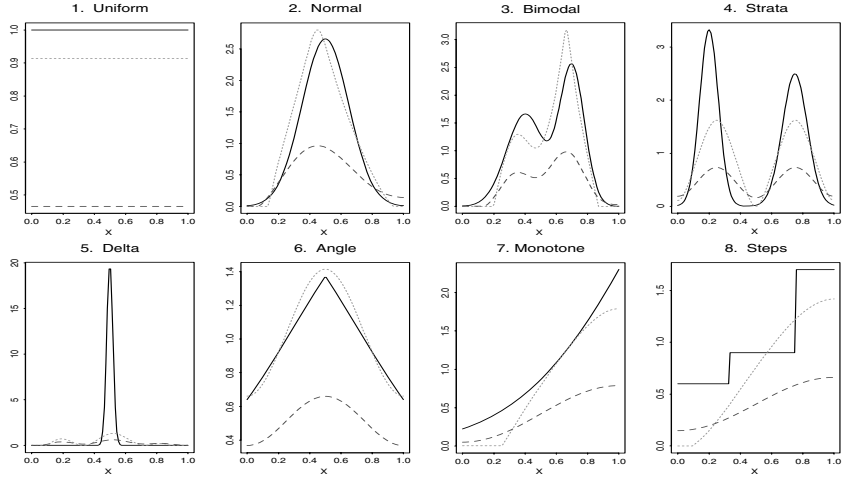

FIGURE 4.20. The estimates for categorical data sets generated according to Figure 4.19. Dashed and dotted lines correspond to estimates $\hat{p}$ and $\hat{f}$ of the binary probabilities and the regression functions. The underlying regression functions are shown by solid lines. [*n=100, desden=7, sigma=1, a=.005, b=.995, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6,cT =4, cB=2, r=2*]

Step 2. Use the estimator $\hat{p}(x)$ of Section 4.5 to estimate the probability of a success. If no information about the additive error ε is given, this is the last step. If the distribution of ε is given, then go to step 3.

Step 3. This step is based on the assumption that a categorical data set is generated by grouping responses of an additive regression $Y_l = f(X_l) + \varepsilon_l$ where the distribution of a continuous error ε is given. Let Y_l belong to the success group iff $Y_l \ge a$. Then

$$\hat{f}(x) = a - Q^{\varepsilon}\big(1 - [\hat{p}(x)]_b^c\big), \tag{4.9.3}$$

where $[z]_b^c = \max(b, \min(z, c))$ and the last truncation allows one to avoid infinite values for $\hat{f}$. The "default" values of b and c are 0.005 and 0.995.

Let us check how this procedure performs. In Figure 4.20 the estimates $\hat{p}(x)$ and $\hat{f}(x)$ are shown by dashed and dotted lines, respectively. The data sets are simulated according to Figure 4.19. The estimates are relatively good even in comparison to the case of direct observations. Of course, in some case, like the Delta, we simply cannot restore the magnitude of an underlying regression curve because all responses larger than 3 look alike. On the other hand, for all the other corner functions, categorical data allow us to restore the underlying curves.

The suggested estimator is not optimal because it is based only on partial information. Nevertheless, asymptotically the suggested estimator is rate optimal (the notion is defined in Section 7.1), it is a good choice for the case of small sample sizes where typically several categories contain a majority

of responses, and it is simple. An asymptotically optimal estimator, where the suggested $\hat{f}$ is used as a pilot estimate, is discussed in Section 7.7.

4.10 Case Study: Learning Machine for Inverse Problems with Unknown Operator

In many applications, ranging from traditional signal communication to medical imaging, one is interested in solving an operator equation.

Consider an operator equation $g = Hf$, $f \in \mathcal{F}$, where a function g, a linear operator H, and a class $\mathcal{F}$ of estimated functions f are given. The problem is to restore f. Such a problem has been one of the main topics in mathematics over the centuries. Classical examples of linear operators are the differential operator, like $Hf(x) := df(x)/dx$, or the integral operator, like $Hf(x) := \int h(x,u)f(u)du$. Typically, it is assumed that an operator is known and g is observed directly. Then there exist numerous theoretical and numerical methods on how to *invert* the problem and restore f. The problem becomes essentially more complicated when g is observed in additive noise, i.e., one observes realizations of $Y := Hf(X)+\varepsilon$. This problem resembles classical regression, only here one should combine classical regression technique with inversion. Recall that we considered a similar problem for a density estimation in Section 3.5.

The aim of this section is more ambitious. We would like to consider a problem where observations of g are noisy and the operator H is unknown. In this case we use a traditional approach of learning theory based on making additional observations that allow us to estimate H (the procedure of estimating H is often referred to as training or study). To do this, a set of *training functions* $e_1, \ldots, e_m$ from the class $\mathcal{F}$ is chosen, and then these training functions are used in place of an unknown f, that is, a training data set is given that is a set of noisy observations of He_l, $l = 1, \ldots, m$. Then, based on a noisy observation of Hf, which is called the main observation, and the training set, a learning machine (i.e., a data-driven estimator) should recover f.

Let us explain both the problem and a suggested solution via a classical example of heat flow on an interval. The mathematical setting is as follows. The temperature $u(f,t,x)$ in a rod of length 1 (here $f = f(x)$ is the initial temperature, t is the time, and x is the coordinate) with ends held at temperature 0 is described by the heat equation,

$$\partial u(f,t,x)/\partial t - (1/2)\partial^2 u(f,t,x)/\partial x^2 = 0\,, \quad t > 0, \quad 0 < x < 1\,, \tag{4.10.1}$$

subject to $u(f,t,x) = 0$ for $t \ge 0$, $x = 0$ and $x = 1$, and $u(f,0,x) = f(x)$.

Note that (4.10.1) may be written as $u = Hf$, and the problem is to estimate the initial temperature $f(x)$, $0 \le x \le 1$, based on noisy equidistant measurements of the temperature $u(f,t_0,x)$ at some moment $t_0 > 0$. Nei-

ther the operator H nor the time t_0 is known to a practitioner. Why is this setting interesting? The reason is that (4.10.1) is a mathematical simplification of real practical problems, where, for instance, a possible cavity in a studied rod may lead to drastic changes in the operator. Thus, operators (here a rod) should also be studied.

Now let us formulate the statistical setting. With respect to the temperature $u(f, t_0, x)$ at a moment t_0, we have a classical regression problem where one observes

$$Y_l = u(f, t_0, l/(n+1)) + \sigma\varepsilon_l, \quad l = 1, 2, \dots, n\,. \tag{4.10.2}$$

There is a dramatic difference between (4.10.2) and the model (4.1.1) discussed in Section 4.1. In the model (4.10.2) the function of interest is an initial temperature $f(x)$, while in Section 4.1 we were interested in estimating the current temperature $u = Hf$. Thus, we should estimate u and then solve the operator equation. Because H is unknown, it should be estimated. Learning theory suggests to use a training data set where some known initial temperatures are used and observations of the corresponding temperatures at the moment t_0 are made. In other words, the training observations are

$$Z_{jl} = u(e_j, t_0, l/(n+1)) + \nu\varepsilon_{jl}, \quad l = 1, 2, \dots, n, \quad j = 1, \dots, m, \tag{4.10.3}$$

where e_j, $j = 1, 2, \dots, m$, are known initial temperatures. The parameters σ and ν may be different. In short, the results of an additional series of m training experiments are given, and they may be used to estimate an underlying operator. Based on these training data sets and the main observation (4.10.2), a learning machine should recover an unknown initial temperature f.

Because we discuss orthogonal series estimators, it is natural to use the first elements of a basis as the training functions. Here these functions are temperatures with the boundary conditions $e_j(0) = e_j(1) = 0$. Among the bases discussed in Chapter 2, only the sine basis with elements $\{e_j = \sqrt{2}\sin(j\pi x), j = 1, 2, \dots\}$ has this property. Thus, we shall use these elements as training functions. The first two elements are shown in the diagrams (e) and (g) of Figure 4.21. A choice of the training functions is the only manual procedure of the learning machine (4.10.8) defined below.

Before describing this machine, let us try it in action. Namely, let us numerically model the heat equation in a rod and then analyze the recovery of an initial temperature. The measurement errors ε_l and ε_{jl} are simulated as iid standard normal, $\sigma = \nu = 0.2$ and $t_0 = 0.05$. Also, the learning machine is allowed to make $m = 6$ training experiments.

Figure 4.21 illustrates both the problem and the solution. The initial (unknown) temperature $f(x) = 10x(x-1)(x-0.3)$ is shown in diagram (a). Then, over time the temperature in the rod is changing according to the heat equation (4.10.1), and at a particular time t_0, the temperature $u(f, t_0, x)$ is shown in diagram (b). Were the heat equation and the time

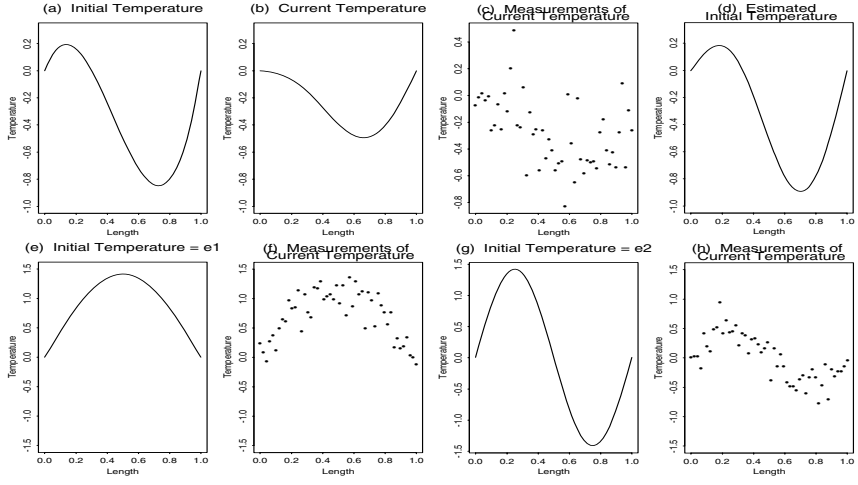

FIGURE 4.21. Recovery of initial temperature in a rod by learning machine. The top row illustrates the main experiment and the estimated temperature. The bottom row illustrates the training experiments. {It is assumed that $\sigma = \nu$ and the parameter ν is controlled by the argument *nu* (ν is the Greek letter "nu"). The time t_0 is controlled by the argument *t0*. The number m of training experiments is controlled by the argument *m*.} [*n=50, nu=.2, t0=.05, m=6*]

t_0 known, then the problem of calculating the initial temperature would become the classical ill-posed problem of solving the operator equation, that is a very complicated mathematical problem by itself.

Our problem is essentially more complicated than the classical one. First, the current temperature $u(f, t_0, x)$ is not known; instead, its noisy measurements at $n = 50$ equidistant points are given. These measurements are shown in diagram (c). Recall that the scattergram (c) is called the main observation because here we see observations related to the estimated initial temperature. Second, neither the heat operator nor the time t_0 is given. Instead, as we discussed earlier, the learning machine is allowed to make several active experiments with the rod. These active experiments are as follows. The first experiment is based on using the initial temperature $e_1(x)$ and noisy measurements of the corresponding current temperature $u(e_1, t_0, x)$ shown in diagrams (e) and (f), respectively. Absolutely similarly, the learning machine gets the results of the experiment where e_2 is used as the initial temperature. The information about the second training experiment is shown in diagrams (g) and (h). And this training continues until all m training experiments are performed.

Thus, if for instance $m = 2$, then the learning machine has at hand the measurements shown in diagrams (c), (f), and (h). No other information is available.

The initial temperature recovered by the learning machine is shown in diagram (d). Note that it remarkably resembles the underlying initial temperature (a) and in no way resembles the current temperature (b). Moreover, for this particular data set the learning machine decided to use only the first two training experiments, that is, the estimated initial temperature (d) is based only on the data shown in diagrams (c), (f), and (h). The result is truly impressive.

Now let us formulate the general setting and describe the learning machine. The so-called main set of observations is

$$Y_l := Hf(X_l) + \sigma\varepsilon_l, \quad l = 1, 2, \dots, n. \tag{4.10.4}$$

Here X_l are either equidistant fixed or uniformly distributed predictors on $[0, 1]$, and H is a linear operator. It is also known that f belongs to some function class $\mathcal{F}$.

Because the operator H is unknown, m additional training sets are given,

$$Z_{jl} := He_j(X_{jl}) + \nu\varepsilon_{jl}\,, \quad l = 1, 2, \dots, n, \quad j = 1, \dots, m, \tag{4.10.5}$$

where these sets are generated similarly to (4.10.4) only with the known functions e_j in place of an estimated f. Thus, it is assumed that $e_j \in \mathcal{F}$. Also, in general, $\sigma \neq \nu$. It is assumed that the stochastic terms ε_l and ε_{jl} are iid zero-mean and unit-variance.

Thus, a learning machine has at hand $m + 1$ scatter plots (regressions). Let us additionally assume that $e_1, e_2, \dots$ are elements of a basis that is "reasonable" for approximation of $f(x)$, $Hf(x)$, and $He_j(x)$. The meaning of a reasonable basis is that a partial sum of m elements may give us a reasonable approximation of all these functions.

Then we may write $f(x) = \sum_{j=1}^{\infty} \theta_j e_j(x)$ and $He_j(x) = \sum_{k=1}^{\infty} \mu_{jk} e_k(x)$, where θ_j and μ_{jk} are the corresponding Fourier coefficients (with respect to the basis $\{e_1, e_2, \dots\}$). Because H is a linear operator, we may write

$$Hf(x) = H \sum_{j=1}^{\infty} \theta_j e_j(x) = \sum_{j=1}^{\infty} \theta_j He_j(x) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \theta_j \mu_{jk} e_k(x).$$

Thus,

$$Hf(x) = \sum_{k=1}^{\infty} \Big(\sum_{j=1}^{\infty} \theta_j \mu_{jk} \Big) e_k(x) =: \sum_{k=1}^{\infty} a_k e_k(x).$$

For each particular regression, the Fourier coefficients can be estimated by the estimator (4.1.3). Denote the estimated Fourier coefficients of $Hf(x)$ by $\{\hat{a}_k, k = 1, 2, \dots\}$ and the estimated Fourier coefficients of $He_j(x)$ by $\{\hat{\mu}_{jk}, k = 1, 2, \dots\}$.

Then, for every $J = 1, \dots, m$ we may write

$$\hat{a}_k =: \sum_{j=1}^{J} \hat{\theta}_{Jj} \hat{\mu}_{jk}, \quad k = 1, 2, \dots, J. \tag{4.10.6}$$

Here $\hat{\theta}_{Jj}$, $j = 1, \dots, J$, are solutions of this system of J equations, and they are natural estimates of the Fourier coefficients θ_j.

Then, as in the previous sections, the optimal cutoff $\hat{J}$ is defined as

$$\hat{J} := \operatorname{argmin}_{1 \le J \le m} \sum_{j=1}^{J} (2\hat{\sigma}^2 \hat{\lambda}_{Jj}^{-2} - \hat{\theta}_{Jj}^2). \tag{4.10.7}$$

Here $\hat{\lambda}_{J1}, \dots, \hat{\lambda}_{JJ}$ are eigenvalues of the $J \times J$ matrix with the entries $\hat{\mu}_{jk}$, $\hat{\sigma}^2 = (3J_n)^{-1} \sum_{j=3J_n+1}^{6J_n} \hat{a}_j^2$, and J_n is the rounded-down $1 + 0.5 \ln(n)$.

The learning machine is defined as

$$\hat{f}(x) := \sum_{j=1}^{\hat{J}} \hat{\theta}_j e_j(x). \tag{4.10.8}$$

It is important to stress that the success of the learning machine crucially depends on a choice of training functions. Also, machine learning is one of the most challenging problems in nonparametric curve estimation theory.

4.11 Case Study: Measurement Errors in Predictors

Let us return to the simplest model of additive homoscedastic regression,

$$Y_l = f(X_l) + \sigma \varepsilon_l, \qquad l = 1, 2, \dots, n, \tag{4.11.1}$$

with predictors X_l uniformly distributed on $[0, 1]$. In Section 4.1 we discussed the problem of estimation of f based on pairs $\{(X_l, Y_l), l = 1, \dots, n\}$ of observations. Here we consider a case where both predictors and responses are noisy, that is, instead of underlying predictors X_l only their noisy measurements

$$U_l = X_l + \xi_l, \qquad l = 1, 2, \dots, n, \tag{4.11.2}$$

are given. It is assumed that X, ε, and ξ are independent and X_l, ε_l and ξ_l are their iid realizations.

The problem is to estimate f based on a set of data $\{(U_l, Y_l),\ l = 1, 2, \dots, n\}$.

There is both bad and good news, based on results of the asymptotic theory, about the regression with errors in predictors. The bad news is that the setting is ill-posed, and it is similar to one discussed in Section 3.5. In other words, the errors in predictors drastically affect MISE convergence. The good news is that the series data-driven estimator is the best among all possible estimators, and a case of small sample sizes is only the onset of the ill-posed problem.

We restrict our attention only to the case of a normal $N(0, \sigma_\xi^2)$ measurement error ξ in (4.11.2). The reason is that it is the most complicated and

(unfortunately) most common case. Even the case of a Cauchy measurement error is much better, and errors with distributions like Gamma or Double Exponential are just "peanuts" in comparison with a Normal error.

To understand a necessary modification of the cosine estimator suggested in Section 4.1, let us calculate the expectation of the estimate (4.1.3) of θ_j if we use noisy observations U_l in place of unobserved X_l, i.e., we ignore measurement errors in predictors. Recall that

$$\hat{\theta}_j = n^{-1} \sum_{l=1}^{n} Y_l \varphi_j(U_l), \tag{4.11.3}$$

and write

$$\begin{aligned} E\{\hat{\theta}_j\} &= E\Big\{n^{-1} \sum_{l=1}^{n} Y_l \varphi_j(U_l)\Big\} = E\{Y \varphi_j(U)\} \\ &= E\{f(X)\varphi_j(U)\} + \sigma E\{\varepsilon \varphi_j(U)\} \\ &= \sqrt{2} E\{f(X) \cos(j\pi(X + \xi))\}. \end{aligned} \tag{4.11.4}$$

In the last equality we used the assumption that ε and U are independent, ε is zero-mean, and $\{\varphi_j\}$ is the cosine basis.

To analyze (4.11.4) we use the elementary trigonometric equality

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta). \tag{4.11.5}$$

Then, using the independence of the predictor X and the measurement error ξ, we get

$$\begin{aligned} E\{f(X)\cos(j\pi(X + \xi))\} &= E\{f(X)\cos(j\pi X)\}E\{\cos(j\pi\xi)\} \\ &\quad + E\{f(X)\sin(j\pi X)\}E\{\sin(j\pi\xi)\} \\ &= E\{f(X)\cos(j\pi X)\}E\{\cos(j\pi\xi)\}. \end{aligned}$$

In the last equality we used the fact that for a normal zero-mean random variable ξ the identity $E\{\sin(j\pi\xi)\} = 0$ holds.

Now recall (see details in Section 3.5) that for a zero-mean normal random variable ξ the expectation $E\{\cos(j\pi\xi)\}$ is equal to the value $h_j^\xi := E\{e^{ij\pi\xi}\} = e^{-(j\pi\sigma_\xi)^2/2}$, which is the value of the characteristic function of ξ at the point $j\pi$.

Combining the results, we obtain

$$E\{Y\varphi_j(U)\} = \theta_j h_j^\xi. \tag{4.11.6}$$

There are three straightforward conclusions from (4.11.6). First, if predictors are polluted by errors and this fact is unknown, then any estimator is inconsistent. Thus, it is not wise to ignore such a possibility. Second, because h_j^ξ decreases exponentially in j, a small error in estimation $E\{Y\varphi_j(U)\}$ causes large deviations in an estimate of θ_j. This is the reason why this type of problem is called *ill-posed*. Finally, finding a good unbiased estimate of θ_j is not difficult. Indeed, we may simply set (recall that

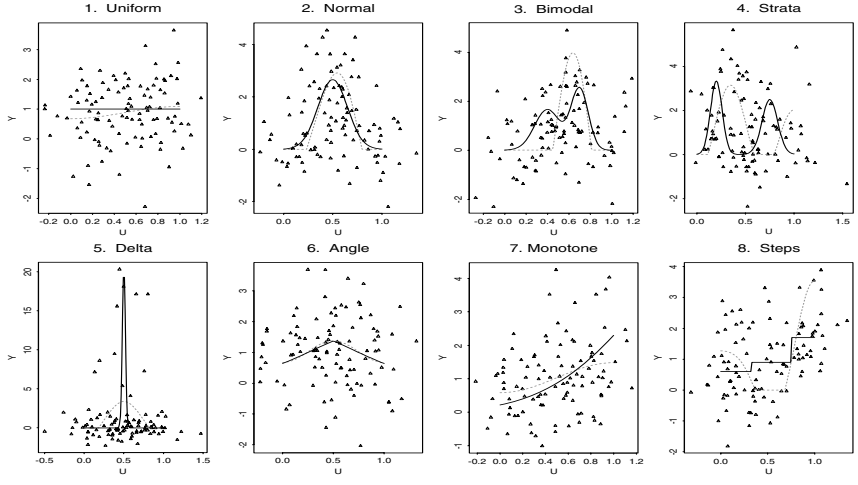

FIGURE 4.22. Regression (4.11.1)–(4.11.2) with normal $N(0, \sigma_\xi^2)$ measurement errors in predictors. Scatter plots of observations $\{(Y_l, U_l), l = 1, \dots, n\}$ are overlaid by the underlying regression functions (solid lines) and the estimates (dotted lines). {The coefficients σ, σ_ξ, c_b, d_0, d_1, d_2, and c_H of the estimator (4.11.9) are controlled by the arguments *sigma, sigma.xi, cb, d0, d1, d2,* and *cH*, respectively.} [*n=100, sigma=1, sigma.xi=.2, cb=8, d0=2, d1=.5, d2=10, cH=1*]

$h_0^\xi = 1$)

$$\tilde\theta_0 = \hat\theta_0, \quad \tilde\theta_j := \bigl(nh_j^\xi\bigr)^{-1} \sum_{l=1}^{n} (Y_l - \tilde\theta_0)\varphi_j(U_l). \tag{4.11.7}$$

Also, a straightforward calculation shows that

$$E\{(\tilde\theta_j - \theta_j)^2\} = n^{-1}(h_j^\xi)^{-2}\Bigl(\int_0^1 (f(x) - \theta_0)^2 dx + \sigma^2\Bigr)(1 + r_{nj}), \tag{4.11.8}$$

where r_{nj} vanishes as j and n increase. This formula explains why this setting is so complicated. The asymptotic theory tells us that MISE decays at an extremely slow logarithmic rate. (The formula (4.11.8) also shows that the more slowly the characteristic function of a measurement error decreases, the better is the estimation of a regression function.)

Because a good unbiased estimator of Fourier coefficients is suggested, we can employ the data-driven estimator (3.5.12) of Section 3.5 suggested for a similar ill-posed problem,

$$\tilde f_n(x) := \sum_{j=0}^{J_n} (1 - \hat\theta_j^{-2}\hat\sigma^2 n^{-1})_+ \, \tilde\theta_j I_{\{|h_j^\xi| > c_H \hat\sigma n^{-1/2+b_n}\}} \varphi_j(x). \tag{4.11.9}$$

Here $\hat\theta_j$ are defined in (4.11.3), $\hat\sigma := 1.48\,\mathrm{median}(\{|Y_l - \tilde\theta_0|, l = 1, \dots, n\})$ is the normed sample median, $b_n = 1/c_b \ln(\ln(n + 20))$, J_n is the rounded-up

$d_0+d_1[\ln(n+20)]^{1/d_2 b_n}$, and the default coefficients are $c_b = 8$, $d_0 = 2$, $d_1 = 0.5$, $d_2 = 10$, and $c_H = 1$. Finally, if a regression function is nonnegative, then the bona fide projection is used.

Figure 4.22 explains the setting (4.11.1)–(4.11.2). Here we see an extremely difficult particular case where $\sigma_\xi = 0.2$. It is apparent from the scatter plots that they can no longer be an inspiration for a manual search after a regression function because everything is blurred and a majority of the scatter plots look alike. Can one see the Delta or the Bimodal in the corresponding scattergrams? The answer is "no." Thus, here we may rely only on an estimator. Recall the discussion in Section 3.5 that for the case of small sample sizes we see only the onset of the ill-posed problems. This is what one may hope for.

Finally, note that the asymptotic theory shows that the standard deviation σ of the additive noise in the responses, see (4.11.1), affects neither constant nor rate of MISE convergence. This is an interesting asymptotic phenomenon, and Exercise 4.11.8 explains how to explore this issue.

4.12 Practical Seminar

The aim of this seminar is to gain experience in using the universal estimator of Section 4.2 for real data sets.

In Figure 4.23 (please look only at the top row of diagrams) four different data sets are shown by plotting the pairs of observed predictors and responses in the xy-plane (recall that this diagram is called a *scattergram*). The corresponding sample sizes n are shown in the subtitles. Do you see any pronounced relationship between X and Y in each of these 4 scattergrams?

All these data sets are challenging, so the answer "no" is okay. Let us see whether a classical parametric regression may help us to gain some understanding of these data sets. The most widely used parametric estimate is *linear* regression. It is assumed that the regression function is $f(x) := \beta_0 + \beta_1 x$, and then the problem is to estimate the *y-intercept* β_0 and the *slope* β_1 by minimizing the *least-squares error* $\sum_{l=1}^n (Y_l - \beta_0 - \beta_1 X_l)^2$. The result leads to the familiar least-squares linear regression. Exercise 4.12.1 discusses the underlying idea of this linear regression.

The linear least-squares regression lines are shown in the middle row of the diagrams (again, please do not look at the bottom row). Do the regression lines help you to realize the relationships between X and Y? Do you see any interesting structures in the data sets highlighted by the linear regression lines? For the data set (a) the linear regression probably helps to visualize a possible relationship between X and Y, but for the other data sets their structures are still puzzles.

Now let us consider a nonparametric approach based on the universal estimate of Section 4.2 (it is apparent that all these regressions are het-

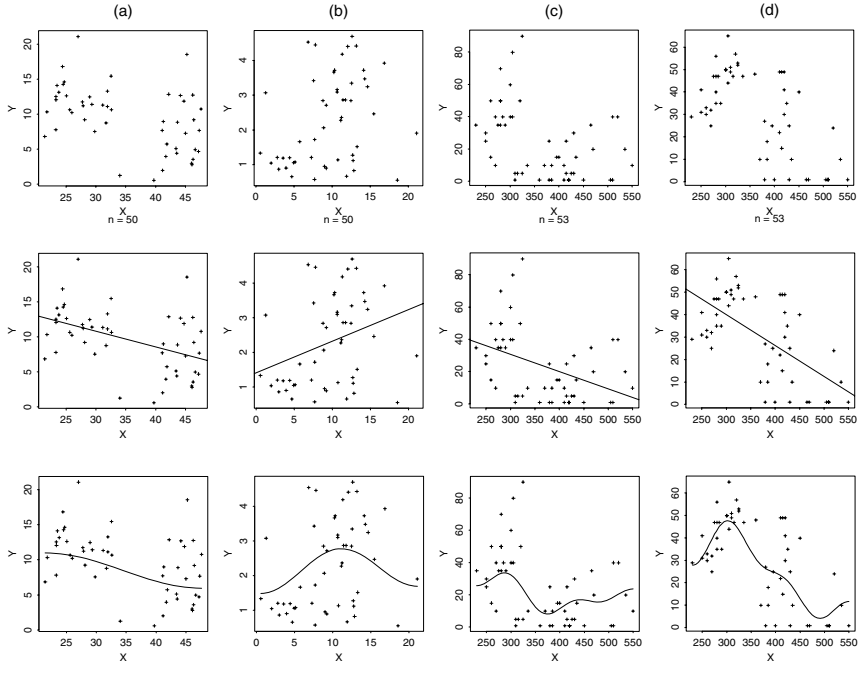

FIGURE 4.23. Four data sets. The scattergrams are shown in the first row, the scattergrams overlaid by the linear regression lines are shown in the second row, and the scattergrams overlaid by the universal estimates are shown in the third row. {The arguments of the estimator are reviewed in the caption to Figure 4.5.} [*X1 =saving.x[, 1], Y1 =saving.x[, 5],X2 =saving.x[, 5], Y2 = saving.x[, 2],X3 =chernoff2[, 1],Y3 =chernoff2[, 4],X4 = chernoff2[, 1], Y4 = chernoff2[, 3], s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2, r=2*]

eroscedastic). The nonparametric estimates are shown in the bottom row. As we see, only for the scattergram (a) do the linear and the nonparametric regression graphs resemble each other. For the data set (b) the nonparametric estimate reveals an absolutely unexpected (in comparison with the linear regression) structure of the data, and the nonparametric curve helps us to see this bell-shaped relationship between X and Y. For the data set (c) the nonparametric estimate again "opens our eyes" to the structure of the data, which is rather complicated. Finally, the curve in diagram (d) looks absolutely natural; it is now probably surprising that we were unable to realize this pronounced structure from the scattergram.

Now let us explain the data sets. The scattergram (a) shows family savings as a percentage of disposable income in the 1960s (the Y variable) versus the percentage of population younger than 15 years old (the X variable) for 50 countries. The data file is **saving.x**. (All the data sets are from the standard S–PLUS distribution.)

For this data set the linear regression line clearly indicates that the youth population diminishes savings, and such a conclusion has common sense. The statement of the nonparametric estimate is just "softer" within the two distinct clusters of countries with smaller and larger young populations. Note that there is a significant gap between these clusters. Moreover, let us note that the two countries on each side of the boundaries between the clusters have the smallest savings among all the 50 countries; these countries are Iceland ($X = 34.03, Y = 1.27$) and Chile ($X = 39.74, Y = 0.6$). These two countries are apparently outliers due to their specific geopolitical situation during the 1960s when the data were collected. On the other hand, it is possible to change the arguments of the universal estimate in such a way that it will pronouncedly indicate these two clusters and highlight these two outliers (Exercise 4.12.2).

Let us continue the analysis of the diagrams in Figure 4.23. In column (b) we again use the data file **saving.x**, only here savings are the X variables and the percentage of population older than 75 are the Y variables. Analyzing this data set, we would like to understand how welfare and prosperity of nations (supposedly measured in units of savings) affect the length of life of their citizens. The conclusion of the linear regression is straightforward, and it definitely has common sense: Savings do not harm and help to live longer. The conclusion of the nonparametric estimate is not so straightforward. Moreover, it is controversial (but this is typically the case with nonparametric methods because they address "small things" that otherwise may be easily overlooked, so when using nonparametric estimates be ready for nonstandard outcomes). The nonparametric regression tells us that while moderate savings are necessary to increase the length of life, large savings (per family) are not healthy for a nation. But please, do not rush off with some kind of "left-wing" sociological and political conclusions. The message of this nonparametric estimate should be read as follows. Suppose that one plays a game to guess in which of two countries (among the 50 countries considered) the percentage of senior citizens in the 1960s was larger when the only information about these two countries is that their average levels of savings per family were about 12% and 21%. Then, the nonparametric regression curve tells us that it is better to bet on the first country. Clearly, the answer, based on linear regression or rational sociological reasoning, should be different.

Now let us try to understand why the nonparametric curve suggested such a contradictory conclusion and why we have no reason to use the bell shape of the curve to condemn large savings. In the scattergram (b) we have a peculiar combination of two factors. The first one is the low density of countries with the largest savings (do you see that there are only 4 (among 50) nations with savings more than 15%?). The second factor is that the two countries with the largest savings per family are Zambia ($X = 18.56$, $Y = .56$), which has the lowest percentage of senior citizens among all the nations, and Japan ($X = 21.1$, $Y = 1.91$), which also has a relatively low

percentage of seniors. Clearly, in the 1960s these two countries were outliers for certain geopolitical and historical reasons. Curiously, only the example of Denmark ($X = 16.85$, $Y = 3.93$) allows us to believe that nice savings (per family) may prolong life.

Finally, a "small" detail about the notion "savings" should be added. Using the command > **help(saving.x)** we can get some information about the data set **saving.x**. It is explained that "savings" means aggregate personal saving divided by disposable income. Thus, the "prosperity" of a nation is proportional to aggregate savings and inversely proportional to income. This explanation sheds light on the message of the bell-shaped nonparametric estimate.

This example shows that nonparametric regression is a good tool to attract our attention to unusual structures in data sets. Then, if necessary, a data analyst should discuss the meaning of "messages" with specialists. It is also worthwhile to repeat that it is not the aim of nonparametric regression to be the only judge in solving scientific or political questions.

The last two scattergrams in Figure 4.23 are based on the famous mineral contents data set **chernoff2**, which is a 53 by 12 matrix representing the mineral analysis of a 4500 foot core drilled from a Colorado mountainside. Twelve variables (columns) represent assays of seven mineral contents. Fifty-three equally spaced specimens (rows) along the core were assayed. What we see in the diagrams is that a nonparametric regression curve can dramatically change the visualization of a data set. Also, you may notice the excellent flexibility of nonparametric curves.

Figure 4.24 allows us to shed further light on a data set and at the same time be trained in using coefficients of the universal estimate. Here the default data set is the set shown in Figure 4.23(d). The scattergram is very complicated and deserves further investigation. The sample size $n = 53$ is shown in the subtitle. Here we try to assess this structure via the coefficient c_T. Recall that this is the coefficient used in the procedure of hard thresholding high-frequency components. Thus, increasing c_T implies fewer high-frequency components, while decreasing c_T keeps more high-frequency components. The particular values of c_T may be seen in the titles.

The first diagram corresponds to $cT = 8$, which is essentially larger than the default value $cT = 4$ used for the second diagram. As we see, there is no high-frequency component with extremely large power, so using $c_T = 4$ and $c_T = 8$ implies the same estimate. If we are decreasing c_T, then more and more high-frequency components with "moderate" power are included. In particular, the third diagram with $c_T = 0.1$ looks very interesting and informative; it apparently sheds light on the data at hand. The smallest $c_T = 0.01$ does not help us to see anything new, and the corresponding estimate apparently undersmooths the data because we see too many high-frequency oscillations.

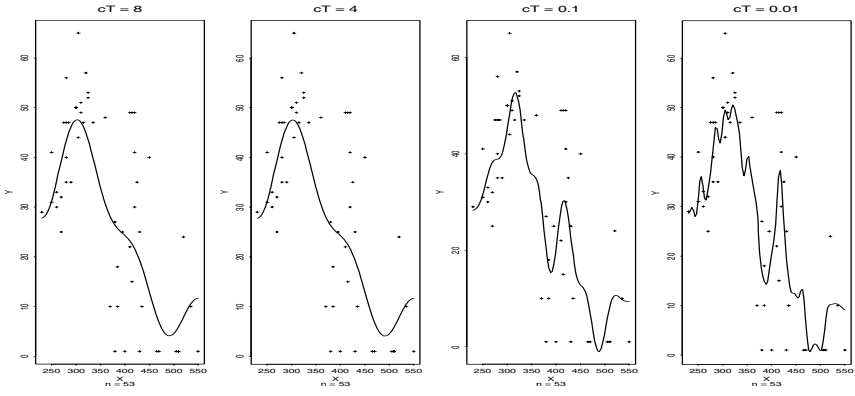

FIGURE 4.24. The effect of a particular "running" coefficient on the universal estimate. {Two data sets, *DATAX* and *DATAY*, are the predictors and the corresponding responses. The running argument is chosen by *arg*, and its values by *set.arg*.} [*arg= "cT", set.arg = c(8, 4, .1, .01), DATAX = chernoff2[,1], DATAY = chernoff2[,3], s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2, r=2*]

Such an approach, where a data set is analyzed via a spectrum of estimates, allows us to see different frequency components of a data set at hand. It is a convenient method to shed new light on the data.

Finally, let us solve a problem that will help us to realize a difference between probability density estimation and nonparametric regression problems. At first glance, it is difficult to be misled, but keep in mind that many practical settings may be rather confusing.

Consider the problem of estimating the probability of a fatal injury after a car accident as a function of the speed of the car at the time of the accident. Is this a probability density estimation or a regression problem?

This is one of the particular scenarios where estimation of a probability becomes a regression problem because here the question of interest is how the speed (predictor) affects the probability of a fatal injury (response) and the answer is to be a function that gives you the probability of a fatal injury for a given speed. Thus, this is a regression problem, more specifically, the binary regression discussed in Section 4.5.

4.13 Exercises

4.1.1 Repeat Figure 4.1 with different n (for instance, choose $n = 25$, 100, 300), and for each n and each corner function find a largest value of the argument *sigma* such that the underlying corner function is still visually recognizable. Rank the corner functions according to ascending *sigma* and explain the obtained ranks.

4.1.2 Explain why (4.1.3) may be considered as a naive numerical integration formula for calculation of θ_j defined at (4.1.2.). Hint: Use $1/(n+1)$ in place of $1/n$ and then assess the difference.
4.1.3 Verify (4.1.4). Hint: Use (3.1.7).
4.1.4 Verify (4.1.5). Hint: Use (3.1.7).
4.1.5 Find the expectation of the estimate (4.1.6) and show that it is a consistent estimate of the coefficient of difficulty d as $n, j \to \infty$.
4.1.6 Explain the underlying idea of (4.1.7). Hint: See Sections 3.1–3.3.
4.1.7 Using Figure 4.3, choose 3 particular sample sizes and try to find arguments (coefficients) of the estimator that are optimal for a set of 2 and then a set of 4 corner functions. Hint: Begin with a description of the arguments and write down what particular characteristics of the estimator are affected by them. Divide the arguments into 2 groups of more and less important ones for improving the estimation of the chosen corner functions. Begin to play around with more important arguments and then polish your estimator with the help of other arguments.
4.1.8 As in the previous exercise, what values of coefficients (or values of arguments of the S-function) would you recommend for all the corner functions and the sample size $n = 100$? Also, does a smaller σ, say $\sigma = 0.1$, affect your choice?
4.2.1 Explain the difference between a homoscedastic and a heteroscedastic regression. Give two practical examples.
4.2.2 Repeat Figure 4.4 ten times and count the number of scattergrams where the underlying regression functions are recognizable. Then repeat this experiment with $sigma = 0.5$. Analyze the results.
4.2.3 Verify (4.2.4).
4.2.4 Explain the underlying idea of the estimator (4.2.5): (a) heuristically; (b) mathematically by finding its bias and variance.
4.2.5 What is the difference between the estimates (4.2.3) and (4.2.5)? When would you recommend using each of them?
4.2.6 Repeat Figure 4.5 with different values of σ. Explain the results. Choose a particular n and find a particular σ for which estimates "reasonably" fit the underlying regression functions in, say, about 90% of realizations.
4.2.7 Show that $nE\{(\hat{\theta}_j - \theta_j)^2\} \to d$ as j increases (the coefficient of difficulty d is defined at (4.2.7)).
4.2.8 Establish (4.2.8). Hint: Use the following Cauchy–Schwarz inequality for square integrable functions g and p,

$$\Big| \int_0^1 g(x)p(x)dx \Big|^2 \le \int_0^1 g^2(x)dx \int_0^1 p^2(x)dx, \tag{4.13.1}$$

and the fact that $h(x)$ is the density supported on $[0, 1]$.
4.2.9 Set $\hat{d}_V = n\sum_{l=1}^n [(Y_{(l)} - \tilde{f}_J(X_{(l)}))\hat{D}_{0ls}]^2$, where $\hat{D}_{0ls}$ are defined in (4.2.6). Consider the mean squared error $\delta_n = E\{(\hat{d}_V - d)^2\}$ of estimation

of the coefficient of difficulty d defined at (4.2.7) and (a) discuss possible assumptions that imply $\delta_n \to 0$ as $n \to \infty$. (b) What can be said about the rate of decrease of δ_n?

4.2.10 (a) Would you recommend any changes in the values of coefficients of the estimator used in Figure 4.5? (b) Would you recommend any changes in these values to make the estimator more robust to changes in σ and n?

4.3.1 Find the expectation and the variance of the estimates $\bar{\theta}$ and $\hat{g}$ defined at (4.3.1).

4.3.2 Set $Z_l = (Y_l - \bar{\theta})^2$, where $\bar{\theta}$ is defined in (4.3.1). Explain why one may write $Z_l \approx g + g(\varepsilon_l^2 - 1)$.

4.3.3 Using the previous exercise, explain the motivation of (4.3.3).

4.3.4 Let $\hat{g}(x)$ be an estimate of a nonnegative function $g(x)$, $0 \le x \le 1$. This estimate may take on negative values. Suggest a projection of $\hat{g}$ in L_1 and L_2 on a class of nonnegative functions.

4.3.5 Using Figure 4.6, suggest optimal values of coefficients of the estimator for any set of 2 and then 3 corner functions. Is your recommendation robust to changes in n?

4.3.6 Use Figure 4.6 with $\sigma = 0.5$ and then with $\sigma = 2$. Would you recommend any changes in the default values of coefficients of the estimator?

4.3.7 Use Figure 4.6 with the Normal being the scale function. Would you recommend any changes in the default values of coefficients of the estimator?

4.3.8 Repeat Figure 4.6 with a different design density. Would you recommend any changes in the default values of coefficients of the estimator?

4.3.9 Suppose that each predictor may be generated according to a desired distribution. Suggest a sequential procedure that leads to the optimal design (4.2.8) for the case of an unknown scale function.

4.4.1 Repeat Figure 4.8 and find cases where Universal catches false spikes created by noise. Then find values for arguments of Universal such that these false spikes appear extremely rarely but the underlying spikes are still shown. Consider the cases of signal-to-noise ratios equal to 3, 5, 10 and sample sizes 512, 1024, and 2048.

4.4.2 Use Figure 4.9 and find values of coefficients of Universal such that it "kills" all wavelet coefficients at the finest scale **s1**. Conversely, find values of coefficients for which all finest noisy wavelet coefficients are kept.

4.4.3 Use Figure 4.9 and signals from Table 4.1 to find optimal values of coefficients of the universal estimator for the cases of signal-to-noise ratios 3, 5, 10 and sample sizes 512, 1024, and 2048.

4.4.4 Use Figure 4.10 and find optimal values of coefficients of Universal for the cases of signal-to-noise ratios 3, 5, 10 and sample sizes 512, 1024, and 2048.

4.4.5 Find an expression for MISE of the universal estimator. Explain how its coefficients affect the MISE.

4.4.6 Write down the main differences between the universal and SureShrink estimators. Explain them using Figure 4.9.
4.4.7 Repeat Figures 4.8–10. Find particular cases where SureShrink outperforms the universal estimator and discuss the outcomes. Then, suggest optimal coefficients for the Universal.
4.5.1 Let Z be a Bernoulli random variable with the probability p of a success. Find the expectation and the variance of Z. Draw a plot of the variance as a function of p.
4.5.2 Explain why the models considered in Sections 4.1–4.2 can be considered as particular cases of (4.5.1).
4.5.3 Explain why the estimator of Section 4.2, developed for a particular case of an additive regression, may also be used for the general model (4.5.2).
4.5.4 Find $f(x)$ that maximizes the coefficient of difficulty (4.5.3).
4.5.5 Verify (4.5.4).
4.5.6 Using Figure 4.11, would you recommend any changes in the values of coefficients of the universal estimator? Is your conclusion robust to changes in the design density?
4.5.7 Verify that the given formulae for the probability mass function, the expectation and the variance of a Poisson random variable are correct. Hint: Use Taylor's formula $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$.
4.5.8 Explain why Poisson regression is a particular case of the model (4.5.1).
4.5.9 Verify (4.5.6).
4.5.10 Repeat Figure 4.12 several times, make hard copies, and explain the scatter plots using the definition of Poisson regression.
4.5.11 Using Figure 4.12, would you recommend any changes in the values of coefficients of the estimator? Is your recommendation robust to changes in the sample size n?
4.6.1 Let $X_1, \ldots, X_n$ be iid with distribution F and density f. Let the ordered X's be denoted by $X_{(1)} \le \cdots \le X_{(n)}$. Prove that the density of $X_{(k)}$ is given by

$$f^{X_{(k)}}(x) = \frac{n(n-1)!}{(k-1)!(n-k)!} F^{k-1}(x)(1-F(x))^{n-k} f(x).$$

4.6.2 Let $\tilde{X}_n$ be the sample median of n iid realizations of a Cauchy random variable. Show that $E\{\tilde{X}_n^2\} < \infty$ when $n \ge 5$, while $E\{\tilde{X}_n^2\} = \infty$ for $n < 5$. Hint: Use Exercise 4.6.1.
4.6.3 Let $X_1, \ldots, X_n$ be iid according to a Cauchy distribution. Show that $E\{X_{(k)}^2\} < \infty$ if and only if $3 \le k \le n-2$, where the notation of Exercise 4.6.1 is used.
4.6.4 A family of symmetric distributions with greatly varying heaviness in the tails is a family of Student's t distributions with ν degrees of freedom,

whose density is

$$f_\nu(x) := \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1+x^2\nu^{-1})^{-(\nu+1)/2}.$$

Here $\Gamma(x)$ is the gamma function. For $\nu = 1$ this reduces to the Cauchy distribution for which $E\{|X|\} = \infty$. For $\nu = 2$ the expectation exists but the variance is still infinite. For $\nu \geq 3$ the variance is finite. As $\nu \to \infty$ the distribution tends to normal. Show that:
a. For $\nu = 2$ the sample median has a finite variance for $n \geq 3$.
b. $E\{X^2\} = \nu/(\nu-2)$ for $\nu \geq 3$.

4.6.5 Another family of distributions that is often used to test robustness of a procedure to a deviation from a normal distribution is a Tukey model $T(\varepsilon, \tau)$, where

$$F^X(x) := (1-\epsilon)\Phi(x) + \epsilon\Phi(x/\tau)$$

and Φ is a standard normal cumulative distribution function. Find the expectation and the variance of X. Also, explain why Tukey errors may be a good test for robustness.

4.6.6 Use Figure 4.13 to find a case where scatter plots in Figure 4.13.1 and Figure 4.13.5 are similar.

4.6.7 Use Figure 4.13.1 to find a scatter plot that strongly indicates a presence of false oscillations created by Cauchy errors.

4.6.8 Using Figure 4.14, for each corner function find a minimal sample size that allows a reasonable and reliable estimation.

4.6.9 Using Figure 4.14, find optimal values of coefficients of the estimator for a set of 4 corner functions.

4.6.10 Explain how a quantile regression can be used to indicate a heteroscedastic regression.

4.6.11 How many observations can be expected to fall within an interquartile band?

4.6.12 Use Figure 4.15 to find optimal values of coefficients of the estimator.

4.6.13 Write a short report about the effect of parameter h on Huber estimate. Support your conclusion by using Figure 4.16.

4.6.14 Use Figure 4.16 and analyze how parameters of Tukey errors affect the estimates.

4.7.1 Verify (4.7.2).

4.7.2 Consider the parametric case (4.7.1) where $\mu_\zeta = \mu_\xi$ but $\mathrm{Var}(\zeta) \neq \mathrm{Var}(\xi)$. Suggest a consistent estimator of θ.

4.7.3 Under the condition of Exercise 4.7.2, suggest a nonparametric estimate of $f(x)$ for the model (4.7.3).

4.7.4 Using Figure 4.17, find values of coefficients of the estimator that lead to showing a pronounced mode for the Delta and, at the same time, to a fair estimation of the other corner functions.

4.8.1 Let $X_1, \ldots, X_n$ be iid realizations of a random variable X. What is a sufficient assumption for the sample mean estimate $\bar{X} = n^{-1} \sum_{l=1}^{n} X_l$ to be an unbiased estimate of the expectation of X?
4.8.2 What assumption is sufficient for $\bar{X}$ to be a consistent estimate of the expectation of X?
4.8.3 Verify (4.8.2).
4.8.4 Show that for the case of long-memory errors of order α, the variance $\mathrm{Var}(\hat{\theta}_0)$ of this sample mean estimate is proportional to $n^{-\alpha}$.
4.8.5 Explain each step in obtaining (4.8.3).
4.8.6 Explain how (4.8.4) has been obtained.
4.8.7 Does (4.8.4) imply that the variance of $\hat{\theta}_j$ decreases at the parametric rate n^{-1}?
4.8.8 Consider the case of a heteroscedastic regression $Y_l = f(X_l) + \sigma(X_l)\varepsilon_l$, $l = 1, 2, \ldots, n$. Note that in this case the additive errors and predictors are dependent. Nevertheless, show that even in this case a random design may lead to more robust estimation than a fixed design.
4.8.9 Use Figure 4.18 and find corner functions that are least and most affected by long-memory errors. Explain the outcome.
4.8.10 Would you recommend any changes in the arguments of the estimator based on the analysis of Figure 4.18?
4.9.1 Use Figure 4.19 with σ equal to 2, 1, 0.5, 0.25, and 0.1. Also consider several underlying regression functions. Does a decrease in the standard deviation of the error help to recognize an underlying curve based on the categorical data?
4.9.2 Using Figure 4.20, would you recommend any changes in the default values of coefficients of the estimator?
4.9.3 Find the expectation and the variance of the estimate (4.9.2).
4.9.4 Suppose that for the particular set of data shown in Figure 4.19 one suggested to combine the data into the following two groups: $\{(1), (2, 3, 4)\}$. What outcome can be expected in this case? Would you recommend such a combination?
4.9.5 Consider the case of censored responses at a level C such that one observes $Z_l = \min(C, Y_l)$ in place of responses Y_l. Suggest a data-driven estimator for this setting.
4.10.1 Is the sine basis a good choice for the example of heat flow on the interval? What is a good basis for a case where temperature at the right end of a rod is not fixed?
4.10.2 Under what circumstances would you recommend using a learning machine? Give several examples.
4.10.3 Why do we use a learning machine for a problem like the heat flow on an interval? Is it possible to solve this problem using another approach (without training sets)?

4.10.4 Suppose that the standard deviations σ and ν of additive error terms in (4.10.2) and (4.10.3) are different. How does this affect the estimation? Also, if you have a choice, which standard deviation should be smaller?
4.10.5 Use Figure 4.21 to find a maximal standard deviation ν such that a reliable restoration of the initial temperature is still possible.
4.10.6 Use Figure 4.21 to find a maximal time t_0 such that a reliable restoration of the initial temperature is still possible.
4.10.7 Use Figure 4.21 to find a minimal sample size n such that a reliable restoration of the initial temperature is still possible. Also, find how the standard deviation ν affects this sample size.
4.11.1 Explain all steps in obtaining (4.11.4).
4.11.2 Is (4.11.7) a sample mean estimate of θ_j?
4.11.3 Verify (4.11.8).
4.11.4 Calculate the MISE of the estimate (4.11.9).
4.11.5 Consider the case of a double exponential measurement error ε, where $p^\varepsilon(x) = b^{-1}e^{-|x-\mu|/b}$, $-\infty < x < \infty$, and suggest a truncated estimate that is based on minimization of the MISE.
4.11.6 Explain the underlying idea of the estimator (4.11.9).
4.11.7 Repeat Figure 4.22 and try to recognize the underlying curves. Then reduce σ_ξ and find a maximal value where the corner functions are visualized from the scatter plots.
4.11.8 Repeat Figure 4.22 with smaller standard deviations σ. Does this help for the case of $\sigma_\xi = 0.2$? Does this help for the case of $\sigma_\xi = 0.05$? Also, compare and explain the answers.
4.11.9 Explain why asymptotically the standard deviation of the error in responses affects neither the constant nor the rate of MISE convergence. Hint: Use the result of Exercise 4.11.4. Then compare the variance and the integrated squared bias (ISB) terms of the MISE. Show that the variance is negligibly small in comparison with the ISB.
4.11.10 Choose a particular pair (σ, σ_ξ) and suggest optimal arguments of the estimator using Figure 4.22.
4.11.11 Suggest a data-driven estimator for the case of a heteroscedastic regression.
4.11.12 Suggest a data-driven estimator for the case of an arbitrary measurement error in predictors.
4.12.1 Explain the underlying idea of least-squares linear regression. Hint: Consider the model $Y = f(X) + \varepsilon$ where $f(X)$ is a linear function and the error ε is a random variable with zero mean and finite variance. Then show that for both random- and fixed-design regressions the relation $f(x) = E\{Y|X = x\}$ holds. Finally, recall (and prove) that $f(x)$ minimizes the conditional *mean squared error* $\text{MSE} := E\{(Y - f(x))^2|X = x\}$.
4.12.2 Use Figure 4.23 and find values of coefficients of the universal estimator such that a nonparametric estimate separates the two clusters in diagram (a) and thus highlights the two countries with the smallest savings. Also explain how this change affects the other diagrams.

4.12.3 Use Figure 4.23 and find values of coefficients that make the nonparametric estimates optimal according to your own judgment of the structure of these data sets.
4.12.3 Use arguments Yj and Xj, $j = 1, 2, 3, 4$, to choose any other 4 data sets and analyze them using Figure 4.23.
4.12.4 Using Figure 4.24, explore the effect of other arguments of the estimator.
4.12.5 Analyze the first 3 data sets of Figure 4.23 using Figure 4.24. To do this, use the arguments *DATAX* and *DATAY*. For instance, to explore the data set (a) just set *DATAX = saving.x[,1]* and *DATAY = saving.x[,5]*.

4.14 Notes

There are many good books where different applied and theoretical aspects of nonparametric regression are discussed. These books include Eubank (1988), Müller (1988), Nadaraya (1989), Härdle (1990, 1991), Wahba (1990), Green and Silverman (1994), Wand and Jones (1995), and Simonoff (1996), among others. A chapter-length treatment of orthogonal series estimates may be found in Eubank (1988, Chapter 3).

4.1 Asymptotic justification of the universal estimator for the regression model is given in Efromovich (1986), where it is established that for smooth functions a data-driven series estimator outperforms all other possible data-driven estimators. Practical applications are also discussed.

4.2 The heteroscedastic regression was studied in Efromovich (1992) and Efromovich and Pinsker (1996), where it is established that asymptotically a data-driven orthogonal series estimator outperforms any other possible data-driven estimators whenever an underlying regression function is smooth. Also, Efromovich and Pinsker (1996) give results of numerical comparison between the universal and local linear kernel estimators discussed in Section 8.4.

4.3 The textbook by Lehmann and Casella (1998, Chapter 3) gives a comprehensive treatment of the problem of estimation of a scale parameter. The book by Carroll and Ruppert (1988) discusses regression settings where estimation of the scale (variance) function becomes the central issue. It also reviews many different useful techniques and approaches.

4.4 The books by Ogden (1997), Mallat (1998), and Vidacovic (1999) are relatively simple and give a nice discussion of wavelets and their use in regression problems. The article by Donoho and Johnstone (1995) introduces and discusses the data-driven estimator SureShrink. This estimator is considered as a benchmark for all other wavelet estimators. The asymptotic justification of the Universal is given in Efromovich (1997c, 1999a).

4.5 The asymptotic justification of the universal estimator for the considered settings is given in Efromovich (1996a) and Efromovich and Thomas

(1996). In the latter article a discussion of parametric and nonparametric methods of binary regression may be found, and application of the universal estimator to a real data set is discussed.

4.6 A discussion of the problem of robust regression, including a review of different robust methods and quantile estimators, may be found in the book by Fan and Gijbels (1996). A rigorous mathematical discussion of robust parametric estimation is given in Serfling (1980) and Huber (1981).

4.7 Parametric mixtures models are discussed in the book by Lehmann and Casella (1998, p. 456). The model of mixture of distributions is discussed in the book by Prakasa Rao (1983, Chapter 10). The asymptotic justification of using the universal estimator is given in Efromovich (1996a).

4.8 The book by Beran (1994) covers diverse statistical methods and applications for dependent data. The fact that dependent errors may dramatically slow down MISE convergence for fixed design regression was established in Hall and Hart (1990). Asymptotic analysis of the problem and other aspects, including the dependency between predictors and the dependency between predictors and error terms, are discussed in Efromovich (1997c, 1999b). The book by Dryden and Mardia (1998) discusses statistical shape analysis.

4.9 A chapter-length discussion of the problem of nonparametric estimation for ordered categorical data may be found in Simonoff (1996, Chapter 6). The asymptotic justification of using the universal estimator and other related examples may be found in Efromovich (1996a).

4.10 The discussion of ill-posed problems and operator equations arising in statistical applications may be found in the books by Wahba (1990) and Vapnik (1995). The latter book also discusses the fundamentals of learning theory. Mathematical justification of the learning machine, as well as the asymptotic theory of learning machines for solving operator equations with unknown operator, is given in Efromovich and Koltchinskii (1997).

4.11 The book by Carrol, Ruppert, and Stefanski (1995) is devoted to measurement errors in nonlinear models. Optimal estimation for both regular and irregular settings is discussed in Efromovich (1994c).

5

Nonparametric Time Series Analysis for Small Samples

In this chapter we shall discuss some basic topics of time series analysis, including the classical decomposition of a time series into deterministic trend and seasonal components and a random component, as well as spectral density estimation. Special topics include cases of missing observations, hidden additive components, and bivariate time series.

5.1 Estimation of Trend and Seasonal Components and Scale Function

A time series (process) is a set of pairs of observations (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) where each response Y_l has been recorded at a specific time X_l, and traditionally $X_1 < X_2 < \cdots < X_n$. Then, the simplest classical *decomposition model* of a time series is

$$Y_l := f(X_l) + S(X_l) + \sigma(X_l)\varepsilon_{X_l}, \tag{5.1.1}$$

where $f(x)$ is a slowly changing function known as a trend component; $S(x)$ is a periodic function with period T (that is, $S(x + T) = S(x)$ for all x), known as a seasonal (cyclical) component (it is also customarily assumed that the integral or sum of the values of the seasonal component over the period is zero); $\sigma(x)$ is called a scale function (it is also often referred to, especially in finance and econometrics literature, as a volatility); and ε_{X_l} are random components that may be dependent, and in this case the responses Y_l become dependent as well. Recall that the familiar phrase "a random

walk down Wall street" is motivated by this type of classical decomposition, and a primary argument in the literature is about the presence or absence of a deterministic part and about the type of a random walk.

A typical feature of a time series is that predictors X_l are equidistant integers. Thus, without loss of generality, we may set $X_l = l$. Then a time series is completely described by the responses $\{Y_l,\ l = 1, 2, \ldots\}$, which may be treated as a sequence of regular observations in time, and this explains why such a sequence is called a *time series*. Of course, many practical examples are indeed sequences in time, but there are plenty of other examples; for instance, data may be collected in space. In the latter case the data are often referred to as *spatial data*, and there is even a special branch in statistics, known as geostatistics, that is primarily concerned with the analysis of such data. A particular example will be considered in Section 6.7. In this chapter, for the sake of clarity, we shall use only time series terminology and assume that data are collected sequentially in time.

Another typical feature of a time series is that the errors $\{\varepsilon_1, \varepsilon_2, \ldots\}$ in (5.1.1) may be dependent. Moreover, the case of dependent errors is the main topic in time series analysis. Thus, we begin our discussion with a short introduction to a class of ARMA processes, which are a good tool to model series of dependent random variables. Then we shall discuss methods of estimation of a trend, seasonal component, and scale function.

• **Causal ARMA Processes.** The main assumption about the class of time series (5.1.1) that we wish to consider is that the noise ε_l is a realization of a so-called second-order zero-mean stationary time series $\{\varepsilon_l\}$ $= \{\ldots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots\}$ such that (i) $E\{\varepsilon_l^2\} < \infty$ for all l, that is, the second moment is finite; (ii) $E\{\varepsilon_l\} = 0$ for all l, that is, the expectation is zero; (iii) the *autocovariance function* $\gamma(l, s) := E\{\varepsilon_l \varepsilon_s\}$ satisfies the relation $\gamma(l, s) = \gamma(l + h, s + h)$ for all l, s, and h, that is, a translation in time does not affect the autocovariance function.

Note that property (iii) implies that $\gamma(l, s) = \gamma(l - s) = \gamma(s - l)$. To see this just set $h = -s$ and $h = -l$. Thus a second-order zero-mean stationary time series is characterized by its autocovariance function $\gamma(h)$ at the *lag* h. Also note that no assumptions about higher moments or about distributions of the errors are made.

The simplest kind of second-order stationary error is one in which the random variables $\{\varepsilon_l\}$ are uncorrelated (that is, $\gamma(h) = 0$ for $h \neq 0$), with mean 0 and variance 1. Let us denote such time series by $\{Z_l\}$ and call it a *standard white noise*. A classical example is a time series of iid standard Gaussian random variables, which is the white noise that we shall use in all the following simulations, and we call it a *standard Gaussian white noise*.

Then a wide variety of dependent second-order stationary processes can be generated by using a white noise and a set of linear difference equations. This leads us to the notion of an *autoregressive moving average process of orders p and q*, an ARMA(p, q) process for short. By definition, the process $\{X_t, t = \ldots, -1, 0, 1, \ldots\}$ is said to be an ARMA(p, q) process if $\{X_t\}$ is

second-order stationary and for every t,

$$X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p} = \sigma(Z_t + b_1 Z_{t-1} + \cdots + b_q Z_{t-q}), \quad (5.1.2)$$

where $\{Z_t\}$ is a standard white noise, $\sigma > 0$, the orders p and q are nonnegative integers, and $a_1, \ldots, a_p, b_1, \ldots, b_p$ are real numbers. For the case of a Gaussian white noise we shall refer to the corresponding ARMA process as a *Gaussian ARMA process*.

Two particular classical examples of an ARMA process are a *moving average* MA(q) process, which is a moving average of $q+1$ consecutive realizations of a white noise,

$$X_t = \sigma(Z_t + b_1 Z_{t-1} + \cdots + b_q Z_{t-q}), \quad (5.1.3)$$

and an *autoregressive* AR(p) process satisfying the difference equation

$$X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p} = \sigma Z_t. \quad (5.1.4)$$

Each of these examples plays an important role in the analysis of time series. For instance, prediction of values $\{X_t,\ t \geq n+1\}$ in terms of $\{X_1, \ldots, X_n\}$ is relatively simple and well understood for an autoregressive process; see Exercise 5.1.16. Also, for a given autocovariance function it is simpler to find an AR process with a similar autocovariance function. More precisely, if an autocovariance function $\gamma(j)$ vanishes as $j \to \infty$, then for any integer k one can easily find an AR(k) process with the autocovariance function equal to $\gamma(j)$ for $|j| \leq k$. The "negative" side of an AR process is that it is not a simple issue to find a stationary solution for (5.1.4), and moreover, it may not exist. For instance, the difference equation $X_t - X_{t-1} = \sigma Z_t$ has no stationary solution, and consequently there is no AR(1) process with $a_1 = 1$. The discussion of such tricky things is beyond the scope of this book, and in what follows a range for the coefficients that "keeps us out of trouble" will always be specified.

The advantages of a moving average process are its simple simulation, the given expression for a second-order stationary solution, and that it is very close by its nature to white noise, namely, while realizations of a white noise are uncorrelated, realizations of an MA(q) process are uncorrelated whenever the lag is larger than q. The "minus" of MA processes is that, surprisingly, they are not so easy for prediction and estimation as AR processes. Thus, among the two, typically AR processes are used for modeling and prediction. Also, AR processes are often used to approximate an ARMA process.

Now we are in a position to define a causal (future-independent) ARMA process (or more specifically, a causal process with respect to an underlying white noise $\{Z_t\}$). The idea is that it is quite natural to expect that an ARMA time series $\{X_t\}$ depends only on current and previous (but not future!) realizations of the white noise. Thus, we say that an ARMA process $\{X_t\}$ generated by a white noise $\{Z_t\}$ is *causal* if $X_t = \sum_{j=0}^{\infty} c_j Z_{t-j}$, where the coefficients c_j are absolutely summable. Clearly, MA(q) processes are

causal, but not all AR(p) processes are; for instance, a stationary process corresponding to the difference equation $X_t - 2X_{t-1} = Z_t$ is not causal. We shall not elaborate more on this issue and note only that below, we consider simulations of only Gaussian ARMA(1, 1) processes corresponding to the difference equation $X_t - aX_{t-1} = \sigma(Z_t + bZ_{t-1})$ with $|a| < 1$ and $-a \neq b$. It may be directly verified (Exercise 5.1.17) that for such a this equation has a stationary and causal solution $X_t = \sigma Z_t + \sigma(a+b)\sum_{j=1}^{\infty} a^{j-1} Z_{t-j}$.

This ends our brief discussion of ARMA processes.

The aim of the next subsections is to explain methods of estimation of the deterministic components $f(x)$, $S(x)$, and $\sigma(x)$ in (5.1.1) where $X_l = l$ and noise $\{\varepsilon_l\}$, $l = 1, \dots, n$, is zero-mean and second-order stationary. A comprehensive example that combines all the steps is postponed until Section 5.3 because finding periods of seasonal components is based on estimation of the spectral density, which is discussed in Section 5.2.

• **Estimation of a Trend.** There is no surprise that time series analysis customarily uses methods of estimation of a trend that are also used by regression analysis, namely, methods such as parametric least-squares regression or smoothing by means of a moving average. On the other hand, the nonparametric orthogonal series approach, developed in Chapter 4, seems an attractive alternative to these classical methods. Indeed, if a time series has a deterministic term that is written as $\sum_{j=0}^{\infty} \theta_j \varphi_j(x)$, then the low-frequency part of this series,

$$f(x) := \sum_{j=0}^{J_{\max}} \theta_j \varphi_j(x), \quad 0 \le x \le n, \tag{5.1.5}$$

can be referred to as a *trend component* (or simply *trend*). Here $\{\varphi_j\}$ are elements of a basis in $L_2([0, n])$ and θ_j are the Fourier coefficients. The choice of $J_{\max}$ is typically up to the practitioner, who defines the meaning of the trend and seasonal components in the frequency domain.

Then the data-driven universal estimator of Section 4.2 can be used to estimate the trend. (Recall that to use the universal estimator we always rescale data onto $[0, 1]$.) Moreover, the estimator is greatly simplified by the fact that its cutoff should be at most $J_{\max}$. Then, all the examples considered in Chapter 4 can be viewed as some particular time series.

• **Estimation of a Scale Function.** The primary concern of the classical time series theory is that the stochastic term in (5.1.1) should be second-order stationary, that is, the scale function $\sigma(x)$ should be constant. Since this is typically not the case, the usually recommended approach is to transform a data set at hand in order to produce a new data set that can be successfully modeled as a stationary time series. In particular, to reduce the variability (volatility) of data, Box–Cox transformations are recommended when the original positive observations $Y_1, \dots, Y_n$ are converted to $\psi_\lambda(Y_1), \dots, \psi_\lambda(Y_n)$, where $\psi_\lambda(y) := (y^\lambda - 1)/\lambda$, $\lambda \neq 0$, and $\psi_\lambda(y) := \log(y)$, $\lambda = 0$. By a suitable choice of λ, the variability may be significantly reduced.

Apparently, the nonparametric technique of Section 4.3 may be used as well. Firstly, we use the nonparametric estimator of a scale function suggested in Section 4.3. All the examples considered in Section 4.3 illustrate how the approach works. Then the original observations are divided by the estimate, and this should give us a new data set with a nearly constant variability of its stochastic term.

• **Estimation of a Seasonal Component.** Traditional time series analysis assumes that the period T of an underlying seasonal component $S(x)$ is given. (We shall discuss in the next two sections how to find the period with the help of the spectral density; also note that in many practical examples, such as daily electricity demands or monthly average temperatures, periods of possible cyclical components are apparent.) By definition, $S(x+T) = S(x)$ for any x, and if a time series is defined at integer points, then $\sum_{l=1}^{T} S(l) = 0$ (a seasonal component should be zero-mean).

Using these two assumptions, classical time series theory recommends the following method of estimating a seasonal component. First, a given time series is detrended by the formula $\tilde{Y}_l = Y_l - \tilde{f}(l)$, where $\tilde{f}(l)$ is an estimated trend. Then, the natural procedure for estimating $S(j)$ is the sample mean estimate

$$\tilde{S}(j) := \lfloor (n-j)/T \rfloor^{-1} \sum_{r=0}^{\lfloor (n-j)/T \rfloor} \tilde{Y}_{j+rT}, \quad j = 1, 2, \ldots, T. \tag{5.1.6}$$

Recall that $\lfloor a \rfloor$ denotes the rounded-down a. Note that (5.1.6) is a nonparametric estimate because no parametric underlying model is assumed.

To understand how this conventional method performs, let us consider a simple example. Assume that $\tilde{Y}_l = S(l) + \sigma\varepsilon_l$, $l = 1, 2, \ldots, n$, where $n = kT$, k is integer, and $\varepsilon_1, \varepsilon_2, \ldots$ are iid standard normal. Then

$$\tilde{S}(j) = S(j) + \sigma k^{-1} \sum_{r=1}^{k} \varepsilon_{j+rT} = S(j) + \sigma k^{-1/2} \eta_j\,, \;\; j = 1, 2, \ldots, T, \tag{5.1.7}$$

where $\eta_j := k^{-1/2} \sum_{r=1}^{k} \varepsilon_{j+rT}$ are again iid standard normal. Thus, if k is large enough (that is, if n is large and T is relatively small), then the conventional estimator should perform well.

On the other hand, if k is small and, respectively, the period T is large (and this is a rather typical case in many applications), then another approach may be used. It is apparent that (5.1.7) is an equidistant nonparametric regression model with $S(x)$ being the regression function and the period T being the sample size. Thus, the universal nonparametric estimator of Section 4.2 (or Section 4.2) may be used straightforwardly to estimate $S(j)$ based on T observations (5.1.7). Note that the nonparametric estimator smoothes the conventional estimate (5.1.6).

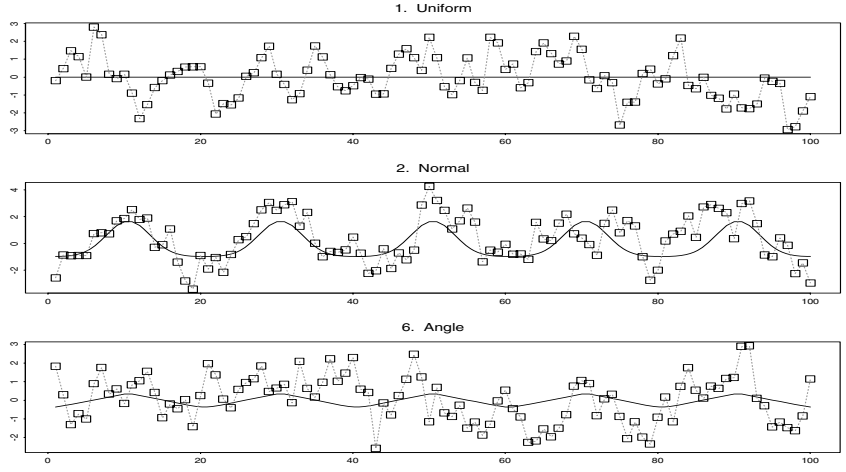

FIGURE 5.1. Three simulated time series shown by squares connected by the dotted lines, with no trends and different seasonal components of period $T = 20$. The underlying seasonal components, shown by solid lines, are the Uniform, the Normal, and the Angle corner functions minus 1. Stochastic terms are generated by a Gaussian ARMA(1, 1) process $\varepsilon_t - a\varepsilon_{t-1} = \sigma(Z_t + bZ_{t-1})$, where $\{Z_t\}$ are iid standard normal, $a = 0.4$, $b = 0.3$, and $\sigma = 1$. {The length n of the realizations is controlled by the argument n. The period T of seasonal components is controlled by the argument *Per*. The parameters σ, a, and b of the ARMA(1,1) noise are controlled by the arguments *sigma, a,* and *b*. Use $|a| < 1$. The argument *set.seas* allows one to choose any 3 corner functions as the underlying seasonal components.} [*n=100, Per=20, sigma=1, a=.4, b=.3, set.seas=c(1,2,6)*]

Figure 5.1 shows 3 simulated time series (with no trends) of length $n = 100$ where seasonal components have period $T = 20$ and they are the corner functions Uniform, Normal, and Angle minus 1. Noise terms are generated by a Gaussian ARMA(1, 1) process $\varepsilon_t - 0.4\varepsilon_{t-1} = Z_t + 0.3Z_{t-1}$.

Let us consider these particular time series. It is known that the time series in Figure 5.1.1 has neither trend nor seasonal component, while the two others do have seasonal components, but is it apparent from the data? And what do we mean here by a trend and a seasonal component?

Let us begin the discussion with the second question. According to (5.1.5), the trend and seasonal components are separated in the frequency domain. Because it is easier to think about the frequency domain in terms of periods, let a deterministic periodic component with period less than $T_{\max}$ be referred to as a seasonal component, and as a trend component otherwise. For instance, if we set $T_{\max} = 40$, then no pronounced trend with this or larger period is visible in Figure 5.1.1, while a seasonal component with period between 10 and 20 is a likely bet. Note that such an illusion of the presence of a seasonal component is a trademark of ARMA processes. Moreover, long-memory processes, considered in Section 4.8, may create

even an illusion of a trend component. Now assume that $T_{\max} = 5$. In this case apparently no seasonal component is present, but a slightly smoothed dotted line that connects the observations may be a possible bet on an underlying trend.

Now let us look at the second diagram with the underlying Normal seasonal component. The fact that this is very pronounced seasonal component makes it easier to conclude that a seasonal component does exist. On the other hand, it is not an easy task to estimate it; just look at the time between 70 and 90 where the shape of this component is completely lost due to the noise.

The third diagram illustrates another typical challenge caused by dependent observations. Look at the first half of the observations; here a majority of observations are above the seasonal component. The situation changes for the second part of the observations. This is what may cause great confusion in any estimate, and this is what the dependency means.

Now let us return to our discussion of the separation of seasonal component and trend. Here all depends on the choice of $T_{\max}$, or in other words, on what we mean by a slowly changing trend component. Fortunately, typically this is a clear-cut issue for practical applications. For instance, for a long-term money investor, $T_{\max}$ is about several years, while for an active stock trader it may be just several days or even hours.

If $T_{\max}$ is specified, then $J_{\max}$ in (5.1.5) is defined as the minimal integer such that $\varphi_{J_{\max}}(x + T_{\max}) \approx \varphi_{J_{\max}}(x)$ for all x. For instance, for the cosine basis on $[0, n]$ with the elements $\varphi_0(t) := n^{-1/2}$, $\varphi_j(t) := (n/2)^{-1/2} \cos(\pi j t/n)$, $j = 1, 2, \ldots$, $0 \le t \le n$, we get

$$J_{\max} = \lfloor 2n/T_{\max} \rfloor. \tag{5.1.8}$$

Now let us return to the first question, namely, can we visualize any trend or seasonal component in the particular realizations shown in Figure 5.1? Assume that $T_{\max}$ is defined approximately correctly, say $T_{\max} = 30$. In other words, if we detect a deterministic cyclical component with period less than 30, then it is a seasonal component; otherwise it is a trend. Then, even in this case of correctly chosen $T_{\max}$, it is not easy (or even impossible) to correctly realize the underlying seasonal components. The issue, of course, is that the stochastic term is relatively large.

Now let us look at how the conventional estimator (5.1.6) and the nonparametric estimator perform for time series generated as in Figure 5.1.

Figure 5.2 exhibits estimated values of seasonal components for 8 different seasonal components, which are our familiar corner functions minus 1. What we see is an example of how to choose an optimal smoothing. Recall that the nonparametric estimator smoothes estimates calculated by the conventional estimator, and it performs well only if the period of a seasonal component is relatively large. In all the cases, with the apparent exception of the Delta (and maybe the Strata), the nonparametric estimator performs

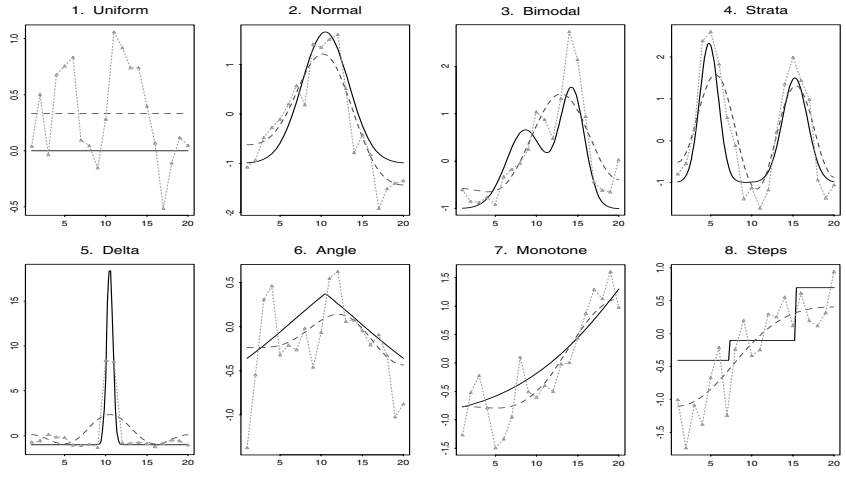

FIGURE 5.2. Seasonal components computed by the conventional method (5.1.6) (shown by triangles connected by dotted lines) and by the nonparametric universal estimator (dashed lines), which smoothes the conventional estimates. The underlying seasonal components are shown by solid lines. Time series are generated similarly to ones shown in Figure 5.1. {The first 5 arguments allow one to change the simulated time series; they are explained in the caption of Figure 5.1. The rest of the arguments control the coefficients of the universal estimator explained in the caption of Figure 4.5.} [*n=100, Per=20, sigma=1, a=.4, b=.3, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2*]

well. In other words, it smoothes correctly. For the cases of the Strata and apparently the Delta, the conventional method is better. Exercise 5.1.14 is devoted to choosing optimal values of coefficients of the nonparametric estimator, and Exercise 5.1.15 to the cases where it is worthwhile to employ this estimator. In short, we should keep in mind that in the regression model (5.1.7) the period T of an estimated seasonal component (regression function) plays the role of the sample size. Thus the case of the period $T = 20$ is a challenging problem for a nonparametric estimator.

5.2 Estimation of Spectral Density

There are two rather distinct approaches to the analysis of stationary time series: the spectral (frequency) domain approach and the time domain (dynamic) approach. The particular strength of the spectral approach is the simplicity of visualization of periodicities and separation long-term and short-term effects, whereas the time domain approach with its explicit equations for an underlying time series, and an important particular case of ARMA(p, q) processes, is easy for predictions and describing the dynamics

of time series. This section is concerned with the first approach, namely, with nonparametric spectral density estimation, but with eyes open to the possibility that an underlying process is an ARMA(p, q) time series.

Analysis of time series is customarily based on the assumption of second-order stationarity after removing the trend and seasonal components and (if necessary) rescaling the original data. This explains why both the study and estimation of second-order characteristics is the most important topic in the analysis of such time series. Let X_t, for $t = \ldots, -1, 0, 1, \ldots$, be a second-order stationary time series with mean 0 and autocovariance function $\gamma(j) := E\{X_{t+j}X_t\}$. Then the second-order properties of a time series are completely described by its autocovariance function, or, equivalently, under mild conditions (for instance, a sufficient condition is $\sum_{j=-\infty}^{\infty} |\gamma(j)| < \infty$), by its Fourier transform, which is called the *spectral density* function,

$$f(\lambda) := (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma(j) \cos(j\lambda) \tag{5.2.1}$$

$$= (2\pi)^{-1}\gamma(0) + \pi^{-1} \sum_{j=1}^{\infty} \gamma(j) \cos(j\lambda), \quad -\pi < \lambda \le \pi. \tag{5.2.2}$$

Here the frequency λ is in units radians/time, and to get (5.2.2) we used the relation $\gamma(-j) = \gamma(j)$.

Because the autocovariance function is symmetric, the spectral density is also symmetric in λ about 0, i.e., the spectral density is an even function. Thus, it is customary to consider a spectral density on the interval $[0, \pi]$. The spectral density is also a nonnegative function (like the probability density), and this explains why it is called a density.

Formula (5.2.1) shows why the spectral density is such a good tool for searching for periodicities; indeed, a peak in $f(\lambda)$ at frequency $\lambda = \lambda^*$ indicates a possible periodic phenomenon with period

$$T^* = \frac{2\pi}{\lambda^*}. \tag{5.2.3}$$

This formula explains why spectral domain analysis is the main tool in searching after periods of seasonal components. The next section gives us an example of how to use this formula.

Now let us explain how to estimate the spectral density. Let a finite realization $X_1, \ldots, X_n$ of a second-order stationary time series (recall that we always assume that its mean is zero) be given. The classical *sample autocovariance estimator* is defined as

$$\hat{\gamma}(j) := n^{-1} \sum_{l=1}^{n-j} X_{l+j}X_l, \quad j = 0, 1, \ldots, n-1. \tag{5.2.4}$$

Note that the divisor n is not equal to the number $n - j$ of terms in the sum. Thus, the sample autocovariance is a biased estimator. On the other

hand, this divisor ensures that an estimate corresponds to some second-order stationary series. (For all our purposes the divisor $n-j$ may be used as well.)

Then, according to (5.2.2), if one wants to estimate a spectral density, a natural step is to plug in the sample autocovariance function in place of an unknown autocovariance function. The resulting estimator (up to the factor $1/2\pi$) is known as a *periodogram*,

$$I(\lambda) := \hat\gamma(0) + 2\sum_{j=1}^{n-1}\hat\gamma(j)\cos(j\lambda) = n^{-1}\Big|\sum_{l=1}^{n} X_l e^{-il\lambda}\Big|^2. \tag{5.2.5}$$

Here i is the imaginary unit, i.e., $i^2 := -1$, $e^{ix} = \cos(x) + i\sin(x)$, and the periodogram is defined at the so-called *Fourier frequencies* $\lambda_k := 2\pi k/n$, where k are integers satisfying $-\pi < \lambda_k \le \pi$. Examples of periodograms are given below.

This simple tool for spectral-domain analysis, invented in the late nineteenth century, has been both the glory and the curse of this analysis. The glory, because many interesting practical problems were solved at a time when no computers were available. The curse, because the periodogram, which had demonstrated its value for locating periodicities, proved to be an erratic and inconsistent estimator.

The reason for the failure of the periodogram is clear from the point of view of nonparametric curve estimation theory discussed in Chapter 3. Indeed, based on n observations, the periodogram estimates n Fourier coefficients (values of an underlying autocovariance function) and then just plugs them in. This explains the erratic performance and inconsistency.

Thus, it is no surprise that in the 1940s interest in frequency-domain inference was reawakened by ideas of averaging (smoothing) the periodogram in the neighborhood of each Fourier frequency (today known as kernel smoothing, discussed in Chapter 8) and by procedures of orthogonal series estimation, in which the sample autocovariance function is smoothed. In particular, the latter approach led to *lag-window* Tukey estimators

$$\tilde f(\lambda) := (2\pi)^{-1}\hat\gamma(0) + \pi^{-1}\sum_{j=1}^{J} w(j/J)\hat\gamma(j)\cos(j\lambda), \tag{5.2.6}$$

which are the cosine series estimators familiar from the previous chapters. Here the lag window function $w(x)$ is such that $|w(x)| \le 1$ and $w(x) = 0$ for $x > 1$, and J is called the window width or cutoff. For instance, the simplest lag window function is rectangular, where $w(x) = 1$ for $x \le 1$, and this implies a truncated estimator.

This series estimator is the most apparent application of the orthogonal series approach, since the spectral density is defined via the cosine series.

Thus, for the problem of estimation of the spectral density, the universal data-driven estimator (3.1.15) of Section 3.1 may be employed

straightforwardly with $\hat{\gamma}(j)$ used in place of $\hat{\theta}_j$ and where the coefficient of difficulty,

$$d := 2\pi \int_{-\pi}^{\pi} f^2(\lambda) d\lambda = \gamma^2(0) + 2 \sum_{j=1}^{\infty} \gamma^2(j), \tag{5.2.7}$$

is estimated by

$$\hat{d}_n := \hat{\gamma}^2(0) + 2 \sum_{j=1}^{J_n} \hat{\gamma}^2(j). \tag{5.2.8}$$

Here the sequence J_n is the same as in Section 3.1.

Exercise 5.2.8 shows that if an underlying time series is a causal ARMA process with bounded fourth moments, then

$$E\{(\hat{\gamma}(j) - \gamma(j))^2\} = dn^{-1}(1 + r_{nj}), \tag{5.2.9}$$

where $r_{nj} \to 0$ as both n and j increase. Relation (5.2.9) explains formula (5.2.7) for the coefficient of difficulty of estimation of the spectral density.

Figure 5.3 illustrates the performance of the estimator for an underlying Gaussian ARMA(1, 1) time series $Y_t - 0.4Y_{t-1} = 0.5(Z_t + 0.5Z_{t-1})$. The top diagram shows a particular realization that "slowly" oscillates over time. This is because here the covariance between Y_t and Y_{t-1} is positive. This follows from the following formula for calculating the autocovariance function of the causal ARMA(1, 1) process $Y_t - aY_{t-1} = \sigma(Z_t + bZ_{t-1})$, $|a| < 1$:

$$\gamma(0) = \frac{\sigma^2[(a+b)^2 + 1 - a^2]}{(1-a^2)}, \quad \gamma(1) = \frac{\sigma^2(a+b)(1+ab)}{(1-a^2)},$$
$$\gamma(j) = a^{j-1}\,\gamma(1), \quad j \ge 2. \tag{5.2.10}$$

See sketch of the proof in Exercise 5.2.9. Note that if $a > 0$ and $b > 0$, then $\gamma(1) > 0$, and a realization will "slowly" change over time. On the other hand, if $a + b < 0$ and $1 + ab > 0$ (for instance, consider a moving average MA(1) process $Y_t = \sigma(Z_t + bZ_{t-1})$ with negative b), then a realization may change its sign almost every time. Thus, depending on a and b, we may see either slow or fast oscillations in a realization of an ARMA(1, 1) process.

Figure 5.3.2 shows the underlying theoretical spectral density of the ARMA(1, 1) process. As we see, because here both a and b are positive, in the spectral domain low frequencies dominate high frequencies. (To look at the inverse situation, the MA(1) process mentioned earlier may be considered.) The formula for calculating the spectral density is $f(\lambda) = \sigma^2|1 + be^{i\lambda}|^2/[2\pi|1 - ae^{i\lambda}|^2]$, and it is a particular case of the

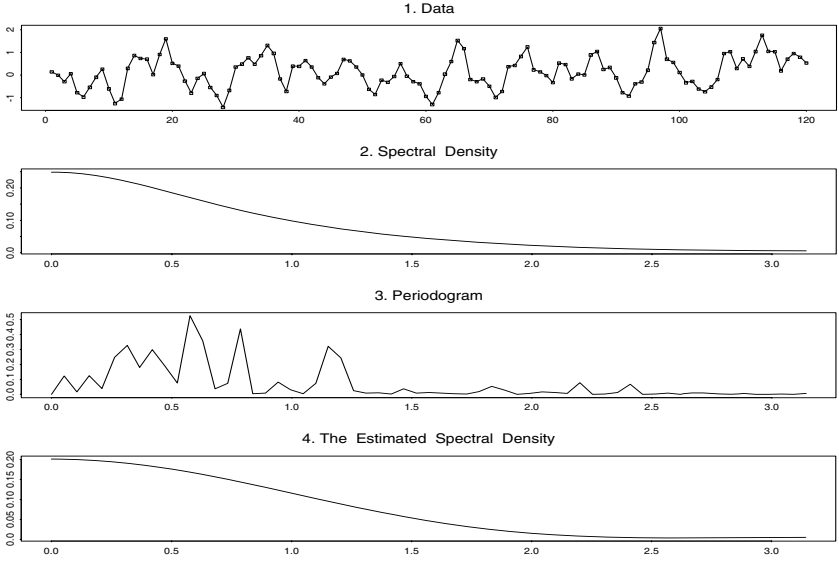

FIGURE 5.3. The top diagram shows a particular realization of a Gaussian ARMA(1, 1) time series $Y_t - aY_{t-1} = \sigma(Z_t + bZ_{t-1})$, $t = 1, 2, \ldots, n$, where $a = 0.4$, $b = 0.5$, $\sigma = 0.5$, and $n = 120$. The diagram below shows the spectral density of this ARMA process. The two bottom diagrams show the periodogram estimate and the universal spectral density estimate. {The length n of a realization is controlled by the argument n. The parameters of an ARMA(1, 1) process are controlled by the arguments *sigma, a*, and *b*. Use $|a| < 1$. All the other arguments control the coefficients of the universal estimator (3.1.15), and they are explained in the caption of Figure 3.2. Note that the string *sp* is added to these arguments to indicate that they control the coefficients of the universal spectral density estimator.} [*n=120, sigma=.5, a=.4, b= .5, cJ0sp=4, cJ1sp=.5, cJMsp=6, cJTsp=4, cBsp=2*]

following formula for a causal ARMA(p, q) process defined at (5.1.2),

$$f(\lambda) = \frac{\sigma^2 \left|1 + \sum_{j=1}^{q} b_j e^{-ij\lambda}\right|^2}{2\pi \left|1 - \sum_{j=1}^{p} a_j e^{-ij\lambda}\right|^2}. \tag{5.2.11}$$

Now let us see how the periodogram (5.2.5) and the universal nonparametric estimator show us the underlying spectral density. (The interesting feature of positively correlated time series is that it may create an illusion of a seasonal component; see Figure 5.3.1. Thus, it will be of a special interest to watch how the nonparametric estimates handle such a realization.) The periodogram is shown in Figure 5.3.3. As we see, it does not resemble the underlying spectral density, and moreover, its mode at frequency $\lambda^* \approx 0.55$ indicates the possibility of a seasonal component with period

11 (the formula (5.2.3) was used to find this period). It is easy to believe in this conclusion after visualizing the series in Figure 5.3.1, but we know that this is just an illusion and there is no seasonal component. This is why a periodogram cannot be used as a reliable tool for searching for cyclical components.

The bottom diagram exhibits the universal estimate, which correctly shows the absence of any seasonal component (there are no modes in the frequency region of possible seasonal components). Also, the estimate nicely resembles the underlying spectral density.

Now let us discuss the following important question, which always arises when one uses a nonparametric estimator for the case of a parametric underlying model. Assume that we have some information about an underlying time series, for instance, that it is an ARMA(p, q) process. Then, is it worthwhile to use a nonparametric estimator that ignores this information?

To answer this question, let us make some preliminary comments about parametric spectral density estimators. For the case of a Gaussian AR time series, a well-known parametric adaptive spectral density estimator, supported by S–PLUS, is an estimator based on Akaike's information criterion (AIC). In short, this is a parametric penalized maximum likelihood estimator of the order p; see more in Chapter 8 about the method. We do not discuss this parametric estimate in more detail because it is supported by S–PLUS and we can use it as a given tool. (If an underlying process is ARMA(p', q), then S–PLUS recommends approximating it by an AR(p) process, that is, again use that parametric estimate.) The only information that is required by this estimator is the largest possible value of p.

Thus, let us explain how to compare our universal nonparametric estimator with this parametric one. We perform a Monte Carlo study that should be absolutely favorable to the parametric estimate; here this means that an underlying time series is a Gaussian AR(p) process with $p \le 7$, and this maximal order 7 is given to the parametric estimator. Then, the parametric estimator is used as an oracle (because it knows the underlying model) for the nonparametric one, and their performances are compared. (Note that the idea of this experiment resembles the experiments with oracles discussed in Sections 3.2–3.3.)

Our particular experiment is as follows. For each pair (p, n) of $p \in \{1, 2, 3, 4, 5, 6, 7\}$ and $n \in \{30, 50, 100, 300, 500, 1000\}$, 1,000 independent Monte Carlo simulations of a causal Gaussian AR(p) time series are performed where roots of the autoregressive polynomials are iid uniform with absolute values between 2 and 10. Then, for each realization from the set of 1,000 simulations, the parametric oracle's estimate and the universal estimate are calculated, and the ratios of their integrated squared errors (ISE) are computed. Table 5.1 displays the sample medians of these ratios. If the ratio is larger than 1, then the oracle (Akaike's parametric estimator, which knows that the underlying model is AR(p) and $p \le 7$) is better than the universal estimator, and vice versa.

Table 5.1. *Median Ratios of Sample ISE: Universal/Oracle*

p	$n = 30$	$n = 50$	$n = 100$	$n = 300$	$n = 500$	$n = 1000$
1	1.2	1.5	1.5	1.7	2.8	2.1
2	1.3	1.3	1.5	1.6	1.7	1.8
3	0.9	1.3	1.3	1.3	1.5	1.5
4	0.9	1.2	1.3	1.3	1.4	1.4
5	1.1	1.0	1.0	1.1	1.1	1.5
6	1.0	0.9	1.0	1.2	1.1	1.5
7	1.0	1.0	1.0	1.2	1.1	1.5

There is no surprise that the outcome is favorable to the parametric oracle, especially for large n and small p; after all, the oracle "knows" the underlying model up to a fixed number of parameters, whereas the nonparametric estimates are based only on data. Nonetheless, the outcome of the experiment is promising, because the ratios are very reasonable, especially for the smallest sample sizes (which we are primarily interested in) and larger orders (more complicated models). In short, even if one knows the underlying AR process up to several parameters, the best parametric estimator does not significantly outperform the universal data-driven estimator for the case of small sample sizes.

To shed further light on the issue, consider a similar experiment only for some specific Gaussian AR(1) processes $X_t - aX_{t-1} = Z_t$ with a equal to 0.5, 0.1, 0.05, and 0. For $n = 100$, the ratios are 1.6, 0.95, 0.92, and 0.87. Thus for the case of small a (including a white noise time series), the nonparametric estimator outperforms the parametric ones. Also note that if an underlying process is not AR(1) but, for instance, a Gaussian MA(1) process $X_t = Z_t + 0.5Z_{t-1}$, then the ratio is 0.41, that is, the nonparametric estimator dramatically outperforms the parametric one.

Thus, it is fair to conclude that only for the cases where a practitioner is absolutely sure in an underlying parametric dynamic model is there an incentive to use only a parametric estimator. Otherwise, it is wiser to begin with a nonparametric estimate as a "first look at the data at hand" and then, if it confirms a prior opinion about an underlying parametric model, use a parametric estimator. Such a conservative approach allows one to avoid inconsistent estimation due to a wrong prior assumption.

5.3 Example of the Nonparametric Analysis of a Time Series

Let us combine all the earlier steps in the nonparametric analysis of a time series and explore them together via an example. The example and all steps are illustrated by Figure 5.4. Because everything in this section is about this figure, this is the only section where it is also discussed how to repeat

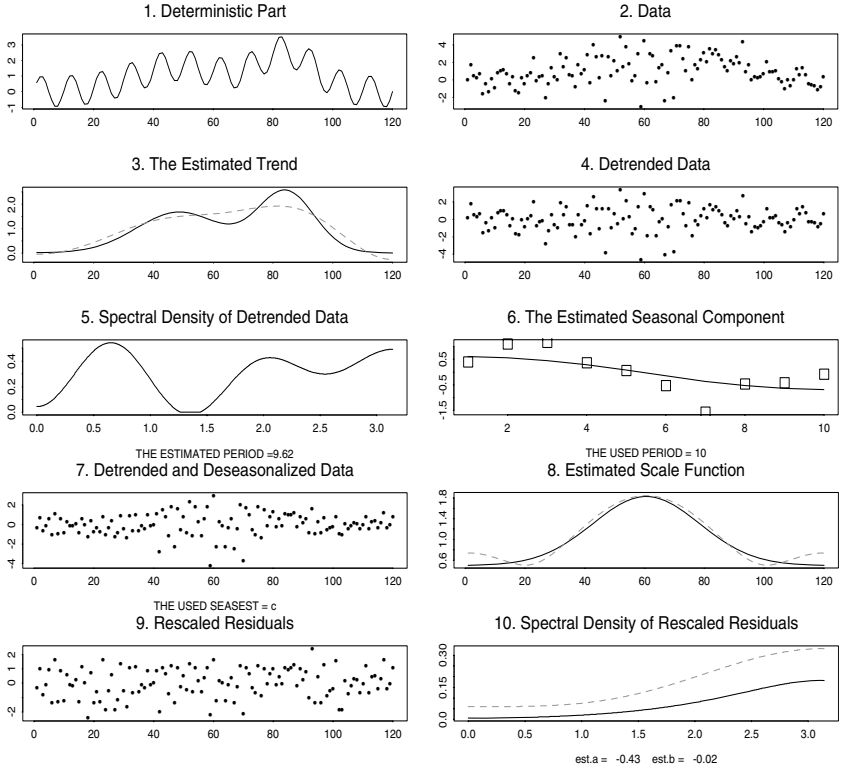

FIGURE 5.4. The comprehensive nonparametric analysis of a time series. In diagrams 3, 8, and 10 the estimates are shown by dashed lines and the underlying functions by solid lines. In diagram 6 the conventional estimate is shown by squares, and its smoothing by the universal estimate is shown by the solid line. The subtitle to diagram 7 shows which estimate, c - conventional or u - universal, was used. The subtitle to diagram 10 shows the coefficients of the ARMA(1, 1) process that give the best fit to the time series of rescaled residuals. {Use $|a| < 1$.} [*n=120, trendf=3, scalef=2, sigmasc=.5, ss=1, sc=1, a = -.3,b= -.5, TMAX=35, Tseas=10, ManualPer=F, seasest= "c",set.period=c(8,12), set.lambda=c(0,2), lbscale=.1, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2, cJ0sp=4, cJ1sp=.5, cJMsp=6, cJTsp=4, cBsp=2*]

this figure using the software (it is simply impossible to discuss all details in the caption).

The underlying deterministic part $f(t)+S(t)$, $1 \le t \le n$, is shown in Figure 5.4.1, and it resembles many practical examples. Here the trend $f(t)$ is the Bimodal corner function (with domain $[1, n]$), and the seasonal component is a trigonometric function $S(t) := s_{\rm s} \sin(2\pi t/T_{\rm seas}) + s_{\rm c} \cos(2\pi t/T_{\rm seas})$ with the period $T_{\rm seas}$. The length of observations n is controlled by the argument n with the default value 120. The trend component is chosen by the argument *trendf*, and the seasonal component is set by the arguments

ss, *sc*, and *Tseas*; the default values of these arguments are *trendf* = 3, $ss = 1$, $sc = 1$, and $Tseas = 10$.

The stationary stochastic term is generated by a normed ARMA(1, 1) process $\varepsilon_t = \varepsilon'_t/(E\{\varepsilon_t^2\})^{1/2}$, where $\varepsilon'_t - a\varepsilon'_t = Z_t + bZ_{t-1}$ and $\{Z_t\}$ is a standard Gaussian white noise. The default values are $a = -0.3$ and $b = -0.5$. We discussed an ARMA(1, 1) process earlier; thus we may predict that this particular stochastic component will be highly oscillatory and its spectral density should monotonically increase in frequency. Then this stationary stochastic term is multiplied by a scale function. The scale function is a coefficient $\sigma_{\rm sc}$ times 1 plus the Normal corner function with the domain $[1, n]$, i.e., the scale function is $\sigma_{\rm sc}(1 + f_2(l/n))$, where $f_2(x)$ is the Normal corner function. The choice of a corner function, used in the scale function, is controlled by the argument *scalef* with the default value 2, and the factor $\sigma_{\rm sc}$ is controlled by the argument *sigmasc* with the default value 0.5.

Data are generated by adding the scaled stochastic term to the deterministic one. A particular realization is shown by dots in Figure 5.4.2, and this is the data set (time series) at hand. Can you realize the underlying trend, seasonal component, scale function, and the structure of the noise from the data? The answer is probably "no," so let us see how the nonparametric data-driven procedures discussed earlier handle this data set.

The first step is the nonparametric estimation of the trend. Recall that according to Section 5.1, the trend and seasonal components are separated in the frequency domain, see (5.1.5), and the boundary $J_{\rm max}$ is defined via a manually chosen $T_{\rm max}$. For this data set we choose the default $T_{\rm max} = 35$, which according to (5.1.8) implies $J_{\rm max} = 7$; the choice of $T_{\rm max}$ is controlled by the argument $TMAX$. By choosing this default value we assume that a cosine approximation (5.1.5) with $J_{\rm max} = 7$ may approximate well an underlying trend and, at the same time, does not touch a possible seasonal component, which, by the assumption, has a period less than 35.

The nonparametric estimate of the trend (the dashed line) is shown in Figure 5.4.3. It clearly oversmooths the underlying trend, but it is necessary to be fair toward this estimate. Yes, this estimate is much worse than Binomial's best estimates, which we saw in the previous chapters. On the other hand, now the problem is much more complicated: The setting is heteroscedastic with the pronounced scale function, the errors are dependent, and there is a significant seasonal component whose period and magnitude are comparable with the distance and the difference between the modes of the underlying Bimodal trend shown by the solid line. This is what causes the trend estimate to be essentially smoother than the underlying Bimodal trend. Actually, even by visualizing the first diagram where the deterministic part is shown, it is not an easy task to realize the modes of the underlying Bimodal model, and the situation becomes much more complicated with the noisy time series exhibited in the second diagram.

The next step is to detrend the data (subtract the estimated trend from the original data), and the result is shown in the fourth diagram (Figure

5.4.4). Now, based on this time series, one must recover the underlying seasonal component. Can you recognize the seasonal component in this detrended data? Because we know that the seasonal component is a trigonometric function, we can see it in the right tail, less so in the left tail, but in the main middle part of the time series the seasonal component is absolutely blurred. This is what makes the heteroscedastic setting so complicated. (One of the options is to use a Box–Cox transformation discussed in the subsection "Estimation of Scale Function" of Section 5.1; we do not use it here because we would like to see how the "pure" nonparametric methods will perform.)

Now let us see how our nonparametric analysis performs. The nonparametric spectral density estimate of the detrended data is shown in Figure 5.4.5. {Recall that as in Section 5.2, arguments of the spectral density estimate have the attached string *sp*, for instance, *cJ0sp* is the argument that controls the coefficient c_{J0} of this estimate. This allows one to use separate arguments for the regression estimate, which recovers the trend and scale functions, and the spectral density estimate.}

Diagram 5 indicates that the detrended data have a spectral density with a pronounced mode at the frequency about 0.6. The period 9.62 (the estimated period) calculated according to (5.2.3) is given in the subtitle. The corresponding rounded (to the nearest integer) period is 10, and this is exactly the underlying period.

While for these particular data the rounded estimated period has been determined correctly, this is not always the case. The small sample sizes and large errors may take their toll and lead to an incorrect estimate of the period. We shall discuss such a case a bit later.

Then the rounded estimated period is used to estimate the underlying seasonal component. Squares in Figure 5.4.6 show the conventional estimate (5.1.6); the solid line shows how the universal nonparametric estimate smooths the conventional estimate. As we see, the conventional estimate is not perfect, but it is fairly good for the setting considered. Note that its magnitude is correct, and the phase is shown absolutely correctly. Keep in mind that each point is the average of just 12 observations, so even for a parametric setting this would be considered a small sample size. The nonparametric estimate apparently oversmooths the data because the period 10 is too small; recall the discussion in Section 5.1.

The next step is to deseasonalize the detrended data, that is, to subtract an estimate of the seasonal component. The argument *seasest* (which is shorthand for seasonal estimate) allows one to use either the conventional or the universal nonparametric estimate of the seasonal component by setting $seasest = "c"$ or $seasest = "u"$, respectively. The data obtained are shown in Figure 5.4.7, and the argument used is given in the subtitle. Note that at this step the data may be referred to as the time series of residuals because the original data set is detrended and deseasonalized (the estimated deterministic part is removed).

The detrended and deseasonalized time series is clearly not stationary, since its variability in the middle part is essentially larger than in the tails. This conclusion is supported by the estimate of the underlying scale function (which is $\sigma_{\rm sc}(1 + f_2(t/n)$ with $f_2(x)$, $0 \le x \le 1$, being the Normal corner function and $\sigma_{\rm sc} = 0.5$). The scale function is shown by the solid line in Figure 5.4.8. The estimate (dashed line) is almost perfect in the middle part, but the tails "spoil" the outcome. This is explained by the cumulative effect of a not perfect estimate of the deterministic part, the small sample size, and dependent errors. Nevertheless, under the circumstances and because the range of the underlying scale function is shown just perfectly, it is fair to rate this particular scale estimate as a good one.

The next step is to rescale the residuals shown in Figure 5.4.7 to obtain a stationary noise. The rescaled residuals are simply the residuals divided by the estimated scale function. To avoid a zero divisor, the estimate is truncated from below by the argument *lbscale*; the default value is 0.1.

Thus, here we divide the detrended and deseasonalized data shown in Figure 5.4.7 by the estimated scale function shown in Figure 5.4.8. The result is shown in Figure 5.4.9. The hope is that these data are stationary and that they correspond to a simple stochastic process like an ARMA(p, q) with small orders p and q. Visual analysis shows that there is no apparent trend, or a seasonal component, or a scale function. Thus, our final step is to look at the spectral density estimate of the rescaled residuals. The estimate (the dashed line) is shown in Figure 5.4.10. As we see, the estimate exhibits no pronounced modes (which can indicate the presence of deterministic periodic components), and we see that in this time series high frequencies dominate low frequencies. Thus, this time series looks like a stationary one, and with the help of the experience gained from Figure 5.3, we may conjecture that an ARMA(1, 1) process $\varepsilon_t - a\varepsilon_{t-1} = \sigma(Z_t + bZ_{t-1})$ with negative a and b may be a good bet on an underlying stochastic term. Indeed, the underlying spectral density (the solid line) has a similar shape, and the fact that it is below the estimate tells us that the rescaled residuals have a larger variance than a typical realization from the underlying ARMA(1, 1) process where $a = -0.3$ and $b = -0.5$. Also, the subtitle shows us the estimated parameters of the ARMA(1, 1) process that gives the best fit to the data. They are obtained using the S–PLUS function (parametric maximum likelihood estimate) **arima.mle**.

This finishes our analysis of this particular time series.

Now let us return to Figure 5.4.5. Here the frequency of the mode correctly defines the period by the formula (5.2.3), but this is not always the case. First, there may be several local modes created by both a seasonal component and a stochastic component, and large errors may produce a wrong global mode. As a result, the period will be estimated incorrectly. One of the possibilities to avoid such a complication is to use prior information about the domain of possible periods. To play around with this possibility, two arguments are added to Figure 5.4, namely, *set.period* and

set.lambda. The first one, *set.period = c(T1,T2)*, allows one to skip estimation of a seasonal component whenever an estimated period is beyond the interval $[T1, T2]$. The second argument, $set.lambda = c(\lambda_1, \lambda_2)$, allows one to restrict the search for the mode to this particular frequency interval. While these two arguments do a similar job, they are good tools for gaining the necessary experience in dealing with the time and frequency domains. {Graphics 6 and 7 are skipped if the estimated period is beyond the interval [T1,T2], in which case a warning statement is issued.}

The second reason for the failure of the estimation of the period is that due to large noise and small sample size, the mode of an estimated spectral density may be relatively flat. As a result, even if a spectral density estimate is close to an underlying density in the sense of integrated squared error, locations of the estimated mode and the underlying mode may differ significantly. To understand why, consider, as an example, frequencies $\lambda_1^* = 0.6$, $\lambda_2^* = 0.59$, and $\lambda_3^* = 0.54$. Then, according to (5.2.3), the corresponding periods are $T_1^* = 2\pi/0.6 = 10.47$, $T_2^* = 2\pi/0.59 = 10.64$, and $T_3^* = 2\pi/0.54 = 11.63$, which imply the rounded periods 10, 11, and 12, respectively. Thus, due to the rounding a relatively small error in the location of a mode may imply a significant error in the estimated period.

Two questions immediately arise: how to detect such a case and how to correct the mistake. To answer the first question, let us look at another realization of Figure 5.4 (i.e., another realization of the noise term), shown in Figure 5.5. As we see, here the estimated period (see the subtitle for Figure 5.5.5) is 10.87, and this leads to the wrong period, 11. Let us assess the consequences of using this wrongly estimated period. First, the estimated seasonal component in no way resembles the underlying one. While this rather chaotic estimate cannot be the indicator of a wrongly estimated period, it should raise a flag of suspicion. Then, we see that the rescaled residuals in Figure 5.5.9 apparently exhibit a cyclical component. This is a one more reason to suspect the mistake. The estimated scale function (the dashed line in diagram 8) is dramatically oversmoothed because now the subtracted estimated seasonal component plays the role of an extra additive noise.

Finally, the estimated spectral density (the dashed line) in Figure 5.5.10 indicates that the seasonal component was not removed. Indeed, we see that the shape of this estimate resembles the shape of the spectral density of detrended data shown in Figure 5.5.5. This spectral density of rescaled residuals is the most reliable indicator of a wrongly estimated period.

The obvious method to cure such a mistake is to try a different period for estimation of a seasonal component, and here the apparent choice of the period is $T = 10$, which is the rounded-down estimated period. {To do this, set the argument *ManualPer = T* (in S–PLUS "T" stands for "True" and "F" for "False"). This stops the calculations at diagram 5, and the first 5 diagrams are displayed. Then the program prompts for entering a period from the keyboard. At the prompt 1: enter a period (here it should

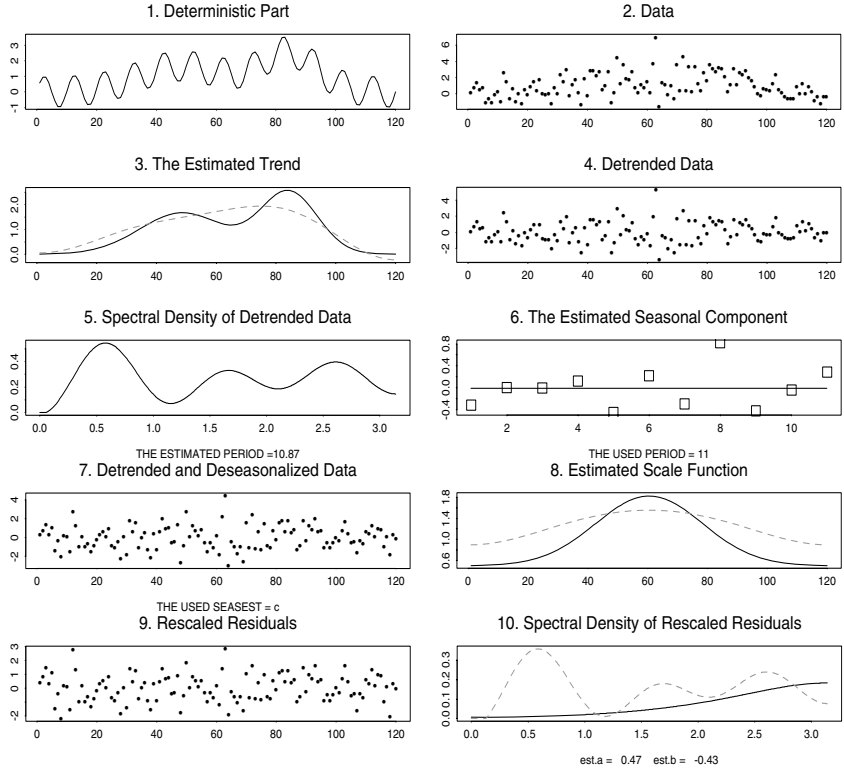

FIGURE 5.5. Another realization of Figure 5.4 where period of the seasonal component is estimated incorrectly.

be 10, but any integer period may be tried) from the keyboard and then press Return; then at the prompt 2: just press Return. This completes the procedure, and the seasonal component will be calculated with the period entered. The period will be shown in the subtitle of diagram 6.} We do not illustrate this procedure by a separate figure because the result is similar to diagrams shown in Figures 5.4.6–5.4.10.

Another useful practical comment is as follows. In many cases a spectral density estimate of rescaled residuals has a relatively large left tail while an underlying theoretical spectral density does not. A particular example will be given in the next section. One of the typical reasons for such a mistake is a poorly estimated trend. Unfortunately, for the cases of small sample sizes and relatively large errors there is no cure for this "disease," but knowledge of this phenomenon may help in explaining a particular outcome.

5.4 Case Study: Missing Observations

In the previous section we have considered the case of a classical realization $X_1, X_2, \ldots, X_n$ of a time series $\{X_t\}$. In many practical situations some of the observations may be skipped (missing). This happens due to stochastic circumstances or because there is no way to obtain realizations at some particular moments. Classical examples of the second case are as follows. Suppose that a researcher should collect some daily data about a group of students at a particular school. Since schools are closed on weekends, every sixth and seventh observation will be missed. Another example is observations of an object from a satellite that periodically "loses" the object. We shall see that a case of deterministically skipped data may be essentially worse than a case of data skipped at random. Thus, the more difficult case of deterministically skipped data will be of our primary interest.

Another interesting practical interpretation of the setting is the case of spatial data (recall the discussion in Section 5.1) that are not collected at a regular grid but may be approximated by a model of data at a regular grid with skipped (missing) observations. For such a setting, for instance, geostatistics considers the problem of interpolation of an underlying trend at skipped points as one of the most important.

What are the necessary changes in our estimates to consider a case of missing observations? Let us again consider the general problem illustrated by Figure 5.4 and discuss a similar analysis for the particular example where every sixth and seventh observation is missed. This mimics a time series of weekly observations with missing weekends.

Figure 5.6 illustrate this setting. The first diagram shows by dots the unobserved deterministic part, which is the same as in Figure 5.4.1, only here every sixth and seventh observation ("weekends") is skipped. Figure 5.6.2 shows observed noisy observations (data at hand). Note that in this time series of length 120 only 86 observations are available and 34 are missing. Thus, the quality of estimation should be worse than for the case considered in Figure 5.4 simply because the sample size is smaller.

The smaller number of observations is not the only issue to worry about. Let us consider a case where every other realization is skipped. Then there is no way to estimate an underlying autocovariance function (just think about estimation of $\gamma(1) = E\{X_{t+1}X_t\}$). Similarly, if a seasonal component has a period equal (or for small samples even close) to the period of missing realizations, then a problem of estimating such a seasonal component becomes impossible. As an example, if for the case of missing weekends a seasonal component has period equal to 7 (a weekly seasonal component), then there is no way to estimate values of this seasonal component at weekends because there are no such observations. Note that this is not the case if the periods are different. For instance, if a seasonal component has period 10, then during the first 10 days the sixth and seventh observations of the seasonal component are missing, during the second 10 days the third and

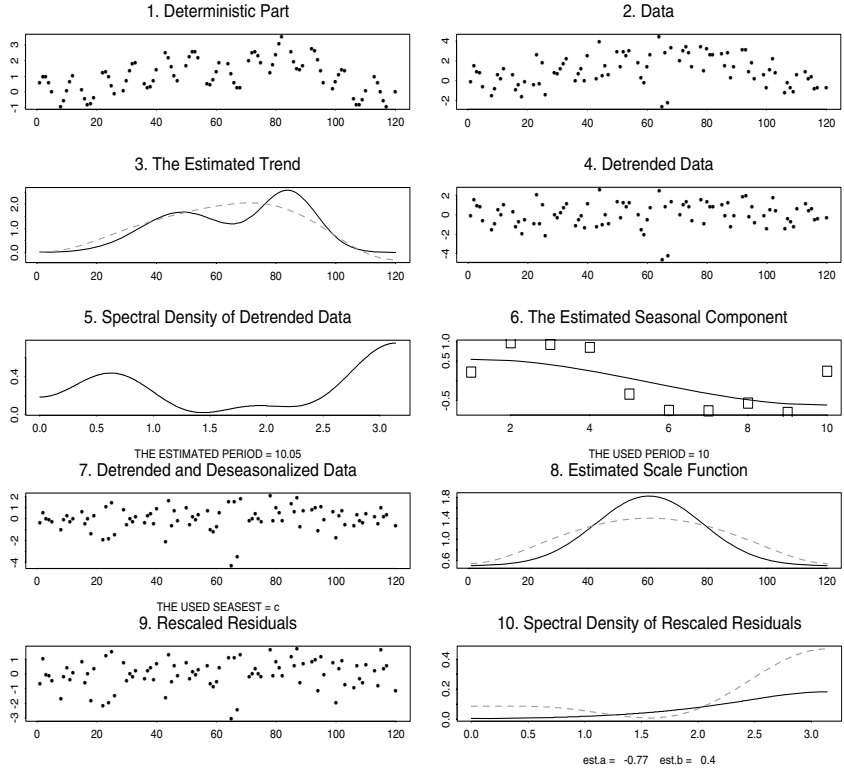

FIGURE 5.6. Nonparametric time series analysis for the case of missing observations. The structure of this figure is the same as that of Figure 5.4. {The sequence of available and missing observations is controlled by the argument *set.obs* with the default value *set.obs=c(1,1,1,1,1,0,0)*, which implies weekly observations with missing weekends.} [*n=120, set.obs=c(1,1,1,1,1,0,0), trendf=3, scalef=2, sigmasc=.5, ss=1, sc=1, a = -.3,b= -.5, TMAX=35, Tseas=10, ManualPer=F, seasest= "c", set.period=c(8,12), set.lambda=c(0,2), lbscale=.1, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2, cJ0sp=4, cJ1sp=.5, cJMsp=6, cJTsp=4, cBsp=2*]

fourth observations of that decade are missing, etc. Also, if observations are missing at random, then only the reduced sample size is the issue. Thus, in general a case of deterministically skipped observations may be essentially worse than a case with stochastically missing data.

Keeping in mind these warnings about possible complications, let us continue the discussion of Figure 5.6. Our next step is to estimate the trend. Since we use the nonparametric estimator of Section 4.2, that is, the estimator for a heteroscedastic regression model, no changes are needed at this step because a heteroscedastic regression allows any spacing of predictors. In other words, the universal estimator is robust to missing observations.

The estimate is shown in Figure 5.6.3, and it is relatively good. Note that here the estimated trend is based on a smaller sample size than the estimates in Figures 5.4.4 and 5.5.4, and furthermore, the observations are regularly missing.

The detrended time series is shown in Figure 5.6.4. Note that here again every sixth and seventh realization is skipped.

The next step is to estimate the spectral density of the detrended data. The estimator of Section 5.2 cannot be used here because the sample autocovariance estimator (5.2.4) requires all n observations. However, it is not difficult to find a reasonable substitution for the sample autocovariance estimator. Recall that $\gamma(j) := E\{X_{t+j}X_t\}$, so an unbiased estimator is

$$\tilde{\gamma}(j) := (1/\hat{m}_j) \sum_{l \in \hat{M}_j} X_{l+j}X_l, \tag{5.4.1}$$

where $\hat{M}_j$ is a random set of $l \in \{1, 2, \ldots, n\}$ such that pairs (X_{l+j}, X_l) are observed (i.e., both X_{l+j} and X_l are not missing) and $\hat{m}_j$ is the number of such pairs. Then the estimator of Section 5.2 may be used straightforwardly with $\tilde{\gamma}(j)$ in place of $\hat{\gamma}(j)$ and the number $\hat{m}_0$ of available observations in place of n.

A particular spectrum estimate is shown in Figure 5.6.5. Here it depicts the location of the mode correctly (it implies the period 10.05), but look at the huge right tail. Here the estimator ignores it because, as we discussed in the previous section, the argument *set.lambda* restricts the search after the period of seasonal component to the frequencies $[0, 2]$.

When the period is found, all the following steps until Figure 5.6.10 are the same as in Figure 5.4. In particular, in Figure 5.6.8 we see that the estimate of the scale function (the dashed line) oversmooths the underlying scale function (the solid line). On the other hand, this estimate correctly exhibits the symmetric shape as well as the correct minimal value of the scale function. In Figure 5.6.10 we use the same modified spectrum estimator as in Figure 5.6.6. The particular spectral density estimate is not perfect, but it indicates that there is no pronounced seasonal component. Note that the left tail, which shows the presence of low-frequency harmonics in the rescaled residuals, is due to imperfect estimation of the trend.

5.5 Case Study: Hidden Components

In the previous sections we discussed the problem of estimating a trend component and a seasonal component in the deterministic part of a time series. There was no problem in separating these two components, since by definition, they have different spectrum domains.

Here we would like to consider a more complicated case where a trend is a linear combination of several low-frequency components. We begin with

a rather simple model of a time series with no seasonal component or a nuisance hidden component (the general case will be considered later)

$$Y_l := f(l) + \sigma(l)\varepsilon_l, \quad l = 1, 2, \dots, n, \tag{5.5.1}$$

where the trend $f(l)$ is a weighted sum of K hidden additive components,

$$f(l) := \sum_{k=1}^{K} w_k \psi'_k(l). \tag{5.5.2}$$

The problem is to estimate either the hidden additive components $\psi'_k(t)$, $k = 1, \dots, K$, when the weights $\{w_k\}$ are given or the weights w_k, $k = 1, \dots, K$, when $\{\psi'_k\}$ are given. Below we consider both these problems.

• **Estimation of Hidden Components.** First of all, it is apparent that in general to estimate the hidden components one needs at least K realizations like (5.5.1) with different weights. Thus, let us assume that K realizations like (5.5.1) are given with K different vectors of weights $W_s := (w_{s1}, \dots, w_{sK})$, $s = 1, \dots, K$. Thus, we observe K different noisy combinations of additive components,

$$Y_{sl} := \sum_{k=1}^{K} w_{sk} \psi'_k(l) + \sigma(l)\varepsilon_{sl}, \quad l = 1, 2, \dots, n, \quad s = 1, 2, \dots, K, \tag{5.5.3}$$

where $\{\varepsilon_{sl},\ l = 1, 2, \dots, n\}$, $s = 1, \dots, K$, are K independent realizations of a second-order stationary time series.

Let us apply our orthogonal series approach to solve this problem. Define $\psi_k(x) := \psi'_k(xn)$ where $\psi_k(x)$ is a function supported on $[0, 1]$, and recall our traditional notation $\{\varphi_j(x)\}$ for elements of the cosine basis on $[0, 1]$. Then the problem of estimating $\psi'_k(x)$ is equivalent to estimating $\psi_k(x)$, which may be solved via estimation of the Fourier coefficients

$$u_{kj} := \int_0^1 \psi_k(x)\varphi_j(x)dx. \tag{5.5.4}$$

Using observations (5.5.3) we can estimate the Fourier coefficients

$$\theta_{sj} := \int_0^1 f_s(x)\varphi_j(x)dx \tag{5.5.5}$$

of the trends

$$f_s(x) := \sum_{k=1}^{K} w_{sk}\psi_k(x) \tag{5.5.6}$$

(rescaled onto the unit interval) by a sample mean estimate

$$\hat{\theta}_{sj} := n^{-1} \sum_{l=1}^{n} Y_{sl}\varphi_j(l/n). \tag{5.5.7}$$

Let us use uppercase letters for denoting K-component column vectors with corresponding lowercase entries, for instance, $\Theta_j := (\theta_{1j}, \theta_{2j}, \ldots, \theta_{Kj})'$, and by W the $K \times K$ matrix with entries w_{sk}. Then, (5.5.4)–(5.5.6) imply the following system of linear equations:

$$\theta_{sj} = \sum_{k=1}^{K} w_{sk} u_{kj}, \quad 1 \le s \le K. \tag{5.5.8}$$

The system of linear equations (5.5.8) may be compactly written as the following matrix equation:

$$\Theta_j = W U_j. \tag{5.5.9}$$

Assume that the matrix W is invertible and denote its inverse by W^{-1}. Then

$$U_j = W^{-1} \Theta_j. \tag{5.5.10}$$

Thus, since the entries of the matrix Θ_j may be estimated by the sample mean estimate (5.5.7), we simply plug the estimates into (5.5.10) and then get estimates $\hat{u}_{kj}$. Recall that

$$\psi_k(x) = \sum_{j=0}^{\infty} u_{kj} \varphi_j(x),$$

and because the Fourier coefficients u_{kj} are estimated, our universal nonparametric estimator may be used straightforwardly.

Figure 5.7 illustrates both the setting and how the estimator performs for the case $K = 3$ and the hidden components being the Normal, the Strata, and the Monotone. First, let us look at the left column of diagrams. The top time series "First Noisy Composition" shows a particular realization of (5.5.1)–(5.5.2) with the weights shown in the subtitle. The error term is a Gaussian ARMA(1, 1) process $\varepsilon_t - 0.4\varepsilon_{t-1} = 0.5(Z_t + 0.3Z_{t-1})$, see examples in Figure 5.1, multiplied by a scale function. The scale function is equal to 1 plus the Normal corner function with the domain $[1, n]$.

Similarly, the second and third time series are shown in the diagrams below. The analysis of these three time series reveals that even the knowledge of the underlying components and the weights does not help to realize them. This is a rather typical situation with linear combinations of functions. So let us see how the estimator, which has at hand only these 3 realizations and the corresponding weights (in other words, the data shown in the left column), solves this puzzle. Estimates of the components (dashed lines) are shown in the right column of Figure 5.7. As we see, the estimates are pretty good and allow us easily to realize the shape of the underlying hidden components (solid lines).

In practical applications the matrix of weights W may not be known exactly. Figure 5.8 allows us to analyze how this can affect the estimates. It is assumed that a given matrix $\tilde{W}$ is equal to an underlying matrix

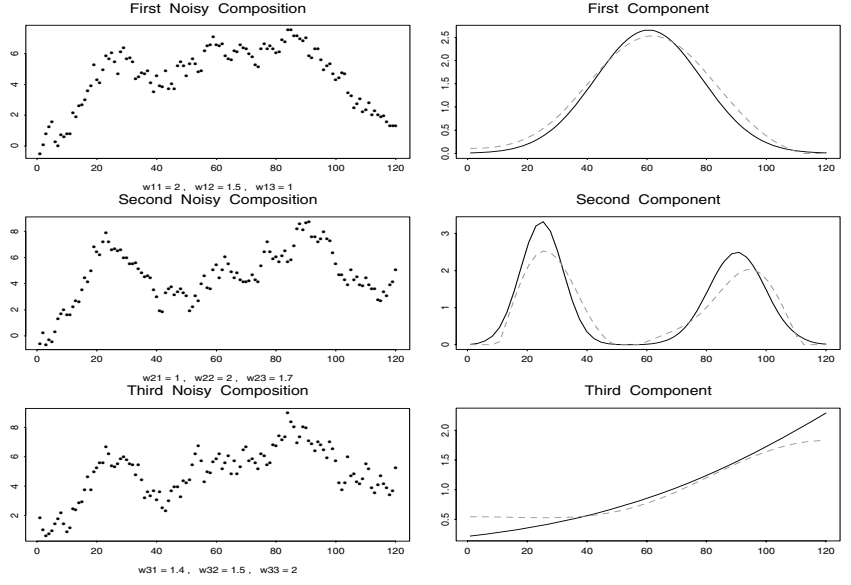

FIGURE 5.7. Recovery of hidden components. The left column of diagrams shows the time series of observations (noisy compositions); the known underlying weights are shown in the corresponding subtitles. The right column of diagrams shows the estimated components (dashed lines) and the underlying components (solid lines). {The default hidden components are the Normal, Strata, and Monotone corner functions with the domain $[1, n]$; their choice is controlled by the argument *set.adc*. It is possible to consider K equal to 2, 3, or 4, and then K components should be chosen using *set.adc*. The underlying weights for the jth composition are controlled by the argument *wj*. The error term is a Gaussian ARMA(1, 1) process $\varepsilon_t - b\varepsilon_{t-1} = \sigma(Z_t + bZ_{t-1})$ multiplied by a scale function. The parameters of the ARMA(1, 1) process are controlled by the arguments a, b, and *sigma*. The scale function is equal to 1 plus a corner function whose choice is controlled by the argument *scalef*. All other arguments control the coefficients of the universal estimator.} [*n=120, set.adc=c(2,4,7), w1=c(2,1.5,1), w2=c(1,2,1.7), w3=c(1.4, 1.5, 2),w4=c(1,1,1,2), scalef=2, a=.4, b= .3, sigma=.5, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2*]

W plus a random matrix with entries being independent standard normal variables multiplied by $\sigma_1/n^{1/2}$. In other words, this mimics the case where the entries are measured with normal $N(0, \sigma_1^2/n)$ additive errors. Figure 5.8 shows three columns of estimates (dashed lines) obtained for different values of σ_1. Each column is obtained similarly to the right column in Figure 5.7, and all the estimates are based on the same data set. Thus, the only difference between the columns is that different noisy matrices $\tilde{W}$ are used. The corresponding σ_1 may be seen in the subtitles. Note that the first column, which corresponds to the case $\sigma_1 = 0$, shows estimates with the correctly known matrix of weights, the two others with noisy matrices.

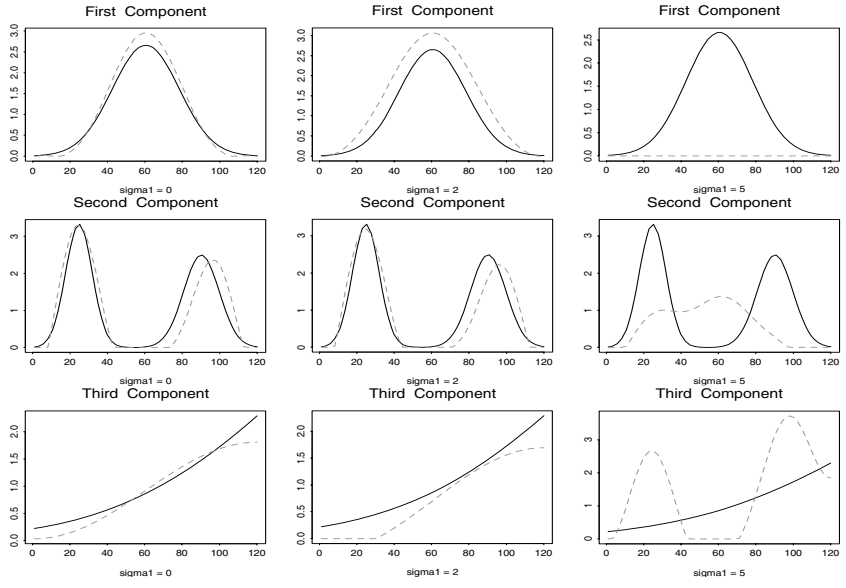

FIGURE 5.8. Recovery of hidden components (solid lines) with a noisy matrix of weights. The estimator is the same as the one used in Figure 5.7, and the estimates are shown by dashed lines. The difference with Figure 5.7 is that here normal $N(0, \sigma_1^2/n)$ errors are added to the underlying weights, and these noisy weights are then used by the estimator. Each column corresponds to a specific noise level shown in the subtitle. Because the first column corresponds to the case $\sigma_1 = 0$ (the weights are known correctly), it allows one to see the effect of noisy weights on the recovery of hidden components. {The set of σ_1 used is controlled by the argument *set.sigma1.*} [*set.sigma1=c(0,2,5), n=120, set.adc=c(2,4,7), w1=c(2,1.5,1), w2=c(1,2,1.7), w3=c(1.4, 1.5, 2),w4=c(1,1,1,2), scalef=2, a=.5, b= .3, sigma=.5, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2*]

As we see, incorrect information about weights may lead to a wrong estimation. Figure 5.8 is a useful tool to get first-hand experience in understanding how random errors in W may affect the recovery of hidden components.

• **Learning Machine for Estimating Weights.** The problem of estimating weights $\{w_k\}$ of a noisy composition (5.5.1)–(5.5.2) arises in many applications where the main issue is not to recover components but to estimate weights. For instance, in applied spectroscopy weights may be considered as concentrations of mixed substances with different spectral profiles.

A typical complication of such a problem is that the components $\psi_k(x)$ are not known as well, so a learning machine should be used (recall the discussion of learning machines in Section 4.10). Here we consider the case where a training data set consists of noisy compositions with known weights (similar to those shown in Figure 5.7), and then a composition with un-

known weights is given (we shall refer to a composition whose weights should be estimated as the composition).

The underlying idea of a learning machine is as follows. As in (5.5.8) we may write for the Fourier coefficients $\{\theta_j\}$ of a trend $f(x)$ of the composition,

$$\theta_j = \sum_{k=1}^{K} w_k u_{kj}, \quad j = 0, 1, \ldots, \tag{5.5.11}$$

where recall that $\{u_{kj},\ j = 0, 1, \ldots\}$ are the Fourier coefficients (5.5.4) of the kth component $\psi_k(x)$.

Note that were the Fourier coefficients θ_j and u_{kj} known, then (5.5.11) implies a classical regression problem with respect to the weights $\{w_k\}$. In our case the Fourier coefficients are unknown, but we may estimate them from given noisy compositions and then plug them into (5.5.11). There are some complications that arise by using such a plugging-in because we get a problem of linear regression with errors in predictors (recall Section 4.11). Here we do not discuss an optimal solution but simply restrict the set of Fourier coefficients in (5.5.11) to $j = 0, 1, 2, \ldots, J_W$ in the hope that the first Fourier coefficients are typically large and thus the errors will be relatively small. (Recall that Figure 5.8 gave us some feeling and experience in dealing with such a situation.) Set $J_W = 5$, and note that it must be at least $K - 1$. Then the corresponding linear regression problem may be solved by standard methods, here the S–PLUS function **lm** is used.

The performance of this learning machine is illustrated in Figure 5.9. The training set of 3 noisy compositions with known weights (but unknown components) is shown in the left column, the data are simulated as in Figure 5.7, and the same notation is used. The top diagram in the right column shows the simulated noisy composition; the corresponding weights are shown in the subtitle and they should be estimated. Thus the learning machine knows the 4 sets of data shown (3 training noisy compositions plus the main one), and it knows the weights for the training compositions shown in the left column.

The diagram "Estimated Components" exhibits the estimated hidden additive components; they are obtained similarly to those shown in the right column of Figure 5.7 and based only on the training sets shown in the left column. As we see, these particular estimates are not perfect but give us a fair impression about the shapes of the underlying Normal, Strata, and Monotone corner functions.

The bottom diagram shows the estimate of the composition. Roughly speaking, the learning machine then tries to fit this estimate by weighted compositions of the estimated components shown above. The important technical detail is that the learning machine does it solely via the first $1 + J_W$ Fourier coefficients of these 4 curves. The estimated weights are shown in the subtitle for the right bottom diagram, and this is the "answer"

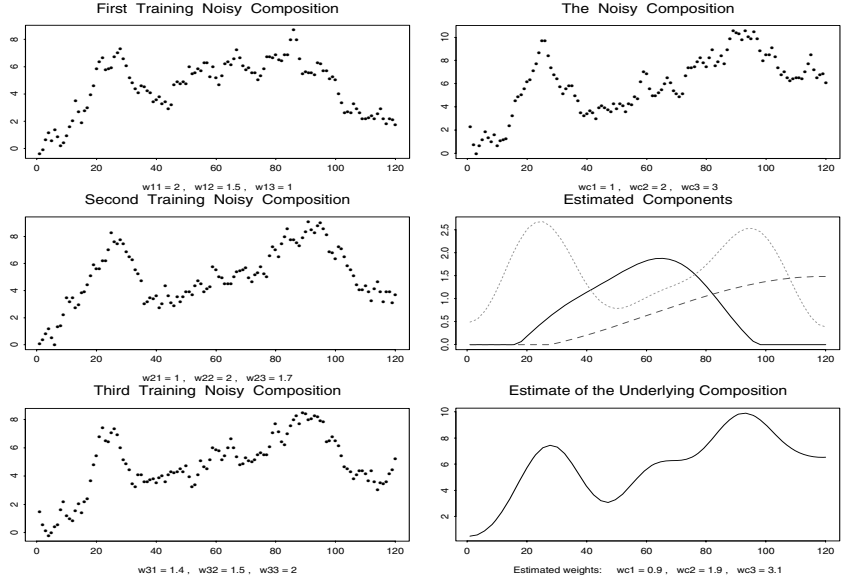

FIGURE 5.9. Learning machine for estimating weights. Training noisy compositions, which are similar to those shown in Figure 5.7, are exhibited in the left column. The noisy composition with unknown weights is shown in the right top diagram. The weights (unknown to the learning machine) are shown in the subtitle of this diagram, and the estimated weights are shown in the subtitle of the right bottom diagram. The two other diagrams in the right column show the "internal" product of the learning machine: estimated components and the estimated underlying regression function for the top scattergram. {The parameter J_W is controlled by the argument JW with the default value $JW = 5$; note that JW must be at least $K - 1$. The choice of underlying estimated weights is controlled by the argument *wc*. The argument *set.adc* controls the triplet of underlying components.} [*n=120, JW=5, set.adc=c(2,4,7), w1=c(2,1.5,1), w2=c(1,2,1.7), w3=c(1.4, 1.5, 2),w4=c(1,1,1,2), scalef=2, a=.5, b= .3, sigma=.5, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2*]

given by the learning machine. For this particular set of data the estimates are 0.9, 1.9, and 3.1, and this is a good outcome for the case of 3 hidden components, the sample size 120, and dependent errors.

• **Extra Nuisance Component.** Consider a more general setting where in model (5.5.1) an extra nuisance additive component $G'(l)$ is presented, namely, when the time series is

$$Y_l = f(l) + G'(l) + \sigma(l)\varepsilon_l, \quad l = 1, 2, \ldots, n, \tag{5.5.12}$$

and (5.5.2) holds.

Our estimators can easily handle this case under the following assumption: $G'(l)$ is the same in all the experiments. In other words, in all the experiments the deterministic components are $f_s + G'$. Under this assump-

tion, the nuisance additive component G' becomes an extra $(K+1)$th additive component with the constant weight equal to 1 for all the experiments. In other words, if we set $\psi'_{K+1}(l) := G'(l)$, $w_{K+1} := 1$, and consider the case of $K+1$ hidden components, then the problem is reduced to the previously discussed ones with just an additional $(K+1)$th experiment. Then, for instance, the problem of estimating the additive components is solved based on $K+1$ experiments with different weights for the first K components. {To simulate the situation, use Figure 5.7 and set the last elements in the vectors w_s to 1; then the Kth component may be considered as a nuisance one. Similar changes are needed in Figures 5.8–5.9.}

Another useful comment is as follows. Suppose that this nuisance component is a seasonal component and its frequency is beyond J_W, that is, $J_{\max} + 1 \geq J_W$. Then this nuisance seasonal component has no effect on our nonparametric estimators because they perform in the low-frequency domain.

5.6 Case Study: Bivariate Time Series

In many practical situations it is necessary to analyze a pair of time series. For instance, the relationship between the price and supply of a commodity is of a central interest for econometrics, and the relationship between the number of police officers on the streets and the level of crime is of a central interest for a government.

We begin our discussion with one simple but very informative example of a bivariate time series $\{(X_t, Y_t)\}$ defined by

$$X_t = \sigma_1 Z_t^{\mathrm{X}}, \qquad t = 1, 2, \ldots, n, \tag{5.6.1}$$

$$Y_t = bX_{t-k} + \sigma_2 Z_t^{\mathrm{Y}}, \quad t = 1, 2, \ldots, n. \tag{5.6.2}$$

Here Z_t^{X} and Z_t^{Y} are independent standard Gaussian white noises (that is, these time series are iid standard normal), the coefficients σ_1 and σ_2 are nonnegative, b is real, and the parameter k, which is called a *delay*, is an integer and may be either positive or negative.

The important feature of this bivariate time series is that for a positive k the time series $\{X_t\}$ leads the time series $\{Y_t\}$, while for negative values of k the situation is reversed. Also, it is apparent that these two univariate time series have a linear relationship, since the second time series is simply a lagged multiple of the first time series with added noise.

Is it possible to realize such a structure of the bivariate time series via just visualizing its realization? Let us check this. The two top diagrams in Figure 5.10 show particular realizations of the first univariate time series $\{X_t\}$ and the second univariate time series $\{Y_t\}$. As we see, it is not an easy task to realize from visualizing these time series that they are related

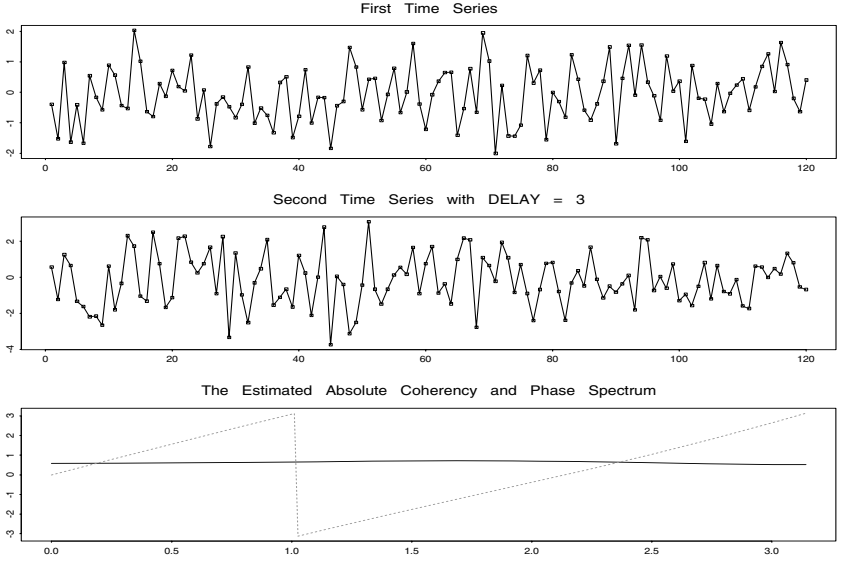

FIGURE 5.10. Realization of a bivariate time series (5.6.1)–(5.6.2) with $b = 1$, $\sigma_1 = \sigma_2 = 1$, $k = 3$, and $n = 120$. The bottom diagram shows the estimates of the absolute coherency (solid line) and phase spectrum (dotted line). {All notations for the arguments are apparent, for instance, *DELAY* controls the delay k in (5.6.2).} [*b=1, sigma1=1, sigma2=1, DELAY=3, n=120, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

in any sense. In short, due to the relatively large additive noise in (5.6.2), visualization is not fruitful.

Thus, let us try to solve this problem using a statistical approach. Recall (see Section 5.2) that for a zero-mean and second-order univariate stationary time series $\{X_t\}$ its main characteristics are the autocovariance function

$$\gamma_{XX}(h) := E\{X_{t+h}X_t\} \tag{5.6.3}$$

and, under a mild assumption like $\sum_{h=-\infty}^{\infty} |\gamma_{XX}(h)| < \infty$, the corresponding spectral density

$$f_{XX}(\lambda) := (2\pi)^{-1} \sum_{h=-\infty}^{\infty} \gamma_{XX}(h) e^{-ih\lambda}, \qquad \lambda \in (-\pi, \pi]\,. \tag{5.6.4}$$

Because the autocovariance function is symmetric, i.e., $\gamma(h) = \gamma(-h)$, the spectral density is a real and even function.

To analyze the relationship between two zero-mean and second-order stationary sequences $\{X_t\}$ and $\{Y_t\}$, we shall use very similar notions of the *cross-covariance*

$$\gamma_{XY}(h) := E\{X_{t+h}Y_t\} \tag{5.6.5}$$

and, under a mild assumption like $\sum_{h=-\infty}^{\infty} |\gamma_{XY}(h)| < \infty$, the *cross spectral density* or simply *cross spectrum*

$$f_{XY}(\lambda) := (2\pi)^{-1} \sum_{h=-\infty}^{\infty} \gamma_{XY}(h) e^{-ih\lambda}, \quad \lambda \in (-\pi, \pi]. \tag{5.6.6}$$

The similarity between the auto-characteristics and cross-characteristics is striking, but there is one very important difference that is necessary to know. While any autocovariance function is always symmetric and thus any spectral density is real, a cross-covariance may be asymmetric, that is, $\gamma_{XY}(h)$ may differ from $\gamma_{XY}(-h)$, and this implies a complex cross spectrum. To see this, let us calculate the cross-covariance function for the example (5.6.1)–(5.6.2). Using the assumption that $Z_t^{\rm X}$ and $Z_t^{\rm Y}$ are independent standard white noises, a simple calculation shows that

$$\gamma_{XY}(h) = b\sigma_1^2 I_{\{h=-k\}}. \tag{5.6.7}$$

Recall that $I_{\{A\}}$ is the indicator function of an event A. Thus, the cross–covariance between the series (5.6.1) and (5.6.2) is not zero only for $h = -k$, thus it is not symmetric in h. As a result, the corresponding cross spectral density (5.6.6) becomes complex and is defined by the formula

$$f_{XY}(\lambda) = (2\pi)^{-1} b\sigma_1^2 (\cos(k\lambda) + i \sin(k\lambda)). \tag{5.6.8}$$

Thus only in the case of the zero delay $k = 0$ is the cross spectrum real.

Since it is not very convenient to analyze a complex function directly, we shall use the following approach. First, let us recall that any complex number may be expressed in polar coordinates. Correspondingly, a complex spectrum $f_{XY}(\lambda) := f_{\rm r}(\lambda) + i f_{\rm im}(\lambda)$ may be written as

$$f_{XY}(\lambda) = \alpha_{XY}(\lambda) e^{i\phi(\lambda)},$$

where $\alpha_{XY}(\lambda) := [f_{\rm r}^2(\lambda) + f_{\rm im}^2(\lambda)]^{1/2}$ is called the *amplitude spectrum* and $\phi(\alpha) := \arg(f_{\rm r}(\lambda) + i f_{\rm im}(\lambda)) \in (-\pi, \pi]$ is called the *phase spectrum*. Note that by definition the phase spectrum lies between $-\pi$ and π.

Second, recall the notion of a *correlation coefficient* between two zero-mean random variables U and V, $\rho_{UV} := E\{UV\}/[E\{U^2\}E\{V^2\}]^{1/2}$. The correlation coefficient varies between -1 and 1 and measures the extent to which these random variables are linearly related, namely, the larger the absolute value of ρ_{UV} the stronger the linear relationship between U and V. For the case of the spectrum, we can introduce the similar notion of the *absolute coherency*

$$\mathcal{K}_{XY}(\lambda) := \frac{\alpha_{XY}(\lambda)}{[f_{XX}(\lambda) f_{YY}(\lambda)]^{1/2}}. \tag{5.6.9}$$

The absolute coherency lies between 0 and 1, and like the coefficient of correlation it measures the extent to which these two series are linearly related at frequency λ.

As an example, let us calculate the above-defined characteristics for the bivariate time series (5.6.1)–(5.6.2). Simple calculations show that the amplitude spectrum, the phase spectrum, and the absolute coherency are defined by the following formulae (recall that the notion of module was introduced in Section 3.5):

$$\alpha_{XY}(\lambda) = (2\pi)^{-1}|b|\sigma_1^2, \tag{5.6.10}$$

$$\phi_{XY}(\lambda) = (k\lambda + \pi)[\mathrm{mod}\ 2\pi] - \pi, \tag{5.6.11}$$

$$\mathcal{K}(\lambda) = |b|\sigma_1/(b^2\sigma_1^2 + \sigma_2^2)^{1/2}. \tag{5.6.12}$$

These results are the key to understanding statistical methods for the analysis of a bivariate time series. First, we see that the delay k is exactly the slope of the phase spectrum $\phi_{XY}(\lambda)$ because this phase spectrum is piecewise linear with constant slope k (like the dotted line in the bottom diagram in Figure 5.10). Of course, this will not be the case for an arbitrary bivariate time series. However, the derivative (slope) $d\phi_{XY}(\lambda)/d\lambda$ of the phase spectrum can still be regarded as a measure of the phase lag of Y_t behind X_t at frequency λ. This explains why the derivative (slope) of a phase spectrum is called the *group delay*. (The derivative may be negative, and this indicates that Y_t leads X_t.) Thus, visualizing the phase spectrum allows one to reveal which time series is the leader and which one is the follower. Second, the absolute coherency (5.6.12) becomes closer to 1 if either $b\sigma_1$ increases or σ_2 decreases. These conclusions are well understood because in both these cases the effect of the additive noise $\sigma_2 Z_t^{\mathrm{Y}}$ on Y_t in (5.6.2) becomes smaller. Thus, the absolute coherency is indeed a notion that is similar to the correlation coefficient, and it shows how strong a linear relationship between $\{X_t\}$ and $\{Y_t\}$ is at frequency λ.

These are the reasons why both the absolute coherency and the phase spectrum are the two primary characteristics used in the spectrum analysis of bivariate time series.

Now let us explain how to estimate these characteristics. Since the only new function here is the cross spectral density (5.6.6), we note that a partial sum for (5.6.6) should be written as

$$f_{XY}(\lambda, J_1, J_2) := (2\pi)^{-1} \sum_{j=-J_1}^{J_2} \gamma_{XY}(j) e^{-ij\lambda}. \tag{5.6.13}$$

In contrast to estimating a spectral density where $J_1 = J_2$ (see (5.2.6)), here J_1 and J_2 may be different because the cross-covariance in general is not symmetric. Apart from this, the estimator of Section 5.2 may be used straightforwardly with the only modification that the estimated cutoffs are

defined by the formula

$$(\hat{J}_1, \hat{J}_2) := \text{argmin}_{j_1,j_2} \left(\sum_{j=-j_1}^{j_2} (2\hat{d}n^{-1} - \hat{\gamma}^2_{XY}(j)), \quad 0 \le j_1, j_2 \le J_n \right), \tag{5.6.14}$$

where

$$\hat{\gamma}_{XY}(j) := n^{-1} \sum_{l=1}^{n-j} X_{l+j} Y_l \tag{5.6.15}$$

is the sample cross-covariance, which is used in place of the sample covariance, and

$$\hat{d} := \sum_{j=-J_n}^{J_n} \hat{\gamma}_{XX}(j)\, \hat{\gamma}_{YY}(j) \tag{5.6.16}$$

is the estimated coefficient of difficulty.

Estimates of the absolute coherency and the phase spectrum, calculated for the bivariate time series shown in Figure 5.10, are exhibited in the bottom diagram of that figure. These estimates are almost perfect. The slope of the estimated phase spectrum is approximately 3 at all frequencies. The estimated absolute coherency also correctly shows that the linear relationship between these two time series is practically the same at all frequencies. Also note that the coherency is far from 1, and this is absolutely right because the variance of the independent additive noise in Y_t is equal to the variance of X_t. To get the absolute coherency close to 1, the coefficients of the model should be changed, as has been discussed above. Also, it is very useful to look at the estimates when the delay is negative. Thus, it is highly recommended to do Exercise 5.6.4.

Now let us apply our methodology and nonparametric universal estimators to an *econometrics model* that defines a bivariate time series with the first component $\{P_t\}$ being the mean corrected price of a commodity and the second component $\{S_t\}$ being the supply of this commodity at time t. The model is defined as

$$P_t = -b_P S_t + \sigma_P Z^{\rm P}_t\,, \quad S_t = b_S P_{t-1} + \sigma_S Z^{\rm S}_t\,, \quad t = 1, \dots, n\,, \tag{5.6.17}$$

where $0 < b_P, b_S < 1$, time series $\{Z^{\rm P}_t\}$ and $\{Z^{\rm S}_t\}$ are independent standard Gaussian white noises, and the initial value of the price is $P_0 = 1$.

Figure 5.11 shows a particular realization of this econometrics model; see the top two diagrams. Here $b_P = 0.4$, $b_S = 0.8$, $\sigma_P = 1$, and $\sigma_S = 0.5$.

It is not an easy task to analyze these realizations manually, so let us see what our nonparametric estimates, shown in the bottom two diagrams, tell us about this bivariate time series. The estimated absolute coherency reveals that a linear relationship between price and supply is strongest at high frequencies. Thus, our next step is to understand who leads whom at high frequencies. This we do with the help of the estimated phase spectrum,

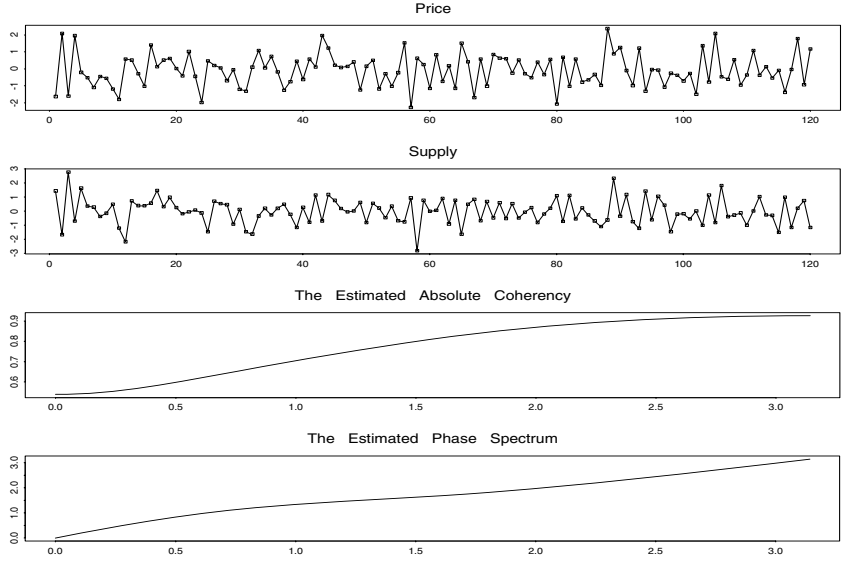

FIGURE 5.11. A case study of econometrics model (5.6.17) for price and supply. The top two diagrams show particular realizations of price and supply, and the bottom two diagrams show the estimated absolute coherency and phase spectrum. [*bP=.4, bS=.8, sigmaP=1, sigmaS=.5, price0=1, n=120, cJ0=4, cJ1=.5, cJM=6, cT=4, cB=2*]

which clearly indicates that price leads supply because the slope of the estimate is positive. Moreover, the slope at high frequencies is about 1, so we see that even the delay may be correctly defined via analyzing the estimated phase spectrum. These conclusions give us some insight into the relationship between price and supply based on this bivariate time series, and they do correspond to the underlying model.

5.7 Case Study: Dynamic Model and Forecasting

Consider the nonlinear dynamic model

$$Y_t := f(Y_{t-1}) + s(Y_{t-1})\varepsilon_t, \quad Y_0 := \xi, \quad t = 1, 2, \ldots, n\,. \tag{5.7.1}$$

Here Y_t is called a *state* of the model, f is called an *iterative map*, and s is called a *scale map*. The noise ε_t is a stationary time series, for instance an ARMA process, and ξ is an initial state of the model. Note that if $s(y) = 0$, then $Y_t = f(Y_{t-1})$, i.e., a current state of this dynamic model is defined solely by its previous state (the states are iterated). This explains the name of f.

At first glance, the dynamic model (5.7.1) may resemble a classical time series model $X_t := f(t) + \sigma(t)\varepsilon_t$, where $f(t)$ is the deterministic part (trend/seasonal component) and $\sigma(t)$ is a scale function. However, these models are absolutely different: The deterministic component in a classical time series decomposition is a function of time, while in a dynamic model the deterministic component is a function of the previous realization.

Dynamic systems naturally arise in applications where one believes that a current state of a model is defined primarily by its previous state and a current "noise." They are also used to approximate stochastic differential equations, for instance the equation for a continuous-in-time *diffusion process* y_t,

$$dy_t = \psi(y_t)dt + \sigma(y_t)dB(t), \quad t \ge 0, \quad y_0 = \xi. \tag{5.7.2}$$

Here $B(t)$ is a Brownian process (the definition will be given in Section 7.2), ψ is called a *drift* function, and σ is called a *volatility* function. A famous example is the *Black–Scholes* model for the stock price S_t,

$$dS_t = (\mu + \nu^2/2)S_t dt + \nu S_t dB(t), \quad t \ge 0, \quad S_0 = \xi. \tag{5.7.3}$$

The parameters μ and ν are the so-called stock drift and volatility.

To explain a relationship between (5.7.1) and (5.7.2), consider equidistant observations of y_t with the sampling interval $\delta := 1/n$. For large n these observations may be approximately written as the *Euler scheme*

$$y_t = n^{-1}\psi(y_{t-\delta}) + y_{t-\delta} + n^{-1/2}\sigma(y_{t-\delta})Z_t, \quad t = \delta, 2\delta, \dots, \ y_0 = \xi, \tag{5.7.4}$$

where Z_t are iid normal random variables. Thus, if we set $f(y) = n^{-1}\psi(y) + y$, $s(t) = n^{-1/2}\sigma(y)$, and consider standard normal ε_t, then the relationship becomes transparent.

Let us explain how the universal estimate may be used for finding iterative and scale maps f and s. Define $X_t := Y_{t-1}$ and rewrite (5.7.1) as

$$Y_t := f(X_t) + s(X_t)\varepsilon_t, \qquad X_1 = \xi, \quad t = 1, 2, \dots, n. \tag{5.7.5}$$

This equation, at least formally, resembles the classical heteroscedastic regression problem discussed in Sections 4.2–4.3, where f was called the regression function and s the scale function. Thus, we may try to use the universal estimates of those sections for estimation of both f and s.

Figure 5.12 illustrates how a dynamic model iterates and how the universal estimator performs. The top diagram shows a particular realization of states simulated by a dynamic model (5.7.1). Here the iterative map is $f(y) = 2y/(1 + 2y^2)$, and the scale map is 2 times a standard normal density. The noise term is a Gaussian ARMA(1, 1) process $\varepsilon_t + 0.3\varepsilon_{t-1} = Z_t + 0.4Z_{t-1}$. The initial state Y_0 is a realization of a uniform random variable ξ on $(0, 1)$.

The analysis of this particular realization, based on methods discussed in Section 5.1, shows that there is no visible trend, and a seasonal component

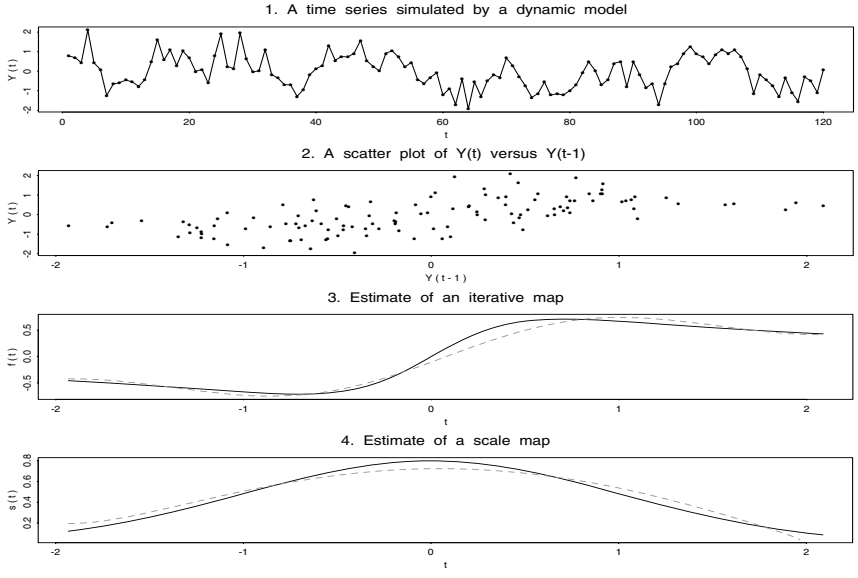

FIGURE 5.12. Analysis of a dynamic model. The underlying maps are shown by the solid lines and their estimates by dashed lines. The underlying iterative map is $f(Y) = AY/(1+BY^2)$, and the underlying scale map is σ times a normal density with zero mean and standard deviation sd. The default values are $A = 2$, $B = 2$, $\sigma = 2$, and $sd = 1$. The initial value is a realization of a uniformly distributed random variable on $[0, 1]$. The noise term ε_t is a Gaussian ARMA(1, 1) process $\varepsilon_t - a\varepsilon_{t-1} = Z_t + bZ_{t-1}$, where Z_t are iid standard normal, $a = -0.3$, and $b = 0.4$. {The length of a time series is controlled by the argument n; all other arguments are explicit.} [*n=120, A=2, B=2, sigma=2, sd=1, a=−.3, b=.4, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2*]

with period about 20 is a remote possibility. In short, it is difficult to see something special in this realization.

The second diagram depicts a scatter plot of current states Y_t versus previous states Y_{t-1} (or Y_t versus X_t). This diagram sheds light on the iterative process, and look how inhomogeneous the scatter plot is. We see that relatively large values of the scale map $s(y)$ for y around zero make the regression problem complicated. Also, only several observations are located beyond the interval $(-1.3, 1.3)$.

The third diagram shows us that the universal estimator does a fair job in recovering f. Note that the shape of the underlying iterative map is clearly recognizable, and even the slopes of the tails are correctly indicated. On the other hand, note that there is no way to estimate this map over a larger interval because no Y_t with large absolute values are given.

The bottom diagram shows the restored scale map. Again, the shape is clearly recognizable, and the quality of estimation is reasonable.

Thus, we see that the universal estimator may be used for analyzing a dynamic model. On the other hand, it is necessary to know that this problem may be extremely complicated for different f, s, and noise.

Interestingly, the role of noise in the estimation of maps f and s is absolutely crucial and resembles the situation discussed in Section 4.9, where a large noise was necessary for a consistent estimation. Indeed, let us assume that $s(y) = 0$ (no noise term) and $f(y) > 0$ for $y > 0$. Then, a positive Y_0 implies positive Y_t, $t = 1, 2, \ldots$, and thus there is no chance to restore $f(y)$ for negative y.

Let us finish this section by introducing one more interpretation of the dynamic model. Assume that $Y_1, Y_2, \ldots$ is a time series, and one would like to make a *forecasting* of an observation Y_{t+1} based on a previous Y_t. This problem is also called a one-step prediction. Then a dynamic model may be used again for the definition of a *nonlinear one-step prediction* $f(Y_t)$.

5.8 Case Study: Change-Point Problem

The change-point problem has a long history in statistics due to important applied settings where abrupt and localized changes are the main concern. They appear to have arisen originally in the context of quality control, where one observes the output of a production process sequentially and wants to signal any departure of the average output from some known target process. Other familiar examples include the analysis of the incidence of a disease in epidemiology, climate change problems, seismology, and performance of stock markets.

In a studied time series context, a change-point may be defined as a point with a discontinuity in at least one component of a time series or its derivative. Most typical situations are as follows: discontinuity (jump) in a trend or its derivative; discontinuity in the period of a seasonal component; change in the noise distribution, for instance a change in parameters of an ARMA process or distribution of an underlying white noise.

We have discussed in Section 4.7 the problem of changing over a time distribution of a noise. The aim of this section is to consider another classical example of estimation of a trend with a jump discontinuity.

The considered approach is identical to one discussed in Section 5.1, only here the cosine basis enriched by a step function is used. This basis was introduced and discussed in the last subsection of Section 2.6. Recall that the corresponding partial sum of $f(x)$, $0 \le x \le 1$, was defined in (2.6.6) as

$$S_J(x, a) := \sum_{j=0}^{J} \theta_j \varphi_j(x) + \kappa(a, J)\, \phi(x, a, J), \quad 0 \le x \le 1,$$

where $\theta_j = \int_0^1 f(u)\varphi_j(u)du$, φ_j are elements of the cosine basis (2.1.3), $\kappa(a,J) = \int_0^1 f(u)\phi(u,a,J)du$, and $\phi(x,a,J)$ is the orthogonal element obtained by applying Gram–Schmidt orthogonalization to a step function $\phi(x,a) := I_{\{x\le a\}}$, $0 \le x \le 1$.

The problem of estimation of these Fourier coefficients is identical to one discussed in Section 4.1, and the recommended estimator may be used directly. On the other hand, it is worthwhile to discuss several related issues.

The first issue is a possibility to use a relatively simple procedure discussed in Section 4.1 of finding a pilot estimator of the variance σ^2 of the noise (that is, d). Let us assume that $Y_t := f(t) + \sigma\varepsilon_t$, where ε_t is a second-order stationary process with $E\{\varepsilon_t\} = 0$ and $E\{\varepsilon_t^2\} = 1$. Then, $Y_{t+1} - Y_t = (f(t+1) - f(t)) + \sigma(\varepsilon_{t+1} - \varepsilon_t)$, which implies

$$E\{(Y_{t+1} - Y_t)^2\} = (f(t+1) - f(t))^2 + \sigma^2 E\{(\varepsilon_{t+1} - \varepsilon_t)^2\}. \qquad (5.8.1)$$

Recall the notion of the quadratic variation of a function; see (2.2.12). Here f is defined at integer points over an interval $[1,n]$. If the corresponding quadratic variation increases more slowly than n, then (5.8.1) implies that $\hat\nu^2 := (n-1)^{-1}\sum_{t=1}^{n-1}(Y_{t+1} - Y_t)^2$ may be a good estimator of $\sigma^2 E\{(\varepsilon_{t+1} - \varepsilon_t)^2\}$. Note that if f has a jump discontinuity of a size S, then this jump affects $\hat\nu^2$ by the term $S^2/(n-1)$. Also, if the noise is white, then $E\{(\varepsilon_{t+1} - \varepsilon_t)^2\} = 2$ and $\hat\nu^2/2$ becomes a consistent estimate of σ^2. Below we shall test the estimate $\hat\nu^2$.

The second issue is how to calculate an optimal cutoff J and a possible location a of a jump. This is done similarly to (4.1.11), namely

$$(\hat J, \hat a) := \operatorname{argmin}_{0\le J\le J_n,\, a\in A_n}\Big[2(J+2)\tilde d - \sum_{j=0}^{J}\hat\theta_j^2 - \hat\kappa^2(a,J)\Big]. \qquad (5.8.2)$$

Here J_n is again the rounded-down $c_{J0} + c_{J1}\ln(n)$, and A_n is a net of integers where a jump may occur; in general this is $\{1,2,\dots,n\}$, but any reasonable net may be used to speed up the calculation. The statistic $\tilde d$ is an estimate of the coefficient of difficulty; in the numerical example $\tilde d = \hat\nu^2$.

Then the estimate is defined as

$$\hat f(t) := \sum_{j=0}^{\hat J}\tilde w_j\hat\theta_j\varphi_j\Big(\frac{t}{n}\Big) + \hat\kappa(\hat a,\hat J)I_{\{(\hat\kappa(\hat a,\hat J))^2 > c_T\tilde d\ln(n)n^{-1}\}}\phi\Big(\frac{t}{n},\hat a,\hat J\Big). \qquad (5.8.3)$$

The shrinkage weights $\tilde w_j$ are defined in (4.1.12). Note that the difference with (4.1.13) is that high-frequency terms are not included because we assume that an estimated trend is smooth except for a possible jump discontinuity.

Figure 5.13 shows how this estimator performs. Two time series are $Y_{s,t} := f_s(t) + \sigma\varepsilon_t$, $t = 1,2,\dots,n$, $s = 1,2$. The same Gaussian ARMA(1,1) noise is added to a smooth trend f_1 and a discontinuous trend f_2. Here

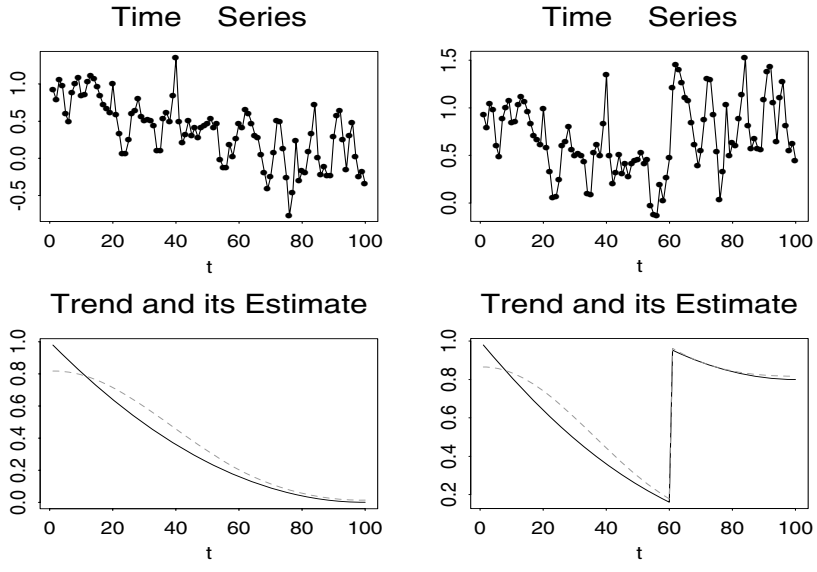

FIGURE 5.13. Estimation of a trend with a possible jump discontinuity. The two columns of diagrams correspond to the cases of continuous and discontinuous trends. The underlying trends and their estimates are shown by solid and dashed lines, respectively. The ARMA$(1,1)$ noise is the same in both time series. {The parameters a, b and the standard deviation of an ARMA$(1,1)$ process are controlled by the arguments a, b, and *sigma*. The size of the jump is controlled by the argument *jump*. The length of a time series is controlled by n. To make the calculations faster, the default is $cJ0 = 1$ and the search after the location of a jump is performed over 17 equidistant points, which are the rounded $0.1n, 0.15n, \ldots, 0.9n$.} [*n=100, jump=.8, sigma=.3, a=.4, b=.4, cJ0=1, cJ1=.5, cT=4*]

$\varepsilon_t = \sigma X'_t$, where $X'_t = X_t/(E\{X_t^2\})^{1/2}$, $X_{t+1} - aX_t = Z_{t+1} + bZ_t$, and Z_t is a standard Gaussian white noise, $\sigma = 0.3$, and $a = b = 0.4$.

The top diagrams depict two particular time series. The left one indicates a decreasing trend, while the second one shows no pronounced trend. The bottom diagrams show us the underlying trends f_1 and f_2 (solid lines) and the corresponding universal estimates (dashed lines). We see that our guess about the trend in the left time series was correct. The right bottom diagram reveals that the underlying trend f_2 is identical to f_1, only at the moment $t = 60$ it has a jump discontinuity. Note that the universal estimator (5.8.3) correctly rejected a jump discontinuity for the left time series and correctly indicated both the location and the size of the change-point for the right time series.

Now, when we know the underlying trends, it is easy to see the change-point in the top right diagram. On the other hand, note that the change in the time series over the time interval $[57, 63]$ is about 1.6, that is, it is

twice as large as the size 0.8 of the underlying jump. Also, can you clearly see the location of a jump? This is what makes the change-point problem so difficult. To solve it manually, one needs to estimate an underlying trend before and after a possible change-point, and this is a complicated problem for the relatively large noise considered in this example.

Finally, let us note that wavelets are another type series estimator that are absolutely natural for solving change-point problems. No changes in the universal wavelet estimator are required, and Figures 4.8-4.10 illustrate the performance.

5.9 Practical Seminar

The aim of this seminar is to use the comprehensive nonparametric analysis developed in Section 5.4 for a real time series.

Let us consider the S–PLUS data file **hstart**, which contains the time series of US monthly housing starts from January 1966 to December 1974. Here we use an analogue of Figure 5.4, where the first two diagrams are identical and show the underlying time series using different formats.

Figure 5.14 shows an analysis of monthly US housing starts. The first two diagrams are the data. Note that the connected points and the points alone give a rather different visualization of the same data set.

To separate a trend from a seasonal component, we use the maximal period $T_{\max} = 35$, i.e., a period of about 3 years. We discussed in Section 5.3 why this was a reasonable choice for the training time series shown in Figure 5.4. Note that the time series of the monthly US housing starts remarkably resembles the training time series.

Diagram 3 shows the estimated trend (the low-frequency change) in the housing starts. We see both the famous boom and the tragic collapse of the housing market. The detrended data show that the residuals are significant, and it looks as if a pronounced seasonal component is present. The spectral density of detrended data supports our conclusion. But look how flat this estimate is near its mode. As we know, this may lead to some troubles in searching for the period of the seasonal component.

The estimated period (according to (5.2.3)) of an underlying seasonal component is 10.93 (it is shown in the subtitle for diagram 5), and this is clearly not the expected 12-month period, which should be the period for such a time series. Nevertheless, let us continue our default analysis and not invoke the option *ManualPer=T*, which allows us to use any period (we shall do this later). At first glance, the estimated seasonal component, depicted in diagram 6, looks nice. (The smoothed nonparametric estimate (the solid line) is not applicable here because the period is too small.) On the other hand, please pay attention to the fact that the seasonal component is smallest during the summer season and largest during the winter season.

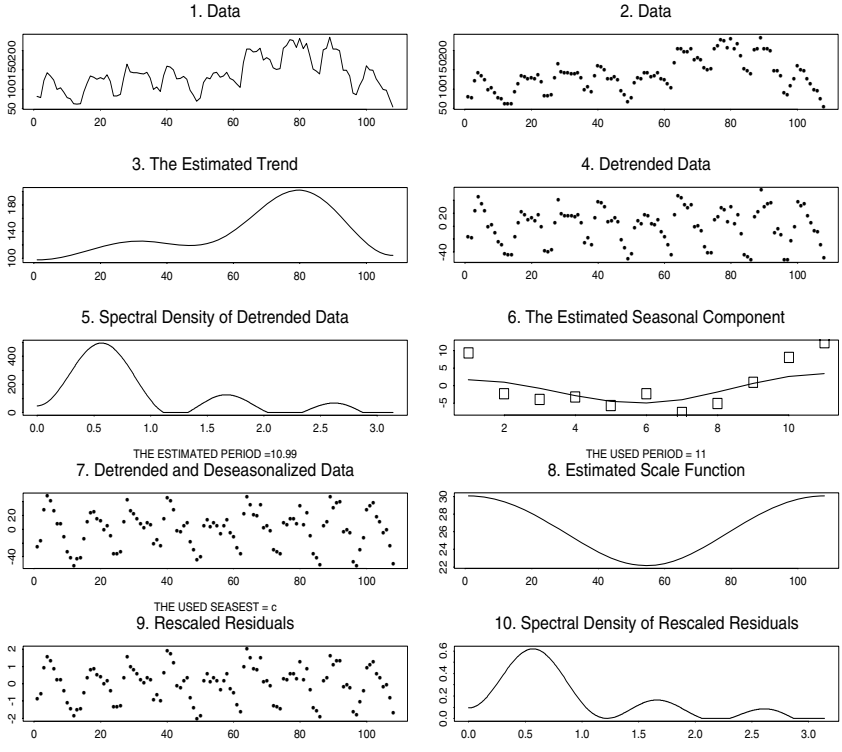

FIGURE 5.14. A comprehensive nonparametric analysis of monthly US housing starts from January 1966 to December 1974. {The choice of a time series is controlled by the argument *DATA*. All other arguments are the same as in Figure 5.4 and are explained in Section 5.3.} [*DATA=hstart, TMAX=35, Tseas=10, ManualPer=F, seasest="c", set.period=c(8,12), set.lambda=c(0,2), lbscale=.1, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=6, cT=4, r=2, cB=2, cJ0sp=4, cJ1sp=.5, cJMsp=6, cJTsp=4, cBsp=2*]

This does not look right and indicates a phase shift due to a wrongly estimated period. Moreover, the spectral density estimate in the last diagram, which is the ultimate judge, apparently indicates that no deseasonalizing has occurred.

We know from Section 5.3 that in this case other periods, which are close to 11, should be tested. Here the nature of the data suggests trying a 12-month period (a year). To do this, we repeat this figure with the argument *ManualPer=T* and, when the figure stops after diagram 5, enter from the keyboard the period 12. The result is shown in Figure 5.15.

Our first glance is at the final diagram. Here all is okay, and the residuals may be modeled by an ARMA model. The seasonal component, shown by squares in diagram 6, looks very reasonable. We see a slow start in January

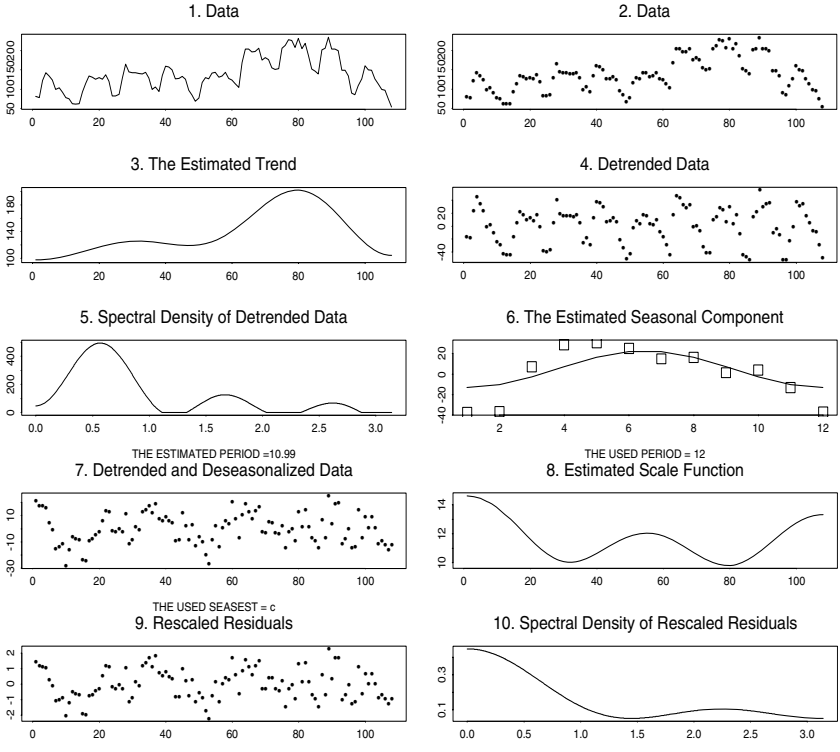

FIGURE 5.15. A comprehensive nonparametric analysis of monthly US housing starts from January 1966 to December 1974. This figure is similar to Figure 5.14, only here the correct period $T = 12$ is used. {To get this figure, call **>ch5(f=14, ManualPer=T)**. The program will stop at the 5th diagram, so you will see the estimated period 10.93. Then the program will ask you to enter a period. Enter 12 from the keyboard. Then the rest of the diagrams will be displayed.}

and February, which is followed by a spurt in March and April. Then only in December do we see a significant slowing down.

The estimated volatility of the detrended and deseasonalized data, shown in diagram 8, looks a bit strange. There is no doubt that the right tail is shown correctly because the volatility increases during the periods of booms and collapses. But the left tail looks strange. Thus, let us examine diagram 7, where the underlying data (residuals) are shown. We see that the residuals decrease almost linearly during the first year and the range is about 40, the largest over all the years. This is what has been indicated by the scale estimate.

5.10 Exercises

5.1.1 Give several practical examples of time series with pronounced trend and seasonal components. Hint: Examples like monthly sales of airline tickets or monthly demand of electricity may help us to think about other possible time series. Also, the history of Wall Street is full of such examples.
5.1.2 Suppose that $X_t = Z_t + bZ_{t-1}$, $t = 1, 2, \ldots, n$, where Z_t are iid standard normal (standard Gaussian white noise). Find the joint cumulative distribution function of $(X_1, \ldots, X_n)$, and prove that this time series is second-order stationary.
5.1.3 Let $\{Z_t\}$ be a standard Gaussian white noise. Which of the following processes are second-order stationary? Also, for each stationary process find the mean and the autocovariance function. (a) $X_t = a + bZ_t$. (b) $X_t = Z_t \cos(wt)$. (c) $X_t = Z_t Z_{t-1}$. (d) $X_t = Z_t \cos(wt) + Z_{t-1} \sin(wt)$.
5.1.4 Assume that $\{X_t\}$ and $\{Y_t\}$ are uncorrelated second-order stationary sequences. Show that their sum is also second-order stationary with the autocovariance function equal to the sum of the autocovariance functions of $\{X_t\}$ and $\{Y_t\}$.
5.1.5 Consider the monthly (or annual) sunspot data discussed in Section 1.4. What choice of $J_{\max}$ would you recommend for separating a seasonal component with period about 10 years?
5.1.6 Check that the autocovariance function satisfies the inequality $|\gamma(h)| \le \gamma(0)$ for any lag h. Hint: Use the Cauchy–Schwarz inequality.
5.1.7 Explain (5.1.7).
5.1.8 Why is it assumed that a seasonal component is summed to zero?
5.1.9 Repeat Figure 5.1 with both positive and negative values of b. How does the sign of b affect the realizations? Make similar experiments with a.
5.1.10 Repeat Figure 5.1 with different a and b, and find the most misleading cases.
5.1.11 Repeat Figure 5.1 with different seasonal components. Which components are most difficult and which simplest for visualization? Why?
5.1.12 Let $\{X_t\}$ have a seasonal component with period T. Consider a new time series $Y_t = X_t - X_{t-T}$. What can be said about this new sequence? Is it plausible to use the differencing to eliminate the seasonal component?
5.1.13 Use the idea of the previous exercise and suggest a method of eliminating a linear trend.
5.1.14 Would you recommend any changes in the default values of coefficients of the estimate used in Figure 5.2?
5.1.15 Use Figure 5.2 to answer the following questions. If σ (controlled by the argument *sigma*) is reduced, then does the conventional estimator outperform the nonparametric one? If the answer is yes, then what is the boundary value of σ when this happens? Also, how do the sample size and the period of the seasonal component affect that boundary value?

5.1.16 Let $X_1, \ldots, X_n$, $n \geq p$, be a realization of a causal AR(p) process (5.1.4). Show that the estimate $\hat{X}_{n+1} := \sum_{j=1}^{p} a_j X_{n+1-j}$ is the *best linear predictor* of X_{n+1} that minimizes the mean squared error $R := E\{(\tilde{X}_{n+1} - X_{n+1})^2\}$ over all linear estimates $\tilde{X}_{n+1} := \sum_{j=1}^{n} c_j X_{n+1-j}$. Also, can the assumption about causality be dropped? Hint: Write the mean squared error as

$$R = E\Big\{\Big(\sum_{j=1}^{p}(c_j - a_j)X_{n+1-j} + \sum_{j=p+1}^{n} c_j X_{n+1-j} - Z_{n+1}\Big)^2\Big\},$$

and also note that due to causality of the AR(p) process, the inequality $R \geq E\{Z_{n+1}^2\}$ holds.

5.1.17 At the end of the subsection *Causal ARMA Processes*, a causal and stationary solution of the ARMA(1, 1) difference equation is suggested. (a) Check it. (b) Does the conclusion hold for $|a| \geq 1$?

5.2.1 Explain why (5.2.1) implies (5.2.2). Also, is it correct to refer to this formula as a cosine orthogonal series expansion?

5.2.2 Show that the spectral density of a second-order stationary time series is always a real, even, and nonnegative function.

5.2.3 Consider the sunspot data discussed in Section 1.3 and find the frequency of the seasonal component. Hint: Use (5.2.3).

5.2.4 Find the mean and variance of the sample autocovariance function (5.2.4). Is it an unbiased estimate? Can you suggest an unbiased estimate? Was the assumption about second-order stationarity sufficient for solving this exercise?

5.2.5 The sample autocovariance function may be computed for any data set, and it is not restricted to realizations of a stationary series. What may be said about this estimate if the data contain a trend or a seasonal component?

5.2.6 Consider an MA(1) process $X_t = Z_t + bZ_{t-1}$. Draw the corresponding autocovariance function and spectral density for the cases of positive and negative b. Discuss the graphics. Use your conclusions to analyze corresponding simulations made by Figure 5.3.

5.2.7 Check the equality in (5.2.5).

5.2.8 Using Exercise (5.2.4), check (5.2.9) for a causal ARMA process with bounded fourth moments.

5.2.9 Establish (5.2.10). Hint: Recall that second-order stationarity implies that the autocovariance function is well defined and depends only on the lag, that is, $\gamma(j) = E\{X_{t+j}X_t\}$ for any t. The causality implies that $X_t = \sum_{j=0}^{\infty} c_j Z_{t-j}$ for some absolutely summable coefficients $\{c_j\}$. Using these facts, obtain 4 different equations via multiplying both sides of the difference equation $X_t - aX_{t-1} = \sigma(Z_t + bZ_{t-1})$ by X_t, X_{t-1}, Z_t, and Z_{t-1}. Then take expectations and get

$$\gamma(0) - a\gamma(1) = \sigma E\{X_t Z_t\} + \sigma b E\{X_t Z_{t-1}\},\ \gamma(1) - a\gamma(0) = \sigma b E\{X_{t-1} Z_{t-1}\},$$

$$E\{X_t Z_t\} = \sigma\,, \quad E\{X_t Z_{t-1}\} - aE\{X_{t-1} Z_{t-1}\} = \sigma b\,.$$

Because $E\{X_{t-1}Z_{t-1}\} = E\{X_t Z_t\}$, one gets a system of 4 linear equations in 4 variables. Its solution will imply (5.2.10).

5.3.1 Explain the underlying idea of formula (5.1.8) and how to use it for separating a trend component from a seasonal component.

5.3.2 Consider the monthly and annual sunspot data discussed in Section 1.3. What $T_{\max}$ and the corresponding $J_{\max}$ would you recommend for these time series?

5.3.3 Using Figure 5.4, find a minimal n such that either detrending or deseasonalizing becomes impossible.

5.3.4 Using Figure 5.4, find most difficult to estimate trend and seasonal components among the set of corner functions.

5.3.5 Choose several different trends, seasonal components, error terms, and sample sizes. Would you recommend any changes in the values of coefficients of estimates used by Figure 5.4?

5.3.6 Use Figure 5.4 and answer the following question. What are the types of ARMA(1, 1) errors that make the problem of detrending more and less complicated? Answer the same question for deseasonalizing.

5.3.7 What changes in the values of coefficients of the spectral density estimate would you recommend to improve the estimation of the period of an underlying seasonal component? Recall that these arguments end with the string *sp*.

5.3.8 Explain why the arguments *set.period* and *set.lambda* duplicate each other.

5.3.9 Let the estimated period be 9.95. What periods would you like to try in this case using the argument *ManualPer=T*, which allows one to choose the period manually?

5.4.1 Give and then discuss several practical examples of time series with missing observations. Then, give several similar examples for spatial data. Also, does the notion of causality have a sense for spatial data? Hint for the last question: Think about spatial data related to a river and lake.

5.4.2 Consider two second-order stationary time series of the same sufficiently large length. In the first one every other observation is skipped, while in the second one half of the observations are skipped at random. Which series would you prefer to deal with to estimate the autocovariance function? Suggest a consistent estimator for that series.

5.4.3 Show that (5.4.1) is an unbiased estimator of $\gamma(j)$ if $m_j > 0$. Also, assuming that fourth moments exist, find its variance.

5.4.4 Use Figure 5.6 and answer the following questions: (a) Let 2 of every 7 observations be missing. Does the location of these missing observations affect the analysis of the time series? (b) Decrease n and find a maximal n^* such that the analysis becomes impossible. (c) What particular values of *TMAX* would you recommend for the cases of the Uniform and the Strata trend components? (d) Would you recommend any changes in the values

of the coefficients of the estimates? (e) Set $Tseas = 7$, run Figure 5.6, and explain the results.

5.5.1 Give several practical examples of time series whose trends have hidden additive components. Discuss a possibility to recover them.

5.5.2 Let the weights $\{w_k\}$ in (5.5.2) be given and let $K > 1$. Explain why knowing only a realization of (5.5.1) is not sufficient for estimating the hidden components.

5.5.3 Explain how (5.5.8) is obtained.

5.5.4 Explain the matrix equation (5.5.9).

5.5.5 Why should the matrix W be invertible? Also, for the case where a nuisance additive component is present, what is a necessary property of W to recover the hidden components?

5.5.6 Explain all the diagrams in Figure 5.7.

5.5.7 Use Figure 5.7 and find most difficult and simplest triplets of hidden components among the set of corner functions.

5.5.8 Explain all steps of the data-driven estimator for the case of a nuisance additive component.

5.5.9 Use Figures 5.7–5.9 to explore the case of a nuisance additive component.

5.6.1 Establish (5.6.7) and (5.6.8).

5.6.2 Consider a bivariate time series $\{(X_t, Y_t)\}$ where $X_t = Z_t$, $Y_t = aX_t + bX_{t-k}$, and Z_t is a standard Gaussian white noise. Find: (a) the covariance function and spectral density for each univariate time series; (b) the cross-covariance and cross spectrum; (c) the absolute coherence; (d) the phase spectrum. Also, what conclusion about this pair of time series may be drawn from the analysis of the absolute coherence and the phase spectrum?

5.6.3 Find formulae for the absolute coherence and the cross spectrum of the econometrics model (5.6.17).

5.6.4 Use Figure 5.10 to analyze outcomes for negative delays and large b. Also, is there a set of parameters in the model (5.6.1)–(5.6.2) that allows one to visualize the linear relationship between these two series via their realizations?

5.6.5 Using Figure 5.11, find parameters of the econometrics model when realizations do not appear to be stationary. Explain the outcome.

5.6.6 Using Figure 5.11, analyze the effect of the parameters b_P and b_S on the coherence and phase spectrum.

5.6.7 Would you recommend any changes in the default values of coefficients of the estimates used by Figures 5.10–5.11?

5.7.1 Consider a dynamic model (5.7.1) for the particular $f(y)$ used in Figure 5.12 and $s(y) = 0$ (no noise term). Draw a typical series of states.

5.7.2 Give several examples in which a dynamic model may be used as an approximation of a real time series.

5.7.3 Is the time series (5.7.1) second-order stationary?

5.7.4 Repeat Figure 5.12 several times. Are the outcomes stable?

5.7.5 Repeat Figure 5.12 with different values of the coefficients A and B. Discuss the results.
5.7.6 Repeat Figure 5.12 with different values of the coefficients σ, a, and b that define the noise term. Discuss the results.
5.7.7 Suggest optimal coefficients of the estimator used in Figure 5.12.
5.8.1. Verify (5.8.1).
5.8.2. Give an example where $\hat{\nu}^2/2$ is an asymptotically unbiased estimate of σ^2.
5.8.3. Use Figure 5.13 to analyze how the size of a jump and the variance of a noise affect the estimation.
5.8.4. Use Figure 5.13 and analyze coefficients of an ARMA(1, 1) process which are more and less favorable to estimation of a trend with a jump change-point.
5.9.1 Make a comprehensive analysis of a time series available in the S–PLUS time series data-files. Write a short report about the analysis.

5.11 Notes

The main practical message of this chapter is similar to the previous ones: nonparametric methods should always be used as a first look at the data at hand. Even if someone is absolutely sure that an underlying model is a parametric one, say a linear trend plus AR(1) noise, it is worthwhile to check this assumption using nonparametric methods. This conservative approach costs nothing but may prevent inconsistent conclusions.

5.1 A relatively simple introduction to the topic may be found in the textbooks by Brockwell and Davis (1991, Chapter 1), Diggle (1990), and Schumway (1988). The book by Ripley (1988) is devoted to spatial processes. Fan and Gijbels (1996, Chapter 6) review and discuss nonseries estimates for nonparametric analysis of time series.

5.2 The asymptotic justification of a series spectral density estimator has been explored by Bentkus (1985), Efromovich (1984, 1998b), Efromovich and Pinsker (1981, 1986), Levit and Samarov (1978), Rudzkis (1985), Rudzkis and Radavicius (1993), and Samarov (1977) among others.

5.3 A rigorous mathematical discussion of estimation of parameters of ARMA processes may be found in Brockwell and Davis (1991, Chapter 8).

5.4 Parametric state-space models with missing observations are discussed in the book by Brockwell and Davis (1991, Section 12.3).

5.5 A related topic of testing on the presence of hidden periodicities (seasonal components) is discussed in the book by Brockwell and Davis (1991, Section 10.2).

5.6 The classical parametric theory of multivariate time series, including the discussion of the econometrics model, may be found in the book by Brockwell and Davis (1991, Chapter 11).

5.7 The books by Prakasa Rao (1983, Chapter 6) and Doukhan (1994, Section 2.4) give a mathematical discussion of this and related models. Applications in mathematical finance as well as the relation to the famous Black–Scholes model are discussed in the book by Baxter and Rennie (1996).

5.8 Using Gibbs phenomenon and wavelets for finding change-points is discussed in Pawlak (1994) and Wang (1995), respectively.

5.9 Applications of the universal estimator to real time series may be found in Efromovich (1998b), where also the historical overview of time- and frequency-domain approaches is given.

6

Estimation of Multivariate Functions for Small Samples

In this chapter we shall discuss several topics in multivariate function estimation with applications to density and regression estimation.

The chapter begins with a discussion of a natural extension of approximation methods discussed in Chapter 2. We shall see that using tensor–product bases makes the problem of series approximation of multivariate functions similar to the univariate case. Nonetheless, several technical difficulties arise. First, apart from bivariate functions (surfaces), there is no simple tool to visualize a multidimensional curve. Second, we have seen in Section 2.1 that to approximate fairly well a smooth univariate function, about 5 to 10 Fourier coefficients are needed. For the case of a d-dimensional curve this translates into 5^d to 10^d Fourier coefficients. Since these coefficients must be estimated, this makes the estimation problem complicated for the case of small samples. Third, suppose that $n = 100$ points are uniformly distributed over the five dimensional unit cube $[0, 1]^5$. What is the probability of having some points in a neighborhood of reasonable size, say a cube with side 0.2? Since the volume of such a cube is $(0.2)^5 = 0.00032$, the expected number of points in this neighborhood is n times $(0.2)^5$, i.e. 0.032. As a result, no averaging over that neighborhood can be performed. For this example, to get on average of 5 points in a cube, its side should be 0.55, that is, more than a half of the range along each coordinate. This shows how sparse multivariate observations are. Fourth, the notion of a small sample for multivariate problems mutates. Suppose that for a univariate regression a grid of 50 points is considered sufficient. Then this translates into 50 points along each axis, i.e., into 50^d data points.

These complications present a challenging problem, which is customarily referred to as the *curse of dimensionality*. However, in no way does this curse imply that the situation is hopeless.

We begin the discussion with some classical series approximation results and their effect on risk convergence. Then classical settings of density estimation and nonparametric regression are discussed. The universal estimator can be used for any dimension, and it allows a straightforward extension to all the more complicated models discussed in Chapters 3 and 4. In some multivariate regression settings an additive model for a regression function may be a fair assumption that drastically simplifies estimation. We discuss this approach in Section 6.5.

6.1 Series Approximation of Multivariate Functions

We begin with the case of bivariate functions. The classical examples are surfaces and images. This case explains all complications of approximation of multivariate functions. On the other hand, its relative simplicity and the availability of good methods for visualizing bivariate functions make this case an excellent introduction into the world of multivariate functions.

Denote by $L_2(A \times B)$ the space of square integrable bivariate functions $f(x,y)$ such that $\int_A \int_B f^2(x,y)dx\,dy < \infty$.

Let $\{\phi_n, n = 0,1,\ldots\}$ and $\{\psi_m, m = 0,1,\ldots\}$ be two bases in the one-dimensional spaces $L_2(A)$ and $L_2(B)$, respectively. Then products of elements from these two bases,

$$\{\varphi_{nm}(x,y) := \phi_n(x)\psi_m(y), \quad n,m = 0,1,\ldots\}, \tag{6.1.1}$$

constitute a basis in $L_2(A \times B)$. This basis is called a *tensor-product* basis. Thus all the univariate bases discussed in Chapter 2 may be used to create the corresponding tensor-product bases for two-dimensional spaces.

As an example, the *cosine tensor-product basis* in $L_2([0,1]^2)$ (here $[0,1]^2 := [0,1] \times [0,1]$ is the unit square) has the elements

$$\begin{aligned}&\varphi_{00}(x,y) = 1,\ \varphi_{01}(x,y) = \sqrt{2}\cos(\pi y),\ \varphi_{02}(x,y) = \sqrt{2}\cos(2\pi y), \ldots,\\ &\varphi_{10}(x,y) = \sqrt{2}\cos(\pi x),\ \ \varphi_{11}(x,y) = 2\cos(\pi x)\cos(\pi y), \ldots.\end{aligned} \tag{6.1.2}$$

A corresponding partial sum with cutoffs J_1 and J_2, respectively relative to the variables x and y, is

$$f_{J_1J_2}(x,y) := \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2} \theta_{j_1j_2}\varphi_{j_1j_2}(x,y), \tag{6.1.3}$$

where the Fourier coefficients $\theta_{j_1j_2}$ are defined by the formula

$$\theta_{j_1j_2} := \int_0^1\int_0^1 f(x,y)\varphi_{j_1j_2}(x,y)dx\,dy. \tag{6.1.4}$$

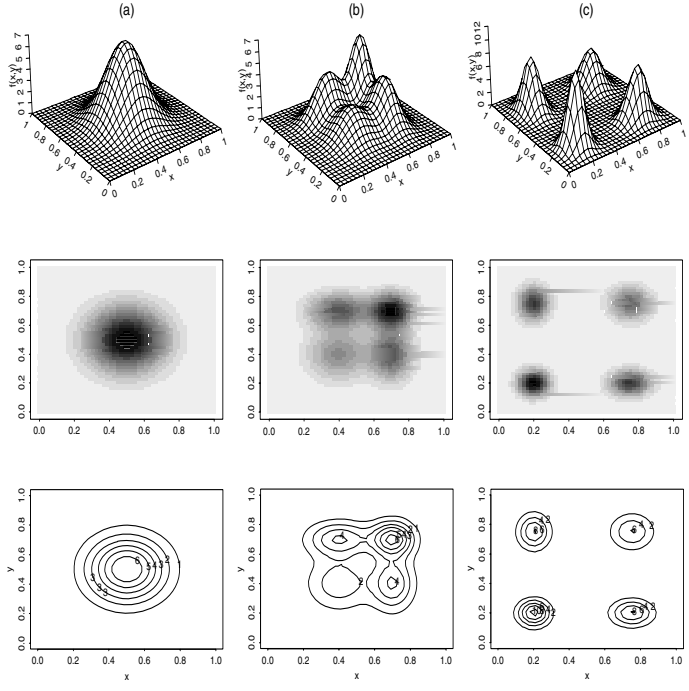

FIGURE 6.1. Three bivariate corner functions: (a) the Normal by the Normal; (b) the Bivariate by the Bivariate; (c) the Strata by the Strata. These functions are shown by perspective plots (upper row), image plots (middle row), and counter plots (bottom row). {The arguments *cij* allow one to visualize a product of any two corner functions. For instance, by choosing $c11 = 3$ and $c12 = 5$ the product of the Bivariate by the Delta can be visualized in column (a), by choosing $c21 = 7$ and $c22 = 8$ the product of the Monotone by the Steps can be visualized in column (b), etc. Precaution: It takes a relatively long time to print a hard copy of an image plot, and it also takes a lot of memory to store an image. Therefore, in all other figures with images we shall use a relatively small number of pixels controlled by the argument *num.pel*. This argument is equal to the square root of the number of pixels used.} [c11=2, c12=2, c21=3, c22=3, c31=4, c32=4]

A bivariate polynomial (Legendre) tensor-product basis is defined absolutely similarly.

For performance assessment we choose a set of bivariate corner functions that are products of our corner functions shown in Figure 2.1. Overall this set contains 64 bivariate functions. In Figure 6.1, three bivariate corner functions are shown: Diagram (a) is the product of the Normal by the Normal; (b) is the product of the Bivariate by the Bivariate; (c) is the product of the Strata by the Strata.

The top row of diagrams shows the perspective plots created by the S–PLUS function **persp**. Perspective plots give a three-dimensional view of

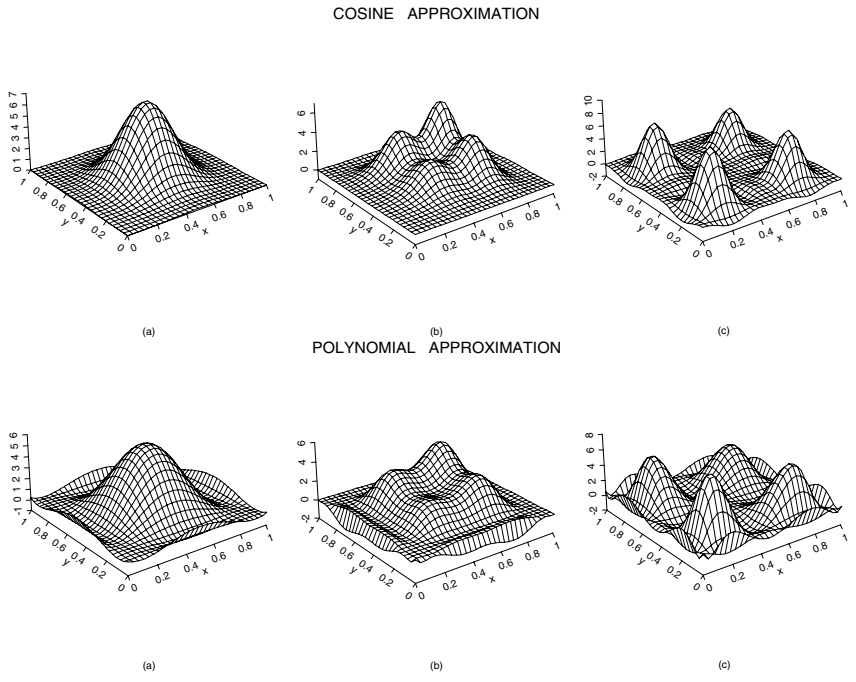

FIGURE 6.2. Approximations by cosine and polynomial partial sums (6.1.3) of the bivariate functions shown in Figure 6.1. {The arguments Jkj control cutoffs; for instance, $J12$ controls J_2 for column (a) and $J31$ controls J_1 for column (c). Thus, these indices are the same as in the arguments ckj.} [c11=2, c12=2, c21=3, c22=3, c31=4, c32=4, J11=5, J12=5, J21=8, J22=8, J31=8, J32=8]

bivariate curves in the form of a matrix of heights on an evenly spaced grid. The heights are connected by line segments to produce the mesh appearance of such plots. A perspective plot can be modified by choosing a different "eye" location. Figure 6.4 in Section 6.3 illustrates this possibility.

The second row of diagrams in Figure 6.1 shows image plots of the same bivariate functions. Image plots are produced by the S–PLUS function **image**. Although the gray-scale images are not impressive, they look very attractive on a color monitor, and color prints are very informative. Also, even the gray-scale images exhibit bivariate functions nicely.

The bottom row shows the counter plots of the same bivariate functions created by the S–PLUS function **counter**. A counter plot shows a surface as a set of counter lines on a grid representing the other two variables.

Figure 6.2 shows partial trigonometric and polynomial sums (approximations) of the 3 bivariate functions shown in Figure 6.1. The diagrams (a) correspond to cutoffs $J_1 = J_2 = 5$, the diagrams (b) to $J_1 = J_2 = 8$, and the diagrams (c) to $J_1 = J_2 = 8$.

Now let us explain how theoretical results on approximation of a univariate function are extended to the bivariate case. A standard theoretical result looks as follows. Fix $y = y_0$ and assume that $f(x, y_0)$ as a function of x is smooth, say it is Lipschitz $Lip_{r_1,\alpha_1,L}$ uniformly over y_0 (see the definition of these functions in (2.4.13)). Denote by $\beta_1 := r_1 + \alpha_1$ the parameter of smoothness corresponding to x. Let us also assume that a similar assumption holds for $f(x_0, y)$ as a function in y with the parameter of smoothness β_2. Then the integrated squared bias $\text{IBS}_{J_1J_2}$ of the partial sum (6.1.3) satisfies

$$\text{ISB}_{J_1J_2} := \int_0^1 \int_0^1 (f(x,y) - f_{J_1J_2}(x,y))^2 dx\, dy \le C[J_1^{-2\beta_1} + J_2^{-2\beta_2}]. \tag{6.1.5}$$

Recall that C denotes finite and in general different constants.

Comparison of (6.1.5) with the univariate approximation result (2.4.18) shows that the integrated squared bias (the error of approximation) for a bivariate function is at most a factor times a sum of integrated squared biases for approximations in each argument. By itself this is a good outcome, since the errors are simply added. The problem is that to get (6.1.5), the number of Fourier coefficients should be of order J_1J_2. This is what makes a multivariate statistical problem essentially more complicated than a univariate one. Indeed, in statistical applications every Fourier coefficient that is used in a partial sum should be estimated, and as we know from previous chapters, an extra estimated Fourier coefficient usually adds the value Cn^{-1} to the variance term. Thus, in a univariate case the mean integrated squared error (MISE) is proportional to $Jn^{-1} + J^{-2\beta}$, where β is the parameter of smoothness of an estimated univariate curve. In the bivariate case, MISE is proportional to $J_1J_2n^{-1} + J_1^{-2\beta_1} + J_2^{-2\beta_2}$.

As a result, a straightforward calculation (Exercises 6.1.3–6.1.4) shows that in the univariate case the optimal cutoff J^* and the corresponding MISE are proportional to

$$J^* \asymp n^{1/(2\beta+1)}, \qquad \text{MISE} \asymp n^{-2\beta/(2\beta+1)}, \tag{6.1.6}$$

whereas for the bivariate case (set $\rho := \beta_1\beta_2/(\beta_1 + \beta_2)$)

$$J_1^* \asymp n^{\rho/\beta_1(2\rho+1)}, \quad J_2^* \asymp n^{\rho/\beta_2(2\rho+1)}, \qquad \text{MISE} \asymp n^{-2\rho/(2\rho+1)}. \tag{6.1.7}$$

There are several important conclusions from these simple results that shed light on the problem of estimation of multivariate functions. First, let $\beta_1 \le \beta$; that is, a bivariate function, as a function in x, is not smoother than a univariate function. Then, regardless of how smooth this bivariate function is in y (that is, regardless of how large β_2 is), the bivariate function cannot be estimated more accurately than the univariate one. This conclusion plainly follows from the inequality $\beta_2/(\beta_1 + \beta_2) < 1$.

Second, these results show that if $\beta = \rho$, then a bivariate function may be estimated with the same accuracy (up to a constant factor) as the univariate

one. This allows us to appreciate the complexity of estimating a bivariate function via our experience of estimating univariate functions.

Note that we have discussed the complexity of the problem via the orthogonal series approach. It is possible to show that the conclusion holds for any other method, in short, there is no other method that can outperform a series approach.

The case of bivariate functions is straightforwardly extended to the case of d-variate functions $f(x_1, x_2, \ldots, x_d)$. For instance, the complex trigonometric expansion for $f \in L_2([0,1]^d)$ will be

$$f(x_1, x_2, \ldots, x_d) = \sum_{j_1, j_2, \ldots, j_d = -\infty}^{\infty} \theta_{j_1 j_2 \cdots j_d} e^{i2\pi(j_1 x_1 + j_2 x_2 + \cdots + j_d x_d)},$$

where the Fourier coefficients are

$$\theta_{j_1 j_2 \cdots j_d} := \int_0^1 \cdots \int_0^1 f(x_1, x_2, \ldots, x_d) e^{-i2\pi(j_1 x_1 + j_2 x_2 + \cdots + j_d x_d)} dx_1 \cdots dx_d.$$

Also, a d-dimensional analogue of formula (6.1.5) for integrated squared bias will have d terms, each corresponding to the smoothness of an underlying function in a particular coordinate. Thus, all the issues discussed earlier can be straightforwardly extended to higher dimensions (Exercise 6.1.5).

6.2 Density Estimation

We begin with the estimation of a bivariate density $f(x, y)$ of a pair of random variables (X, Y), which, in general, can be dependent. Assume that $f(x, y)$ should be estimated over the unit square $[0, 1]^2$ based on n iid realizations (X_l, Y_l), $l = 1, 2, \ldots, n$, of (X, Y). (The estimators discussed below are similar to ones suggested in Sections 3.1–3.3, so a review of those sections is recommended.)

Consider a series estimator based on the bivariate cosine tensor-product basis (6.1.2). According to (6.1.3), a projection series estimator should look like

$$\tilde{f}_{J_1 J_2}(x, y) := \sum_{j_1=0}^{J_1} \sum_{j_2=0}^{J_2} \hat{\theta}_{j_1 j_2} \varphi_{j_1 j_2}(x, y). \tag{6.2.1}$$

Here $\hat{\theta}_{j_1 j_2}$ is an estimate of the Fourier coefficient

$$\theta_{j_1 j_2} = \int_0^1 \int_0^1 \varphi_{j_1 j_2}(x, y) f(x, y) dx\, dy = E\{I_{\{(X,Y) \in [0,1]^2\}} \varphi_{j_1 j_2}(X, Y)\}.$$

Recall that $I_{\{A\}}$ is the indicator of an event A.

Since $\theta_{j_1 j_2}$ is the expectation of $I_{\{(X,Y)\in[0,1]^2\}}\varphi_{j_1 j_2}(X,Y)$, it is natural to estimate the Fourier coefficient by a sample mean estimate

$$\hat{\theta}_{j_1 j_2} := n^{-1}\sum_{l=1}^{n} I_{\{(X,Y)\in[0,1]^2\}}\varphi_{j_1 j_2}(X_l, Y_l). \tag{6.2.2}$$

Then, as in (3.1.8) we may obtain that

$$E\{(\hat{\theta}_{j_1 j_2} - \theta_{j_1 j_2})^2\} = n^{-1}(\theta_{00} + r_{n,j_1,j_2}), \tag{6.2.3}$$

where r_{n,j_1,j_2} decays as n, j_1, and j_2 increase. This implies that the coefficient of difficulty is $d := \theta_{00}$, and if $[0,1]^2$ is the support, then $d = 1$. Thus, as in the univariate case the estimation of a bivariate density may be considered as a basic model for all other problems.

Moreover, we may straightforwardly use the universal estimator suggested in Section 3.1. Let us repeat the steps of this estimator using the same notation as in Section 3.1.

Step 1. Fourier coefficients $\theta_{j_1 j_2}$, $0 \le j_1, j_2 \le c_{JM} J_n$, are estimated by the sample mean estimator (6.2.2), and $\hat{d} := \hat{\theta}_{00}$.

Step 2. As in (3.1.10), optimal cutoffs $\hat{J}_1$ and $\hat{J}_2$ are calculated by the formula

$$(\hat{J}_1, \hat{J}_2) := \operatorname{argmin}_{0\le J_1\le J_n,\, 0\le J_2\le J_n}\Big\{\sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}[2\hat{d}n^{-1} - \hat{\theta}^2_{j_1 j_2}]\Big\}. \tag{6.2.4}$$

Step 3. Smoothing weights are calculated:

$$\hat{w}_{00} := 1 \quad \text{and} \quad \hat{w}_{j_1 j_2} := (1 - \hat{d}/n\hat{\theta}^2_{j_1 j_2})_+, \quad j_1 + j_2 > 0. \tag{6.2.5}$$

Step 4. The universal estimate is calculated:

$$\begin{aligned}\tilde{f}(x,y) := &\sum_{j_1=0}^{\hat{J}_1}\sum_{j_2=0}^{\hat{J}_2}\hat{w}_{j_1 j_2}\hat{\theta}_{j_1 j_2}\varphi_{j_1 j_2}(x,y)\\ &+ \sum_{(j_1,j_2)\in D} I_{\{\hat{\theta}^2_{j_1 j_2} > c_T \hat{d}\ln(n)/n\}}\hat{\theta}_{j_1 j_2}\varphi_{j_1 j_2}(x,y),\end{aligned} \tag{6.2.6}$$

where D is the set of indices (j_1, j_2) such that $0 \le j_1, j_2 \le c_{JM} J_n$ with deleted indices considered in the first term of (6.2.6).

Step 5. A bona fide series estimate is defined as

$$\hat{f}(x,y) := (\tilde{f}_{J_1 J_2}(x,y) - c)_+, \tag{6.2.7}$$

where $(x)_+ = \max(0,x)$ denotes the positive part of x, and the constant c is chosen in such a way that $\hat{f}(x,y)$ is a bona fide density on $[0,1]^2$, that is,

$$\int_0^1\int_0^1 \hat{f}(x,y)dx\,dy = 1. \tag{6.2.8}$$

Also, small bumps are removed, as discussed in Section 3.1.

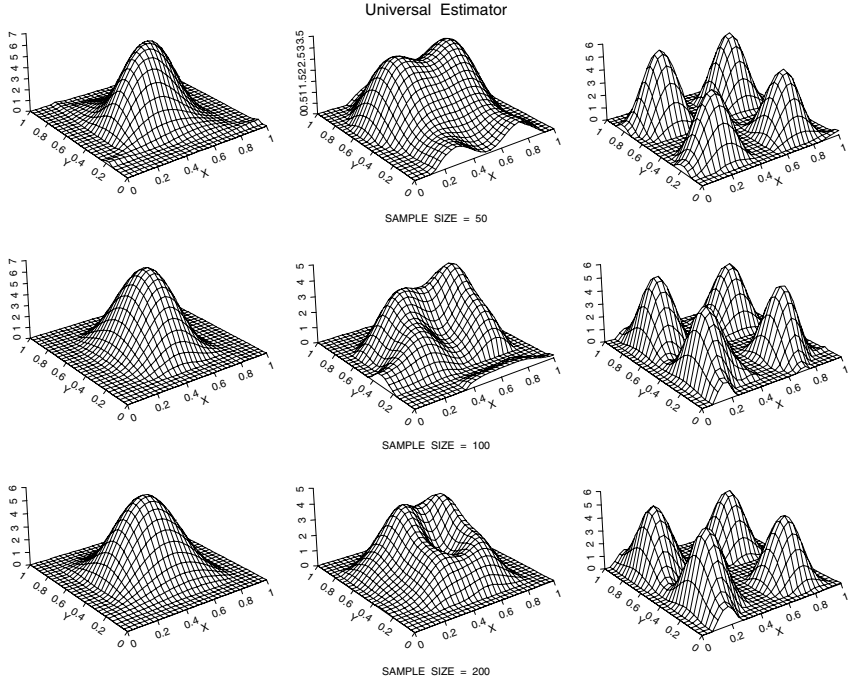

FIGURE 6.3. Estimation of three bivariate densities shown in Figure 6.1. The top, middle, and bottom rows correspond to sample sizes 50, 100, and 200. The estimator is universal. {Here the "new" arguments are *set.n*, which controls the choice of sample sizes; *estimate*, which allows one to use either the default universal estimate or the hard-threshold estimate; in this case set *estimate* = $''h''$ and you will see the change in the title as well. The arguments *cJ0, cJ1, cJM, cT, cB* control the coefficients of the estimator. The coefficients of the universal estimator are reviewed in the caption of Figure 3.2, and for the hard-threshold estimator check with (6.2.9) and recall that here the default is *cT=2.*} [*set.n = c(50,100,200), c11=2, c12=2, c21=3, c22=3, c31=4, c32=4, cJ0=4, cJ1=.5, cJM=2, cT=4, cB=1, estimate* = $''u''$]

These 5 steps define the universal bivariate estimator. Note that an extension to the d-variate case is straightforward.

To evaluate the performance of this estimator, consider the following Monte Carlo experiment. We study estimation of bivariate densities shown in Figure 6.1 for particular samples of sizes 50, 100, and 200. Recall that for the univariate corner densities and the same sample sizes particular realizations are shown in Figure 3.2.

Figure 6.3 shows particular estimates. As we see, the data-driven estimator does a decent job even for the smallest sample sizes. A striking similarity with the univariate cases is that larger samples may sometimes lead to worse estimates. This is clearly the case for the Strata by Strata

density. A remark is needed about the procedure of removing small bumps. For the bivariate case it is probably more correct to refer to this procedure as removing small "hills," and some training with different coefficients c_B is recommended. Figure 6.3 uses the default value $c_B = 1$. Also, to make the calculations faster, the default value of the coefficient c_{JM} is 2. (Recall that for a d-variate function the universal estimator calculates $(1+c_{JM}J_n)^d$ Fourier coefficients.) Repeated simulations show that estimation of bivariate densities is a reasonable task even for the smallest sample sizes.

The universal estimate is relatively simple; nevertheless, for multivariate settings it is also worthwhile to consider a simpler hard-threshold series estimator (recall the discussion in Sections 3.2–3) of a d-variate density. The estimator is defined as

$$\tilde{f}(x_1, x_2, \ldots, x_d) :=$$

$$\sum_{j_1,j_2,\ldots,j_d=0}^{c_{JM}J_n} I_{\{\hat{\theta}^2_{j_1j_2\cdots j_d} > c_T \hat{d} \ln(n)/n\}} \hat{\theta}_{j_1j_2\cdots j_d} \varphi_{j_1j_2\cdots j_d}(x_1, x_2, \ldots, x_d). \quad (6.2.9)$$

Then the bona fide projection (Step 5) is performed. Since thresholding is applied to all the estimated Fourier coefficients, it may be worthwhile to decrease the default c_T from 4 to 2. The reason for the decrease is that the universal estimator uses thresholding only at high frequencies for searching for extraordinary large Fourier coefficients, while the hard-threshold estimator applies the same thresholding to all Fourier coefficients.

The simplicity of this estimator is extremely important for multivariate settings. Note that no minimization problems like (6.2.4) should be solved; one just calculates sample mean estimates of Fourier coefficients and then thresholds them. We know from Sections 3.2–3.3 that this data-driven estimator performs relatively well for estimating univariate densities. Simulations show that the quality of estimation is reasonable for bivariate functions as well.

In Figure 6.3 we analyzed cases of independent random variables. Now let us test the hard-threshold estimate for the case of dependent variables. Figure 6.4 exhibits the case where X is distributed according to the Monotone corner density, and Y is distributed according to the Normal corner density if $X \le 0.5$ and according to the Strata corner density if $X > 0.5$. The particular sample size is $n = 100$.

The shape of this estimate, except for some minor boundary "wings," nicely mimics the shape of the underlying bivariate density. Indeed, for small x this estimate, as a function in the y coordinate, mimics a normal density, and for x close to 1 it resembles the Strata. On the other hand, for any fixed y the estimate, as a function of x, resembles the Monotone.

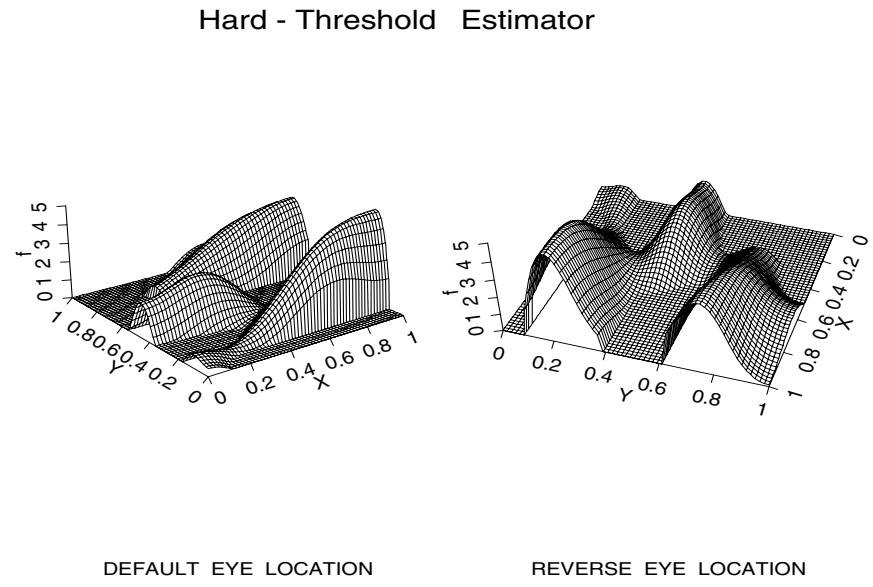

FIGURE 6.4. Estimation of the bivariate density for the dependent X and Y based on $n = 100$ observations. Two perspective plots of the universal estimate with two different "eye" locations are shown. {This figure allows one to use the universal estimate as well; to do this, set *estimate* $= "u"$, and then the argument cT will be automatically reset from 2 to the default value 4. The title always indicates which estimate has been used. The sample size is controlled by the argument n.} [*n = 100, cJ0=4, cJ1=.5, cJM=6, cT=2, cB=1, estimate* $= "h"$]

6.3 Density Estimation in Action: Discriminant Analysis

A basic problem of discriminant analysis is as follows. An observation can belong to a population 1 or population 2. Distributions of these populations are unknown. We must decide to which population that particular observation belongs based on given *training* sets (samples) from the first and the second population.

Such a problem is at the core of the theory of learning machines whose decisions are typically based on using additional training sets. As an example, assume that one wants to create a learning machine that gives a recommendation to buy or not to buy a used car. This machine, using results of specific tests of a used car, should give a specific recommendation: to buy or not to buy. To "train" the learning machine, results of tests for two sets of good and bad cars are available. Based on these training sets, a learning machine should develop an algorithm that separates good cars from bad ones. A similar example is automated medical diagnosis. Here an observation is a list with results of specific medical tests for a patient

whose diagnosis should be determined, and a training set is a collection of files with results of tests for patients with known diagnoses.

To put ourselves in a statistical mood, let us recollect a classical solution of this problem in which distributions of populations are supposed to be known. Let $f_1(x^d)$ and $f_2(x^d)$ be densities for a d-dimensional data-vector x^d from the first and second populations. For instance, the first population is a population of good cars and the second is of bad cars; or the first population is a population of healthy patients and the second is of patients with the flu. (In these particular examples x^d represents a file with results of available tests). A familiar maximum likelihood approach would allocate an observation z^d to the first population if

$$f_1(z^d) \geq f_2(z^d) \tag{6.3.1}$$

and to the second population otherwise. A more general hypothesis testing known as a Bayesian approach, where the probability p of z^d to be from the first population is assumed to be known, leads to the rule of allocating z^d to the first population if

$$f_1(z^d) \geq q f_2(z^d), \tag{6.3.2}$$

where q is a given constant. For instance, for the Bayesian approach $q = (1-p)/p$; see Problem 6.3.2.

A rule that defines the allocation of an observation to a particular population (like (6.3.1) or (6.3.2)) is called a *discrimination rule*.

In a majority of practical applications the densities f_1 and f_2 are unknown, so the discrimination rule must be estimated from training sets. In this case parametric discrimination theory assumes that the unknown densities come from some parametric family. There is no surprise, then, that typically this family is chosen to be a multivariate normal with respective mean vectors μ_1^d and μ_2^d and common variance matrix V. Then training sets can be used to estimate these parameters. Denote sample mean estimates by $\bar{\mu}_1^d$, $\bar{\mu}_2^d$, and the pooled sample variance matrix by $\bar{V}$. Then a simple calculation shows that the maximum likelihood discrimination rule (6.3.1) becomes a familiar Fisher's *linear discrimination rule*

$$(z^d - (\bar{\mu}_1^d + \bar{\mu}_2^d)/2)' \bar{V}^{-1} (\bar{\mu}_1^d - \bar{\mu}_2^d) \geq 0. \tag{6.3.3}$$

Also recall that if the normal populations have different variance matrices, then each variance matrix should be estimated based on the corresponding training set, and then (6.3.1) leads to a *quadratic discrimination rule* where the allocation depends on the value of a quadratic form in the observed data.

Now it is easy to appreciate the beauty and simplicity of the nonparametric approach. Indeed, because the densities f_1 and f_2 are unknown but samples according to these densities are given, they may be estimated by a nonparametric estimator, for instance, by the data-driven universal estimator of Section 6.2. Then, according to (6.3.2), the ratio of these estimates

defines a nonparametric discrimination rule. Note that training sets from f_1 and f_2 may have different sample sizes; in the numerical example they will be of the same size denoted by n.

There is one specific detail in using this plug-in method. In areas where both f_1 and f_2 are smaller than or equal to the accuracy of the nonparametric density estimation, the ratio of the estimates cannot be a good estimate of the ratio f_1/f_2. To avoid this complication, let us make no recommendations for such areas. In other words, if both densities are small, then no discrimination is made and this fact is declared.

As a result, the nonparametric discrimination rule (nonparametric learning machine) divides the domain into 3 areas. If an observation belongs to the first area, then it is declared to belong to population 1. If an observation belongs to the second area, then it is declared to belong to population 2. If it belongs to the third area, then no decision is made, since values of both estimated densities are too small in that area. In the following example they are to be smaller than $t(\ln(n+3)/n)^{1/2}$, where t is a coefficient with default value 3.

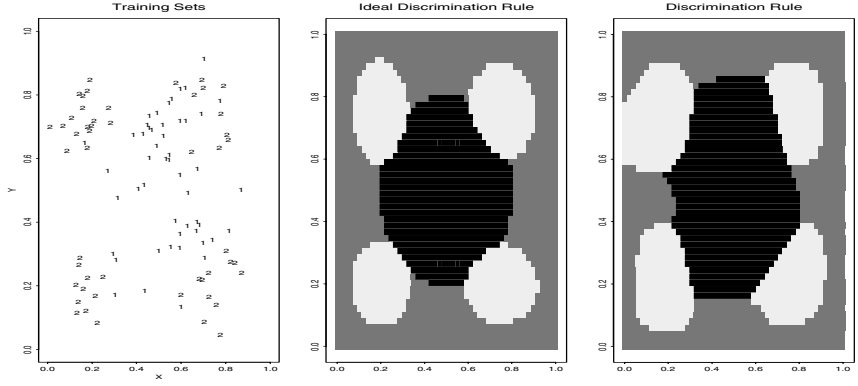

FIGURE 6.5. Nonparametric discriminant analysis based on training sets of size $n = 50$. The first and second populations have densities shown in Figures 6.1(a) and (c), respectively. The left diagram shows two particular training sets. Realizations from the first and second populations are shown by the symbols 1 and 2. The ideal discrimination rule (6.3.2) with $q = 1$, based on the underlying densities, is shown in the middle diagram. The grey area corresponds to the points where both these densities are smaller than $t(\ln(n)/n)^{1/2}$, $t = 3$. The nonparametric discrimination rule is shown in the right diagram. {As in Figure 6.1, the arguments $c1j$, $j = 1, 2$, control the choice of an underlying bivariate density for the first population and $c3j$, $j = 1, 2$, for the second population. The argument *num.pel* controls the number of pixels. It may be worthwhile to increase this number while visualizing this figure on a monitor, but for making a hard copy it is better to keep this number smaller to avoid a long printing time. All other arguments are apparent.} [*n = 50, q = 1, t=3, num.pel=50, c11=2, c12=2, c31=4, c32=4, cJ0=4, cJ1=.5, cJM=2, cT=4, cB=2, estimate = "u"*]

The result of a Monte Carlo simulation for $n = 50$ is shown in Figure 6.5. The left diagram shows a training set from a population 1 that corresponds to the bivariate density shown in Figure 6.1(a), and another training set from a population 2 that corresponds to the bivariate density shown in Figure 6.1(c). The discriminant analysis recalls a puzzle for youngsters in kindergarten: One should paint the area where the 1's are predominant in black, the area where the 2's are predominant in white, and in grey the area where no conclusive decision can be made.

Let us see how an oracle, who knows the underlying densities, and the nonparametric estimator, here based on the universal estimate, perform. The ideal discrimination rule, "painted" by the oracle, is shown in the middle diagram. This is a rather complicated discrimination rule. Note that all the boundaries should be smooth, but here we use a relatively small number of pixels, which implies rough boundaries.

The discrimination rule calculated by the nonparametric learning machine is shown in the right diagram. Based just on 50 observations from each population, this rule does a good job. It correctly shows 4 spots where the second population is predominant. The only incorrect detail is that the top left spot touches the boundary but this particular detail is supported by the data at hand. Also note that the outliers from the first population are correctly ignored. So, overall this result is good for such a small sample size.

6.4 Nonparametric Regression

Consider the heteroscedastic bivariate regression model

$$Y = f(X1, X2) + \sigma(X1, X2)\,\varepsilon, \tag{6.4.1}$$

where $(X1, X2)$ are predictors (a pair of random variables that may be dependent, and in what follows we shall often refer to them as *covariates*) with a design joint density $h(x1, x2)$ supported on the unit square $[0, 1]^2$ and that is bounded from zero on this square, $\sigma(x1, x2)$ is a bivariate scale function, and ε is a zero-mean and unit-variance error that is independent of the predictors.

The problem is to estimate the bivariate regression function (surface) $f(x1, x2)$ based on n iid realizations $\{(Y_l, X1_l, X2_l),\ l = 1, 2 \dots, n\}$ of the triplet of random variables $(Y, X1, X2)$.

The underlying idea of a series estimation of a bivariate regression function is absolutely the same as in the univariate case discussed in Section 4.2. First, we choose a convenient basis, for instance, the cosine tensor-product basis $\{\varphi_{j_1 j_2}(x1, x2)\}$ defined in (6.1.2). Then, according to Section 6.1, a bivariate regression function $f(x1, x2)$ can be approximated by a partial

sum,

$$f_{J_1J_2}(x1,x2) := \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\theta_{j_1j_2}\varphi_{j_1j_2}(x1,x2), \tag{6.4.2}$$

where the Fourier coefficients $\theta_{j_1j_2}$ are defined as

$$\theta_{j_1j_2} := \int_0^1\int_0^1 f(x1,x2)\varphi_{j_1j_2}(x1,x2)dx1\,dx2. \tag{6.4.3}$$

The main statistical issue is how to estimate the Fourier coefficients. The key idea is to rewrite the right-hand side of (6.4.3) as the expectation of a function of the triplet $(Y, X1, X2)$, and then use a sample mean estimate. Using the assumption that the design density $h(x1,x2)$ is bounded from below from zero on $[0,1]^2$ and that the error ε is zero-mean and independent of the predictors, we write,

$$\begin{aligned}\theta_{j_1j_2} &= \int_0^1\int_0^1 [f(x1,x2)\varphi_{j_1j_2}(x1,x2)/h(x1,x2)]h(x1,x2)dx1\,dx2\\ &= E\{f(X1,X2)\varphi_{j_1j_2}(X1,X2)/h(X1,X2)\}\\ &= E\{[f(X1,X2)+\sigma(X1,X2)\,\varepsilon]\varphi_{j_1j_2}(X1,X2)/h(X1,X2)\}\\ &= E\{Y\varphi_{j_1j_2}(X1,X2)/h(X1,X2)\}.\end{aligned} \tag{6.4.4}$$

Thus, the natural estimate of $\theta_{j_1j_2}$ is the sample mean estimate

$$\hat\theta_{j_1j_2} := n^{-1}\sum_{l=1}^{n} Y_l\varphi_{j_1j_2}(X1_l,X2_l)/h(X1_l,X2_l). \tag{6.4.5}$$

If a design density is unknown, and this is the typical case, then a density estimate $\tilde h$ should be plugged in. As an example, in this section we shall use the hard-threshold estimate of Section 6.2. Because a density estimate is used as the divider, it is truncated from below. Thus, the plugged-in density estimate is defined as

$$\hat h(x1,x2) := \max(c_D/\ln(n), \tilde h(x1,x2)). \tag{6.4.6}$$

Here c_D is a coefficient with the default value 1.

Then, as in Section 6.2, either a universal estimator or hard-threshold estimator may be used. Again, as an example, here we consider a hard-threshold estimator whose simplicity makes it so attractive for multivariate settings. A hard-threshold data-driven estimator is defined as

$$\hat f(x1,x2) := \sum_{j_1,j_2=0}^{c_{JM}J_n} I_{\{\hat\theta^2_{j_1j_2} > c_T\hat\sigma^2\ln(n)n^{-1}\}}\hat\theta_{j_1j_2}\varphi_{j_1j_2}(x1,x2), \tag{6.4.7}$$

where $\hat\sigma^2$ is the sample variance of responses. Then, if it is known that an underlying regression surface is larger than a given constant or integrated to a given constant, a projection like (6.2.7) should be used to get a bona fide estimate.

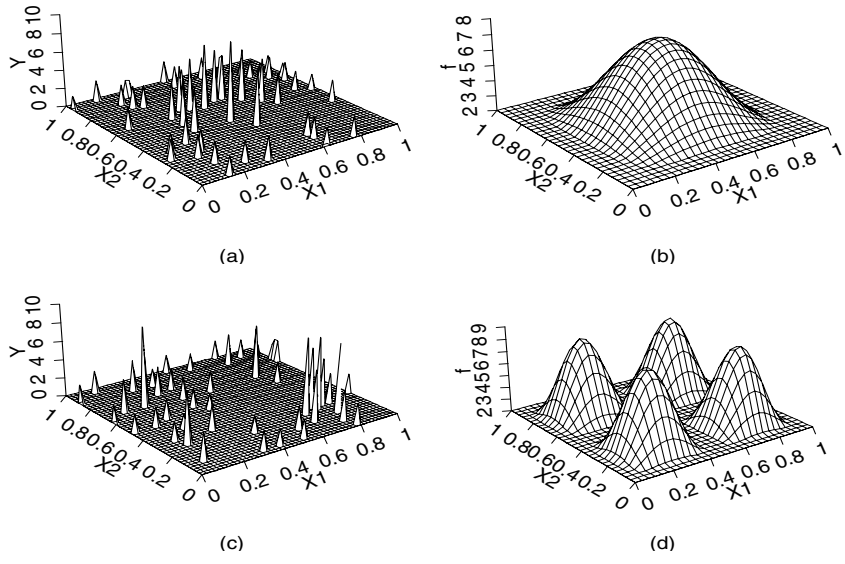

FIGURE 6.6. Scatter diagrams and data-driven hard-threshold estimates for two regression surfaces. The underlying surfaces are the surfaces shown in Figures 6.1(a) and 6.1(c), only here we increase their level by 2 (this is done to see all the spikes in the scattergrams). The sample size is $n = 50$. {The new arguments are *cD*, which controls the coefficient c_D used in (6.4.6), and *sigma*, which controls the value of the scale coefficient σ.} [*c11=2, c12=2, c21=4, c22=4, n = 50, sigma=.3, cJ0=4, cJ1=.5, cJT=2, cJM=2, cB=2, cD=1*]

Figure 6.6 illustrates both the problem and how the data-driven estimate performs. In diagram (a) a scatter plot is shown that was created by a Monte Carlo simulation of (6.4.1) based on 50 predictors uniformly distributed on the square $[0, 1]^2$; a regression function is 2 plus the Normal by the Normal corner functions (here we add 2 to a regression surface to make all responses positive and therefore visible in the scattergram), $\sigma(x1, x2) = 0.3$, and ε is a standard normal error. The heights of the drawn vertical spikes show values of responses, and their locations in the X1-X2 plane show the predictors.

First of all, look at how sparse these 50 observations are, and compare with the univariate scattergrams with the same sample size in Figure 4.1. At first glance, it is even difficult to believe that you see 50 observations; it looks as though the number is at most several dozen. Also, there are huge empty spaces on the square with no observations at all, and therefore no information about an underlying regression surface for these spots is available. Thus, such a scattergram explains all the complexities of a multivariate setting better than any words or theorems.

Diagram (b) shows how the estimator (6.4.7) performs for this data set. As we see, the estimate nicely resembles the Normal by the Normal surface, and this particular outcome is truly impressive because the sample size is very small for a bivariate problem. On the other hand, other simulations may lead to worse estimates; after all, even for univariate random design regressions the sample size $n = 50$ is very small for a reliable estimation.

The bottom row of diagrams in Figure 6.6 shows a scatter plot and estimate for the case of the underlying regression surface the Strata by the Strata shown in Figure 6.1(c), and here again 2 is added to that corner surface to visualize all the responses. This scattergram again illustrates the complexity of the bivariate setting, and again the sparsity of the data is striking. It is also fair to say that a manual fitting of a surface to the data is extremely complicated, while it is typically not a problem for a univariate setting with 50 observations and coefficient of difficulty 0.09; just recall Figure 4.2.

The nonparametric estimate in diagram (d) correctly shows the number of hills and their separation and location. However, there is no chance to see the correct magnitudes for a sample of this size.

6.5 Additive Regression Model

The classical linear regression model for the case of a d-dimensional predictor (covariate) assumes that $f_L(x_1, \dots, x_d) := \beta_0 + \sum_{k=1}^d \beta_k x_k$ is an underlying regression function. Note that by this assumption a regression function is both linear and additive in the predictors. If we drop the assumption on the linearity and preserve the additivity, then we get an *additive model*,

$$Y = f_A(X_1, \dots, X_d) + \sigma\varepsilon := \beta + \sum_{k=1}^d f_k(X_k) + \sigma\varepsilon. \tag{6.5.1}$$

Here Y is a response that corresponds to d possibly dependent predictors (covariates) $X_1, \dots, X_d$ with a joint d-variate design density $h(x_1, \dots, x_d)$; $f_k(x)$, $k = 1, \dots, d$, are unknown univariate functions; ε is a random variable that is independent of the predictors and has zero mean and unit variance; and σ is a scale parameter.

The problem is to estimate a regression function f_A and its additive univariate components f_k based on n iid realizations $\{(Y_l, X_{1l}, \dots, X_{dl}), l = 1, 2, \dots, n\}$.

Note that the problem of estimating additive univariate functions is interesting on its own merits because these functions show additive contributions of covariates.

To use a series approach, let us additionally assume that a known design density $h(x_1, \dots, x_d)$ is supported on the d-dimensional unit cube $[0, 1]^d$

and that it is bounded away from zero on this cube. Also, to make the additive functions unique, we assume that

$$\int_0^1 f_k(x)dx = 0, \qquad k = 1, 2, \dots, d. \tag{6.5.2}$$

To estimate the univariate components $f_k(x)$, we choose a univariate basis, for instance the cosine basis $\{\varphi_j,\ j = 0, 1, \dots\}$ defined in (2.1.3). Then the partial sums

$$f_{kJ}(x) := \sum_{j=1}^{J} \theta_{kj}\varphi_j(x) \tag{6.5.3}$$

may be used as approximations of $f_k(x)$ for $x \in [0, 1]$. Here θ_{kj} denotes the jth Fourier coefficient of f_k (the kth additive component), that is,

$$\theta_{kj} := \int_0^1 f_k(x)\varphi_j(x)dx. \tag{6.5.4}$$

Note that due to (6.5.2) we have $\theta_{k0} = 0$.

These partial sums lead us to the following approximation of a regression function f_A:

$$f_{AJ}(x_1, \dots, x_d) := \beta + \sum_{k=1}^{d}\sum_{j=1}^{J} \theta_{kj}\varphi_j(x_k). \tag{6.5.5}$$

Now we are in a position to figure out how to estimate these Fourier coefficients. As usual, we try to write them as an expectation of a function of observations and then use a corresponding sample mean estimate.

For the constant term β we write using (6.5.1)–(6.5.2) that

$$E\{Y/h(X_1, \dots, X_d\} = \int_0^1 \cdots \int_0^1 f_A(x_1, \cdots, x_d)dx_1 \cdots dx_d = \beta. \tag{6.5.6}$$

This implies the use of a sample mean estimate

$$\hat{\beta} := n^{-1}\sum_{l=1}^{n} Y_l/h(X_{1l}, \dots, X_{dl}). \tag{6.5.7}$$

Now let us discuss a possible estimation of θ_{kj}. Write for $j \ge 1$,

$$\theta_{kj} = \int_0^1 f_k(x_k)\varphi_j(x_k)dx_k = \sum_{s=1}^{d}\int_0^1 f_s(x_s)\varphi_j(x_k)dx_k. \tag{6.5.8}$$

In (6.5.8) the second equality holds because for any $j \ge 1$ and $s \ne k$,

$$\int_0^1 f_s(x_s)\varphi_j(x_k)dx_k = f_s(x_s)\int_0^1 \varphi_j(x_k)dx_k = 0.$$

Continuing (6.5.8) we obtain for $j \ge 1$,

$$\begin{aligned}\theta_{kj} &= \int_0^1 \sum_{s=1}^{d} (\beta + f_s(x_s))\varphi_j(x_k)dx_k = \int_0^1 f_A(x_1,\dots,x_d)\varphi_j(x_k)dx_k \\ &= \int_0^1 \cdots \int_0^1 f_A(x_1,\dots,x_d)\varphi_j(x_k)dx_1\cdots dx_d \\ &= E\big\{\big(f_A(X_1,\dots,X_d)+\sigma\varepsilon\big)\varphi_j(X_k)/h(X_1,\dots,X_d)\big\} \\ &= E\{Y\varphi_j(X_k)/h(X_1,\dots,X_d)\}. \end{aligned} \tag{6.5.9}$$

Thus, each Fourier coefficient is written as an expectation. (Note that covariates may be dependent.) Then a natural sample mean estimate is

$$\hat\theta_{kj} := n^{-1}\sum_{l=1}^{n} Y_l\varphi_j(X_{kl})/h(X_{1l},\dots,X_{dl}). \tag{6.5.10}$$

Set J^* to be the rounded-up $c_{JM}(c_{J0}+c_{J1}\ln(n))$. Then, a hard-threshold estimator of the kth additive component is

$$\hat f_k(x) := \sum_{j=1}^{J^*} I_{\{\hat\theta_{kj}^2 > c_T\hat\sigma^2\ln(n)n^{-1}\}}\hat\theta_{kj}\varphi_j(x), \tag{6.5.11}$$

where $\hat\sigma^2$ is a sample variance of responses. The steps for calculating a universal estimator are left as Exercise 6.5.8.

A corresponding data-driven series estimate for f_A is

$$\hat f_A(x_1,\dots,x_d) := \hat\beta + \sum_{k=1}^{d}\hat f_k(x_k). \tag{6.5.12}$$

In the case of an unknown design density h, its estimate is plugged in. Recall a similar approach and the discussion in Section 6.4.

Figure 6.7 shows how the series estimator (6.5.11) recovers additive components. Here $d = 4$, and the underlying additive functions are the Uniform, the Normal, the Strata, and the Monotone (recall that 1 is always subtracted from these functions to satisfy (6.5.2)), and the predictors are iid according to the Uniform distribution on $[0, 1]$. The sample size is $n = 500$, the noise ε is standard normal, $\sigma = 0.2$ and $\beta = 1$.

The particular estimates are reasonably good, keeping in mind the high dimensionality. On the other hand, the sample size $n = 500$ is a moderate one. Repeated simulations show that essentially smaller sample sizes do not allow a stable recovery of these additive components. Thus, while the problem is not so difficult as a multivariate 4-dimensional one, the case of 4 components does take its toll in terms of the necessity to use larger sample sizes than ones used for univariate and bivariate settings.

Recall that no assumption about independence of covariates has been made to obtain the formula (6.5.9). The following Monte Carlo experiments show that a possible dependence of predictors is indeed not an issue.

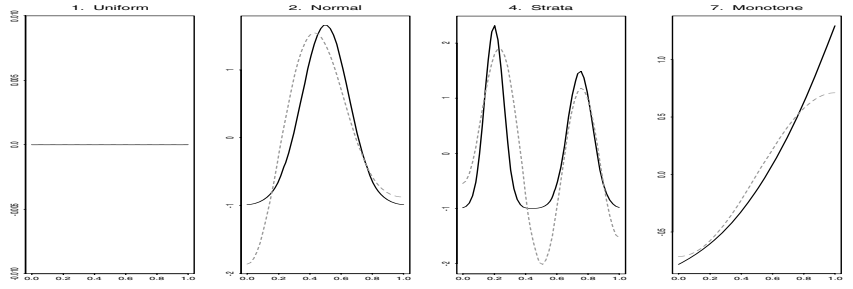

FIGURE 6.7. Data-driven estimates of 4 additive components for the case of the additive model (6.5.1) with iid uniform predictors and normal errors. The underlying components are shown by solid lines and the estimates by dashed lines. Here $n = 500$, $\sigma = 0.2$, and $\beta = 1$. {The set of underlying additive functions is chosen by the argument *set.k*, whose cardinality is *d*.} [*set.k =c(1,2,4,7), n = 500, sigma=.2, cJ0=4, cJ1=.4, cJM=.5, cT=2*]

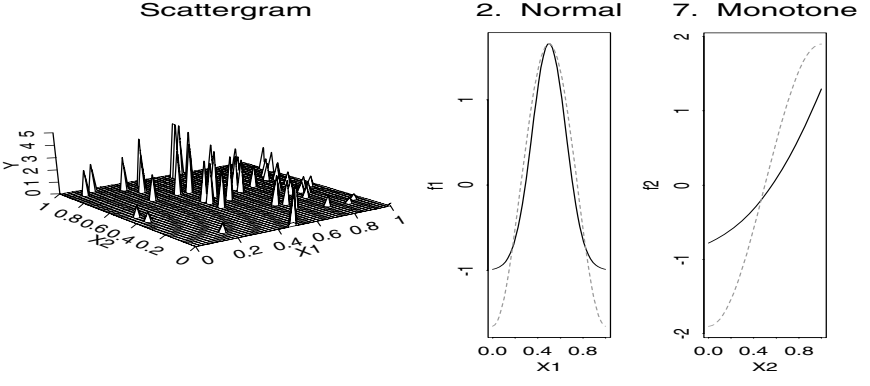

FIGURE 6.8. Scatter plot of 50 observations and estimates (dashed lines) of 2 additive components for the case of dependent covariates with unknown design density. The underlying components are shown by solid lines. The noise is normal and $\sigma = 0.2$. {The choice of underlying additive components is controlled by the argument *set.k*.} [*set.k =c(2,7), n = 50, sigma=.2, cJ0=4, cJ1=.5, cJM=.5, cT=2, cB=2, cD=1*]

Consider Figure 6.8. The left diagram shows a scattergram obtained by a Monte Carlo simulation of (6.5.1) with f_1 and f_2 being the Normal and the Monotone, ε being standard normal, $\sigma = 0.2$, and $\beta = 2$. The first predictor is generated by a random variable $X1$, which is distributed according to the Monotone density. The second predictor is generated by a random variable $X2$, which is distributed uniformly if $X_1 < 0.5$ and it is distributed according to the Angle density otherwise.

Can you realize the underlying additive components from this scattergram? It is apparently not a simple puzzle, so let us see how our estimator

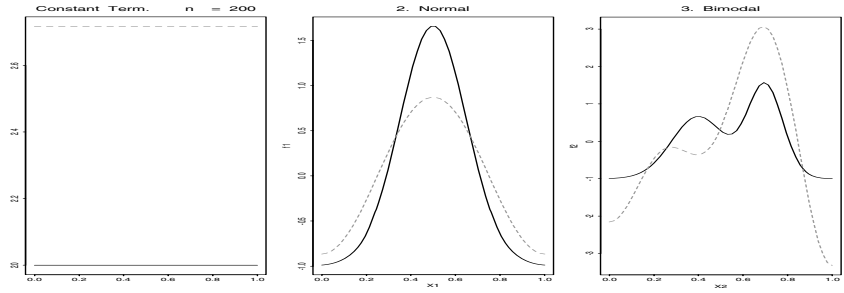

FIGURE 6.9. Estimates of a constant term β and two additive components for the case of dependent covariates with unknown design density. The sample size $n = 200$ is shown in the title of the left diagram. Estimates are shown by dashed lines and underlying components by solid lines. [*set.k=c(2,3), n = 200, sigma=.2 cJ0=4, cJ1=.5, cJM=.5, cT=2, cB=2, cD=1*]

solves it. Estimates of the components are shown in the middle and right diagrams. These particular estimates are very good even in comparison with the univariate regression cases shown in Figure 4.3; see the case $n = 50$. But note that here the standard deviation of errors is five times less.

Figure 6.9 exhibits a similar experiment, only here instead of the scattergram an estimate of the constant term $\beta = 2$ is shown as a horizontal line whose y-intercept is $\hat{\beta}$. (This figure may be used for larger sample sizes where a scattergram is too "overcrowded" by spikes.) We see that several hundred observations allow one to recognize the shape of the Bimodal.

We conclude that if an underlying regression model is additive, then its components can be fairly well estimated even for the case of relatively small (with respect to the corresponding general multivariate setting) sample sizes. Also, even if an underlying regression function is not additive, then such a model may shed some light on the data at hand. This explains why the additive model may be used as a first look at the data.

6.6 Case Study: Conditional Density

We discussed in Section 4.5 that in many cases a regression function f may be defined as the conditional expectation, i.e., $f(x) := E\{Y|X = x\}$. The notion of conditional expectation was introduced in Appendix A; see (A.13) and (A.20) for discrete and continuous random variables, respectively. Here we shall discuss the case of continuous random variables; thus recall that the *conditional expectation* of Y, given that $X = x$, is

$$E\{Y|X = x\} := \int_{-\infty}^{\infty} y f^{Y|X}(y|x)dy. \tag{6.6.1}$$

Here $f^{Y|X}(y|x)$ is the *conditional density* of Y, given that $X = x$; if the marginal density $f^X(x) := \int_{-\infty}^{\infty} f^{XY}(x,y)dy$ is positive, then

$$f^{Y|X}(y|x) := f^{XY}(x,y)/f^X(x). \tag{6.6.2}$$

The relation (6.6.1) implies that the regression function $f(x)$ is a functional of the conditional density, and thus the conditional density may reveal more about the relationship between X and Y than the regression function. Also, recall that if the conditional density is either multimodal or highly skewed or has heavy tails (like a Cauchy density discussed in Section 4.6), then the regression function (6.6.1) is no longer a reasonable characterization of the relationship between X and Y.

In short, a conditional density may give a more detailed and correct picture of the relationship between X and Y. Also, note that apart from the case of a uniform X, a joint density $f^{XY}(x,y)$ may not resemble the corresponding conditional density (6.6.2).

If $f^X(x)$ is positive over a set where a conditional density $f^{Y|X}(y|x)$ should be estimated, then the ratio $\hat{f}^{XY}(x,y)/\hat{f}(x)$ of the estimates discussed in Sections 6.2 and 3.1 may be used. This is a reasonable approach if it is also given that $f^X(x)$ is bounded from below by a known positive number and the sample size is relatively large. Otherwise, all the problems discussed in the previous sections arise, and in that ratio a relatively small denominator may "spoil" this plug-in estimate.

Thus, let us consider our "universal" approach based on the analysis of Fourier coefficients of an estimated function.

Consider the problem of a series approximation of a conditional density over a unit square $[0,1]^2$. Then a partial sum (6.1.3) with the cosine tensor-product basis (6.1.2) may be used, where the Fourier coefficients are

$$\begin{aligned}\theta_{j_1 j_2} &= \int_0^1 \int_0^1 f^{Y|X}(y|x)\varphi_{j_1 j_2}(x,y)dx\,dy \\ &= \int_0^1 \int_0^1 \big(f^{YX}(y,x)/f^X(x)\big)\varphi_{j_1 j_2}(x,y)dx\,dy \\ &= E\{I_{\{(X,Y)\in[0,1]^2\}}\varphi_{j_1 j_2}(X,Y)/f^X(X)\}. \end{aligned}\tag{6.6.3}$$

Thus, the Fourier coefficients of a conditional density may be written as expectations of random variables $I_{\{(X,Y)\in[0,1]^2\}}\varphi_{j_1 j_2}(X,Y)/f^X(X)$, and they may be estimated by a sample mean estimate.

The only issue here is that again the marginal density $f^X(x)$ is unknown. On the other hand, a sample mean estimate uses this density only at values equal to observations, and this makes the situation much simpler. For instance, the universal estimate may be plugged in. Another approach was discussed in Section 4.3, which was to estimate $f^X(X_{(l)})$ via a normed spacing $n\hat{D}_{0ls}$. Recall that $X_{(l)}$ are ordered predictors, $\hat{D}_{0ls} = (2s)^{-1}(X_{(l+s)} - X_{(l-s)})$, and s is the rounded-up $s_0 + s_1 \ln(\ln(n+20))$,

$s_0 = s_1 = 0.5$. This implies the following estimate of the Fourier coefficients:

$$\hat{\theta}_{j_1 j_2} = \sum_{l=1}^{n} I_{\{(X_{(l)}, Y_{(l)}) \in [0,1]^2\}} \hat{D}_{0ls} \varphi_{j_1 j_2}(X_{(l)}, Y_{(l)}) \,. \tag{6.6.4}$$

Then either a universal or hard-threshold series bivariate estimator may be used for estimation of $f^{Y|X}$.

In Figure 6.10 the hard-threshold estimator is studied. The left column of diagrams shows two scatter plots with the same predictors and underlying "regression functions" but different noises. The top one is a classical case of a zero-mean normal noise whose variance depends on X. The right top diagram shows the corresponding estimated conditional density over the unit square. Let us explore what this estimate "tells" us about the data set. First, we see that the shown ridge is approximately linear in X with a positive slope. This is consistent with the analysis of the scatter plot. For smaller X's the estimate also reveals that the noise is unimodal (just look at any slice of this surface given X). We cannot draw this conclusion for the largest values of X because only part of the conditional density is shown, but this part perfectly resembles a normal density with the mean about 0.9 and the standard deviation about 0.1. Note that for the larger X's the crest is clearly observed. Also, we see that the height of the ridge depends on X.

Now let us return to the top scattergram. First of all, we see that only a few observations have values of X less than 0.15 and larger than 0.85. Thus, there is no way to use any local method of estimation of the conditional density over approximately 30% of the square. Also, note that any vertical strip of width, say, 0.05 contains an insufficient number of points for estimating $f^{Y|X}(y|x)$ as a univariate density. Moreover, can you say that points located near any vertical slice resemble normally distributed points? And on top of this, look at the main body of points with $X \in [0.4, 0.8]$. If you ignore points beyond this strip, then it is even difficult to be sure that the body of points slopes upward.

Thus, keeping in mind that the estimator shows us an underlying conditional density over the unit square, the estimate depicts a classical linear relationship between predictor and response over this area. But is the estimator correct? To answer this question, let us explain how the data set was generated. The regression curve is a linear one, $0.4 + 0.5X$. The noise is normal with zero mean and standard deviation $0.2 - 0.15X$. Thus, we see that the estimate shows a correct positive trend in the data. For $X = 0$ the estimate is a bit skewed, and its mode is at $Y = 0.55$ instead of the correct $Y = 0.4$. But can an estimate do better? The perfect shape of a normal density with mean 0.4 and standard deviation 0.2 implies a domain with range about 1.2, but there are no data to indicate this.

For $X = 1$ the underlying conditional density is normal with mean 0.9 and standard deviation 0.05. By analyzing the corresponding vertical slice

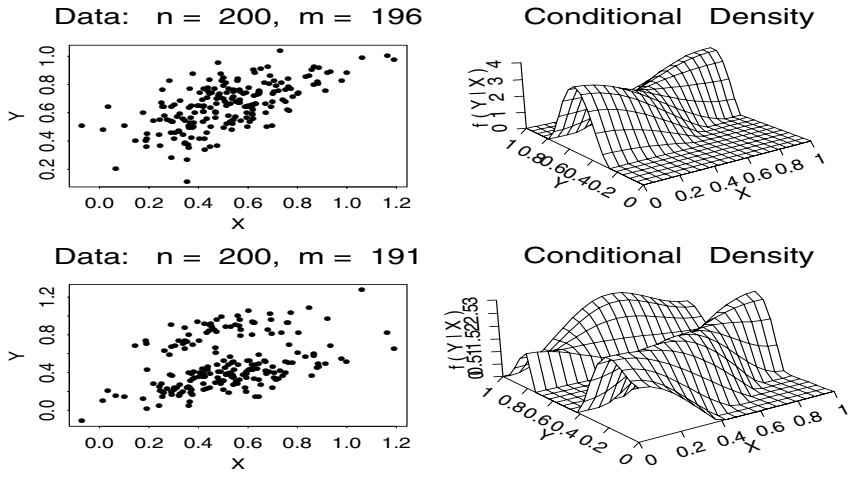

FIGURE 6.10. Estimation of a conditional density $f^{Y|X}(y|x)$ for two different data sets. {In both scatter plots the X's are the same iid realizations of a normal random variable $N(mX, sdX^2)$ with $mX = 0.5$ and $sdX = 0.2$. In both examples the Y's are generated according to a linear regression model $Y = m_1 + m_2X + \varepsilon$, $m_1 = 0.4$, $m_2 = 0.5$, with specific noise terms for the top and bottom diagrams. For the top diagram the noise is normal with mean zero and standard deviation $sd_1 + sd_2X$, $sd_1 = 0.2$, and $sd_2 = -0.15$. For the bottom diagram the noise is more complicated. It is a mixture $\varepsilon = (1-\theta)\xi_1 + \theta\xi_2$ where ξ_1 and ξ_2 are independent normal $N(mY_1, sdY_1^2)$ and $N(mY_2, sdY_2^2)$, $mY_1 = -0.3$, $sdY_1 = 0.1$, $mY_2 = 0.2$, and $sdY_2 = 0.1$. The Bernoulli random variable θ takes on values 1 and 0 with the probabilities pr and $1-pr$, $pr = 0.3$. The sample size $n = 200$ is shown in the titles of the left diagrams, where also the corresponding numbers m of observations within the unit square $[0,1]^2$ are indicated.} [*n = 200, m1=.4, m2=.5, sd1=.2, sd2=−.15, mX=.5, sdX=.2, mY1=−.3, sdY1=.1,mY2=.2, sdY2=.1, pr = .3, s0=.5, s1=.5, cJ0=4, cJ1=.5, cJM=.5, cT=4, cB=2*]

of the estimate, we conclude that the mean is shown almost correctly and the shape is also almost perfect. Thus, if the estimate were twice as wide for smaller X, then it could be declared as a perfect one.

Now let us look at the left bottom scatter plot. If we ignore points to the right of $X = 1$ and to the left of $X = 0.15$ (overall 9 points from 200), then no pronounced trend (slope) in the data may be visualized. A closer look at the data reveals that points are denser in the bottom part of the scattergram, but this also may be an illusion. You can easily imagine something similar in the top scattergram.

The estimate in the right bottom diagram tells us a lot about the data and relationship between X and Y. We see a canyon created by two ridges sloped upward. The valley between the ridges is pronounced, and now we can also recognize it in the left bottom scattergram. Also, vertical (with constant X) slices of the estimate reveal that the ridges are unimodal.

Now it is time to "reveal" the underlying relationship between X and Y. This is a model $Y = 0.4 + 0.5X + (1 - \theta)\xi_1 + \theta\xi_2$, where θ is a Bernoulli random variable that takes on the value 1 with probability 0.3 and 0 with probability 0.7. The random variables ξ_1 and ξ_2 are independent normal with means -0.3 and 0.2 and equal standard deviations 0.1. Thus, the underlying conditional density indeed has two ridges of different heights that slope upward.

Note that for the bottom example the line $0.4 + 0.5X$ cannot be called a regression function because $E\{Y|X = x\} = 0.25 + 0.5x$. Also, even this expectation does not describe the relationship between X and Y because it is much more involved. On the other hand, the estimated conditional density sheds light on and explains this complicated relationship.

6.7 Practical Seminar

The main objective of this seminar is to apply the density estimators of Section 6.2 to real data sets and then analyze their performance. Then a spatial data set, analyzed by the regression estimator of Section 6.4, is considered.

We begin with a bivariate density estimation. The particular data set is a pair of random variables from the data file **state.x77**. This data set is a matrix whose columns contain various statistics about the 50 states of the United States. The first variable $X1$ is "Income," and this is per capita income in 1974. The second variable $X2$ is "Illiteracy" in 1970 given as a percentage.

The first diagram in Figure 6.11 allows us to look at the data (pairs of observations are shown as a scattergram, and the sample size is given in the title). We see that the data are sparse and heavily concentrated near low Illiteracy and moderate Income. Also, there is a clear tendency for a state with smaller Income to have a higher Illiteracy. The state with the largest Income, \$6315, is Alaska, which has a rather moderate Illiteracy of 1.5. This state is a clear outlier, since there is no other state with Income larger \$5500. Another extreme state is Louisiana, which has the highest Illiteracy, 2.8, combined with one of the smallest Incomes, but in no way is this state an outlier because you can see several other states nearby. On the other hand, the top right corner is empty, that is, there are no states with large Incomes and high Illiteracy. Also, there is a smaller empty spot in the bottom left corner, which tells us that the smallest Incomes imply high Illiteracy.

Now we would like to get a bivariate density (surface) that reflects all the observations. Here again, as in the previous seminars, we simultaneously look at a spectrum of estimates with different values of a particular "running" argument. As an example, we check the influence of the coeffi-

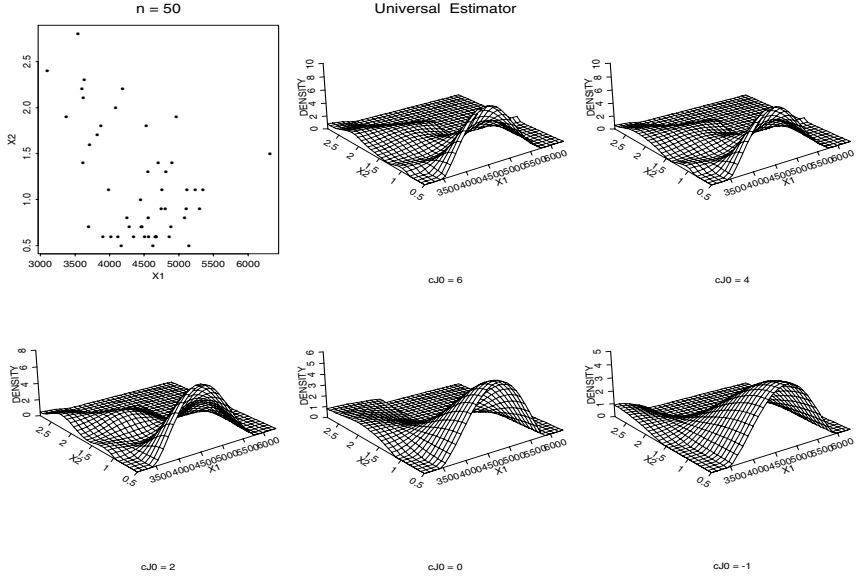

FIGURE 6.11. Data set of Income ($X1$) and Illiteracy ($X2$) and the sequence (spectrum) of universal bivariate density estimates (multiplied by the product of ranges of the two variables) with different values of coefficients c_{J0} shown in the subtitles. The sample size $n = 50$ is shown in the title. {Figure 6.3 may be used as a prototype. To analyze this particular data set, we set $DATA = state.x77[, c(2, 3)]$, that is, we use an $n \times 2$ matrix where the first column is "Income" and the second one is "Illiteracy." Any other data set should have the same format of an $n \times 2$ matrix. To get more information about this data set, call > **help(state.x77)**. The running coefficient is chosen by *arg*, and its values by *set.arg*. The default estimator is universal; the hard-threshold estimator may be used by setting *estimate= "h".*} [*DATA=state.x77[,c(2,3)], estimate= "u", arg= "cJ0", set.arg=c(6,4,2,0,-1), cJ0=4, cJ1=.5, cJM=2, cT=4, cB=1*]

cient c_{J0}. The values are shown in the subtitles. Also, recall that both the universal and hard-threshold estimates may be used, and the fact that the universal estimator was used is highlighted by the title.

As we see, the values $c_{J0} = 6$ and the default $c_{J0} = 4$ imply approximately the same estimates with the tiny difference between the left corners. The estimates correctly show us: the main areas of concentration of the points; that the density is positive at and near the corner of smallest Incomes ($X1$) and high Illiteracy ($X2$) due to Louisiana and several other "neighbors"; that Alaska made its presence apparent by a small ridge that runs along the line $X2 = 1.5$ for largest Incomes $X1 > \$6000$ (you can see it behind the largest hill); that the density vanishes at the "correct" corners, and even boundaries of the empty spaces are shown almost perfectly. The only caveat is the small hill near the Illiteracy $X2 = 1.2$ and small

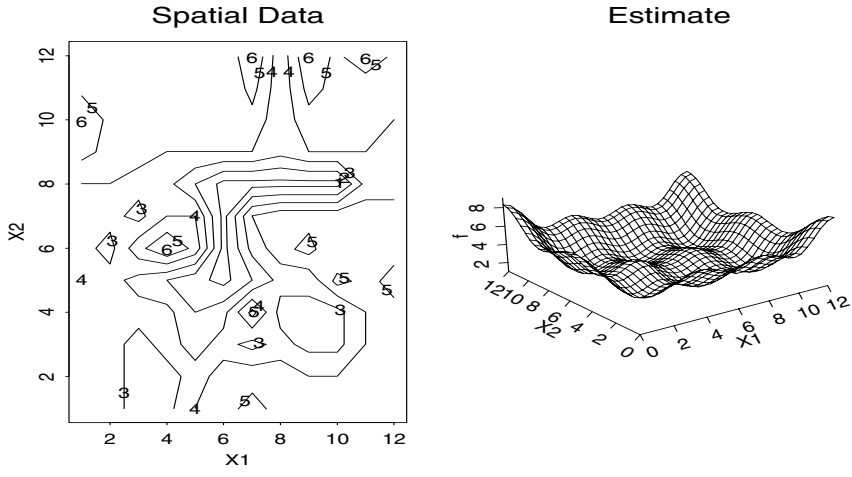

FIGURE 6.12. Contour plot of a rescaled spatial data set "switzerland" and the corresponding regression estimate. {Any spatial data (a matrix *DATAS*) defined on a regular grid may be analyzed. Otherwise a data set should be given by vectors *X1, X2*, and Y. Arguments of the estimate are the same as in Figure 6.6.} [*DATAS=switzerland, X1=NA, X2=NA, Y=NA, cJ0=4, cJ1=.5, cJT=2, cJM=2, cB=2, cD=1*]

Incomes; it is doubtful and may be explained by the nature of the series approximation. Overall, for just 50 observations, the result is fabulous.

If the number of Fourier coefficients is reduced by using a smaller c_{J0}, then this changes the estimate. As should be expected, the estimate becomes smoother and fewer fine details are shown. The case of $c_{J0} = 2$ is still pretty good. We do not see here the ridge showing Alaska, but on the other hand the estimate for smallest Incomes is more realistic. The two last estimates with the smallest c_{J0} also show us the main features, but look at the increased domain of the main hill. For instance, for $c_{J0} = -1$ the estimate vanishes only for Incomes ($X1$) larger than \$6000, while the estimate with $c_{J0} = 4$ vanishes for incomes larger than \$5500.

Figure 6.12 is devoted to analyzing spatial data sets where one of the main issues is interpolation and smoothing. The considered data set "switzerland" is measurements of topological heights of Switzerland on 12 by 12 grid. Accuracy of the data is questionable, that is, the data set is probably contaminated by errors. Thus using the estimator of Section 6.4 is quite appropriate. The left diagram shows the contour plot of the rescaled data, and the right one shows the corresponding estimate. The estimate correctly shows the main inner peak with coordinates ($X1 = 4$, $X2 = 6$) but heights of some peripheral peaks are probably exaggerated.

6.8 Exercises

6.1.1 Repeat Figure 6.1 with different corner functions. Try to "read" and compare the three different methods of presentation of surfaces using a color monitor (do not make hard copies, since it takes too much time). Which method do you prefer and why?
6.1.2 Choose 4 different triplets of bivariate functions and repeat Figure 6.2 for them. What cutoffs are sufficient for a fair visualization of these bivariate functions? Where do you prefer to use a cosine approximation and where a polynomial one?
6.1.3 Verify (6.1.6).
6.1.4 Verify (6.1.7).
6.1.5 Consider the case of a d-variate function that is approximated by a partial sum with cutoffs $J_1, \dots, J_d$. As in the bivariate case, it is possible to show that the MISE of a series estimate is proportional to $n^{-1} \prod_{s=1}^{d} J_s + \sum_{s=1}^{d} J_s^{-2\beta_s}$. Find optimal cutoffs and the corresponding MISE for this d-variate case.
6.2.1 Show that the estimate (6.2.2) satisfies the relation $E\{\hat{\theta}_{j_1 j_2}\} = \theta_{j_1 j_2}$. Then, how do we refer to such an estimate?
6.2.2 Verify (6.2.3).
6.2.3 Suggest an extension of the bivariate estimate (6.2.6) to the d-variate case.
6.2.4 Repeat Figure 6.3 with different values of coefficients of the estimator. Answer the following questions: (a) What are the most complicated and the simplest corner densities for estimation? (b) Choose 3 particular corner densities and find minimal sample sizes that give a "stable" visualization of the shape of these particular densities.
6.2.5 Draw a sketch of the underlying density used in Figure 6.4. Then repeat Figure 6.4 with different sample sizes and draw a conclusion about a minimal sample size that gives a "stable" visualization of this density.
6.3.1 Let P_1 be the probability of a wrong decision given that z^d is from a population 1, and let P_2 be the probability of a wrong decision given that z^d is from a population 2. Show that the rule (6.3.1) minimizes the total probability $P_1 + P_2$.
6.3.2 Using the notation of Problem 6.3.1, show that the rule (6.3.2) minimizes the Bayes error $pP_1 + (1-p)P_2$, where p is the probability that an observation is from the population 1.
6.3.3 Verify (6.3.3).
6.3.4 Repeat Figure 6.5 with different arguments n, q, t, and *estimate*. Then answer the following questions: (a) What are the minimal sample sizes of training sets for which the learning machine "reliably" mimics the ideal discrimination rule? (b) How does the parameter q affect the discriminant analysis? (c) What is the role of the parameter t? (d) Which density estimator would you recommend to use?

6.4.1 Explain all 4 steps in establishing (6.4.4).
6.4.2 Is the estimate (6.4.5) unbiased for the parameter $\theta_{j_1 j_2}$? Then suppose that the error ε depends on predictors. Does this change the answer?
6.4.3 Explain the motivation of the hard-threshold rule used in (6.4.7).
6.4.4 Repeat Figure 6.6 with different arguments, and then answer the following questions: (a) What corner functions may be "reliably" estimated based on samples of size 50, 100, 200 given $\sigma = 0.1$? (b) Consider $\sigma = 1$ and answer question (a). (c) What is the most complicated regression function for visualizing its shape among the corner ones for the case of sample size $n = 200$? (d) Choose 4 different underlying regression surfaces and find for them minimal sample sizes that imply a reliable estimation.
6.5.1 Where is the assumption (6.5.2) used?
6.5.2 Verify (6.5.6).
6.5.3 Does the relation (6.5.6) hold if the error ε depends on the predictors?
6.5.4 Verify (6.5.8).
6.5.5 Verify and then explain all the steps in obtaining (6.5.9). Also explain why the assumption $j \ge 1$ is important.
6.5.6 Explain why the suggested procedure of a data-driven estimation does not require the independence of covariates.
6.5.7 Repeat Figure 6.7 with different arguments including different numbers of additive components. Then answer the following questions. (a) For a chosen set of additive components, what is a minimal sample size that allows a "reliable" visualization of all the additive functions? (b) Consider the case of 3 additive functions and the sample size 100. Among the corner functions, what are the simplest and most difficult triplets for estimation?
6.5.8 Write down and then explain all the steps of a universal estimate. Hint: Use Sections 4.1, 4.2, and 6.2.
6.5.9 Repeat Figure 6.8 with different additive components. Are there components that may be realized from the scattergrams? What sample size is optimal for such a visualization?
6.5.10 Would you recommend any changes in the values of coefficients of the estimator used by Figure 6.8?
6.6.1 Explain the notion of a joint, conditional, and marginal density.
6.6.2 Let X and Y be independent. What is $f^{Y|X}$?
6.6.3 Let $Y = X + (1 + X^2)Z$ where Z is a standard normal. Find $f^{Y|X}$ and describe it.
6.6.4 Explain (6.6.4).
6.6.5 Write down a universal estimate for a conditional density.
6.6.6 Choose any two coefficients used in Figure 6.10 and then analyze their effect on the estimates.
6.6.7 Find boundary values of parameters of the noise terms which make the estimation impossible. How do these values depend on n?

6.7.1 Explore the influence of all the other coefficients of the universal estimate. Hint: Changing default arguments may be beneficial for such an analysis.
6.7.2 Consider the hard-threshold estimate (set *estimate* = $''h''$) and explore it using Figure 6.11.
6.7.3 Choose any other data set (for instance, the other pair of observations from **state.x77** or a pair from **air**) and analyze the estimates.
6.7.4 Try to find optimal coefficients of the estimate used in Figure 6.12.
6.7.5 Consider another spatial data set and analyze an estimate.

6.9 Notes

Although generalization of most of the univariate series estimators to multivariate series estimators appears to be feasible, we have seen that serious problems arise due to the *curse of multidimensionality*, as it was termed by Bellman (1961). The curse is discussed in the books by Hastie and Tibshirani (1990), Scott (1992), and Silverman (1986). Many approaches have been suggested aimed at a simplification and overcoming the curse: additive and partially linear modeling, principal components analysis, projection pursuit regression, classification and regression trees (CART), multivariate adaptive regression splines, etc. Many of these methods are supported by S–PLUS and briefly discussed in the book by Venables and Rippley (1997), where further references may be found.

6.1 Approximation theory is discussed in the books by Nikolskii (1975), Temlyakov (1993), and Lorentz, Golitschek, and Makovoz (1996). Donoho (1997) discusses the case of anisotropic smoothness.

6.2 A book-length discussion of multivariate density estimates (with a particular emphasis on kernel estimators) is given by Scott (1992). The asymptotic justification of the series approach is given in Efromovich (1994b), where spherical data are considered as an example.

6.3 The formulation of the problem and the terminology are due to Fisher (1936, 1952). The problem is a particular case of a more general theory of pattern recognition, see the books by Rippley (1996) and Vapnik (1995).

6.4 A review of nonseries estimators may be found in the book by Fan and Gijbels (1996, Chapter 7). The asymptotic justification of the series approach is given in Efromovich (1994b).

6.5 A book-length treatment of additive models may be found in Hastie and Tibshirani (1990). Hart (1997, Section 9.4) discusses an additivity test that checks the correctness of additive models.

6.6 A kernel estimator of a conditional density is discussed in the book by Fan and Gijbels (1996, Section 6.2.3).

7

Filtering and Asymptotics

This chapter is primarily devoted to a discussion of asymptotics when the size of a sample tends to infinity. Nowadays it is customary to study the asymptotics via a filtering model thanks to the equivalence principle, which basically says that under certain conditions an asymptotic result proved for a filtering model also holds for corresponding density, regression, and spectral density models.

Thus, we begin this chapter with an introduction to a filtering model, which is also interesting on its own merits and is of primary interest in many applications including communication systems and econometrics. Both the cases of large noise (which is equivalent to small sample sizes) and small noise (which is equivalent to large sample sizes) are discussed in detail, and all the main asymptotic results are proved. As a result, sifting through some mathematics, especially in Sections 7.1, 7.3 and 7.4, is required. This effort will be rewarded by understanding methods for finding asymptotically sharp and rate optimal estimates for functions and their derivatives. Section 7.4 discusses methods of adaptive estimation. The multivariate case is considered in Section 7.5. All other sections are devoted to special topics.

7.1 Recovery of a Signal Passed Through Parallel Gaussian Channels

In this section we discuss the ideas and results of asymptotic nonparametric curve estimation theory using the model of the recovery of a signal trans-

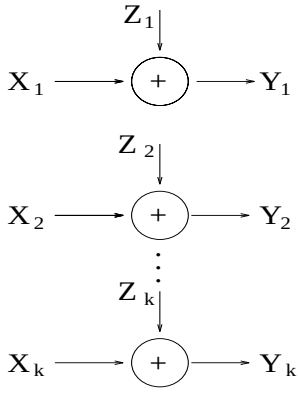

FIGURE 7.1. Parallel Gaussian channels.

mitted via k independent continuous parallel Gaussian channels depicted in Figure 7.1. The output Y_j of the jth channel is equal to the sum of the input X_j and the Gaussian (normal) noise Z_j,

$$Y_j := X_j + Z_j, \quad j = 1, 2, \ldots, k. \tag{7.1.1}$$

Here $Z_1, \ldots, Z_k$ are independent identically distributed (iid) realizations of a normal random variable Z with zero mean and variance σ^2. We shall refer to Z_j as noise. The noise is assumed to be independent of the input signal. Note that in many channels, including radio and satellite links, additive noise may be due to a variety of causes. Also recall that by the central limit theorem (see Appendix A) the cumulative effect of a large number of small random effects should be approximately normal. Thus the Gaussian assumption is reasonable in a large number of situations, and this explains why the model of parallel Gaussian channels is a key example in information and communication theories.

The relation of this communication system to a curve estimation setting is as follows. Assume that a continuous-in-time t signal $f(t)$, $0 \le t \le 1$, can be written as a partial sum

$$f(t) := \sum_{j=1}^{k} X_j g_j(t), \quad 0 \le t \le 1. \tag{7.1.2}$$

Here $\{g_j(t),\ j = 1, 2, \ldots, k\}$ is a set of k finite functions from $L_2([0,1])$, that is, $|g_j(t)| < \infty$ and $\int_0^1 g_j^2(t)dt < \infty$. Recall that we use the notation $g^a(t) := (g(t))^a$. We shall refer to $g_j(t)$, $j = 1, \ldots, k$, as *coding functions* because they allow one to code a continuous-in-time function (7.1.2) by the k-dimensional vector $(X_1, \ldots, X_k)$. Then the communication system in Figure 7.1 allows one to transmit such a function (signal). It is always assumed that the set $\{g_j(t)\}$ is known at the receiving end of this communication system. Then the problem is to recover the input signal f or its sth derivative $f^{(s)}$ based on k noisy observations $Y_1, \ldots, Y_k$ and the coding

functions $\{g_j(t)\}$. Define $f^{(0)}(t) := f(t)$, and then we can always refer to this problem as estimation of the sth derivative $f^{(s)}$, where s is a natural number, that is, $s = 0, 1, \ldots$.

Probably the most familiar practical example of the expansion (7.1.2) is the case where $f(t)$ is band-limited in the frequency domain and $g_j(x) = \varphi_{j-1}(x)$ are elements of the classical trigonometric basis defined in (2.4.1). In this case each channel represents a different frequency and phase. A similar example is the cosine basis, in which case each channel represents a different frequency. A more modern example is the case of a wavelet basis, where each channel corresponds to a specific scale–location.

Two familiar criteria to measure the quality of the restoration of $f^{(s)}$ by an estimate $\hat{f}_s$ are as follows. The first one is global, where the recovery of $f^{(s)}(t)$ for all $t \in [0, 1]$ is taken into account. Here we consider the *mean integrated squared error* (MISE) defined as

$$\mathrm{MISE}(\hat{f}_s, f^{(s)}) := E\Big\{ \int_0^1 (\hat{f}_s(t) - f^{(s)}(t))^2 dt \Big\}. \tag{7.1.3}$$

The second criterion is pointwise, where the recovery of $f(t)$ only at a point $t_0 \in [0, 1]$ is of interest. Here we consider the *mean squared error* (MSE) defined as

$$\mathrm{MSE}(\hat{f}_s(t_0), f^{(s)}(t_0)) := E\{(\hat{f}_s(t_0) - f^{(s)}(t_0))^2\}. \tag{7.1.4}$$

It is apparent that in general both these risks should decrease as the noise level becomes smaller. The primary aim of this section is to explore how fast they may decrease. We shall see that the rate of convergence of these risks dramatically depends on the smoothness of estimated functions and, of course, s. In this section we shall study two familiar function classes discussed in detail in Chapter 2: analytic and Lipschitz.

We begin with the case of analytic functions defined via their Fourier coefficients $\theta_j = \int_0^1 f(t)\varphi_j(t)dt$. Here $\{\varphi_j(t),\ j = 0, 1, \ldots\}$ is the classical sine–cosine trigonometric basis defined in (2.4.1). An analytic function space is defined as

$$A_{\gamma,Q} := \{f :\ |\theta_0| \le Q, |\theta_{2j-l}| \le Qe^{-\gamma j},\ l = 0, 1,\ \ j = 1, 2, \ldots\}. \tag{7.1.5}$$

This class has been introduced and discussed in Section 2.4; see (2.4.20). Here let us just briefly recall that there are plenty of interesting examples of such functions. For instance, if a circular random variable is contaminated by a normal error, the noisy random variable has an analytic density. Another example is the spectral density of a causal ARMA process. Also recall that these functions are extremely smooth and infinitely differentiable.

Below we formulate the main results about estimation of analytic functions. Recall that the notions of a supremum, infimum, and minimax risk are explained in Appendix A, and the minimax approach is also discussed below in Remark 7.1.1; $o_\sigma(1)$ denotes (in general different) sequences that tend to 0 as $\sigma \to 0$, and $\lfloor x \rfloor$ denotes the rounded-down x.

Theorem 7.1.1 *Let a signal $f(t) = \sum_{j=0}^{\infty} \theta_j \varphi_j(t)$, $0 \le t \le 1$, be transmitted by setting $X_j := \theta_{j-1}$, $j = 1, 2, \dots, k$, via the parallel Gaussian channels (7.1.1) where Z_j are iid normal $N(0, \sigma^2)$, $0 < \sigma < 1$. Then, for any $s = 0, 1, \dots$ and $t_0 \in [0, 1]$, and regardless of how large the number k of parallel channels is, the following lower bounds for the minimax risks hold:*

$$\inf_{\tilde{f}_s} \sup_{f \in A_{\gamma,Q}} \mathrm{MSE}(\tilde{f}_s(t_0), f^{(s)}(t_0)) \ge P_{s,\gamma} (\ln(\sigma^{-1}))^{2s+1} \sigma^2 (1 + o_\sigma(1)), \quad (7.1.6)$$

$$\inf_{\tilde{f}_s} \sup_{f \in A_{\gamma,Q}} \mathrm{MISE}(\tilde{f}_s, f^{(s)}) \ge P_{s,\gamma} (\ln(\sigma^{-1}))^{2s+1} \sigma^2 (1 + o_\sigma(1)), \quad (7.1.7)$$

where

$$P_{s,\gamma} := 2(2\pi)^{2s} (2s+1)^{-1} \gamma^{-2s-1}, \quad (7.1.8)$$

and the infimum is taken over all possible estimators $\tilde{f}_s$ based on both the output signals $Y_1, \dots, Y_k$ and the parameters σ, γ, and Q.

Moreover, set $J_\gamma := 2\lfloor \gamma^{-1} \ln(\sigma^{-1}) \rfloor + 1$, and let the number of available channels k be at least J_γ. Then the projection estimator,

$$\hat{f}(t) := \sum_{j=1}^{J_\gamma} Y_j \varphi_{j-1}(t), \quad (7.1.9)$$

is a versatile sharp minimax estimator, that is,

$$\sup_{f \in A_{\gamma,Q}} \mathrm{MSE}(\hat{f}^{(s)}(t_0), f^{(s)}(t_0)) = P_{s,\gamma} (\ln(\sigma^{-1}))^{2s+1} \sigma^2 (1 + o_\sigma(1)), \quad (7.1.10)$$

$$\sup_{f \in A_{\gamma,Q}} \mathrm{MISE}(\hat{f}^{(s)}, f^{(s)}) = P_{s,\gamma} (\ln(\sigma^{-1}))^{2s+1} \sigma^2 (1 + o_\sigma(1)). \quad (7.1.11)$$

In other words, the lower bounds (7.1.6)–(7.1.7) are asymptotically (as $\sigma \to 0$) sharp and attainable by the sth derivative of the projection estimate.

This proposition is an example of what the asymptotic theory is about, namely, it allows one to find best estimators and understand how well functions (and their derivatives) from large function classes may be estimated (recovered). Note that both rates and optimal constants for MSE and MISE convergences are established.

Remark 7.1.1 Game Theory and Minimax. It may be convenient to think about both the setting of Theorem 7.1.1 and the minimax approach in terms of concepts of the game theory. We may think that nature (player I) chooses a function $f(t) = \sum_{j=1}^{k} X_j \varphi_{j-1}(t)$, and the statistician (player II), based on the noisy observations $Y_1, \dots, Y_k$, tries to estimate $f(t)$. Nature's strategies or choices are $\{X_j\}$ (and in some cases coding functions as well), while the statistician's strategies or choices are $\hat{f}(t)$. The lower bounds (7.1.6)–(7.1.7) tell us that nature's choice may be such that best strategies of the statistician cannot lead to smaller risks. On the other hand, regardless

of nature's choices, the statistician can use the projection estimate, which is the best strategy against smartest plays made by nature, and guarantee the accuracy (7.1.10)–(7.1.11). In other words, a minimax approach is based on the greatest respect to nature (player I) by assuming that nature employs only optimal strategies and never makes mistakes.

Proof of Theorem 7.1.1 We begin with establishing the lower bounds (7.1.6)–(7.1.7). First, let us recall one classical result of parametric estimation theory; see the sketch of proof in Exercise 7.1.3.

Lemma 7.1.1 *Let $Y := \theta + \sigma\xi$, where $\theta \in [-c\sigma, c\sigma]$ is an estimated parameter, c and σ are positive constants, and ξ is a standard normal random variable. Then there exists a random variable Θ independent of ξ with the density supported on $[-c\sigma, c\sigma]$ such that*

$$\inf_{\tilde\theta} \sup_{\theta\in[-c\sigma,c\sigma]} E\{(\tilde\theta-\theta)^2\} \ge \inf_{\tilde\theta} E\{(\tilde\theta-\Theta)^2\} \ge \frac{\mu(c)c^2}{1+c^2}\,\sigma^2. \tag{7.1.12}$$

Here the infimum is taken over all possible estimates $\tilde\theta$ based on the triplet (Y, c, σ), and the Ibragimov–Khasminskii function $\mu(c)$ is such that $\mu(c) \ge 0.8$ and $\mu(c) \to 1$ as $c \to 0$ or $c \to \infty$.

We use this lemma to establish the following proposition. Define a function class $D(c,k) := \{f : f(t) = \sum_{j=1}^k X_j g_j(t), \quad X_j \in [-c\sigma, c\sigma],\ j = 1,\dots,k\}$. Note that this class includes all possible signals transmitted via k parallel channels whose entries may have absolute values not larger than $c\sigma$. Also, recall that the coding functions g_j are finite and square integrable.

Lemma 7.1.2 *Consider the communication system (7.1.1) depicted in Figure 7.1 and satisfying the assumptions of Theorem 7.1.1. Suppose that all the coding functions $\{g_j(t),\ j = 1, 2, \dots, k\}$ are s-fold differentiable and all the derivatives are finite. Also suppose that a function $f(t) \in D(c,k)$, $0 \le t \le 1$, is transmitted as explained in Theorem 7.1.1. Then*

$$\inf_{\tilde f_s(t_0)} \sup_{f\in D(c,k)} \mathrm{MSE}(\tilde f_s(t_0), f^{(s)}(t_0)) \ge \frac{\mu(c)c^2\sigma^2}{1+c^2} \sum_{j=1}^k (g_j^{(s)}(t_0))^2, \tag{7.1.13}$$

$$\inf_{\tilde f_s} \sup_{f\in D(c,k)} \mathrm{MISE}(\tilde f_s, f^{(s)}) \ge \frac{\mu(c)c^2\sigma^2}{1+c^2} \sum_{j=1}^k \int_0^1 (g_j^{(s)}(t))^2 dt. \tag{7.1.14}$$

Here $\mu(c)$ is the Ibragimov–Khasminskii function introduced in Lemma 7.1.1, and the infimum is taken over all possible estimates $\tilde f_s(t)$ based on output signals $(Y_1, \dots, Y_k)$ and parameters c and σ.

Proof of Lemma 7.1.2 It suffices to look after best estimates among a class of the estimates $\hat f_s(t) = \sum_{j=1}^k \hat X_j g_j^{(s)}(t)$, where $\hat X_j$ are statistics based on all the output signals and parameters c and σ. The last sentence looks quite reasonable in light of (7.1.1)–(7.1.2); nevertheless, it should be proved, and we do this at the end of this proof.

Then, to establish (7.1.13) we write

$$\mathrm{MSE}(\tilde{f}_s(t_0), f^{(s)}(t_0)) = E\Big\{\Big[\sum_{j=1}^{k}(\hat{X}_j - X_j)g_j^{(s)}(t_0)\Big]^2\Big\}. \tag{7.1.15}$$

Using the fact that a minimax risk is not smaller than a corresponding Bayes risk and then the equality (A.42) in Appendix A, we may write for iid Θ_j, $j = 1, 2, \ldots, k$, distributed as the random variable Θ in Lemma 7.1.1,

$$\sup_{f\in D(c,k)} \mathrm{MSE}(\hat{f}_s(t_0), f^{(s)}(t_0)) \geq E\{[\sum_{j=1}^{k}(E\{\Theta_j|(Y_1, \ldots, Y_k)\} - \Theta_j)g_j^{(s)}(t_0)]^2\}. \tag{7.1.16}$$

Recall that if Θ_j is the jth input signal, then $Y_j = \Theta_j + \sigma\xi_j$ is the corresponding outcome. Here ξ_j are iid standard normal and independent of $\Theta_1, \ldots, \Theta_k$. This implies that (Θ_j, Y_j) and $\{Y_l,\ l = 1, \ldots, k,\ l \neq j\}$ are independent. Thus $E\{\Theta_j|(Y_1, \ldots, Y_k)\} = E\{\Theta_j|Y_j\}$ and

$$E\{[E\{\Theta_l|Y_l\} - \Theta_l][E\{\Theta_m|Y_m\} - \Theta_m]\} = 0, \quad l \neq m.$$

In the last line we used the relation $E\{E\{\Theta_l|Y_l\}\} = E\{\Theta_l\}$ based on the definition of conditional expectation; see (A.14) and (A.20) in Appendix A. Thus we get that

$$\begin{aligned}
&\inf_{\tilde{f}(t_0)} \sup_{f\in D(c,k)} \mathrm{MSE}(\hat{f}_s(t_0), f^{(s)}(t_0)) \\
&\geq \sum_{j=1}^{k} E\{(E\{\Theta_j|Y_j\} - \Theta_j)^2\}(g_j^{(s)}(t_0))^2 \\
&\geq \sum_{j=1}^{k} \inf_{\tilde{\theta}_j} E\{(\tilde{\theta}_j - \Theta_j)^2\}(g_j^{(s)}(t_0))^2.
\end{aligned} \tag{7.1.17}$$

Then Lemma 7.1.1 yields (7.1.13). A lower bound for the minimax MISE is established following the same lines of the proof because the MISE is just the MSE integrated over $[0, 1]$. Following (7.1.15)–(7.1.17) we write

$$\begin{aligned}
\sup_{f\in D(c,k)} \mathrm{MISE}(\hat{f}_s, f^{(s)}) &= \sup_{f\in D(c,k)} \int_0^1 E\Big\{\Big[\sum_{j=1}^{k}(\hat{X}_j - X_j)g_j^{(s)}(t)\Big]^2\Big\}dt \\
&\geq \int_0^1 E\Big\{\Big[\sum_{j=1}^{k}(E\{\Theta_j|(Y_1, \ldots, Y_k)\} - \Theta_j)g_j^{(s)}(t)\Big]^2\Big\}dt \\
&\geq \sum_{j=1}^{k} \inf_{\tilde{\theta}_j} E\{(\tilde{\theta}_j - \Theta_j)^2\} \int_0^1 (g_j^{(s)}(t))^2 dt.
\end{aligned}$$

Then (7.1.12) yields (7.1.14).

To complete the proof of Lemma 7.1.2, let us explain why we can restrict our attention to the specific estimates $\hat{f}_s$. For the pointwise approach this is elementary because if, for instance, $g_j^{(s)}(t_0) \neq 0$ (if there is no such $j \in \{1, 2, \dots, k\}$, then the assertion (7.1.13) is apparent), then we may set $\hat{X}_j = \tilde{f}_s(t_0)/g_j^{(s)}(t_0)$ and all other $\hat{X}_l = 0$. The case of MISE is also simple. The projection of an estimate $\tilde{f}_s$ on the closed linear span of $\{g_j^{(s)}, j = 1, 2, \dots, k\}$ either does not change or decreases the MISE (recall Section 2.3 and the projection theorem), and this explains why it suffices to consider only estimates $\hat{f}_s$ that belong to that span. Lemma 7.1.2 is proved.

Lemma 7.1.2 is the key tool to establish lower bounds for minimax risks over different function spaces because it suffices to find an appropriate class $D(c, k)$ that belongs to a given function class.

Introduce $J := 2\max(1, \lfloor \gamma^{-1} \ln(\sigma^{-1})(1 - 1/\ln(\ln(\sigma^{-1})))\rfloor)$, which is a "bit" smaller than J_γ and consider $X_j \in [-\ln(\sigma^{-1})\sigma, \ln(\sigma^{-1})\sigma]$ for $j = 1, \dots, J$. Then a direct calculation together with definition (7.1.5) of the class $A_{\gamma,Q}$ of analytic functions shows that if σ is sufficiently small (more precisely if $\ln(\sigma^{-1})\sigma < Qe^{-\gamma J/2}$), then all the signals $f_J(t) = \sum_{j=1}^J X_j \varphi_j(t)$ belong to $A_{\gamma,Q}$. Since the number k of channels is arbitrarily large, we assume that the k is at least J, and therefore we may transmit all the elements of f_J. Thus we can use (7.1.13) and write

$$\begin{aligned}
&\inf_{\tilde{f}_s(t_0)} \sup_{f \in A_{\gamma,Q}} \mathrm{MSE}(\tilde{f}_s(t_0), f^{(s)}(t_0)) \\
&\geq \inf_{\tilde{f}_s(t_0)} \sup_{\{X_j \in [-\ln(\sigma^{-1})\sigma, \ln(\sigma^{-1})\sigma],\, j=1,\dots,J\}} \mathrm{MSE}\Big(\tilde{f}_s(t_0), \sum_{j=1}^J X_j \varphi_j^{(s)}(t_0)\Big) \\
&\geq \mu(\ln(\sigma^{-1}))\big[(\ln(\sigma^{-1}))^2/(1 + (\ln(\sigma^{-1}))^2)\big]\sigma^2 \sum_{j=1}^J [\varphi_j^{(s)}(t_0)]^2. \qquad (7.1.18)
\end{aligned}$$

A calculation based on using the familiar relation $\cos^2(\alpha) + \sin^2(\alpha) = 1$ shows that uniformly over $t \in [0, 1]$,

$$\sum_{j=1}^J [\varphi_j^{(s)}(t)]^2 = P_{s,\gamma}\ (\ln(\sigma^{-1}))^{2s+1}(1 + o_\sigma(1)). \qquad (7.1.19)$$

Then (7.1.18), (7.1.19) and the property $\mu(c) \to 1$ as $c \to \infty$ imply the lower bound (7.1.6). The lower bound (7.1.7) for the minimax MISE is established absolutely similarly, and the proof is left as an exercise.

Now let us establish the upper bounds for risks of the estimator (7.1.9). Denote by $\theta_j := \int_0^1 f(t)\varphi_j(t)dt$ the jth Fourier coefficient of a signal $f(t)$, and recall that $Y_j := \theta_{j-1} + Z_j$ (we count the elements of the trigonometric basis beginning from 0 and the channels from 1, and this causes that minor

complication in indices). Recall that Z_j are zero-mean and write

$$\begin{aligned}
&\mathrm{MSE}(\hat f^{(s)}_{J_\gamma}(t_0), f^{(s)}(t_0))\\
&= E\Big\{\Big(\sum_{j=1}^{J_\gamma}(Y_j-\theta_{j-1})\varphi^{(s)}_{j-1}(t_0)+\sum_{j>J_\gamma}\theta_{j-1}\varphi^{(s)}_{j-1}(t_0)\Big)^2\Big\}\\
&= E\Big\{\Big[\sum_{j=1}^{J_\gamma}Z_j\varphi^{(s)}_{j-1}(t_0)\Big]^2\Big\}+\Big(\sum_{j>J_\gamma}\theta_{j-1}\varphi^{(s)}_{j-1}(t_0)\Big)^2. \qquad (7.1.20)
\end{aligned}$$

In (7.1.20) the MSE is written as the sum of the variance and the squared bias terms. To estimate the variance term we use the elementary relation

$$E\Big\{\Big(\sum_{j=1}^m\eta_j a_j\Big)^2\Big\}=\mathrm{Var}\Big(\sum_{j=1}^m\eta_j a_j\Big)=\sigma^2\sum_{j=1}^m a_j^2, \qquad (7.1.21)$$

which holds for any iid $\eta_1,\dots,\eta_m$ with zero mean and variance σ^2 and any finite constants $a_1,\dots,a_m$. Together with (7.1.19) this implies

$$\begin{aligned}
E\Big\{\Big[\sum_{j=1}^{J_\gamma}Z_j\varphi^{(s)}_{j-1}(t_0)\Big]^2\Big\}&=\sigma^2\sum_{j=1}^{J_\gamma}\big[\varphi^{(s)}_{j-1}(t_0)\big]^2\\
&=P_{s,\gamma}(\ln(\sigma^{-1}))^{2s+1}\sigma^2(1+o_\sigma(1)). \qquad (7.1.22)
\end{aligned}$$

The squared bias term is estimated by the following lines:

$$\begin{aligned}
\sup_{f\in A_{\gamma,Q}}\Big[\sum_{j\ge J_\gamma}\theta_j\varphi^{(s)}_j(t_0)\Big]^2&\le C\Big[\sum_{j\ge J_\gamma}j^s e^{-\gamma j/2}\Big]^2\\
&\le CJ_\gamma^{2s}e^{-\gamma J_\gamma}\le C(\ln(\sigma^{-1}))^{2s}\sigma^2. \qquad (7.1.23)
\end{aligned}$$

Here and in what follows C denotes (in general different) positive constants.

Using the results in (7.1.20) we get

$$\sup_{f\in A_{\gamma,Q}}\mathrm{MSE}(\hat f^{(s)}_{J_\gamma}(t_0),f^{(s)}(t_0))\le P_{s,\gamma}(\ln(\sigma^{-1}))^{2s+1}\sigma^2(1+o_\sigma(1)). \qquad (7.1.24)$$

Absolutely similarly (just by integration of the right-hand side of (7.1.20) and then using (7.1.21)–(7.1.23)) we get

$$\sup_{f\in A_{\gamma,Q}}\mathrm{MISE}(\hat f^{(s)}_{J_\gamma},f^{(s)})\le P_{s,\gamma}\left(\ln(\sigma^{-1})\right)^{2s+1}\sigma^2(1+o_\sigma(1))\,. \qquad (7.1.25)$$

Since the upper bounds (7.1.24) and (7.1.25) asymptotically coincide with the lower bounds (7.1.6) and (7.1.7), we conclude that the estimator $\hat f^{(s)}_{J_\gamma}(t)$ is sharp minimax. This also shows the versatility of this estimator. In short, if the parameter γ is known, then $\hat f_{J_\gamma}(t)$ is an ideal estimator for analytic functions because both the rate and the constant of the risks' convergences are optimal, derivatives of the optimal estimator of an underlying input signal are optimal estimates of the corresponding derivatives of

the input signal, and the optimal estimator is both pointwise and globally optimal. This constitutes a beautiful bouquet of asymptotic properties that are difficult to beat. This remark concludes the proof of Theorem 7.1.1.

Theorem 7.1.1 shows that analytic functions and their derivatives may be estimated with accuracy just a bit worse (in terms of a logarithmic factor) than a single parameter. So now let us consider a function class where nonparametrics takes a much larger toll. We shall discuss a familiar class of Lipschitz functions with r continuous derivatives such that

$$|f^{(r)}(u) - f^{(r)}(v)| \le L|u-v|^{\alpha}\,. \tag{7.1.26}$$

Here $0 < \alpha \le 1$, and the Lipschitz constant is $L < \infty$; note that here we do not assume that f is periodic.

To get lower bounds we can use the characterization of (generalized) Lipschitz classes via wavelets discussed in Section 2.5. We leave exploring this approach as Exercise 7.1.9 (it also will be used in Section 7.5). Instead, here we shall use a self-contained method of the proof.

First of all, let us begin with some notions. We say that f is locally Lipschitz at a point t_0 if (7.1.26) holds for u and v in some vicinity of t_0. This function space we denote by $Lip_{r,\alpha,L}(t_0)$. If (7.1.26) holds for all $u, v \in A$, then we denote this space by $Lip_{r,\alpha,L}(A)$.

Theorem 7.1.2 *Consider a signal $f(t) := \sum_{j=1}^{\infty} \theta_j g_j(t)$, $0 \le t \le 1$, where $\{g_j\}$ are coding functions. Suppose that this signal is transmitted via the parallel Gaussian channels (7.1.1) by setting $X_j := \theta_j$. It is assumed that the noise in channels is normal $N(0, \sigma^2)$ with $0 < \sigma < 1$. Consider any $s = 0, \dots, r$ and $t_0 \in [0,1]$. Then, regardless of how large the number k of the channels, the following lower bounds for the minimax risks hold:*

$$\inf_{\tilde f_s} \sup_{f \in Lip_{r,\alpha,L}(t_0)} \mathrm{MSE}(\tilde f_s(t_0), f^{(s)}(t_0)) \ge C\sigma^{4(\beta-s)/(2\beta+1)}, \tag{7.1.27}$$

$$\inf_{\tilde f_s} \sup_{f \in Lip_{r,\alpha,L}([0,1])} \mathrm{MISE}(\tilde f_s, f^{(s)}) \ge C\sigma^{4(\beta-s)/(2\beta+1)}. \tag{7.1.28}$$

Here $\beta := r + \alpha$, the supremum over f means that the supremum over the corresponding $\{\theta_j\}$ and $\{g_j\}$ is considered, and the infimum is taken over all possible estimators $\tilde f_s$ based on outputs $(Y_1, \dots, Y_k)$, the coding functions $\{g_j\}$, and parameters r, α, L, and σ.

Estimators that attain these lower bounds will be considered in Section 7.3.

Proof of Theorem 7.1.2 Consider a function $m(t) = e^{-1/(1-4x^2)} I_{\{x \in [-\frac{1}{2}, \frac{1}{2}]\}}$ infinitely differentiable on the real line; see Figure 7.2. Such exceptionally smooth functions with bounded support are called *mollifiers*.

We shall use this mollifier as a building block to construct coding functions. We do this, as in the construction of a wavelet basis, by dilation and translation, only here all will be essentially simpler because either one

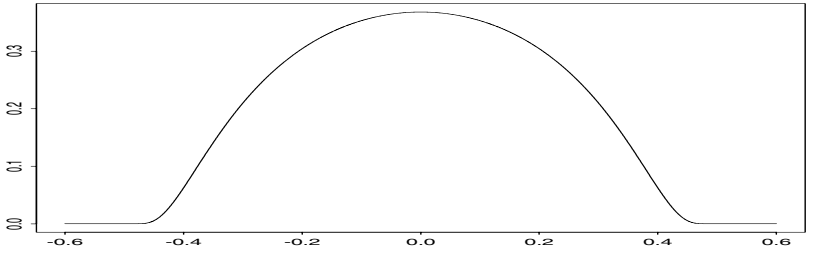

FIGURE 7.2. A mollifier.

coding function is used for the analysis of MSE or a single resolution level of coding functions is used for the analysis of MISE.

First, consider the lower bound for the minimax MSE. Set $g_1(t) := 2^{J/2}\, m((t - t_{s,J})2^J)$, where $J := \lfloor (2\beta+1)^{-1} \log_2(\sigma^{-2}) \rfloor$ and the sequence of points $t_{s,J}$ is such that $|m^{(r)}((t_0 - t_{s,J})2^J)| \ge c^* > 0$. Note that, for instance, $t_{1,J} = t_0$ cannot be used, because $m^{(1)}(0) = 0$, but if t_s is a point such that $|m^{(s)}(t_s)| \ge c^* > 0$, and such a point always exists for a sufficiently small c^*, then we may set $t_{s,J} = t_0 - t_s 2^{-J}$.

Then, let us check that $f_1(t) := X_1 g_1(t)$ belongs to $Lip_{r,\alpha,L}(t_0)$ if $X_1 \in [-c\sigma, c\sigma]$ with a sufficiently small positive constant c. First, note that the mollifier $m(t)$ belongs to the Lipschitz space $Lip_{r,\alpha,L_r}(\,(-\infty,\infty)\,)$ with some finite Lipschitz constant L_r. Second, recall the chain rule for differentiation of a composite function, which implies that the lth derivative of the function $m(c(t-a))$ may be calculated by the formula $d^l m(c(t-a))/dt^l = c^l m^{(l)}(c(t-a))$. This rule yields

$$g_1^{(l)}(t) = 2^{(l+1/2)J} m^{(l)}((t - t_{s,J})2^J), \quad l = 0, 1, \dots. \tag{7.1.29}$$

Using these results we write

$$\begin{aligned} &|f_1^{(r)}(u) - f_1^{(r)}(v)| \le |X_1||g_1^{(r)}(u) - g_1^r(v)| \\ &\le |X_1| 2^{(r+1/2)J} |m^{(r)}((u - t_{s,J})2^J) - m^{(r)}((v - t_{s,J})2^J)| \\ &\le c\sigma 2^{(r+1/2)J} L_r |(u-v)2^J|^\alpha \le [c L_r \sigma 2^{(2\beta+1)J/2}] |u - v|^\alpha. \end{aligned} \tag{7.1.30}$$

For $\sigma < 1$ we get $\sigma 2^{(2\beta+1)J/2} \le 1$; thus any $c \le L/L_r$ implies $f_1 \in Lip_{r,\alpha,L}(t_0)$. Then, using (7.1.13) with $k = 1$, (7.1.29), and definitions of J and $t_{s,J}$, we get

$$\begin{aligned} &\inf_{\tilde f_s(t_0)} \sup_{f \in Lip_{r,\alpha,L}(t_0)} \mathrm{MSE}(\tilde f_s(t_0), f^{(s)}(t_0)) \\ &\ge \inf_{\tilde f_s(t_0)} \sup_{f \in D(c,1)} \mathrm{MSE}(\tilde f_s(t_0), f^{(s)}(t_0)) \ge C\sigma^2 (g_1^{(s)}(t_0))^2 \\ &= C\sigma^2 \big[2^{(2s+1)J/2}\, m^{(s)}((t_0 - t_{s,J})2^J) \big]^2 \\ &\ge C\sigma^2 \sigma^{-(4s+2)/(2\beta+1)} = C\sigma^{4(\beta-s)/(2\beta+1)}. \end{aligned}$$

The lower bound (7.1.27) is established. Note that only one channel has been used in the proof, so, using the terminology of Remark 7.1.1, nature may make its best game by using only one coding function whose magnitude is unknown to the statistician.

Now let us consider a lower bound for the minimax MISE. Here we use $k := 2^J - 1$ coding functions and

$$f_k(t) := \sum_{j=1}^{k} X_j g_j(t), \quad \text{where} \quad g_j(t) := 2^{J/2} m(2^J t - j). \tag{7.1.31}$$

If $X_j \in [-c\sigma, c\sigma]$ with a sufficiently small c, then absolutely similarly to (7.1.30) we verify that $f_k(t) \in Lip_{r,\alpha,L}([0,1])$ (Exercise 7.1.13). Also, using (7.1.29) we get that $\int_0^1 (g_j^{(s)}(t))^2 dt \ge C\sigma^{-4s/(2\beta+1)}$. Then,

$$\begin{aligned}
&\inf_{\tilde f_s} \sup_{Lip_{r,\alpha,L}([0,1])} \mathrm{MISE}(\tilde f_s, f^{(s)}) \ge \inf_{\tilde f_s} \sup_{f \in D(c,k)} \mathrm{MISE}(\tilde f_s, f^{(s)}) \\
&\ge C\sigma^2 \sum_{j=1}^{k} \int_0^1 (g_j^{(s)}(t))^2 dt \ge C\sigma^2 2^J \sigma^{-4s/(2\beta+1)} \\
&\ge C\sigma^2 \sigma^{-2/(2\beta+1)} \sigma^{-4s/(2\beta+1)} = C\sigma^{4(\beta-s)/(2\beta+1)}.
\end{aligned}$$

Theorem 7.1.2 is proved.

Remark 7.1.2 Special Features. When a lower bound for minimax risks is studied, it is often desirable to reduce an underlying function space to incorporate some additional restrictions. For instance, signals may be positive, bounded, periodic, or monotone. The above-proposed proofs allow us to consider such settings rather straightforwardly. Indeed, in the proofs all the function spaces $D(c,k)$ were spaces of 1-periodic and bounded functions. Only some minor modifications are needed for the case of positive or monotone functions. For instance, consider the case of positive functions (like densities) on $[0,1]$ and the last proof for Lipschitz functions and the MSE. We considered a subclass of functions $f(t) = X_1 g_1(t)$ such that $\max_t |f(t)| = o_\sigma(1)$. Thus, if instead of using only 1 channel we add a second one and choose $X_2 = 1$ and a corresponding sufficiently smooth coding function $g_2(t)$ bounded from below on $[0,1]$, then $X_1 g_1(t) + g_2(t)$ will be positive on $[0,1]$ whenever σ is sufficiently small. Absolutely similarly, for the case of monotone functions one extra monotone coding function should be introduced. In short, if f satisfies a restriction, then in the course of finding a lower bound all the $f_k(t) = \sum_{j=1}^{k} X_j g_j(t) \in D(c,k)$ should satisfy this restriction. Because both $\{X_j\}$ and the coding functions $\{g_j\}$ may be typically chosen with great flexibility, natural restrictions like periodicity or monotonicity are easily incorporated.

Remark 7.1.3 Local Minimax. This remark continues the discussion of the previous one, only here the notion of a local minimax is the objective. All minimax risks considered in this section were global over a function

space, i.e., the supremum was taken over the whole function space. Using the terminology of Remark 7.1.1, this implies that nature may choose any function from that space to beat the statistician. In practice it is often more reasonable to think that an underlying function belongs to some vicinity of a particular function f^*, for instance, we consider only f such that $\max_t |f(t) - f^*(t)| < \delta_\sigma$ where $\delta_\sigma \to 0$ as $\sigma \to 0$. In this case the minimax approach is called *local*. The only new issue in the analysis of a local minimax is the lower bound. Here the approach of Remark 7.1.2 works perfectly: Introduce a new first coding function $g_1(t) := f^*(t)$ and set $X_1 := 1$. Then the proofs may be straightforwardly repeated. In particular, if $\delta_\sigma > \sigma^{2/3}$, then the result of Theorem 7.1.1 holds for the local minimax, where the supremum is taken over $f \in A_{\gamma,Q} \cap \{\psi : \max_{t\in[0,1]} |\psi(t) - f^*(t)| < \delta_\sigma\}$. The proof is left as Exercise 7.1.16.

Remark 7.1.4 Efficient Estimation. The assertion of Theorem 7.1.2 is about rates of risk convergence, while the assertion of Theorem 7.1.1 is about both rates and sharp constants of risk convergence. As in classical parametric theory, an estimate whose risk attains both optimal rate and a sharp constant is called (asymptotically) *efficient*. It is possible to go further and explore a subclass of efficient estimates whose risks converge with optimal constant and rate of the second order in σ. As a result, a best estimator among efficient estimators may be found. Let us consider a particular example of estimation of the integral functional $F(t) := \int_0^t f(x)dx$ (note that F is the cumulative distribution function for a density model). The claim is that under the assumption of Theorem 7.1.1 for any $t_0 \in (0, 1)$,

$$\inf_{\tilde F} \sup_{f\in A_{\gamma,Q}} \mathrm{MSE}(\tilde F(t_0), F(t_0)) \ge \sigma^2\Big(t_0 - \frac{\gamma + o_\sigma(1)}{\pi^2 \ln(\sigma^{-1})}\Big), \tag{7.1.32}$$

and the estimate $\hat F(t) := \int_0^t \hat f(x)dx$, where $\hat f$ is defined in (7.1.9), is *second-order efficient*, i.e.,

$$\sup_{f\in A_{\gamma,Q}} \mathrm{MSE}(\hat F(t_0), F(t_0)) = \sigma^2\Big(t_0 - \frac{\gamma + o_\sigma(1)}{\pi^2 \ln(\sigma^{-1})}\Big). \tag{7.1.33}$$

The proof of this assertion is left as Exercise 7.1.17, where a detailed hint is given. Note that this example adds one more nice property to the projection estimate (7.1.9), namely, it is not only versatile in the sense that its derivatives are efficient estimates of the derivatives, but its integral is a second-order efficient estimate of the corresponding integral functional. The case of MISE is left as Exercise 7.1.18.

Remark 7.1.5 Bayesian Approach. Let us assume that in (7.1.1) the input signals are independent random variables. Then, according to (A.40) (see Appendix A) a Bayes estimator,

$$\hat f_B(t) := \sum_{j=1}^k E\{X_j|Y_j\}g_j(t)\ , \tag{7.1.34}$$

minimizes both Bayes MISE and Bayes MSE, in short, BMISE and BMSE. Here $\mathrm{BMISE}(\hat{f}, f) := E\{\int_0^1 (\hat{f}(t) - \sum_{j=1}^k X_j g_j(t))^2\}$, and BMSE is defined similarly. Note that this assertion holds for any independent parallel channels (not necessarily Gaussian).

If the channels are Gaussian and X_j are normal $N(0, \nu_j^2)$, then according to Example A.27 the Bayes estimate $\hat{f}_B$ becomes the *Wiener filter*,

$$\hat{f}_W(t) = \sum_{j=1}^k \frac{\nu_j^2}{\nu_j^2 + \sigma^2} Y_j g_j(t). \tag{7.1.35}$$

Moreover, it is not difficult to calculate the risks of $\hat{f}_W$. For instance, let $\{g_j\}$ be the sine-cosine basis, $\nu_{2j+1}^2 = \nu_{2j}^2$, $j = 1, 2, \ldots$ and k is odd. Then

$$\mathrm{BMISE}(\hat{f}_W, f) = \mathrm{BMSE}(\hat{f}_W(t_0), f(t_0)) = \sigma^2 \sum_{j=1}^k \frac{\nu_j^2}{\nu_j^2 + \sigma^2}\,. \tag{7.1.36}$$

Suppose that σ is sufficiently small, and let us give two examples of ν_j^2 that explain the relationship between the minimax and Bayesian approaches. First, let ν_j^2 be proportional to $e^{-2\gamma j}$; then an appropriate choice of k implies that the BMISE decreases as the right-hand side of (7.1.11). Second, let ν_j^2 be proportional to $j^{-2\beta-1}$; then an appropriate choice of k implies that the BMISE decreases proportionally to the right-hand side of (7.1.28). Thus, asymptotically these prior distributions "mimic" the minimax estimation of analytic and Lipschitz functions. Verification of the calculations is left as Exercise 7.1.20.

7.2 Filtering a Signal from White Noise

The objective of this Section is to discuss a mathematically fruitful generalization of the communication system (7.1.1) in which the number k of parallel channels becomes infinity. There are two particular benefits from this generalization. The first one is that we will be able to introduce an important notion of a Brownian motion and a continuous-in-time filtering model. The second one is that we will be able to introduce the principle of equivalence between that communication system and statistical models discussed in the previous chapters.

We begin the discussion with a filtering model and a Brownian motion.

Consider the system depicted in Figure 7.1 where Z_j are iid normal $N(0, \sigma^2)$. Because in this section we use only the classical sine–cosine trigonometric basis $\{\varphi_j,\ j = 0, 1, \ldots\}$, we assume that the channels are numerated from 0 to $k-1$ instead of 1 to k. Set $f_k(t) := \sum_{j=0}^{k-1} \theta_j \varphi_j(t)$, $0 \le t \le 1$, and note that $\theta_j = \int_0^1 \varphi_j(t) f_k(t) dt$ are the Fourier coefficients of

f_k. This continuous-in-time signal can be transmitted by the system (7.1.1) if we set $X_j := \theta_j$, $j = 0, 1, \ldots, k-1$.

Then the corresponding continuous-in-time output signal $y_k(t)$ may be defined as

$$y_k(t) := \sum_{j=0}^{k-1} Y_j \varphi_j(t). \tag{7.2.1}$$

Such an idea of thinking about k discrete outputs $\{Y_j\}$ as the Fourier coefficients of a continuous-in-time output signal $y_k(t)$ looks rather attractive; after all, $y_k(t)$ is an unbiased estimate of the transmitted $f_k(t)$.

However, a serious complication arises as k increases. Write

$$y_k(t) = f_k(t) + \sum_{j=0}^{k-1} Z_j \varphi_j(t) =: f_k(t) + W_k^*(t),$$

and consider the stochastic term $W_k^*(t)$. For any particular moment in time t this term is a normal random variable with mean zero and (if k is odd) variance $\sigma^2 k$ because $\sum_{j=0}^{k-1} \varphi_j^2(t) = k$ for the classical trigonometric elements (recall that $\varphi_{2j-1}(t) = 2^{1/2}\sin(2\pi jt)$ and $\varphi_{2j}(t) = 2^{1/2}\cos(2\pi jt)$, so $\varphi_{2j-1}^2(t) + \varphi_{2j}^2(t) = 2$ for $j = 1, 2, \ldots$). Thus, if k increases, then this stochastic term just blows up.

To avoid this complication, we use the following simple trick. Let us instead of matching a continuous-in-time input signal $f_k(t)$ match its integral $\int_0^t f_k(u)du$. This integral is again a continuous-in-time signal, and the corresponding continuous-in-time output signal is $Y_k(t) := \int_0^t y_k(u)du$. To see why this approach is better, write

$$Y_k(t) = \int_0^t f_k(u)du + \sum_{j=0}^{k-1} Z_j \int_0^t \varphi_j(u)du =: \int_0^t f_k(u)du + B_k^*(t). \tag{7.2.2}$$

Since Z_j are zero-mean random variables, we get that $Y_k(t)$ is an unbiased estimate of $\int_0^t f_k(u)du$. Also, the stochastic term $B_k^*(t)$ for a given time t is a normal random variable with mean zero and variance

$$\mathrm{Var}\big(B_k^*(t)\big) = \sigma^2 \sum_{j=0}^{k-1} \Big(\int_0^t \varphi_j(u)du \Big)^2. \tag{7.2.3}$$

This variance, as a sequence in k, has two nice properties: It is always less than $\sigma^2 t$, and it tends to $\sigma^2 t$ as $k \to \infty$. Let us verify these properties using the Parseval identity (2.3.11). Set $p_t(u) := I_{\{0 \le u \le t\}}$ and write

$$t = \int_0^1 p_t^2(u)du = \sum_{j=0}^{\infty} \Big(\int_0^1 p_t(u)\varphi_j(u)du \Big)^2 = \sum_{j=0}^{\infty} \Big(\int_0^t \varphi_j(u)du \Big)^2$$

$$= \sum_{j=0}^{k-1} \Big(\int_0^t \varphi_j(u)du \Big)^2 + \sum_{j \ge k} \Big(\int_0^1 p_t(u)\varphi_j(u)du \Big)^2.$$

This implies the above-formulated properties of $\mathrm{Var}(B_k^*(t))$.

As we see, for a fixed time t the stochastic term $B_k^*(t)$ converges to a normal random variable $N(0, \sigma^2 t)$ as $k \to \infty$. Thus, at least formally, we can consider the limit of $B_k^*(t)$ as $k \to \infty$. Denote this limit by $B^*(t)$ and call it a *Brownian motion on* $[0, 1]$. A so-called *standard Brownian motion* is obtained by using standard normal Z_j, i.e., $\sigma^2 = 1$, and we denote it by $B(t)$. It is possible to show that a Brownian motion (the limit) does not depend on an underlying basis in $L_2(0, 1)$; see Exercise 7.2.4. Properties of a Brownian motion are formulated in Exercise 7.2.3. (These properties are customarily used to define a Brownian motion, and then our approach is used as an example that shows how this process may be generated.)

If we denote by $Y(t)$ the formal limit of $Y_k(t)$ as $k \to \infty$, then we can compactly write that the input signal $f(t)$ satisfies any of the following two stochastic equations:

$$Y(t) = \int_0^t f(u)du + \sigma B(t), \quad 0 \le t \le 1, \tag{7.2.4}$$

or

$$dY(t) = f(t)dt + \sigma dB(t), \qquad 0 \le t \le 1. \tag{7.2.5}$$

Here $Y(t)$, $0 \le t \le 1$, is called an observed (continuous-in-time) signal. Also, just formally, the derivative $W(t) := dB(t)/dt$ is called a *standard white Gaussian noise*.

The white noise is a pure mathematical notion. Indeed, a white noise $W^*(t) := \sum_{j=0}^{\infty} Z_j \varphi_j(t)$ has the same power at all frequencies (this explains the name "white"). Thus its total power is infinity, and no physical system can generate a white noise. On the other hand, its frequency-limited version $W_k^*(t) = \sum_{j=0}^{k-1} Z_j \varphi_j(t)$ has a perfect physical sense, and at least theoretically, $W_k^*(t)$ may be treated as $W^*(t)$ passed through an ideal low-pass rectangular filter. This explains why a white noise is widely used in communication theory.

A mathematical model where a continuous-in-time input signal $f(t)$ satisfies stochastic equation (7.2.4) or (7.2.5) is called an observation of a signal in a white Gaussian noise, and the problem of estimation (recovery) of a signal f is called *filtering a signal from a white Gaussian noise*. Also, the stochastic equations (7.2.4)–(7.2.5) mean that outputs Y_j of parallel Gaussian channels can be written as

$$Y_j = \int_0^1 \varphi_j(u)dY(t) = X_j + \sigma Z'_j, \quad j = 0, 1, \dots, \tag{7.2.6}$$

where here Z'_j are iid standard normal. And conversely, if $Y_0, Y_1, \dots$ are outputs of the channels, then $Y(t) = \sum_{j=0}^{\infty} Y_j \int_0^t \varphi_j(t)dt$.

Let us look at some realizations of a truncated (in the frequency domain) Brownian motion (denote iid standard normal variables by ξ_j)

$$B^*(t,k,n,d) := (d/n)^{1/2} \sum_{j=0}^{k-1} \xi_j \int_0^t \varphi_j(u)du, \tag{7.2.7}$$

and the corresponding truncated white Gaussian noise

$$W^*(t,k,n,d) := (d/n)^{1/2} \sum_{j=0}^{k-1} \xi_j \varphi_j(t). \tag{7.2.8}$$

These stochastic processes are often referred to as *frequency-limited*; the notion is clear from the fact that no high frequencies are present. Note that we use both the parameter d and the parameter n in the definition instead of just $\sigma = (d/n)^{1/2}$. The reason is that, as we shall explain later, for the equivalent models of density estimation and nonparametric regression, d plays the role of the coefficient of difficulty and n plays the role of the sample size. Moreover, if we repeat n times the transmission of a signal via k parallel Gaussian channels with $\sigma^2 = d$ and then average the outputs, the corresponding mathematical model may be written using a frequency-limited Brownian motion (7.2.7) or a frequency-limited white noise (7.2.8).

The top row of diagrams in Figure 7.3 shows three particular realizations of a frequency-limited ($k = 100$) Brownian motion (7.2.7) with $d = 1$ and n shown in the subtitles. The bottom row of diagrams shows the corresponding frequency-limited white noise. As we see, realizations of a Brownian motion may have very different shapes and create an illusion of pronounced trends or seasonal components (recall the discussion in Chapter 5). Thus, for instance, it is not a surprise that many Wall Street pundits think that Brownian motion is an excellent (and the only realistic) model for stock and bond prices. Indeed, we see a bear market (the left diagram), a market in transition (the middle diagram), and a bull market. However, all these realizations are purely stochastic, and here we know this for sure.

Below each Brownian motion the corresponding white noise is depicted. Note that here we consider a continuous-in-time white noise (a stochastic process with continuous time), whereas in Chapter 5 we discussed a discrete white noise (a time series with discrete time). It is extremely beneficial to spend some time playing around with this figure and getting used to possible patterns of Brownian motions and white noise.

Now let us return to the mathematical problem of filtering a signal from a white noise for models (7.2.4) or (7.2.5). The Fourier coefficients of the signal f may be estimated by

$$\hat\theta_j := \int_0^1 \varphi_j(t)dY(t) = \theta_j + \sigma\xi_j \quad . \tag{7.2.9}$$

If we knew the parameter σ, then we could use the universal estimator of Section 3.1 by setting the sample size n equal to the rounded-up σ^{-2}.

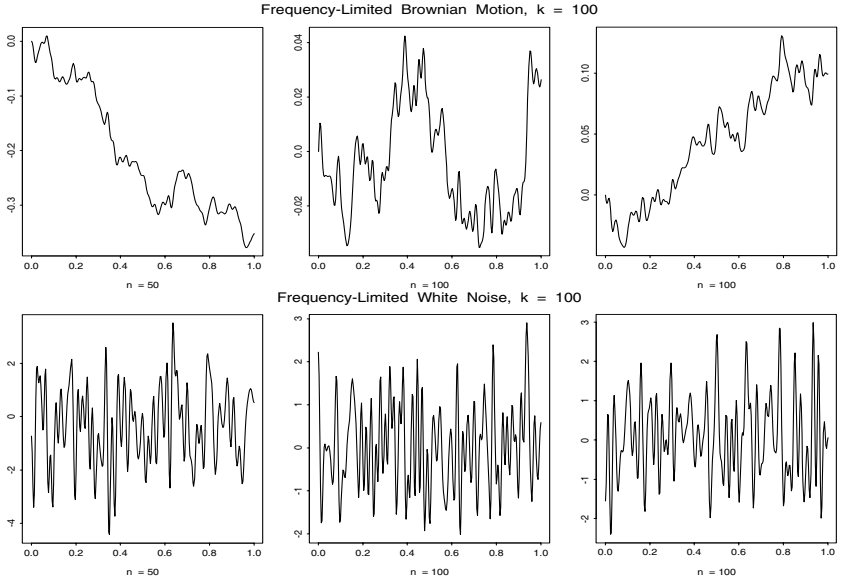

FIGURE 7.3. Realizations of a frequency-limited Brownian motion (7.2.7) and the corresponding white noise (7.2.8). {The argument *set.n* controls the set of *n*.} [*set.n = c(50,100,100), k=100, d=1*]

(Recall that for the general setting of a density estimation with coefficient of difficulty d and sample size n the formal equality $d/n = \sigma^2$ holds.) There is no way to estimate σ^2 without some additional assumptions. Probably one of the most reasonable assumptions, based on the results of Chapter 2, is that an underlying signal $f(t)$ has small power at high frequencies, that is, θ_j^2 are small for large j. Then, if (7.2.9) holds for all $j \le k$ and k is sufficiently large (recall that (7.2.7)–(7.2.8) are mathematical approximations of some real stochastic processes that are always frequency-limited, so (7.2.9) typically holds only for low frequencies), then a sample variance

$$\hat{\sigma}^2 := m^{-1} \sum_{j=k-m+1}^{k} \hat{\theta}_j^2 \tag{7.2.10}$$

may be used as an estimate of σ^2 by choosing a reasonable m.

Figure 7.4 illustrates the performance of this universal data-driven estimator. The underlying transmitted signals are our corner functions. The noise is generated by a frequency-limited Brownian motion defined at (7.2.7) and illustrated in Figure 7.3, so we know what this motion and the corresponding white noise look like. The filtering model may be considered as an analogue to all the other statistical models, and to get a "feeling" of those models just choose a corresponding coefficient of difficulty d and a sample size n. In particular, in this figure the case of $d = 1$ (which is the

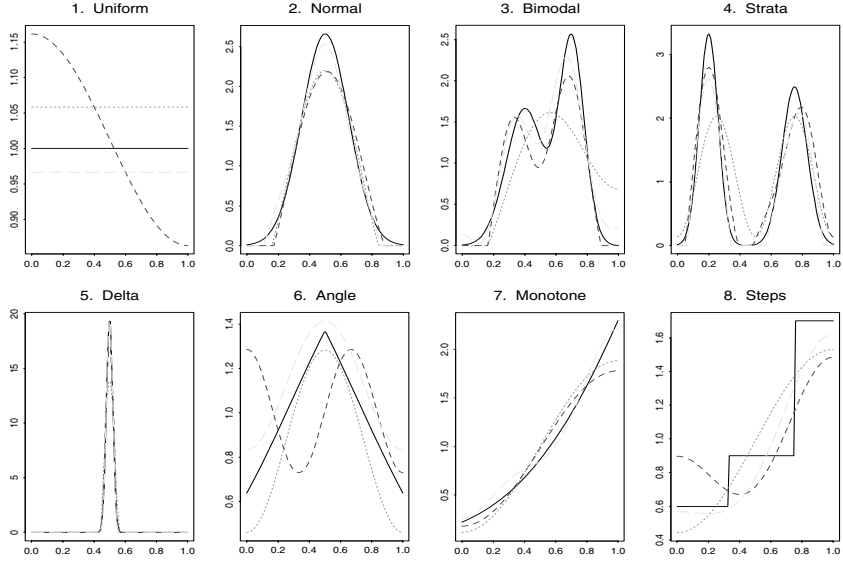

FIGURE 7.4. Filtering a signal from a frequency-limited ($k = 100$) white Gaussian noise by the universal estimate. The underlying signals are shown by solid lines. Dotted, short-dashed, and long-dashed lines correspond to the sample sizes 50, 100, and 200. The variance $\sigma^2 := d/n$ is estimated by (7.2.10) with $m = 30$; the parameters $d = 1$ and $k = 100$. {Recall that the caption of Figure 3.2 reviews the coefficients of the universal estimate.} [*set.n=c(50,100,200), d=1, k=100, m=30, cJ0 = 4, cJ1 = .5, cJM = 6, cT = 4, cB = 2*]

coefficient of difficulty for the classical density and homoscedastic regression models with additive standard normal errors) and our traditional set 50, 100, and 200 of sample sizes are considered.

As we see, the particular estimates resemble those obtained for the equivalent settings in the previous chapters. The case of the Uniform signal and $n = 100$ is especially interesting. Note that here all θ_j are equal to 0, except that $\theta_0 = 1$. And look at that peculiar "signal" (the short-dashed line) created by the white noise. Here $\sigma = 1/\sqrt{100} = 0.1$, and this looks like a small standard deviation, but as we see (and as we know from Figure 7.3) a Brownian motion may present surprises. Also note that outcomes depend on an underlying function. The same level of noise causes no problems in recognizing the "bright" signals, like the Delta, whose low-frequency Fourier coefficients are huge in comparison to this noise, and at the same time it causes problems in recognizing "dull" functions like the Uniform or the Angle whose low-frequency coefficients are relatively small.

Now let us return to the asymptotics. First, note that the asymptotic lower and upper bounds for the minimax MSE and MISE obtained in Section 7.1 are also valid for the filtering model because no assumption on

the boundedness of k has been made. On the other hand, since equations (7.2.4)–(7.2.5) are the same for any basis $\{g_j\}$ in $L_2([0,1])$ (see Exercise 7.2.4), we may use the approach of Section 7.1 for any system of coding functions that is a basis. This is one of many examples that shows why a mathematical generalization may be so fruitful.

Second, a so-called *principle of equivalence* says that if a risk with a bounded loss function is considered, then, under mild assumptions on the smoothness of f, all results that are valid for the filtering model are also valid for other statistical models including probability density estimation, regression, and spectral density estimation. We discussed earlier the technical part of the equivalence, namely, that $\sigma^2 = d/n$. In other words, the power of a noise in a particular channel mimics the ratio between the coefficient of difficulty and the sample size. There is no need for us to go into more detail about this principle, because we shall use (without proof) only one of its corollaries. Namely, under the assumption that an underlying function f is Sobolev or Lipschitz with the parameter of smoothness $\beta = r + \alpha > .5$ or under the assumption that an underlying function is analytic, the lower bounds obtained in Section 7.1 hold for the other equivalent statistical models as well, except for a local MSE result discussed below.

Remark 7.2.1 For the case of a local minimax, discussed in Remark 7.1.3, the corresponding density model is locally equivalent to the filtering model $dY(t) = f(t)dt + (f^*(t)/n)^{1/2}dB(t)$, $0 \le t \le 1$. Assuming that $f^*(t) > C > 0$, no changes occur for the MISE, but an extra factor $f^*(t_0)$ appears in the sharp constant of MSE convergence (Exercise 7.2.10).

In the next section we shall discuss simultaneously the upper bounds for three statistical models, and this will be a convincing example to support the principle of equivalence.

7.3 Rate Optimal Estimation When Smoothness Is Known

In this section we simultaneously study three statistical models: (i) filtering model where one observes a continuous-in-time output signal $Y(t)$,

$$Y(t) = \int_0^t f(u)du + n^{-1/2}B(t), \qquad 0 \le t \le 1, \tag{7.3.1}$$

and $B(t)$ denotes a standard Brownian motion; (ii) density estimation model where n iid observations $X_1, X_2, \ldots, X_n$ are drawn from a distribution with a density $f(x)$ supported on $[0,1]$; (iii) random design nonparametric regression where n pairs $\{(X_l, Y_l),\ l = 1, \ldots, n\}$ are observed, the responses $Y_l := f(X_l) + \xi_l$ where f is an estimated regression function, the predictors $X_1, \ldots, X_n$ are iid uniform on $[0,1]$, and additive errors $\xi_1, \ldots, \xi_n$ are iid standard normal.

Suppose that a function f (which may be either a signal, a probability density, or a regression function) belongs to a Lipschitz function space $Lip_{r,\alpha,L}$ of 1-periodic functions defined at (2.4.13) in Section 2.4. Recall that r is a nonnegative integer, $\alpha \in (0,1]$, and $L < \infty$. Then Theorem 7.1.1 together with Remark 7.1.2 and the equivalence principle (introduced at the end of Section 7.2) implies the following lower bound for the minimax risks of estimation of the sth derivative whenever the parameter of smoothness $\beta := r + \alpha$ is greater than 0.5:

$$\inf_{\tilde{f}_s(t_0)} \sup_{f \in Lip_{r,\alpha,L}} \text{MSE}(\tilde{f}_s(t_0), f^{(s)}(t_0)) \ge Cn^{-2(\beta-s)/(2\beta+1)}, \tag{7.3.2}$$

$$\inf_{\tilde{f}_s} \sup_{f \in Lip_{r,\alpha,L}} \text{MISE}(\tilde{f}_s, f^{(s)}) \ge Cn^{-2(\beta-s)/(2\beta+1)}. \tag{7.3.3}$$

The aim of this section is to show that if the parameter of smoothness β is known, then the same estimator is rate optimal (attains these lower bounds with perhaps different constants C) for all the statistical models. Recall that C denotes positive and in general different constants.

Let $\{\varphi_j,\ j = 0, 1, \ldots\}$ be the trigonometric basis (2.4.1), $J := 2\lfloor n^{1/(2\beta+1)} \rfloor$, and denote by $\theta_j := \int_0^1 \varphi_j(u) f(u) du$ the jth Fourier coefficient of f.

Also, let $\hat{\theta}_j$ denote an estimator of θ_j. Specific estimators for each model will be defined later.

For the case of the global risk MISE we may use the projection estimator

$$\hat{f}(x) := \sum_{j=0}^{J} \hat{\theta}_j \varphi_j(x). \tag{7.3.4}$$

To have the same estimator for the cases of both the global risk MISE and the local risk MSE, we "smooth" a projection estimator and use

$$\tilde{f}(x) := \hat{V}_J(x), \tag{7.3.5}$$

where $\hat{V}_J(x)$ is the de la Vallée Poussin sum (2.4.11) with the Fourier coefficients θ_j being replaced by the estimates $\hat{\theta}_j$.

Our first step is to prove the following proposition.

Theorem 7.3.1 *Let the mean squared error of an estimator $\hat{\theta}_j$ decrease with the parametric rate n^{-1} uniformly over $0 \le j \le J$, that is,*

$$E\{(\hat{\theta}_j - \theta_j)^2\} \le Cn^{-1}, \quad 0 \le j \le J. \tag{7.3.6}$$

Then, for any $s = 0, \ldots, r$ the sth derivative of a projection estimate (7.3.4) is globally rate optimal, that is, its MISE *attains (up to a constant) the lower bound (7.3.2), namely,*

$$\sup_{f \in Lip_{r,\alpha,L}} \text{MISE}(\hat{f}^{(s)}, f^{(s)}) \le Cn^{-2(\beta-s)/(2\beta+1)}. \tag{7.3.7}$$

The de la Vallée Poussin estimate (7.3.5) also satisfies (7.3.7), that is, it is globally rate optimal. If additionally the estimate $\hat{\theta}_j$ is unbiased, that is,

$$E\{\hat{\theta}_j\} = \theta_j, \quad 0 \le j \le J, \tag{7.3.8}$$

and for any sequence of uniformly bounded numbers $a_0, a_1, \dots, a_J$

$$E\Big\{\Big(\sum_{j=0}^{J}(\hat{\theta}_j - \theta_j)b_{js}a_j\Big)^2\Big\} \le CJ^{2s+1}n^{-1}, \tag{7.3.9}$$

where

$$b_{js} := \Big[\int_0^1 \big(\varphi_j^{(s)}(t)\big)^2 dt\Big]^{1/2}, \tag{7.3.10}$$

then the sth derivative of the estimate (7.3.5) is pointwise rate optimal, that is, for any $x_0 \in [0,1]$ its MSE *attains (up to a constant) the lower bound (7.3.3), namely,*

$$\sup_{f \in Lip_{r,\alpha,L}} \mathrm{MSE}(\tilde{f}^{(s)}(x_0), f^{(s)}(x_0)) \le Cn^{-2(\beta-s)/(2\beta+1)}. \tag{7.3.11}$$

In words, this proposition tells us that under mild assumptions, namely, if the Fourier coefficients can be estimated with the parametric rate n^{-1} in addition to some minor assumptions for the case of the pointwise approach, the estimator (7.3.5) is both globally and pointwise rate optimal for any underlying model (filtering, probability density, etc.), and the projection estimator (7.3.4) is globally rate optimal for all these models. Also, the estimates are versatile because their derivatives are rate optimal estimates of the corresponding derivatives of an estimated function. This result is exactly in the spirit of the principle of equivalence that essentially says that a good solution for one model should also be a good solution for other models.

Let us first prove this assertion (the proof is plain) and then show how to construct an estimator $\hat{\theta}_j$ that satisfies the formulated conditions.

Proof of Theorem 7.3.1 To verify (7.3.7) we use Parseval's identity and the notation of (7.3.10),

$$E\Big\{\int_0^1 (\hat{f}^{(s)}(x) - f^{(s)}(x))^2 dx\Big\} = \sum_{j=0}^{J} b_{js}^2 E\{(\hat{\theta}_j - \theta_j)^2\} + \sum_{j>J} b_{js}^2\theta_j^2. \tag{7.3.12}$$

Then we note that

$$\sum_{j=0}^{J} b_{js}^2 \le C\sum_{j=0}^{J/2} j^{2s} \le CJ^{2s+1},$$

and that $f^{(s)} \in Lip_{r-s,\alpha,L}$. The last fact, according to (2.4.18), implies $\sum_{j>J} b_{js}^2\theta_j^2 < CJ^{-2(\beta-s)}$. These results together with the assumption

(7.3.6) and $J := 2\lfloor n^{1/(2\beta+1)} \rfloor$ imply

$$\text{MISE}(\hat{f}^{(s)}, f^{(s)}) \le C[n^{-1}J^{2s+1} + J^{-2(\beta-s)}] \le Cn^{-2(\beta-s)/(2\beta+1)}.$$

The upper bound (7.3.7) is verified.

The proof of (7.3.11) is similar. We note that according to (7.3.8) the estimate $\hat{\theta}_j$ of θ_j is unbiased. Thus $E\{\tilde{f}^{(s)}(x)\} = V_J^{(s)}(x)$, where $V_J(x)$ is the de la Vallée Poussin sum (2.4.11). Using this we get

$$\text{MSE}(\tilde{f}^{(s)}(x_0), f^{(s)}(x_0)) = E\{(\tilde{f}^{(s)}(x_0) - V_J^{(s)}(x_0) + V_J^{(s)}(x_0) - f^{(s)}(x_0))^2\}$$

$$= E\{(\tilde{f}^{(s)}(x_0) - V_J^{(s)}(x_0))^2\} + (V_J^{(s)}(x_0) - f^{(s)}(x_0))^2.$$

Note that $\tilde{f}^{(s)}(x_0) - V_J^{(s)}(x_0) = \sum_{j=0}^{4J-1} w_j \varphi_j^{(s)}(x_0)(\hat{\theta}_j - \theta_j)$, where $0 \le w_j \le 1$ are smoothing weights in the de la Vallée Poussin sum (2.4.11). Thus, if we set $a_j := w_j \varphi_j^{(s)}(x_0)/b_{js}$, then a_j are uniformly bounded. This together with (7.3.9) and (2.4.15) yields (7.3.11). Also, by integrating the MSE we establish that this estimator is also globally rate optimal. We leave more detailed calculations as Exercise 7.3.1. Theorem 7.3.1 is proved.

Now we are in a position to suggest an estimator $\hat{\theta}_j$ that satisfies the assumptions (7.3.6), (7.3.8), and (7.3.9).

For the filtering model it is natural to set $\hat{\theta}_j := \int_0^1 \varphi_j(t)dY(t)$. According to (7.2.6) we get $\hat{\theta}_j = \theta_j + n^{-1/2}Z'_j$, where Z'_j are iid standard normal. Then the conditions (7.3.6), (7.3.8), and (7.3.9) obviously hold.

For the density model the natural estimator is a sample mean estimator,

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^{n} \varphi_j(X_l),$$

since $\theta_j = \int_0^1 \varphi_j(x)f(x)dx = E\{\varphi_j(X_1)\}$. Thus the conditions (7.3.6) and (7.3.8) hold. Let us verify (7.3.9). We shall do it assuming that $\beta > 0.5$. Write

$$E\Big\{\Big(\sum_{j=0}^{J}(\hat{\theta}_j - \theta_j)b_{js}a_j\Big)^2\Big\} = \sum_{j,i=0}^{J} E\{(\hat{\theta}_j - \theta_j)(\hat{\theta}_i - \theta_i)\}b_{js}a_j b_{is}a_i. \quad (7.3.13)$$

Since $\hat{\theta}_j$ is an unbiased estimate of θ_j we get

$$E\{(\hat{\theta}_j - \theta_j)(\hat{\theta}_i - \theta_i)\} = E\{\hat{\theta}_j\hat{\theta}_i\} - \theta_j\theta_i \quad (7.3.14)$$

$$= n^{-2}E\Big\{\sum_{l,m=1}^{n} \varphi_j(X_l)\varphi_i(X_m)\Big\} - \theta_j\theta_i$$

$$= n^{-2}\Big[\sum_{l=1}^{n} E\{\varphi_j(X_l)\varphi_i(X_l)\} + \sum_{l\ne m=1}^{n} E\{\varphi_j(X_l)\varphi_i(X_m)\}\Big] - \theta_j\theta_i$$

$$= n^{-1}E\{\varphi_j(X_1)\varphi_i(X_1)\} + n(n-1)n^{-2}\theta_j\theta_i - \theta_j\theta_i$$

$$= n^{-1}\Big[\int_0^1 f(x)\varphi_j(x)\varphi_i(x)dx - \theta_j\theta_i\Big].$$

Now let us recall Bernstein's inequality for Fourier coefficients of Lipschitz functions (this is the place where we use the periodicity of estimated Lipschitz functions and the assumption $\beta > 0.5$):

$$\sup_{f\in Lip_{r,\alpha,L}} \sum_{j=0}^{\infty} |\theta_j| < \infty \quad \text{if} \quad r+\alpha > 0.5. \tag{7.3.15}$$

Exercise 7.3.4 (with detailed hints) is devoted to the proof of (7.3.15).

Thus, $\sum_{j,i=1}^{J} |\theta_j\theta_i| \le C$ if $f \in Lip_{r,\alpha,L}$ and $r+\alpha > 0.5$. Also note that the product $\varphi_j(x)\varphi_i(x)$ can be written as a sum of weighted elements of the trigonometric basis. For instance, $2^{1/2}\cos(2\pi jx)2^{1/2}\cos(2\pi ix) = 2^{-1/2}[2^{1/2}\cos(2\pi(j-i)x) + 2^{1/2}\cos(2\pi(j+i)x)]$. The results imply (7.3.9). The more detailed calculations are left as Exercise 7.3.5.

For the nonparametric regression we choose a sample mean estimate

$$\hat\theta_j := n^{-1}\sum_{l=1}^{n} Y_l\varphi_j(X_l), \tag{7.3.16}$$

since $\theta_j = \int_0^1 f(x)\varphi_j(x)dx = E\{Y_1\varphi_j(X_1)\}$. Then (7.3.6) and (7.3.8) are obvious. The condition (7.3.9) is verified absolutely similarly to the case of the density model, so we leave this step as Exercise 7.3.6.

Is it possible to relax the assumption about the periodicity of an estimated function? The answer is yes, and this may be done by using the cosine basis (for $\beta < 1.5$) or by enriching a trigonometric basis by polynomial elements, as has been explained in Section 2.6.

We conclude that if the smoothness parameter β is known, then a simple series estimator is both globally and pointwise rate optimal. Another important conclusion is that these results support the principle of equivalence because a rate optimal estimator for a particular model is also rate optimal for the other models whenever an estimated function is sufficiently smooth.

7.4 Adaptive Estimation

We have shown in the previous sections that if the parameter of smoothness $\beta := r+\alpha$ is known for a Lipschitz function space $Lip_{r,\alpha,L}$, then there exist rate optimal minimax estimators. Recall that $n^{-2\beta/(2\beta+1)}$ is the minimax rate of convergence for both MISE (global risk) and MSE (pointwise risk) defined in (7.1.3) and (7.1.4).

In a majority of practical applications the parameter of smoothness is unknown. After all, nonparametric estimation is typically a first glance at the data at hand. Thus the aim of this section is to discuss estimates that are adaptive to an unknown smoothness, i.e., data-driven. We shall see that

there exists a difference between adaptive global and pointwise minimax estimation. While the MISE of adaptive estimators can attain the minimax rate, that is, it is not necessary to know the parameter β (in this case the data speak for themselves), no adaptive estimator has MSE that converges with the minimax rate for all Lipschitz functions. More precisely, the MSE of optimal adaptive estimator converges with the adaptive minimax rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$. Thus the absence of information about β slows down the convergence of minimax MSE, and the logarithmic penalty should be paid. On the other hand, it will be explained that only "few" Lipschitz functions must pay that penalty and all others may be estimated with the classical rate $n^{-2\beta/(2\beta+1)}$. The outcome is much better for analytic functions with unknown coefficient γ, where only the sharp constant in MSE convergence may be lost.

Thus in this section we discuss the cases of global and pointwise adaptive estimation separately. Also recall that due to the principle of equivalence it suffices to explain all methods only for one of the statistical models. In this section, if a model is not specified, it is assumed that it is the filtering model (7.3.1).

• **Global Estimation.** Below, several methods are discussed that can be used for optimal asymptotic estimation. Some of them are close "relatives" and some are familiar from the previous chapters, where they were used for the cases of small samples.

1 Universal Thresholding. This method may lead to a loss of a logarithmic factor in the minimax rate of MISE convergence. On the other hand, its simplicity is so appealing that it is worthwhile to begin the discussion of data-driven series estimators with this method.

The underlying idea of universal thresholding is as follows. Consider the case of parallel Gaussian channels shown in Figure 7.1 with Z_j being iid $N(0,\sigma^2)$ and assume that we transmit a function $f(t)=\sum_{j=1}^{k}\theta_j g_j(t)$ by setting $X_j=\theta_j$. Here g_j are elements of a basis in $L_2(0,1)$, for instance, the trigonometric or wavelet basis, and $\theta_j=\int_0^1 g_j(t)f(t)dt$. As we know, such a setting is equivalent to the filtering model (7.3.1), but it is simpler to explain the idea using the model of parallel channels.

Set $\sigma:=n^{-1/2}$. Then the outputs are

$$Y_j:=\theta_j+Z_j=\theta_j+n^{-1/2}Z'_j, \qquad j=1,2,\dots k, \tag{7.4.1}$$

where Z'_j are iid standard normal.

Now let us assume for a moment that $f(t)=0$ for $0\le t\le 1$, that is, all the input signals θ_j are zeros. Then an ideal procedure for the recovery of this signal should realize that the inputs are zeros and that all outputs Y_j are just noise. Under this assumption, let us consider the extreme noise

$$Z'_{(k)}:=\max_{1\le j\le k} Z'_j. \tag{7.4.2}$$

It is a simple exercise to estimate the probability that an extreme element $Z'_{(k)}$ of k iid standard normal random variables is larger than a positive constant z. Indeed, since the probability that this extreme is larger than z is less than the sum of the probabilities that each of the iid standard normal Z'_j is larger than z, we write

$$P\big(Z'_{(k)} > z\big) \le \sum_{j=1}^{k} P(Z'_j > z) = kP(Z' > z).$$

Here Z' is a standard normal random variable. Recall a well-known property of the distribution of a standard normal random variable Z':

$$P\big(Z' > z\big) < \big(z(2\pi)^{1/2}\big)^{-1} e^{-z^2/2}. \tag{7.4.3}$$

Combining the results we get the inequality

$$P\big(Z'_{(k)} > z\big) \le k\big(z(2\pi)^{1/2}\big)^{-1} e^{-z^2/2}. \tag{7.4.4}$$

Because $Z_j = n^{-1/2} Z'_j$, the last inequality implies

$$P\big(Z_{(k)} > zn^{-1/2}\big) \le k\big(z(2\pi)^{1/2}\big)^{-1} e^{-z^2/2}. \tag{7.4.5}$$

We have obtained an inequality that is the key for understanding the idea of universal thresholding. Namely, we know that for Lipschitz and Sobolev (discussed in Exercise 7.1.15) function spaces the optimal number k of channels is proportional to $n^{1/(2\beta+1)}$. Thus, even if nothing is known about β we can say that the number $k^* := \lfloor n/\ln(n) \rfloor$ is asymptotically larger than any optimal k. Then the inequality (7.4.5) implies the following rough inequality:

$$P\big(Z_{(k^*)} > (2x\ln(n))^{1/2} n^{-1/2}\big) < n^{1-x}/\ln(n), \quad x \ge 1. \tag{7.4.6}$$

Note that if $x \ge 1$, then the probability that the extreme noise is larger than $(2x\ln(n))^{1/2} n^{-1/2}$ tends to zero as n increases. Also, due to the symmetry of a standard normal distribution the same inequality holds for $\max_{1\le j\le k^*} |Z_j|$ in place of $Z_{(k^*)}$.

This result shows how to construct a data-driven estimator which performs well for the case of zero input signal. Namely, it "keeps" $\hat\theta_j$ whose absolute values are larger than $(2x\ln(n))^{1/2} n^{-1/2}$ and "kills" the others. As a result, except of an event of a small probability, if an input signal is zero, then the threshold estimate is also zero.

A *universal threshold* estimator is defined as

$$\hat f(t) := \sum_{j=1}^{n/\ln(n)} I_{\{Y_j^2 > 2c_T \hat d \ln(n) n^{-1}\}} Y_j g_j(t), \tag{7.4.7}$$

with the default $c_T = 1$ and $\hat d = 1$.

Note that this is a completely data-driven estimator for the filtering model where $\sigma^2 = n^{-1}$ and thus $d := n\sigma^2 = 1$. On the other hand, as

we explained in Section 7.1, in many cases the parameter σ^2 (the variance of a noise in a channel), or equivalently the coefficient of difficulty d, is unknown. In this case the estimator (7.2.10) may be used. For a filtering model with a normal noise, as an example, the sample variance $\hat{\sigma}^2 := [(2/n)\sum_{n/2<j\le n} Y_j^2]^{1/2}$ may be used for an asymptotic study. Another choice is to use a robust estimator, for instance, a rescaled sample median of absolute deviations discussed in Section 4.1.

Let us now explain why the universal thresholding may imply the loss of a logarithmic factor in the MISE convergence. The reason is that a function f may have Fourier coefficients θ_j with absolute values just a bit less than the minimal universal threshold level $\sqrt{2c_T \ln(n) n^{-1}}$. For instance, consider the case of the Sobolev function class $W_{\beta,Q}$ defined in (2.4.19). As we know from Exercise 7.1.15, the minimax MISE$(\tilde{f}, f)$ over the Sobolev space converges as $n^{-2\beta/(2\beta+1)}$ (this is the same rate as for a Lipschitz class with the same parameter of smoothness). Then a simple calculation (Exercise 7.4.3) shows that for any positive c,

$$\sup_{f\in W_{\beta,Q}} \sum_{j=1}^{k^*} I_{\{\theta_j^2 < c\ln(n)n^{-1}\}}\theta_j^2 \;\ge\; C(n/\ln(n))^{-2\beta/(2\beta+1)}. \tag{7.4.8}$$

As a result, some Sobolev functions will be "killed" by the universal thresholding while their integrated squared bias (and thus MISE) is proportional to the right-hand side of (7.4.8). A similar result holds for Lipschitz functions (Exercise 7.4.4).

Thus, the universal thresholding may imply the loss of a logarithmic factor in MISE convergence, and we have seen in Section 3.3 that a hard threshold estimator does perform slightly worse than a smoothing estimator for small sample sizes. On the other hand, the simplicity of this procedure is so appealing and in some cases the issue of the extra logarithmic factor is so minor that this method deserves to be included in our "tool-box" of adaptive estimators. We shall also see that this method is an optimal one for adaptive minimax pointwise estimation.

2 Empirical Risk Minimization. The underlying idea of this method is as follows: Consider the MISE of an estimator, understand which part of the risk is affected by an unknown parameter, estimate that part of the risk by a statistic, and then choose the value of the parameter that minimizes that statistic.

As an example, consider the case of a projection estimator (7.3.4) where the cutoff J is the only parameter to be adaptively chosen. To make the setting more general, consider the case where $\sigma^2 = dn^{-1}$ and recall that for our basic models $d = 1$.

According to (7.3.12), the MISE of this estimator is the sum of the variance and the integrated squared bias term, that is,

$$\text{MISE}(\hat{f}, f) = n^{-1}d(J+1) + \sum_{j>J} \theta_j^2.$$

Here we used the formula $E\{(\hat{\theta}_j - \theta_j)^2\} = E\{(\theta_j + (d/n)^{1/2}Z'_j - \theta_j)^2\} = dn^{-1}$. Using Parseval's identity $\int_0^1 f^2(t)dt = \sum_{j=0}^J \theta_j^2 + \sum_{j>J} \theta_j^2$, we get

$$\text{MISE}(\hat{f}, f) = \sum_{j=0}^{J}(dn^{-1} - \theta_j^2) + \int_0^1 f^2(t)dt. \tag{7.4.9}$$

Since $\int_0^1 f^2(t)dt$ is a constant for a particular underlying f, we see that an optimal cutoff J that minimizes (7.4.9) is also the one that minimizes $\sum_{j=0}^J (dn^{-1} - \theta_j^2)$. This is the part of the risk that may be estimated. As an example, let us use the unbiased estimate $\hat{\theta}_j^2 - \hat{d}n^{-1}$ of θ_j^2. This implies the following procedure for choosing a data-driven cutoff:

$$\hat{J} := \text{argmin}_{0 \le J \le J_n^*} \sum_{j=0}^{J}(2\hat{d}n^{-1} - \hat{\theta}_j^2). \tag{7.4.10}$$

Here $\hat{d}$ is the estimator discussed earlier, and $\hat{d}$ is equal to 1 for our basic models. Also, J_n^* is the maximal cutoff. For instance, one can always set $J_n^* := \lfloor n/\ln(n) \rfloor$. Also recall that the function $\text{argmin}_{J \in A}\Psi(J)$ returns the value of J from the set A that minimizes $\Psi(J)$.

Then the corresponding adaptive projection estimator is

$$\tilde{f}(t) := \sum_{j=0}^{\hat{J}} \hat{\theta}_j \varphi_j(t). \tag{7.4.11}$$

Recall that this adaptive estimator has been thoroughly studied for samples of small sizes; in Section 3.3 we referred to it as raw truncated. Overall, this estimator performed well.

This finishes our discussion of the empirical risk minimization procedure.

Before considering the next procedure of adaptation, let us pause for a moment and discuss the following question. In the previous method our aim was to find a data-driven cutoff J of a projection estimator. But is a projection estimator, that mimics a partial sum (also called a linear approximation), always rate optimal? We know that this is the case for smooth functions, say Lipschitz. But this may be not the case for nonsmooth functions, for instance, for functions with bounded total variation which have jumps. In this case a *nonlinear approximation*,

$$f_{M(J)}(x) := \sum_{j \in M(J)} \theta_j \varphi_j(x), \tag{7.4.12}$$

may outperform the *linear approximation* $f_J := \sum_{j=0}^{J} \theta_j \varphi_j(x)$. Here $M(J)$ is a set of $J+1$ natural numbers, that is, a set of cardinality $J+1$.

Note that a nonlinear approximation always dominates the linear one because the choice $M(J) = \{0, 1, \dots, J\}$ implies the linear approximation. It is the luck of smooth functions that linear approximations are optimal for them.

We have considered a universal thresholding that mimics a nonlinear approximation. The next adaptive method, which is closely related to empirical risk minimization, allows one to mimic (7.4.12) straightforwardly.

3 Penalization. The relation (7.4.9) implies that including a jth term of a series estimate increases its variance on dn^{-1} and decreases its integrated squared bias on θ_j^2. Thus, the choice of an optimal set M may be achieved by solving the minimization problem

$$\hat{M} := \operatorname{argmin}_{M(J)} \Big\{ \operatorname{pen}(J) \hat{d} n^{-1} - \sum_{j \in M(J)} \hat{\theta}_j^2 \ , \ 0 \le J \le J_n^* \Big\}. \qquad (7.4.13)$$

Here $\operatorname{pen}(J)$ is a *penalty* function, and the minimization is considered over both J and sets $M(J)$.

For instance, if $\operatorname{pen}(J) = C(J+1)$, then we get an analogue of the empirical risk minimization method. In some cases a larger penalty function may be recommended. For instance, choosing $\operatorname{pen}(J) = C \ln(J) J$ makes the penalization similar to the universal thresholding.

It is possible to show that by choosing an appropriate penalization function this adaptation leads to rate optimal estimation over a wide variety of function spaces and bases including wavelets.

Different examples of penalization will be considered in Section 8.6.

4 Cross-Validation. Here the basic idea is to calculate an estimate based on a part of the data and then choose parameters of the estimate that give the best fit to the rest of the data. This is a method that is simpler to explain using the regression model (iii) defined at the beginning of Section 7.3.

Let $\hat{f}(x, \lambda)$ be any estimator of an underlying regression function that depends on a parameter λ. For instance, λ can be a cutoff of a projection estimator, that is, $\lambda = J$. Then denote by $\hat{f}_{-l}(x, \lambda)$ the same estimator, only calculated without using the lth pair (X_l, Y_l) of observations. Then it is natural to expect that $\hat{f}_{-l}(X_l, \lambda)$ should well approximate the response Y_l if λ is close to an optimal value λ^*. This leads to a cross-validation least-squares procedure of choosing the parameter λ (the so-called *leave-one-out method*),

$$\hat{\lambda} := \operatorname{argmin}_{\lambda} \sum_{l=1}^{n} (Y_l - \hat{f}_{-l}(X_l, \lambda))^2. \qquad (7.4.14)$$

Absolutely similarly, a leave-m-out method is defined.

A cross-validation typically implies a rate optimal adaptive estimation. An example of how to apply this method to the case of a density model is discussed in Section 8.10.

5 Efromovich–Pinsker Block Shrinkage. This is a data-driven estimator whose underlying idea is to mimic the linear smoothing oracle discussed in Sections 3.2–3. We shall see that this is possible for the case of smooth functions (like Lipschitz and analytic), and it is also a reasonable approach for estimation of spatially inhomogeneous functions. The theoretical trademark of this data-driven estimator is that it is sharp minimax (efficient) over Sobolev and analytic function classes. Its practical trademark is the simplicity—no optimization problem is involved.

Let us begin the discussion with a parametric setting and review of results obtained in Appendix A and Sections 3.2–3. Suppose that one would like to recover a parameter θ based on an observation $Y := \theta + n^{-1/2}Z'$, where Z' is a standard normal random variable. Also assume that the only allowed method of recovery is a shrinkage estimator $\hat{\theta} := \lambda Y$, where λ is a constant. Write the mean squared error of this estimate,

$$E\{(\lambda Y - \theta)^2\} = \lambda^2 E\{Y^2\} - 2\lambda E\{Y\}\theta + \theta^2. \tag{7.4.15}$$

A shrinkage (smoothing) weight λ^* that minimizes the mean squared error (7.4.15) is

$$\lambda^* \;=\; \frac{E\{Y\}\theta}{E\{Y^2\}} \;=\; \frac{\theta^2}{\theta^2 + n^{-1}}. \tag{7.4.16}$$

As we see, the optimal weight is the ratio of a squared estimated signal θ^2 to the second moment of an observed signal Y, or in other words, λ^* is the ratio between the powers of the input and output signals. Also

$$E\{(\lambda^* Y - \theta)^2\} = n^{-1}\,\frac{\theta^2}{\theta^2 + n^{-1}} = n^{-1}\lambda^*. \tag{7.4.17}$$

Let us compare the optimal shrinkage with a hard-threshold shrinkage where λ may be either 0 or 1 (for instance, this is the idea of a penalization method or a hard-threshold method). If $\lambda = 0$, then the estimate is zero, and the mean squared error is $\theta^2 \ge n^{-1}\lambda^*$ with equality iff $\theta = 0$. If $\lambda = 1$, then the estimate is equal to Y and the mean squared error is $n^{-1} > n^{-1}\lambda^*$. In short, the optimal shrinkage does a superb job.

Let us extend this parametric result to a nonparametric setting. Consider a smoothing filter (shrinkage estimator)

$$\tilde{f}(t) := \sum_{j=0}^{\infty} \lambda_j Y_j \varphi_j(t), \tag{7.4.18}$$

where λ_j are constants (so-called shrinkage weights or coefficients). Then Parseval's identity yields

$$\text{MISE}(\tilde{f}, f) = \sum_{j=0}^{\infty} E\{(\lambda_j Y_j - \theta_j)^2\}. \tag{7.4.19}$$

Thus, according to (7.4.15)–(7.4.16), the optimal shrinkage weights are

$$\lambda_j^* = \frac{\theta_j^2}{\theta_j^2 + n^{-1}}, \tag{7.4.20}$$

and according to (7.4.17), the MISE of the optimal smoothing filter

$$\tilde{f}^*(t) := \sum_{j=0}^{\infty} \lambda_j^* Y_j \varphi_j(t) \tag{7.4.21}$$

can be calculated by the formula:

$$\text{MISE}(\tilde{f}^*, f) = n^{-1} \sum_{j=0}^{\infty} \lambda_j^*. \tag{7.4.22}$$

Recall that (7.4.21) is the analogue of the linear oracle for the density model discussed in Sections 3.2–3, and for the case of small samples we have used a naive mimicking (7.4.20) by estimates $\hat{\lambda}_j := (Y_j^2 - n^{-1})_+ / Y_j^2$.

On the other hand, we explained in Section 3.3 that asymptotically the naive mimicking may lead to inconsistent estimation. The reason for this negative conclusion is that $E\{(Y_j - \theta_j)^2\} = n^{-1}$. Thus, if θ_j^2 is close to n^{-1}, then the naive estimator $\hat{\lambda}_j$ of λ_j^* becomes inconsistent.

Thus, something else should be suggested for mimicking (7.4.21). Here we consider the Efromovich–Pinsker block shrinkage procedure. Let us explain the idea for the case of Lipschitz functions. We know that a projection estimator $\tilde{f}(t) = \sum_{j=0}^{n^{1/(2\beta+1)}} Y_j \varphi_j(t)$ is rate optimal. Note that it uses unit weights for low frequencies and zero weights for high frequencies; in other words, all Fourier coefficients are grouped into two blocks, and then the same shrinkage is applied to all Fourier coefficients from a block. Thus, if we consider a net of blocks that includes (or approximates) those two blocks for any β, then optimal shrinkage within each block may lead to an optimal data-driven estimation.

To explore this idea, let us divide the set of natural numbers (frequencies) $\{0, 1, 2, \ldots\}$ into a sequence of blocks G_m, $m = 1, 2, \ldots$. For instance, one may set $G_1 = \{0\}$, $G_2 = \{1, 2\}$, $G_3 = \{3, 4, 5, 6, 7, 8\}$, etc. It is not necessary that these blocks are clusters (include only neighbors). Also, blocks may depend on n.

Consider an estimator that applies the same shrinkage weight w_m to all Fourier coefficients within a corresponding block G_m, that is,

$$\hat{f}(t) := \sum_{m=1}^{\infty} w_m \sum_{j \in G_m} Y_j \varphi_j(t). \tag{7.4.23}$$

Then a direct calculation shows that an optimal weight w_m^* that minimizes the MISE of $\hat{f}$ is (compare with (7.4.20))

$$w_m^* = \frac{|G_m|^{-1} \sum_{j \in G_m} \theta_j^2}{|G_m|^{-1} \sum_{j \in G_m} \theta_j^2 + n^{-1}}. \tag{7.4.24}$$

Here $|G_m|$ denotes the number of elements in the block G_m (i.e., the *cardinality* of G_m). Note that if $|G_m| = 1$, then (7.4.24) becomes (7.4.20).

The optimal weight (7.4.24) for a block of Fourier coefficients looks similarly to the shrinkage weight (7.4.20) for a singular coefficient, but the crucial difference is that in (7.4.24) the optimal weight depends on the mean value of squared Fourier coefficients from the block (i.e., on the mean power of the input signals). This is the key point of any block procedure because the mean power may be estimated better than a single one. More precisely, according to Exercise 7.4.11,

$$E\Big\{\Big(|G_m|^{-1} \sum_{j \in G_m} (Y_j^2 - n^{-1}) - |G_m|^{-1} \sum_{j \in G_m} \theta_j^2\Big)^2\Big\}$$

$$= |G_m|^{-1} n^{-1} \Big[2n^{-1} + 4|G_m|^{-1} \sum_{j \in G_m} \theta_j^2\Big]. \tag{7.4.25}$$

The larger $|G_m|$ is, the better the estimation of w_m^*. On the other hand, the larger a block is, the farther w_m^* is from the optimal individual shrinkage (7.4.20). Thus the choice of blocks is a tradeoff between mimicking optimal singular shrinkage and the better accuracy of estimating w_m^*.

The Efromovich–Pinsker block shrinkage estimator is defined as

$$\hat{f}(t) := \sum_{m=1}^{M} \hat{w}_m \sum_{j \in G_m} \hat{\theta}_j \varphi_j(t), \tag{7.4.26}$$

where

$$\hat{w}_m := \frac{\hat{\Theta}_m}{\hat{\Theta}_m + \hat{d} n^{-1}} I_{\{\hat{\Theta}_m > t_m \hat{d} n^{-1}\}}. \tag{7.4.27}$$

Here $\hat{\theta}_j = Y_j$, $\hat{d} = 1$, t_m are threshold coefficients, which may depend on n, and M is a sufficiently large sequence in n, as an example, such that $\sum_{m>M} \sum_{j \in G_m} \theta_j^2 \le C n^{-1} / \ln(n)$, and

$$\hat{\Theta}_m := |G_m|^{-1} \sum_{j \in G_m} (\hat{\theta}_j^2 - \hat{d} n^{-1}) \tag{7.4.28}$$

is an unbiased estimate of $\Theta_m := |G_m|^{-1} \sum_{j \in G_m} \theta_j^2$.

Note that the used shrinkage $\hat{w}_m$ is the product of a continuous shrinkage and a hard thresholding. Below we shall explore the effect of these two factors on the estimation.

Let us make several remarks about the choice of blocks G_m and threshold levels t_m. For asymptotically efficient (when both the rate and constant of MISE convergence are asymptotically optimal) estimation of smooth functions, the threshold levels should decay, while for a rate optimal estimation it suffices for them to be bounded. If the noise in channels has a bounded eighth moment, then the boundness of $\sum_{j=1}^{M} |G_m|^{-1} t_m^{-3}$ implies sharp mimicking of the block shrinkage oracle (7.4.23–24). For normal noise this assumption may be relaxed, and for instance, it suffices that the sequence $\sum_{j=1}^{M} \exp(-c|G_m|t_m^2)$ be bounded for a sufficiently small c.

Second, let us check how block shrinkage works for the case of wavelets and small sample sizes (a review of Sections 2.5 and 4.4 is recommended), and let us get some feeling on how two factors in the used shrinkage (7.4.27) affect the estimation. Using the setting and notation of Section 4.4, the block shrinkage estimator may be written as

$$\hat{f}(x) = \sum_{k=1}^{n/2^{j_0}} \hat{s}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=1}^{j_0} \sum_{m=1}^{n/(2^j L_{j,n})} [\hat{D}_{j,m}/(\hat{D}_{j,m} + \hat{\sigma}^2 n^{-1})] \times I_{\{\hat{D}_{j,m} > t_{j,n}\hat{\sigma}^2 n^{-1}\}} \sum_{k \in G_{j,m,n}} \hat{d}_{j,k}\psi_{j,k}(x), \tag{7.4.29}$$

where $G_{j,m,n} := \{k : \ L_{j,n}(m-1) < k \le L_{j,n}m\}$ are the blocks for the jth resolution scale, which have the same length $L_{j,n} := |G_{j,m,n}|$, and $\hat{D}_{j,m} := L_{j,n}^{-1} \sum_{k \in G_{j,m,n}} (\hat{d}_{j,k}^2 - \hat{\sigma}^2 n^{-1})$.

To assess the effect of blocks and threshold levels on estimation, as well as to evaluate the robustness of the Efromovich–Pinsker procedure with respect to the choice of blocks and threshold levels, consider two particular sets of these parameters. The first particular estimator (we shall refer to it as an estimator with "increasing blocks and $t = 0$") has $L_{j,n} = b_n 2^{\lfloor (j_0 - j)/3 \rfloor}$ and $t_{j,n} = 0$; here b_n is the largest dyadic (i.e., $2^k, k = 1, 2, \ldots$) number that is at most $\log_2(n)$. The second particular estimator (we shall refer to it as one with "constant blocks and $t = 5$") has $L_{j,n} = b_n$ and $t_{j,n} = 5$.

Thus, the first set of parameters implies that a continuous smoothing is the primary factor in the shrinkage (7.4.27), and one may expect that the corresponding estimator will perform well, because blocks increase as the scales become finer. The second set implies that a hard threshold is the primary factor, and one may expect that the corresponding estimator will perform well, because the threshold level is sufficiently large (but note that it is a constant while, for instance, the level in the universal thresholding procedure is $2\ln(n)$; see subsection 1).

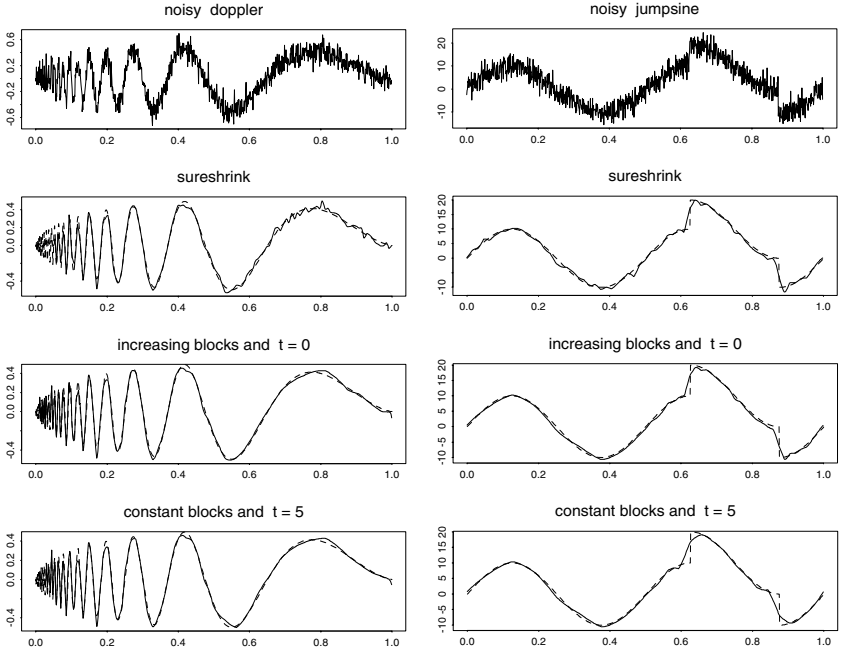

FIGURE 7.5. Performance of two particular block shrinkage wavelet estimators. The SureShrink estimate of the toolkit S+WAVELETS is given as a benchmark. Estimates are shown by solid lines and the underlying signals by dashed lines. The sample size is $n = 1024$, the signal-to-noise ratio is 3, the wavelet is Symmlet 8. {Do not forget to download the wavelet toolkit by calling > **module(wavelets)**. The arguments *n, snr, set. signal, wavelet, t1*, and *t2* control the sample size, signal-to-noise ratio, two underlying signals, the wavelet, and threshold levels for the first and the second estimator.} [*n=1024, snr=3, set.signal=c("doppler", "jumpsine"), wavelet= "s8", t1=0, t2=5*]

Figure 7.5 shows particular estimates of the familiar "doppler" and "jumpsine" signals for the case of $n = 1024$ and signal-to-noise ratio 3. As in Section 4.4, j_0 is equal to 6 and the Symmlet 8 wavelet is used. As a reference estimator, the SureShrink estimate is also shown. As we see, the block shrinkage estimators perform reasonably well for these particular simulations. This shows robustness of the procedure with respect to the choice of its parameters.

Intensive Monte Carlo simulations for different sample sizes and signal-to-noise ratios support this conclusion. The first estimator (the one with zero threshold level) performs exceptionally well in terms of integrated squared errors (ISE) but slightly worse than both the second block shrinkage estimator and SureShrink in terms of data compression (the latter is not a surprise). The second block shrinkage estimator is comparable with

SureShrink in terms of ISE and yields better data compression. Repeated simulations also show that the first estimator (with zero threshold level) periodically produces estimates whose smooth parts are contaminated by blocks of shrunk noise, which ideally should be "killed." The asymptotic theory discussed below explains why a nonzero thresholding is required for optimal estimation.

One more remark is must be made. These two examples show that the choice of blocks and threshold levels is rather flexible, but extremes should be avoided. For instance, the choice of blocks that are whole resolution scales implies poor estimation.

Let us finish the discussion of the Efromovich–Pinsker block shrinkage estimator by presenting an asymptotic proposition that is an "adaptive" version of Theorem 7.1.1. Note that the noise may be non-Gaussian.

Theorem 7.4.1 *Consider the transmission of an analytic function $f(x)$ via k parallel channels shown in Figure 7.1. It is assumed that the noise Z_j is zero mean, $E\{Z_j^2\} = \sigma^2 := n^{-1}$, $E\{Z_j^8\} < Cn^{-4}$, $1 \le j \le k$, and $k > (\ln(n))^2$. Set $\hat\theta_j := Y_{j+1}$, and choose M, blocks G_m, and threshold levels t_m such that elements of G_m are smaller than elements of G_{m+1},*

$$|G_m|^{-1}t_m^{-3} \to 0 \quad \text{and} \quad t_m \to 0 \quad \text{as} \quad m \to \infty, \tag{7.4.30}$$

$$\sum_{m>M}\sum_{j\in G_m} e^{-\gamma j} = o_n(1)\ln(n)n^{-1}, \ \sum_{m=1}^{M} |G_m|^{-1}t_m^{-3} = o_n(1)\ln(n), \tag{7.4.31}$$

and for any positive c there exists a sequence $m(n,c)$ such that the series $\sum_{m=1}^{m(n,c)} |G_m|/\ln(n) \to c$ as $n \to \infty$. Then the Efromovich–Pinsker block shrinkage estimator (7.4.26) is a versatile sharp minimax estimator satisfying (7.1.11) with $\sigma^2 = n^{-1}$, that is, for any $s = 0, 1, \ldots$,

$$\sup_{f\in A_{\gamma,Q}} \text{MISE}(\hat f^{(s)}, f^{(s)}) = P_{s,\gamma}\,(\ln(n^{1/2}))^{2s+1}n^{-1}(1+o_n(1)). \tag{7.4.32}$$

For instance, $M = \lfloor \ln(n) \rfloor$, the set $G_1 = \{0\}$, $G_m = \{(m-1)(m-2)+1, \ldots, m(m-1)\}$, $m > 1$, and threshold levels $t_m = 1/\ln(m+1)$ satisfy the assumption of Theorem 7.4.1.

Proof. The proof consists of several steps whose detailed verification is left as Exercise 7.4.15.

Step 1. A direct calculation, which is similar to verification of (7.1.11), shows that the estimate

$$\tilde f_\gamma(x) := \hat\theta_0 + \sum_{m=1}^{M} w_{\gamma,m} \sum_{j\in G_m} \hat\theta_j \varphi_j(x) \tag{7.4.33}$$

satisfies (7.4.32) if $w_{\gamma,m} = 1$ for $m \le m(n,\gamma^{-1})$ and $w_{\gamma,m} = 0$ otherwise. Note that this estimate depends on γ, i.e., it is a pseudo estimate (it depends on an underlying function space).

Step 2. According to (7.4.23)–(7.4.24), the estimate (7.4.33) is dominated by the oracle

$$\tilde{f}^*(x) := \hat{\theta}_0 + \sum_{m=1}^{M} w_m^* \sum_{j \in G_m} \hat{\theta}_j \varphi_j(x). \tag{7.4.34}$$

The estimate does not depend on γ, but the weights w_m^* depend on an underlying function; thus we refer to it as an oracle.

Step 3. The inequality $(a+b)^2 \le (1+\rho)a^2 + (1+\rho^{-1})b^2$, $\rho > 0$ implies

$$\mathrm{MISE}(\hat{f}^{(s)}, f^{(s)}) \le (1+\rho)\mathrm{MISE}(\tilde{f}^{*\,(s)}, f^{(s)}) + (1+\rho^{-1})\mathrm{MISE}(\hat{f}^{(s)}, \tilde{f}^{*\,(s)}). \tag{7.4.35}$$

Note that the first term satisfies (7.4.32) if $\rho \to 0$ as $n \to \infty$.

Step 4. Consider the second term in (7.4.35) by applying Parseval's identity,

$$\mathrm{MISE}(\hat{f}^{(s)}, \tilde{f}^{*\,(s)}) = \sum_{m=1}^{M} E\Big\{(w_m^* - \hat{w}_m)^2 \sum_{j \in G_m} \hat{\theta}_j^2 \int_0^1 (\varphi_j^{(s)}(x))^2 dx\Big\}. \tag{7.4.36}$$

Step 5. A direct calculation shows that

$$E\Big\{(w_m^* - \hat{w}_m)^2 \sum_{j \in G_m} \hat{\theta}_j^2 \int_0^1 (\varphi_j^{(s)}(x))^2 dx\Big\}$$

$$\le Cn^{-1}\Big(\sum_{l=1}^{m} |G_l|\Big)^{2s} |G_m| [w_m^*(t_m^{1/2} + (|G_m| t_m^3)^{-1/2}) + |G_m|^{-2} t_m^{-3}]. \tag{7.4.37}$$

Step 6. Choose a slowly decreasing $\rho \to 0$ as $n \to 0$, and then steps 1–5 yield (7.4.32). Theorem 7.4.1 is proved.

6 SureShrink Wavelet Estimator. This is an adaptive estimator that is rate optimal over the Besov space B^{σ}_{pqQ}, $p, q \ge 1$, $\sigma - \frac{1}{2} + p^{-1} > 0$, defined in (2.5.4). The key idea of this adaptive procedure is as follows. Consider a wavelet expansion (2.5.2)

$$f(t) = \sum_k \kappa_{j_1,k} \phi'_{j_1,k}(t) + \sum_{j=j_1}^{\infty} \sum_k \theta_{j,k} \psi'_{j,k}(t), \quad 0 \le t \le 1. \tag{7.4.38}$$

Here the sum is over all integer k, and recall that for a wavelet with bounded support the number of nonzero wavelet coefficients on a resolution scale j is at most $C2^j$.

Set J to be the maximal integer such that $2^J < n/\ln(n)$. Denote by $\hat{\kappa}_{j_1,k}$ and $\hat{\theta}_{j,k}$ estimates of the corresponding wavelet coefficients; for instance, the estimates suggested in Section 7.3 can be used with the obvious replacement of the elements of the trigonometric basis by elements of a wavelet basis.

Then it is possible to show that there exist threshold levels $\lambda_{n,j_1}, \lambda_{n,j_1+1}, \dots, \lambda_{n,J}$, depending on parameters of the underlying Besov space, such that

a soft-threshold estimator

$$\hat{f}(t) := \sum_k \hat{\kappa}_{j_1,k}\phi'_{j_1,k}(t) + \sum_{j=j_1}^{J}\sum_k \mathrm{sgn}(\hat{\theta}_{j,k})(|\hat{\theta}_{j,k}| - \lambda_{n,j})_+\psi'_{j,k}(t), \quad (7.4.39)$$

is rate minimax over the Besov space. Here $\mathrm{sgn}(x)$ denotes the sign of x, and recall that $(x)_+ := \max(0, x)$. In other words, a *soft-threshold* shrinkage $\mathrm{sgn}(\hat{\theta}_{j,k})(|\hat{\theta}_{j,k}| - \lambda_{n,j})_+$ may imply a rate minimax estimation.

The important part of this theoretical result is that optimal threshold levels may be the same for all wavelet coefficients from a resolution scale. Thus, these levels may be estimated, for instance, by the empirical risk minimization procedure, for every resolution scale one at a time.

This is the key idea of the adaptive procedure SureShrink. An unbiased estimate of the mean squared error of a soft-threshold shrinkage estimator was developed in the 1980s by Stein. This explains the abbreviation SURE, which stands for Stein's unbiased risk estimation.

For a jth resolution scale and a filtering model with $\hat{d} = 1$, the empirical risk is defined as

$$\mathrm{SURE}(\{\hat{\theta}_{j,k}\}, \lambda) := \sum_k \big[n^{-1} - 2n^{-1}I_{\{|\hat{\theta}_{j,k}|\le\lambda\}} + \min(\hat{\theta}^2_{j,k}, \lambda^2)\big],$$

and then an adaptive threshold $\hat{\lambda}_{n,j}$, which minimizes the SURE, is used in the soft-threshold procedure (7.4.39).

SureShrink is a built-in S-function of the S+WAVELETS toolkit, and we saw its performance for small sample sizes in Section 4.4 and in Figure 7.5. This is a good estimator for both small and large sample sizes, and it is considered as a benchmark for other data-driven wavelet estimators.

7 Universal Wavelet Estimator. The universal data-driven estimator, discussed in detail in Chapters 3–5 for the cosine basis, may be used for a wavelet basis as well. A particular estimator was defined in (4.4.2)–(4.4.5). This estimator matches properties of SureShrink over Besov function spaces. Moreover, if Efromovich–Pinsker block shrinkage is applied to its linear part, then this estimator becomes sharp minimax over Sobolev function spaces. This makes this estimator particularly attractive for estimation of monotone functions (or similar order-restricted functions), since it guarantees both an optimal rate of convergence and, if an underlying function is smooth enough, a sharp minimax convergence.

Now, when we know different methods of adaptation, it is apparent that the adaptive procedure (4.4.5) is just an empirical risk minimization discussed in subsection 2 (to see this, compare (4.4.5) with (7.4.10)).

8 Block Threshold Estimator. This is an estimator (7.4.26) (or its wavelet version (7.4.29)) with weights $\hat{w}_m := I_{\{\sum_{j\in G_m}(\hat{\theta}^2_j - c_T\hat{d}n^{-1})>0\}}$. If we just for a moment forget about blocks, then, using the terminology of Sections 3.2–3.3, this estimator is a hard-threshold one. Also, if a relatively large c_T is used by both block shrinkage and block threshold estimators,

then their performance is almost identical. Thus, the estimates "constant blocks and $t = 5$" in Figure 7.5 show how a wavelet block threshold estimator may perform (note that in our notation $c_T = t + 1$). Overall, this is a simple and reliable estimator with a good data compression property.

To shed light on the block threshold approach, consider the case of a filtering model (7.3.1) with $f(t) = 0$ (no input signal) and constant blocks of a length $L := |G_m|$. In this case the ideal solution is to "kill" signals in all blocks, and this occurs if $\sum_{j \in G_m} \hat{\theta}_j^2 \le c_T |G_m| n^{-1}$ for $m = 1, \dots, M$. For the case of a zero input signal we have $\hat{\theta}_j = Z_j$ where Z_j are iid normal, $E\{Z_j\} = 0$, and $\mathrm{Var}(Z_j) = n^{-1}$. Thus we should explore $\sum_{j \in G_m} Z_j^2$.

Let $\xi_1, \xi_2, \dots$ be iid standard normal. A direct calculation shows that

$$E\Big\{ \exp\Big(\sum_{l=1}^{L} \xi_l^2/4 \Big) \Big\} = \exp(L \ln(2)/2).$$

This relation yields

$$P\Big(\sum_{l=1}^{L} \xi_l^2 > c_T L \Big) \le E\Big\{ \exp\Big(\sum_{l=1}^{L} \xi_l^2/4 - c_T L/4 \Big) \Big\} = \exp(-L(c_T - \ln(4))/4). \tag{7.4.40}$$

Thus,

$$P\Big(\max_{m \in \{1,\dots,M\}} \sum_{j \in G_m} Z_j^2 > c_T L n^{-1} \Big) \le M \exp(-L(c_T - \ln(4))/4).$$

This inequality explains why a block threshold estimator may imply a reliable filtering a pure noise signal. For instance, consider blocks of a logarithmic length, say, $L = \lfloor \ln(n) \rfloor$. Recall that any series estimate is based on at most $n/\ln(n)$ Fourier (wavelet) coefficients, so $M < n/\ln^2(n)$. Thus, if $c_T - \ln(4) > 4$, then the probability of a not ideal estimation decreases as n^{-c}, $c > 0$. This example motivated the choice of the parameters for the "constant blocks and $t = 5$" estimate used in Figure 7.5.

This discussion of the case of a zero input signal reveals a striking similarity between how the universal threshold and block threshold adaptive estimates deal with zero input signals. On the other hand, the difference between the statistical properties of these estimators is also striking: While the minimax MISE of the universal threshold estimate loses the logarithmic factor, the block threshold estimator is rate optimal over a wide spectrum of spatially inhomogeneous and smooth functions.

9 Bias–Variance Tradeoff. Let us begin an explanation of this method via a particular case of two Lipshitz spaces $L^1 := Lip_{r_1,\alpha_1,L_1}$ and $L^2 := Lip_{r_2,\alpha_2,L_2}$ with different smoothness parameters $\beta_1 > \beta_2$ (recall that $\beta := r + \alpha$). In this case the projection estimator $\hat{f}_J$, defined at (7.3.4) and "equipped" with two particular cutoffs $J_1 < J_2$, $J_s := \lfloor n^{1/(2\beta_s+1)} \rfloor$, may estimate Lipschitz functions from L^1 and L^2 with optimal MISE convergence whenever β is given. Thus the only issue is how to choose a right estimate

(cutoff). The estimate $\hat f_{J_1}$ is simpler and implies better data compression; thus let us try to understand when it is worthwhile to use $\hat f_{J_2}$.

The MISE of a projection estimate $\hat f_J$ may be written as a variance term plus an integrated squared bias term,

$$\mathrm{MISE}(\hat f_J, f) = (J+1)n^{-1} + \sum_{j>J} \theta_j^2.$$

As we know, $\sum_{j=J_1+1}^{J_2} \theta_j^2$ may be estimated by $\sum_{j=J_1+1}^{J_2} (\hat\theta_j^2 - n^{-1})$. Thus, it is worthwhile to choose $\hat f_{J_2}$ only if for a sufficiently large C

$$\sum_{j=J_1+1}^{J_2} \hat\theta_j^2 > Cn^{-1}(J_2 - J_1).$$

The $\mathrm{MISE}(\hat f_{J_2}, f)$ is always proportional to $n^{-1}J_2$, i.e., this holds for both $f \in L^1$ and $f \in L^2$. Also, for large n the cutoff J_2 is always significantly larger than J_1. Thus, the term $n^{-1}(J_2 - J_1)$ is proportional to $\mathrm{MISE}(\hat f_{J_2}, f)$. Set $R_2 := n^{-1}J_2$, and note that R_2 is proportional to the MISE.

Also note that by Parseval's identity,

$$\sum_{j=J_1+1}^{J_2} \hat\theta_j^2 = \int_0^1 (\hat f_{J_1}(x) - \hat f_{J_2}(x))^2 dx =: l(\hat f_{J_1} - \hat f_{J_2}),$$

where $l(\cdot)$ is the loss function (integrated squared error) used in MISE.

Combining these facts, we may conclude that the more complicated estimate $\hat f_{J_2}$ *should not be chosen* if

$$l(\hat f_{J_1} - \hat f_{J_2}) < CR_2. \tag{7.4.41}$$

We have obtained an algorithm of a bias–variance tradeoff for the case of two competing estimates.

In the general case of an unknown β, there is a net of m cutoffs $J_1 < J_2 < \cdots < J_m$ that allows one to estimate (approximate) any underlying function whenever the cutoffs and m may depend on n. (It is more accurate to refer to this net as a sequence of nets.) Also, set $R_s := n^{-1}J_s$ for the corresponding variance terms (R_s is proportional to the minimax MISE when J_s is the optimal cutoff). Then the method of bias–variance tradeoff implies a pairwise comparison between the corresponding m projection estimates. A relatively simple algorithm of pairwise comparison is as follows:

$$\hat f := \hat f_{J_{\hat k}}, \ \text{ where } \ \hat k := \min\{k : \ l(\hat f_{J_k} - \hat f_{J_s}) < CR_s, \ k \le s \le m\}. \tag{7.4.42}$$

Note that (7.4.42) coincides with (7.4.41) when $m = 2$. The procedure of a pairwise bias–variance tradeoff is called Lepskii's algorithm.

It is possible to show that a bias–variance tradeoff may be used for a wide class of loss functions.

• **Pointwise Estimation.** The remarkable outcome of the previous subsection on global estimation is that a data-driven estimator may have the same MISE convergence as an estimator that knows that an underlying function is Lipschitz with a given smoothness β. The same outcome holds for analytic, Besov, and many other function spaces.

The situation changes if a pointwise approach is used. Consider, as an example, the case of two Lipschitz spaces $L^1 := Lip_{r_1,\alpha_1,L}$ and $L^2 := Lip_{r_2,\alpha_2,L}$, where $\beta_1 := r_1 + \alpha_1$, $\beta_2 := r_2 + \alpha_2$, and $\beta_1 > \beta_2$. In other words, smoother functions belong to the space L^1, and thus $L^1 \subset L^2$. Let us assume that for some function $f_1 \in L^1$ there exists an estimator $\tilde{f}$ that is rate optimal over L^1, that is,

$$\mathrm{MSE}(\tilde{f}(t_0), f_1(t_0)) < Cn^{-2\beta_1/(2\beta_1+1)}. \tag{7.4.43}$$

Then the following assertion holds (its relatively simple proof for differentiable functions may be found in Brown and Low 1996b). It is always possible to find a function f_2 from a larger space L^2 such that the optimal rate $n^{-2\beta_2/(2\beta_2+1)}$ is not attainable by the estimate $\tilde{f}$. More precisely,

$$\mathrm{MSE}(\tilde{f}(t_0), f_2(t_0)) \ \geq \ C(n/\ln(n))^{-2\beta_2/(2\beta_2+1)}. \tag{7.4.44}$$

Because (7.4.44) holds for any $\tilde{f}$ satisfying (7.4.43), we have established the lower bound for minimax MSE of an adaptive estimator. Below we shall consider two data-driven estimators whose MSE converge at the rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$ for a Lipschitz space with unknown smoothness β. This will imply that the rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$ is the *optimal adaptive* rate for a minimax MSE convergence.

The statement formulated above is an explanation of the slogan that "minimax MSE of adaptive estimators loses a logarithmic factor (pays a logarithmic penalty)." This slogan should be considered only as popular statistical jargon because the mathematical fact is that the lower minimax bound (7.3.2) is not attainable whenever β is unknown.

There is no need to be too pessimistic about this outcome. The reason is that we discuss minimax rates, that is, this result just tells us that one can always find two functions with different parameters of smoothness that cannot be simultaneously estimated with optimal nonadaptive rates. But are there many such functions? It will be explained below that the set of functions where the logarithmic loss must occur is relatively small, and for all other Lipschitz functions the nonadaptive rate $n^{-2\beta/(2\beta+1)}$ is attainable. In other words, the logarithmic loss is inevitable for some functions, but luckily enough this loss occurs only for a small subset of Lipschitz functions.

While it is beyond the scope of this subsection to discuss all the details, it may be worthwhile to recall a familiar example from classical parametric theory. Consider a random variable X distributed according to a binomial distribution $B(n, p)$, that is, X is a number of "successes" in n Bernoulli trials with the probability of "success" p. If p is unknown, then the classical estimate of p is the sample mean $\bar{X} := X/n$. On the other hand, using

the method of finding minimax estimates outlined in Appendix A, it is possible to show that the minimax estimate is $\hat{p} = (X + n^{1/2}/2)/(n + n^{1/2})$. The mean squared error $R(\bar{X}, p) := E\{(\bar{X} - p)^2\}$ was calculated in Appendix A, and it is $n^{-1}p(1-p)$. The mean squared error of the minimax estimate is $R(\hat{p}, p) = (1 + n^{1/2})^{-2}/4$. A comparison of these errors shows that $R(\hat{p}, p) < \max_p R(\bar{X}, p)$, that is, for some p the minimax estimator is better. On the other hand, the inequality $R(\hat{p}, p) < R(\bar{X}, p)$ holds only for $p \in (0.5 - c_n, 0.5 + c_n)$, where $c_n \to 0$ as $n \to \infty$. Thus, the minimax estimator is better than the traditional one only for p in some shrinking (as $n \to \infty$) vicinity of $p = 0.5$. Also, any reasonable prior distribution for p implies that the Bayes risk of $\bar{X}$ is smaller than the Bayes risk of the estimate $\hat{p}$.

The adaptive logarithmic penalty for MSE has the same flavor: Some Lipschitz functions cannot be estimated with the nonadaptive optimal rate, but a "majority" of functions may be estimated with the nonadaptive rate.

Below two methods will be considered that lead to an optimal adaptive pointwise series estimation with the minimax rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$. The first estimator is extremely simple, but it does not allow one to control a subset of functions where no logarithmic penalty is paid. The second one is more involved, but it allows one to control this subset of functions where the nonadaptive rate $n^{-2\beta/(2\beta+1)}$ is attained.

a. Universal Wavelet Threshold Estimator. The optimal adaptive rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$ resembles the rate of the universal threshold estimator (recall (7.4.8) and the discussion), and this is indeed the simplest among rate optimal data-driven estimators.

Here we apply the idea discussed in subsection 1 to a wavelet basis. Set $J := \lfloor \log_2(n/\log_2(n)) \rfloor$ and consider the following *universal hard threshold* wavelet estimator (the notation of Section 4.4 is used):

$$\hat{f}(t) := \sum_k \hat{\kappa}_{j_0,k}\phi'_{j_0,k}(t) + \sum_{j=j_0}^{J}\sum_k I_{\{\hat{\theta}^2_{j,k} > 2\ln(n)\hat{d}n^{-1}\}}\,\hat{\theta}_{j,k}\psi'_{j,k}(t). \quad (7.4.45)$$

The beauty of this fantastically simple data-driven estimator is that under very mild assumptions it is rate optimal, that is, it is possible to show that its MSE converges with the optimal adaptive minimax rate $(n/\ln(n))^{2\beta/(2\beta+1)}$. The proof is left as Exercise 7.4.12.

b. Bias–Variance Tradeoff. It is possible to show that Lepskii's procedure (7.4.42) may be directly applied to this case. The only changes are as follows. The loss function is $l(f_1 - f_2) := (f_1(x_0) - f_2(x_0))^2$; note that this is the loss function used in the MSE; see (7.1.4). The estimates $\tilde{f}_{J_k}$ are de la Vallée Poussin sums (7.3.5) with cutoffs $J_1 < J_2 < \cdots < J_m$; the corresponding risks are $R_k := J_k n^{-1}(\ln(n))^{1-\ln(J_k)/\ln(n)}$. To see that R_k are indeed the adaptive risks corresponding to the cutoffs J_k, note that if $J_k =: \lfloor n^{1/(2\beta_k+1)} \rfloor$, then $R_k = (n/\ln(n))^{-2\beta_k/(2\beta_k+1)}(1 + o_n(1))$.

Then the bias–variance tradeoff estimate (7.4.42) implies MSE convergence at the optimal adaptive minimax rate $(n/\ln(n))^{-2\beta/(2\beta+1)}$.

To control a set of Lipschitz functions where the optimal nonadaptive rate $n^{-2\beta/(2\beta+1)}$ is attainable, a more complicated Efromovich–Low algorithm of a bias–variance tradeoff has been suggested. It is based on a pilot net and a main net of cutoffs. The main net is similar to one used in Lepskii's algorithm, i.e., it is a net of cutoffs $J_1 < J_2 < \cdots < L_m$ that allows one to estimate any Lipschitz function with an optimal rate. The additional net of pilot cutoffs $J_1^* < \cdots < J_m^*$ together with the corresponding risks R_k^* is used only to choose an optimal cutoff from the main net. Then, a bias–variance tradeoff estimator is defined as

$$\tilde{f} := \tilde{f}_{J_{\hat{k}}}, \text{ where } \hat{k} := \min\{k : l(\tilde{f}_{J_k^*} - \tilde{f}_{J_s^*}) < CR_s^*, \ k \le s \le m\}. \quad (7.4.46)$$

A particular example of this estimate is discussed in Exercise 7.4.14.

Using two nets of cutoffs together with a special net of risks $\{R_k^*\}$ makes the algorithm extremely flexible. For instance, as in the example of the minimax estimation of the probability of a success for a binomial experiment, it is possible to show that under mild assumptions a Bayes pointwise risk of the estimate (7.4.46) decreases with the optimal nonadaptive rate $n^{-2\beta/(2\beta+1)}$ (recall that a Bayesian approach was introduced in Remark 7.1.5). This result sheds new light on the issue of the logarithmic penalty and how "often" it should be paid. Another interesting aspect of the Efromovich–Low algorithm is as follows. It is possible to show that

$$\sup_{f \in Lip_{r,\alpha,L}} P(J_{\hat{k}} > bn^{1/(2\beta+1)}) = o_n(1)n^{-1}, \quad \beta := r + \alpha.$$

Here b is a positive constant that may be controlled by coefficients of the estimator, and recall that $n^{1/(2\beta+1)}$ is the optimal (up to a factor) cutoff for the underlying Lipschitz space. Thus, this data-driven estimator implies almost optimal (up to a factor) data compression.

• **The Case Where No Adaptation Is Needed for Optimal Estimation.** Let us finish this section about adaptive estimation with an example of a setting where, surprisingly, no adaptation is needed for optimal estimation of Lipschitz functions with unknown smoothness.

Consider a classical example of a communication system where an input signal f is first passed through a linear filter and then its output is contaminated by a white noise, i.e., an observed signal $Y(t)$ satisfies the differential equation

$$dY(t) = \int_0^1 h(t-x)f(x)dx + n^{-1/2}dB(t), \ 0 \le t \le 1. \quad (7.4.47)$$

Here $h(t)$ is an *impulse response kernel* of a linear filter. This signal transmission model is similar to a density estimation model with measurement errors considered in Section 3.5; another important analogue is a blurred regression model that is a discrete analogue of (7.4.47).

The problem of recovery of an input signal for the model (7.4.47) is also referred to as a *deconvolution* or *ill-posed* problem.

Assume that both the signal and the impulse response kernel are 1-periodic and

$$h(t) := \sum_{j=-\infty}^{\infty} h_j \psi_j(t), \tag{7.4.48}$$

where

$$C_0(|j|+1)^{c_0} e^{-c|j|^\nu} \le |h_j| \le C_1(|j|+1)^{c_1} e^{-c|j|^\nu}, \tag{7.4.49}$$

$\psi_j(t) := e^{-ij2\pi t}$ is the classical complex trigonometric basis discussed in Section 2.4, $C_0 > 0$, $C_1 > 0$, c and ν are some given positive constants, and c_1 and c_2 are some real numbers that will have no effect on the problem.

Then, as in our discussion in Sections 7.1–3, one may show that for $f \in Lip_{r,\alpha,L}$ the optimal minimax estimator is the same for both MSE and MISE risks, and it is

$$\hat{f}(t) := \sum_{j=-J_n}^{J_n} \Big(\int_0^1 \psi_{-j}(x) dY(x) dx \Big) h_j^{-1} \psi_j(t), \tag{7.4.50}$$

where

$$J_n := \lfloor (\ln(n)(1 - 1/\ln(\ln(n)))/2c)^{1/\nu} \rfloor. \tag{7.4.51}$$

The beauty of this estimate is that it depends on neither r nor α, that is, no information about smoothness of a recovered signal is needed for optimal estimation. Thus, no adaptation is needed either.

The explanation of this phenomenon is very simple: The MSE and MISE of the optimal estimate are defined by their squared bias terms, and the variance terms are negligible in comparison to the squared bias terms. To shed light on this statement, let us formulate and then prove it for the case of a global risk MISE. Proof of a similar result for MSE and that the suggested estimator is versatile, namely, that its derivatives are optimal estimates of derivatives, are left as an exercise.

Theorem 7.4.2 *For the convolution filtering model (7.4.47)–(7.4.49) with* 1*-periodic signal and impulse response kernel, the following lower bound for the minimax* MISE *holds:*

$$\inf_{\tilde{f}} \sup_{f \in Lip_{r,\alpha,L}} \text{MISE}(\tilde{f}, f) \ge C\,(\ln(n))^{-2\beta/\nu}, \quad \beta := r + \alpha, \tag{7.4.52}$$

where the infimum is taken over all possible estimators $\tilde{f}$ based on the data, the impulse response kernel, and the parameters r, α, and L. This bound is attained (up to a constant factor) by the estimate (7.4.50), that is,

$$\sup_{f \in Lip_{r,\alpha,L}} \text{MISE}(\hat{f}, f) \le C\,(\ln(n))^{-2\beta/\nu}. \tag{7.4.53}$$

Proof. Let us begin with establishing the lower bound (7.4.52). Set $m := \lfloor (\ln(n)(1 + 1/\ln(\ln(n)))/2c)^{1/\nu} \rfloor$ and note that m is a "bit" larger than J_n. Consider an input signal $f^*(t) := \theta(\psi_{-m}(t) + \psi_m(t))$. It is easy to check that for a real θ this signal is also real, and if $\theta^2 \le C^* m^{-2\beta}$, then $f^* \in Lip_{r,\alpha,L}$ whenever C^* is sufficiently small.

According to (7.2.6), the observed statistics are $Y_{-m} = \theta h_{-m} + n^{-1/2} Z'_{-m}$ and $Y_m = \theta h_m + n^{-1/2} Z'_m$. Note that in general the variables are complex, so to get real numbers we use the traditional "trick" and consider the equivalent real statistics $Y_1 := (Y_m + Y_{-m})/2$ and $Y_2 := (Y_m - Y_{-m})/2i$. Then we may conclude that any estimate of θ based on these two statistics will be dominated by an optimal estimate of θ based on the observation $Y := \theta + a|h_m|^{-1} n^{-1/2} Z'$, where a is a sufficiently small constant and Z' is a standard normal random variable. Then Lemma 7.1.1 together with the directly verified relation $m^{-2\beta} = o_n(1) n^{-1} |h_m|^{-2}$ yields the desired lower bound,

$$\begin{aligned} \inf_{\tilde f} \sup_{f \in Lip_{r,\alpha,L}} \mathrm{MISE}(\tilde f, f) &\ge \inf_{\tilde\theta} \sup_{\theta^2 \le C^* m^{-2\beta}} E\{(\tilde\theta - \theta)^2\} \\ &\ge C m^{-2\beta} \ge C(\ln(n))^{-2\beta/\nu}. \end{aligned}$$

The upper bound is established even more easily. A calculation shows that $|h_{J_n}|^{-2} n^{-1} = o_n(1) J_n^{-2\beta}$. Also, according to (2.4.18) we get that $\sup_{f \in Lip_{r,\alpha,L}} \sum_{j > J_n} \theta_j^2 \le C J_n^{-2\beta}$. These results together with Parseval's identity imply the upper bound (7.4.53). Theorem 7.4.2 is proved.

As we see from the proof, for the deconvolution problem the phenomenon of "no adaptation is needed" is due to the fact that the MISE of the optimal estimate has a variance term that is negligible in comparison to the squared bias term. Thus, no variance–bias tradeoff is needed. Recall that this outcome is just opposite to the estimation of analytic functions discussed in Section 7.1, where the efficiency of the optimal estimate for both MSE and MISE risks was due to the negligible squared biases in comparison to the variance terms. Thus, in both these cases no bias–variance tradeoff is needed for optimal estimation, and this is what makes these two cases so simple. On the other hand, the difference between these two settings is dramatic. Analytic functions may be estimated with an almost parametric rate $\ln(n)/n$, while an optimal deconvolution is possible only with an extremely slow logarithmic rate.

7.5 Multivariate Functions

The objective of this section is to explain what stands behind the expression "the curse of multidimensionality" discussed in Chapter 6 for particular examples.

The issue is that estimation of a multivariate function becomes very complicated, since typically the sample size needed for accurate curve estimation increases dramatically even if the dimensionality increases rather modestly.

The simplest way to shed some theoretical light on the issue is to derive lower bounds for minimax MISE and MSE convergences.

Suppose that we would like to estimate a d-variate function $f(t^d)$ for values of vectors $t^d := (t_1, t_2, \ldots, t_d)$ from the unit d-dimensional cube $[0,1]^d$. As an example, consider functions that are Hölder $H_{0,\beta}(L, L_1)$, $0 < \beta < 1$, that is, $|f(u^d) - f(v^d)| \le L|u^d - v^d|^\beta$ and $|f(t^d)| \le L_1$. Below we skip the parameters 0, L_1, and L_2 in the notation for this space, that is, we simply write H_β.

To establish lower bounds for the minimax MISE and minimax MSE we again use the approach of Section 7.1. We choose a multivariate wavelet basis with bounded support and then transmit wavelet coefficients of a signal $f(t^d)$ via parallel Gaussian channels, as shown in Figure 7.1. As we know, this is an analogue of a filtering model, and to make the setting similar to a multivariate density estimation or a multivariate regression, let us assume that $\sigma^2 := n^{-1}$.

To establish a lower bound for the minimax MISE, consider a signal

$$f(t^d) := \sum_{s_1,s_2,\ldots,s_d=0}^{2^J-1} X_{J,s^d}\, \psi'_{J,s^d}(t^d), \tag{7.5.1}$$

where $J := \lfloor \log_2(n^{1/(2\beta+d)}) \rfloor$, $s^d := (s_1, s_2, \ldots, s_d)$, and the wavelet function $\psi'_{J,s^d}(t^d) := \prod_{l=1}^m \psi'_{J,s_l}(t_l)$ is a a d-variate function created by a product of d univariate wavelet functions $\psi'_{J,s_l}(t_l)$ at the Jth resolution scale.

It is well known, see, for instance, Meyer (1992, Section 6.4), that there exist wavelets such that a function (7.5.1) belongs to H_β if and only if (compare to (2.5.3) where $d = 1$)

$$|X_{J,s^d}| \le C 2^{-J(2\beta+d)/2}. \tag{7.5.2}$$

Thus, if in (7.5.1) all $X_{J,s^d} \in [-cn^{-1/2}, cn^{-1/2}]$ and c is sufficiently small, then $f(t^d)$ defined in (7.5.1) belongs to H_β (Exercise 7.5.1).

Using $k = 2^{Jd}$ channels one can transmit the signal (7.5.1) via the communication system shown in Figure 7.1, and then, according to (7.1.14) (note that to get (7.1.14) we never really used the fact that the input signal is univariate, and the only needed modification in the proof is to use d-variate coding functions and the corresponding integrals),

$$\inf_{\tilde f} \sup_{f \in H_\beta} \mathrm{MISE}(\tilde f, f)$$

$$\ge \inf_{\tilde f} \sup_{\{X_{J,s^d} \in [-cn^{-1/2}, cn^{-1/2}]\}} \mathrm{MISE}\Big(\tilde f, \sum_{s_1,s_2,\dots,s_d=0}^{2^J-1} X_{J,s^d}\psi'_{J,s^d}\Big)$$

$$\ge \frac{\mu(c)c^2}{(1+c^2)n} \sum_{s_1,s_2,\dots,s_d=0}^{2^J-1} \int_{[0,1]^d} (\psi'_{J,s^d}(t^d))^2 dt^d$$

$$\ge Cn^{-1}2^{Jd} = Cn^{-2\beta/(2\beta+d)}. \tag{7.5.3}$$

It is even simpler to establish a lower bound for the minimax MSE. As in Section 7.1 we note that for any $t_0^d \in [0,1]^d$ it is always possible to assume that there exists s_0^d such that

$$|\psi'_{J,s_0^d}(t_0^d)|^2 \ge C2^{dJ}. \tag{7.5.4}$$

Then consider $f(t_0^d) := X_1\psi'_{J,s_0^d}(t_0^d)$. As was explained above, if $X_1 \in [-cn^{-1/2}, cn^{1/2}]$, then $f \in H_\beta$. Thus, we set $k = 1$, transmit only one input signal X_1, and then, according to (7.1.13), get

$$\inf_{\tilde f(t_0^d)} \sup_{f\in H_\beta} \mathrm{MSE}(\tilde f(t_0^d), f(t_0^d))$$

$$\ge \inf_{\tilde f(t_0^d)} \sup_{X_1\in[-cn^{-1/2}, cn^{-1/2}]} \mathrm{MSE}(\tilde f(t_0^d), X_1\psi'_{J,s_0^d}(t_0^d))$$

$$\ge Cn^{-1}2^{dJ} = Cn^{-2\beta/(2\beta+d)}. \tag{7.5.5}$$

We conclude that both the minimax MISE and the minimax MSE for the estimation of d-dimensional Hölder functions cannot converge faster than $n^{-2\beta/(2\beta+d)}$.

To shed light on how this rate affects estimation of multivariate curves and to get a feeling for the necessary sample sizes, consider the following example. Suppose that we would like to estimate a Hölder function with a MISE or MSE not larger than $\delta = 0.1$. Assume that $\beta = 0.5$ and that there exists a rate optimal estimator whose risk is $n^{-2\beta/(2\beta+d)}$. Then, to get this precision of estimation, one needs at least $n^*(d)$ observations, where $n^*(d)$ is the rounded-up δ^{-1-d}. In particular, $n^*(1) = 100$, $n^*(2) = 1000$, $n^*(3) = 10000$, and $n^*(4) = 100000$. This is what defines the curse of multidimensionality, because to get a reasonable precision of estimation, astronomically large sample sizes are needed even for moderate dimensions.

On the other hand, it is a good idea to know that estimation of multivariate functions is not necessarily so complicated. Consider a function space of d-variate analytic functions

$$f(t^d) := \sum_{j_1,\dots,j_d=0}^{\infty} \theta_{j^d}\,\varphi_{j^d}(t^d), \tag{7.5.6}$$

where $\varphi_{j^d}(t^d) := \varphi_{j_1}(t_1)\cdots\varphi_{j_d}(t_d)$ are the elements of the trigonometric tensor-product basis (see Section 6.1) and $\theta_{j^d} = \int_{[0,1]^d} \varphi_{j^d}(t^d)f(t^d)dt^d$ are

Fourier coefficients that satisfy the inequality

$$|\theta_{j^d}| \le Q \exp\Big\{ -\sum_{l=1}^{d} \gamma_l \lfloor (j_l+1)/2 \rfloor \Big\}. \tag{7.5.7}$$

Denote this function class by $A_{\gamma^d Q}$.

Suppose that for a statistical setting one can suggest an estimator $\hat{\theta}_{j^d}$ of Fourier coefficients θ_{j^d} such that

$$E\{(\hat{\theta}_{j^d} - \theta_{j^d})^2\} \le Cn^{-1}. \tag{7.5.8}$$

Note that such an estimator exists for all the models considered.

Define a projection estimator

$$\hat{f}(t^d) := \sum_{j_1=0}^{J_1} \sum_{j_2=0}^{J_2} \cdots \sum_{j_d=0}^{J_d} \hat{\theta}_{j^d} \varphi_{j^d}(t^d), \tag{7.5.9}$$

where $J_l := 2\lfloor \gamma_l^{-1} \ln(n)/2 \rfloor$. Denote by D all natural indices $(j_1, \dots, j_d)$ that are not included in the sum (7.5.9). Then direct calculations, based on Parseval's identity, show that

$$\sup_{f \in A_{\gamma^d Q}} \mathrm{MISE}(\tilde{f}, f) = n^{-1} \prod_{l=1}^{d} (J_l + 1) + \sum_{j^d \in D} \theta_{j^d}^2 \le C\,(\ln(n))^d\, n^{-1}. \tag{7.5.10}$$

Thus, at least asymptotically, estimation of multivariate analytic functions is scarcely worse than estimation of univariate functions.

The conclusion from (7.5.5) and (7.5.10) is that smoothness of multivariate functions plays a far more dramatic role in their estimation than for their univariate counterpart. Thus, optimal adaptive estimation becomes necessary for the multivariate case. Fortunately, almost all methods of adaptation discussed in Section 7.4 may be used for multivariate settings.

7.6 Special Topic: Estimation of Quadratic Functionals

In the previous sections we have discussed estimation of a *linear* functional $f(t_0)$, that is, the value of a function at a given point. In this section we would like to estimate a nonlinear functional, namely, a quadratic functional

$$F_s(f) := \int_0^1 (f^{(s)}(t))^2 dt, \tag{7.6.1}$$

where $f^{(s)}$ is the sth derivative of f.

Let us begin with the case $s = 0$, filtering model (7.3.1), and the assumption that f belongs to the Sobolev function space $W_{\beta,Q}$ defined in (2.4.19).

A natural idea of how to estimate a functional is to plug in an estimate of an underlying function. Let is check this idea. For our function space, a projection estimator

$$\hat{f}(t) := \sum_{j=0}^{J} \hat{\theta}_j \varphi_j(t) \tag{7.6.2}$$

is globally rate optimal if $J = 2\lfloor n^{1/(2\beta+1)} \rfloor$. Here

$$\hat{\theta}_j = \int_0^1 \varphi_j(t) dY(t) = \theta_j + n^{-1/2} Z'_j, \tag{7.6.3}$$

$\{\varphi_j\}$ is the trigonometric basis (2.4.1), $\theta_j = \int_0^1 \varphi_j(t) f(t) dt$ is the jth Fourier coefficient of an underlying f, and Z'_j are iid standard normal; see (7.2.6).

It is clear from Parseval's identity that a plug-in estimator $F_0(\hat{f})$ has a very simple form,

$$F_0(\hat{f}) = \sum_{j=0}^{J} \hat{\theta}_j^2. \tag{7.6.4}$$

Now let us calculate the mean squared error of this plug-in estimator. The calculations are straightforward and based on the fact that odd moments of Z'_j are zero and $E\{(Z')^{2m}\} = (2m-1)(2m-3)\cdots 1$. Write

$$\begin{aligned}
&E\{(F_0(\hat{f}) - F_0(f))^2\} \\
&= E\Big\{\Big[\sum_{j=0}^{J} \big(\theta_j^2 + 2n^{-1/2} Z'_j \theta_j + n^{-1}(Z'_j)^2 - \theta_j^2\big) - \sum_{j>J} \theta_j^2\Big]^2\Big\} \\
&= 4n^{-1} \sum_{j=0}^{J} \theta_j^2 + n^{-2}[(J+1)^2 + 2(J+1)] \\
&\quad - 2n^{-1}(J+1) \sum_{j>J} \theta_j^2 + \Big(\sum_{j>J} \theta_j^2\Big)^2 \\
&= 4n^{-1} \int_0^1 f^2(t) dt + \Big(n^{-1}(J+1) - \sum_{j>J} \theta_j^2\Big)^2 \\
&\quad + 2n^{-2}(J+1) - 4n^{-1} \sum_{j>J} \theta_j^2.
\end{aligned} \tag{7.6.5}$$

For the Sobolev space the absolute value of the second term in (7.6.5) is at most $Cn^{-4\beta/(2\beta+1)}$, and the absolute values of the third and fourth terms are at most $Cn^{-(4\beta+1)/(2\beta+1)}$. Thus for some $|C_n| < C < \infty$,

$$E\{(F_0(\hat{f}) - F_0(f))^2\} = 4F_0(f)n^{-1} + C_n n^{-4\beta/(2\beta+1)} \tag{7.6.6}$$

$$= 4F_0(f)n^{-1}(1 + o_n(1)) \quad \text{if} \quad \beta > 0.5. \tag{7.6.7}$$

We see that the plug-in estimator gives a perfect parametric rate n^{-1} of the mean squared error convergence if an underlying function is sufficiently smooth. It is also possible to show that the factor $4F_0(f)$ in (7.6.7) cannot be improved, that is, this estimator is asymptotically efficient.

Thus, this plug-in estimator is an excellent one. However, there are at least two reasons why it is worthwhile to find an alternative estimator. The first one is that the plug-in estimator is based on β, since $\hat{f}(t)$ is based on β. The second one is that it is of interest to understand how one can efficiently estimate the quadratic functional for the case $\beta \le 0.5$.

As an alternative procedure, consider the data-driven estimator

$$\hat{F}_0 := \sum_{j=0}^{J_0} (\hat{\theta}_j^2 - n^{-1}), \qquad J_0 := 2\lfloor n/\ln(n) \rfloor. \tag{7.6.8}$$

Clearly, this estimator is motivated by the Parseval identity $F_0(f) = \sum_{j=0}^{\infty} \theta_j^2$ and by the unbiased estimator $\hat{\theta}_j^2 - n^{-1}$ of θ_j^2.

Let us calculate the mean squared error of $\hat{F}_0$. Write for $f \in W_{\beta,Q}$,

$$\begin{aligned} E\{(\hat{F}_0 - F_0(f))^2\} &= E\Big\{\Big(\sum_{j=0}^{J_0} (\hat{\theta}_j^2 - n^{-1} - \theta_j^2) - \sum_{j>J_0} \theta_j^2\Big)^2\Big\} \\ &= 2(J_0+1)n^{-2} + n^{-1}4\sum_{j=0}^{J_0} \theta_j^2 + \Big(\sum_{j>J_0} \theta_j^2\Big)^2 \\ &\le 4F_0(f)n^{-1}(1+o_n(1)) + CJ_0^{-4\beta}. \end{aligned} \tag{7.6.9}$$

We see that this simple data-driven estimator outperforms the plug-in estimator, which is based both on data and β, because $J_0^{-4\beta} = (n/\ln(n))^{-4\beta} = o(1)n^{-1}$ whenever $\beta > \frac{1}{4}$ while (7.6.7) holds only for $\beta > \frac{1}{2}$.

Our conclusion from this example is twofold. First, a plug-in idea usually works out for sufficiently smooth underlying functions. Second, it is always worthwhile to look at the specific nature of a functional and then consider (if possible) a simpler estimate that is based on the structure of the functional. As we have seen, such an attempt may pay a dividend.

What will occur if the smoothness parameter β is smaller than $\frac{1}{4}$? It is possible to show that in this case the rate decreases from the parametric n^{-1} to a slower $n^{-8\beta/(4\beta+1)}$. Thus this case is called irregular. There is one more bit of "bad" information about the irregular case: An adaptation penalty, which is again a logarithmic factor (similar to the adaptive pointwise estimation discussed in Section 7.4), should be paid.

It is easy to extend these results to the case of estimating the functionals $F_s(f)$, that is, integrals of squared derivatives. Let us again assume that $f \in W_{\beta,Q}$ and $\beta > 2s + 0.25$. Set $J_s := 2\lfloor n^{1/(4s+1)}/\ln(n) \rfloor$ and define the

estimator

$$\hat{F}_s := \sum_{j=0}^{J_s} (\hat{\theta}_j^2 - n^{-1}) \int_0^1 (\varphi_j^{(s)}(t))^2 dt. \tag{7.6.10}$$

Note that $\int_0^1 (\varphi_j^{(s)}(t))^2 dt = (2\pi\lfloor (j+1)/2 \rfloor)^{2s}$, $j > 0$.

Clearly, this estimator is motivated by the Parseval identity

$$F_s(f) = \sum_{j=0}^{\infty} \theta_j^2 \int_0^1 \left(\varphi_j^{(s)}(t)\right)^2 dt. \tag{7.6.11}$$

Then, according to Exercise 7.6.4, if $\beta > 2s + 0.25$, then for $f \in W_{\beta,Q}$

$$E\{(\hat{F}_s - F_s(f))^2\} = 4F_{2s}(f)\, n^{-1}\, (1 + o_n(1)). \tag{7.6.12}$$

It is again possible to show that the factor $4F_{2s}(f)$ cannot be improved. Thus the data-driven estimator (7.6.10) is asymptotically efficient.

Extension of these results to other statistical models is straightforward: Just use the estimates $\hat{\theta}_j$ of Fourier coefficients θ_j recommended in Section 7.3. On the other hand, we mentioned in Section 7.3 that the principle of equivalence has its limits whenever an estimated function is not smooth enough. Consider an interesting example that explains this limit.

Let us evaluate the possible risk of estimating $F_0(f)$ for the case of the random design regression model (iii) defined at the beginning of Section 7.3. Recall that the responses are $Y_l = f(X_l) + \xi_l$, where predictors X_l are iid uniform on $[0, 1]$ and ξ_l are iid standard normal. Consider the estimator

$$\hat{F} := n^{-1} \sum_{l=1}^{n} (Y_l^2 - 1). \tag{7.6.13}$$

Note that the predictors are not used by this estimator. Also, this is a sample mean estimator, because

$$E\{Y_l^2 - 1\} = E\{(f(X_l) + \xi_l)^2 - 1\} = E\{f^2(X_l)\} = \int_0^1 f^2(x) dx = F_0(f).$$

The mean squared error of a sample mean estimate always decreases proportionally to the inverse sample size. More precisely,

$$E\{(\hat{F} - F_0(f))^2\} = n^{-1} \Big[\int_0^1 (f^2(x) + 2)^2 dx - (F_0(f))^2 - 2 \Big]. \tag{7.6.14}$$

Thus, if for filtering model the rate of the mean squared error convergence, as a function of β, has an elbow at the point $\beta = \frac{1}{4}$, there is no such phenomenon for the regression model, where the rate is always proportional to n^{-1} regardless of the smoothness of an underlying regression function. This shows the *limits of the equivalence principle*.

Note that the elegance of the data-driven estimator (7.6.13) is appealing.

7.7 Special Topic: Racing for Constants

There are many interesting statistical settings where asymptotic constants are of special interest simply by themselves regardless of the fact that someone wants to find a sharp estimator as we did in Theorem 7.1.1. In this section three such statistical problems are discussed.

The first one is the estimation of a monotone density, the second one is estimation of a density based on censored data, and the third one is a general setting of nonparametric regression.

Estimation of a monotone density is probably one of the most beautiful and interesting nonparametric problems. It is known that any monotone density can be estimated with MISE that converges proportionally to $n^{-2/3}$ regardless of how smooth the underlying monotone density. This fact will be discussed in Section 8.6, and here just note that if a monotone function is bounded, then its total variation is bounded, and this implies the rate $n^{-2/3}$. Recall that without monotonicity the rate is $n^{-2\beta/(2\beta+1)}$, where β is the parameter of smoothness for Lipschitz or Sobolev function spaces. Thus, a discontinuous monotone density can be estimated with a precision of estimation of a differentiable density.

What will occur if a density is monotone and $\beta > 1$? Can monotonicity improve the convergence in this case? It took a long time to answer this question, and it was only in the early 1980s that Kiefer gave a negative answer. Thus, monotonicity does not affect the rate of MISE convergence whenever an underlying density is differentiable.

On the other hand, it is clear that monotonicity is important additional information. Can monotonicity affect the sharp constant of MISE convergence for the case of differentiable functions? This question was raised in the famous article by Kiefer (1982).

Why is this problem important? Suppose that monotonicity does affect a sharp constant. Then this implies that a special procedure of estimation should be used that takes into account monotonicity. On the other hand, if monotonicity does not affect this constant, then any sharp minimax estimator can be used. Also, since a monotonic function cannot approximate a nonmonotonic one, an estimate that is based on the assumption of monotonicity is not robust.

It has been established that monotonicity does not affect a sharp constant. Thus, at least asymptotically, there is no need for a special estimator for monotonic functions. Moreover, a universal wavelet estimator allows one to get rate optimal estimation over a wide spectrum of function spaces automatically, that is, if an underlying density is monotone and not differentiable, it has MISE convergence $n^{-2/3}$, and otherwise the rate is $n^{-2\beta/(2\beta+1)}$.

Our second example is estimation of a density based on censored data. We discussed this setting in Section 3.4 for the case of small sample sizes.

Let us briefly recall one of the possible settings. We are interested in an underlying density f of iid unobserved survival times $X_1, \dots, X_n$ of n items or individuals that are censored on the right by iid nonnegative random variables $T_1, \dots, T_n$. Denote the distribution of T's by G^* and set $G := 1 - G^*$ for their survivor function.

The problem is to suggest an estimate of f that is based on right-censored data (Y_l, δ_l), $l = 1, 2, \dots, n$, where $Y_l = \min(X_l, T_l)$ and $\delta_l = I_{\{X_l \le Y_l\}}$.

The question is how the censoring affects MISE convergence, where one is interested in estimating f over a given interval $[a, b]$. For this particular setting we consider $\text{MISE}(\tilde{f}, f) := \int_a^b (\tilde{f}(x) - f(x))^2 dx$.

It was established in the 1980s that under mild assumptions on G and for densities $f \in W_{\beta,Q}$, the rate of the MISE convergence is $n^{-2\beta/(2\beta+1)}$, that is, it is the same as for the case of directly observed $X_1, \dots, X_n$. This result shows that rates of MISE convergence shed no light on the effect of censorship on density estimation. On the other hand, it is apparent that censorship affects the precision of estimation and the rates simply do not reveal this.

The situation changes if the analysis of a sharp constant, which is similar to the analysis of risks for estimation of analytic functions in Section 7.1, is performed. It shows that under a very mild assumption there is an additional factor d' in the sharp constant of the minimax (over the Sobolev function class) MISE convergence, and

$$d' := \Big[\int_a^b (f(x)/G(x)) dx \Big]^{2\beta/(2\beta+1)}. \tag{7.7.1}$$

Recall that in Section 3.4 the coefficient $d = (d')^{(2\beta+1)/2\beta}$ was referred to as the coefficient of difficulty due to censoring.

Finally, let us consider the problem of nonparametric regression in a general setting. This problem is motivated by classical parametric estimation problems, so it is worthwhile to begin with a very brief review of parametric problems.

Let n iid observations $V_1, V_2, \dots, V_n$ be given that are drawn from a distribution with a density $p(v|\theta)$ that depends on a parameter θ. Several familiar examples are as follows. (i) A model with a *location* parameter where $V_l = \theta + \xi_l$, $l = 1, 2, \dots, n$, and ξ_l are errors with a density $p^\xi(v)$. In this case $p(v|\theta) = p^\xi(v - \theta)$. (ii) A model with a *scale* parameter where $V_l = \theta \xi_l$. In this case $p(v|\theta) = (1/\theta) p^\xi(v/\theta)$. (iii) A *mixture* model where

$$p(v|\theta) = \theta g(v) + (1 - \theta) h(v), \quad 0 \le \theta \le 1,$$

g and h are densities of two different random variables (typically it is assumed that their means are different). In other words, with probability θ the observed random variable V is generated from the density g and with probability $1 - \theta$ from the density h. (iv) A model of *censored data* where $V_l = \min(U_l, c)$ and $U_l = \theta + \xi_l$. In other words, unobserved data are gen-

erated by a location model and then censored at the level c. (v) A *binomial* model where $V_1, \dots, V_n$ are observed successes and failures and θ is the probability of a success.

In parametric asymptotic theory we are interested in efficient estimation of the parameter θ in the following sense. If $\hat{\theta}_n$ is an estimate of θ, then we say that it is asymptotically *efficient* if $n^{1/2}(\hat{\theta}_n - \theta)$ is asymptotically normal with zero mean and variance $1/I(\theta)$, where

$$I(\theta) := \int [(p'(v|\theta))^2/p(v|\theta)]dv \tag{7.7.2}$$

is the *Fisher information*. Here $p'(v|\theta) := \partial p(v|\theta)/\partial\theta$.

The definition of an asymptotically efficient estimate is motivated by the famous Rao–Kramer inequality, which states that under mild assumptions the variance of any unbiased estimate cannot be smaller than $1/nI(\theta)$.

Also recall that under mild assumptions a maximum likelihood estimate is asymptotically efficient. Except for some trivial cases, there is no close formulae for this estimate, so in many practical applications the Newton–Raphson one-step approximation of this estimate is used. The approximation is based on using a *pilot* estimate $\tilde{\theta}$ that satisfies the assumption $E\{(n^{1/4}(\tilde{\theta}_n - \theta))^2\} \to 0$ as $n \to \infty$. Then the one-step approximation, which is also called the *scoring* estimate, is used:

$$\hat{\theta}_n := \tilde{\theta}_n + \frac{L^{(1)}(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)}. \tag{7.7.3}$$

Here $L^{(1)}(\theta) = \sum_{l=1}^n p'(V_l|\theta)/p(V_l|\theta)$ is the derivative of the log-likelihood function $L(\theta) := \sum_{l=1}^n \ln(p(V_l|\theta))$. Under mild assumptions, it is possible to show that the scoring estimate (7.7.3) is asymptotically efficient.

Now we are in a position to explain how to use these classical parametric results for nonparametric regression settings.

First, a general parametric model of iid observations drawn from a distribution with a density $p(v|\theta)$ is straightforwardly translated into the following nonparametric model. It is assumed that n iid pairs of observations $(X_1, Y_1), \dots, (X_n, Y_n)$ are generated by a pair of random variables (X, Y) with a joint density $p^{X,Y}(x, y) = h(x)p(y|f(x))$. Here $h(x)$ is the density of the predictor X supported on $[0, 1]$, Y is the response, and $f(x)$ is an estimated regression function.

This generalized model includes all the regression settings considered in Chapter 4. For instance, if $p(y|f(x)) = p^{\xi}(y - f(x))$, we get a classical additive model $Y = f(X) + \xi$, which is an analogue of the parametric location model. If $p(y|f(x)) = (1/f(x))p^{\xi}(y/f(x))$, we get the model of a scale (volatility) regression $Y = f(X)\xi$, etc.

Second, the notion of asymptotically efficient estimation is translated into a sharp (efficient) nonparametric estimation. Under mild assumptions, it is possible to show that for the generalized regression model and $f \in W_{\beta,Q}$

the sharp constant for the minimax MISE convergence is

$$C^* := \Big[\int_0^1 [h(x)I(f(x))]^{-1}dx\Big]^{2\beta/(2\beta+1)} P(\beta, Q), \tag{7.7.4}$$

where

$$P(\beta, Q) := (2\beta/2\pi(\beta+1))^{2\beta/(2\beta+1)}(2\beta+1)^{1/(2\beta+1)}\, Q^{1/(2\beta+1)} \tag{7.7.5}$$

is a so-called Pinsker constant. Note that $C^* = P(\beta, Q)$ for a model of additive regression with standard Gaussian noise and uniformly distributed predictors where $I(\theta) = 1$, $\theta \in (-\infty, \infty)$, and $h(x) = 1$, $x \in [0, 1]$.

Because all the considered models have the same familiar rate $n^{-2\beta/(2\beta+1)}$ of MISE convergence, only the constant (7.7.4) indicates how a particular regression model affects MISE convergence.

If one would like to find an *optimal design* for an experiment, that is, a design density $h(x)$ that minimizes (7.7.4), it is easy to do. Using Cauchy–Schwarz inequality (2.3.4) we get the relation

$$\int_0^1 [h(x)I(f(x))]^{-1}dx \ge \Big[\int_0^1 [I(f(x))]^{-1/2}dx\Big]^2,$$

with equality if and only if

$$h^*(x) := \Big[[I(f(x))]^{1/2}\int_0^1 [I(f(t))]^{-1/2}dt\Big]^{-1}. \tag{7.7.6}$$

Thus, (7.7.6) defines the optimal design density. For instance, since for the additive regression model the Fisher information is always constant (see Exercise 7.7.2), the optimal design for an additive regression is always uniform regardless of the distribution of errors. In the general case a pilot estimate of f is needed to find the optimal design density (7.7.6).

Finally, note that parametric theory allows us to suggest a rather simple sharp minimax procedure of nonparametric series estimation based on the scoring estimator (7.7.3). Indeed, since θ_j are parameters, the scoring estimator can be used for estimating θ_j for all the regression models discussed. It is possible to show that this procedure together with Efromovich–Pinsker block shrinkage implies a sharp minimax estimation where the constant (7.7.4) is attained.

7.8 Special Topic: Confidence Intervals, Confidence Bands, and Hypothesis Testing

We begin with a review of basic results of the parametric theory (see also Appendix A). Let one observe $X = \theta + \sigma\xi$ where θ is an unknown parameter and ξ is a standard normal random variable. A $1 - \alpha$ confidence interval estimate of θ is an interval that encloses θ with the probability at least

equal to the confidence level $1-\alpha$, $0<\alpha<1$. Recall that a well-known confidence interval estimate of θ is

$$\mathrm{CI}(X,\alpha) := [X-\sigma z_{\alpha/2}, X+\sigma z_{\alpha/2}]. \tag{7.8.1}$$

Here $z_{\alpha/2}$ is defined as a function in α such that $P(\xi > z_{\alpha/2}) = \alpha/2$.

Also recall that the confidence interval (7.8.1) is closely related to a two-tailed hypothesis test. Consider a classical Neyman–Pearson problem of testing a null hypothesis $\theta=\theta_0$ versus an alternative hypothesis $\theta\neq\theta_0$ with the level of significance α. The problem is to find a rejection region $R(\theta_0,\alpha)$ such that if Y belongs to this region, then the null hypothesis is rejected; otherwise the null hypothesis is accepted, and the probability of rejection under the null hypothesis should be at most α, that is, $P(Y\in R(\theta_0,\alpha)|\theta=\theta_0)\le\alpha$. In other words, given that the null hypothesis is true, it may be rejected with probability at most α.

Then a customarily used rejection region is

$$R(\theta_0,\alpha) := \{X:\ \ X\notin[\theta_0-\sigma z_{\alpha/2}, \theta_0+\sigma z_{\alpha/2}]\}. \tag{7.8.2}$$

The striking similarity between (7.8.1) and (7.8.2) is not surprising, because the confidence interval estimation and the hypothesis testing are dual problems. A method of finding an interval estimate by inverting a test (and vice versa) is a fairly general technique in parametric statistics. Also, the reader with a major in statistics might recall that the test with the rejection region (7.8.2) is uniformly most powerful unbiased and the confidence interval estimator (7.8.1) is uniformly most accurate unbiased; see more in Exercise 7.8.3.

Direct extension of these classical parametric problems is a nonparametric analogue when one wants to find a confidence interval for $f(t_0)$ (which can be considered as a parameter) or solve a corresponding two-tailed hypothesis testing problem.

Let us explore these two problems for the case of analytic functions $f\in A_{\gamma,Q}$, defined in (7.1.5), and the filtering model (7.3.1). According to (7.1.11), the projection estimate

$$\hat f(t_0,Y) := \sum_{j=0}^{J_\gamma}\hat\theta_j\varphi_j(t_0) \tag{7.8.3}$$

is sharp minimax. Here $\{\varphi_j\}$ is the classical sine–cosine trigonometric basis on $[0,1]$,

$$\hat\theta_j := \int_0^1 \varphi_j(t)dY(t), \qquad J_\gamma := 2\lfloor \gamma^{-1}\ln(n^{1/2})\rfloor. \tag{7.8.4}$$

Then using (7.2.9) with $\sigma=n^{-1/2}$ we get

$$\hat f(t_0,Y) = \sum_{j=0}^{J_\gamma}\theta_j\varphi_j(t_0) + n^{-1/2}\sum_{j=0}^{J_\gamma}\xi_j\varphi_j(t_0)$$

$$= f(t_0) + n^{-1/2} \sum_{j=0}^{J_\gamma} \xi_j \varphi_j(t_0) - \sum_{j>J_\gamma} \theta_j \varphi_j(t_0). \qquad (7.8.5)$$

Here $\theta_j = \int_0^1 \varphi_j(t) f(t) dt$ are Fourier coefficients of f, and $\xi_0, \xi_1, \dots$ are iid standard normal random variables.

The second term in (7.8.5) is a normal random variable $N(0, n^{-1}(J_\gamma+1))$. The third term is the bias, which for large n is negligible in comparison to the standard deviation of the second term, namely,

$$\sup_{f \in A_{\gamma,Q}} \Big| \sum_{j>J_\gamma} \theta_j \varphi_j(t_0) \Big| \le C e^{-\gamma(\gamma^{-1} \ln(n^{1/2}))} \le C n^{-1/2}. \qquad (7.8.6)$$

Recall that we consider the case of large n; thus both the parametric confidence interval estimate (7.8.1) and the parametric rejection rule (7.8.2) can be used for our nonparametric setting with $\hat f(t_0, Y)$ in place of X and $\sigma_n := (n^{-1} J_\gamma)^{1/2}$ in place of σ. This implies the nonparametric confidence interval

$$\mathrm{NCI}(\hat f(t_0, Y), \alpha) := [\hat f(t_0, Y) - z_{\alpha/2}\sigma_n, \hat f(t_0, Y) + z_{\alpha/2}\sigma_n], \qquad (7.8.7)$$

and the nonparametric rejection region is

$$\mathrm{NR}(f_0(t_0), \alpha) := \{Y : \ \hat f(t_0, Y) \notin \mathrm{NCI}(f_0(t_0), \alpha)\}. \qquad (7.8.8)$$

Here $f = f_0$ is the null hypothesis and $\hat f(t, Y)$ is the estimate (7.8.3). The procedures enjoy the property of being asymptotically uniformly most accurate (powerful) unbiased.

Let us show, as a simple exercise, that for any $f \in A_{\gamma,Q}$ the recommended confidence interval (7.8.7) encloses an unknown $f(t_0)$ with probability at least $1 - \alpha + o_n(1)$, where $o_n(1) \to 0$ as $n \to \infty$. Write

$$\begin{aligned}
&\inf_{f \in A_{\gamma,Q}} P\big(f(t_0) \in [\hat f(t_0, Y) - z_{\alpha/2}\sigma_n, \hat f(t_0, Y) + z_{\alpha/2}\sigma_n]\big) \\
&= \inf_{f \in A_{\gamma,Q}} P\big(\big|f(t_0) - \hat f(t_0, Y)\big| \le z_{\alpha/2}\sigma_n\big) \\
&= \inf_{f \in A_{\gamma,Q}} P\Big(\Big|n^{-1/2} \sum_{j=0}^{J_\gamma} \xi_j \varphi_j(t_0) - \sum_{j>J_\gamma} \theta_j \varphi_j(t)\Big| \le z_{\alpha/2}\sigma_n\Big) \\
&\ge P\Big(\Big|n^{-1/2} \sum_{j=0}^{J_\gamma} \xi_j \varphi_j(t_0)\Big| \le z_{\alpha/2}\sigma_n - C n^{-1/2}\Big) \ge 1 - \alpha + o_n(1).
\end{aligned}$$

Here in the third line we used (7.8.5), in the first inequality of the last line we used (7.8.6), and the last inequality is based on the fact that the random variable $n^{-1/2} \sum_{j=0}^{J_\gamma} \xi_j \varphi_j(t_0)$ is normal $N(0, n^{-1}(J_\gamma + 1))$.

The approach discussed so far has been pointwise, that is, for a given point t_0 we have suggested a confidence interval for estimating $f(t_0)$. In many cases it is also desirable to have a global confidence band that shows

that an underlying signal $f(t)$ belongs to that band for all t with at least a probability $1 - \alpha$.

The key idea in finding a confidence band is to use (7.8.5)–(7.8.6) and then study a sequence of random variables

$$Z_J := (J+1)^{-1/2} \max_{0 \le t \le 1} \sum_{j=0}^{J} \xi_j \varphi_j(t), \tag{7.8.9}$$

which is the normed maximum of the second term (the frequency-limited white noise) in (7.8.5).

As in the definition of $z_{\alpha/2}$, let us define $z_{\alpha/2,J}$ as a value satisfying

$$P(Z_J \ge z_{\alpha/2,J}) := \alpha/2. \tag{7.8.10}$$

These values can be found by a Monte Carlo method. For instance, $z_{0.05,4} = 1.97$, $z_{0.05,10} = 2.1$, and $z_{0.01,10} = 2.49$. {The S-function **zalpha(a,J)** of our software toolkit allows one to get values of $z_{a,J}$.} Then a natural confidence band estimator is (Exercise 7.8.4)

$$\mathrm{CB} = \big[\hat{f}(t,Y) - z_{\alpha/2,J_\gamma}\sigma_n), \hat{f}(t,Y) - z_{\alpha/2,J_\gamma}\sigma_n)\big]. \tag{7.8.11}$$

Recall that $\sigma_n = (n^{-1}J_\gamma)^{1/2}$. A corresponding dual problem of hypothesis testing is considered absolutely similarly; see Exercise 7.8.5.

7.9 Exercises

7.1.1 Consider a single Gaussian channel $Y = X + Z$ where the noise Z is normal $N(0, \sigma^2)$ and the input X is an unknown constant (parameter). Let $w^* = X^2/(X^2 + \sigma^2)$. Show that mean squared error of a linear estimate $\hat{X} = wY$, where w is a constant (shrinkage weight), satisfies the relation

$$E\{(wY - X)^2\} \ge E\{(w^*Y - X)^2\} = w^*\sigma^2. \tag{7.9.1}$$

7.1.2 Consider the Gaussian channel of Exercise 7.1.1 and assume that the input X is a normal random variable $N(0, c^2\sigma^2)$ and Z is a normal $N(0, \sigma^2)$ noise independent of X. Show that any estimate $\tilde{X}$ of X based on output Y satisfies

$$E\{(\tilde{X} - X)^2\} \ge E\{(\lambda Y - X)^2\} = \lambda\varepsilon^2, \tag{7.9.2}$$

where $\lambda := c^2/(c^2 + 1)$.

7.1.3 Consider the Gaussian channel of Exercise 7.1.1 and assume that the input $X = \theta$ is a constant and $\theta \in [-c\sigma, c\sigma]$. Let $\hat{X}$ be any estimator of X based on both the output Y and the parameters c and σ. Show that there exists a random variable Θ with a cumulative distribution function $F(x)$ satisfying $F(-c\sigma) = 0$ and $F(c\sigma) = 1$ such that

$$\inf_{\hat{X}} \sup_{X \in [-c\sigma, c\sigma]} E\{(\hat{X} - X)^2\} \tag{7.9.3}$$

$$\geq \inf_{\hat{X}} \int_{-c\sigma}^{c\sigma} E\{(\hat{X}-x)^2\}dF(x) = E\{(E\{\Theta|Y\}-\Theta)^2\} \qquad (7.9.4)$$

$$= \mu(c)c^2\sigma^2/(1+c^2), \qquad (7.9.5)$$

where $\mu(c) \geq 0.8$ and $\mu(c) \to 1$ as $c \to \infty$ or $c \to 0$. Hint: The inequality (7.9.4) states that a minimax risk is not smaller than a corresponding Bayes risk. After working on Exercises 7.1.1–7.1.2, the relations (7.9.4) and (7.9.5) should be intuitively clear. Their rigorous proof is more involved; see Donoho, Liu, and MacGibbon (1990).

7.1.4 Compare the lower bounds of the previous exercises and discuss them.

7.1.5 Verify the lower bound (7.1.7).

7.1.6 Prove the relations (7.1.21).

7.1.7 Consider the variance and the squared bias terms of the MSE in (7.1.20). One of them is negligible in comparison to the other. Which one? Also, is it possible to find a cutoff that makes these two terms comparable? Would you recommend to use this cutoff?

7.1.8 Verify (7.1.25).

7.1.9 Suggest a proof of Theorem 7.1.2 for $0 < \alpha < 1$ using a wavelet basis as a set of coding functions. Hint: Use the characterization (2.5.3).

7.1.10 How does the fact that the sth derivative is the estimand affect the convergence of the minimax risks for Lipschitz and analytic functions? Explain why the difference is so dramatic.

7.1.11 Find the first and second derivatives of the mollifier and draw their graphs.

7.1.12 Consider a set of functions (7.1.31). What is the support of $f_k(t)$? What can be said about $f_k^{(l)}(0)$ and $f_k^{(l)}(1)$, $l = 0, 1, \ldots$?

7.1.13 Let $f_k(t)$ be as defined in (7.1.31). Show that if all the X_j are in the interval $[-c\sigma, c\sigma]$ with a sufficiently small c, then $f_k \in Lip_{r,\alpha,L}([0,1])$. Hint: explore how many coding functions $g_j(t)$ vanish at points $t = u$ and $t = v$; then use (7.1.30).

7.1.14 Use Remark 7.1.2 and indicate necessary changes in the proofs for the case of positive and decreasing $f(t)$ on $[0,1]$ functions. Also, for Theorem 7.1.1 consider the case where underlying functions are probability densities supported on $[0,1]$.

7.1.15 Let $f(t) = \sum_{j=1}^{2\lfloor \sigma^{-2/(2\beta+1)} \rfloor} X_j \varphi_{j-1}(T)$, where $\{\varphi_j\}$ is the classical trigonometric basis (2.4.1). Show that for any combination of values $\{X_j\}$ such that $\{X_j \in [-c\sigma, c\sigma], j = 1, \ldots\}$ the function $f(t)$ belongs to a Sobolev function space $W_{\beta,Q}$, defined in (2.4.19) whenever c is sufficiently small. Then use this result to find the asymptotic minimax MISE over this function class. Hint: Follow along the steps of the proof of Theorem 7.1.1. The answer should be the same as for a Lipschitz space with the same parameter of smoothness β.

7.1.16 Prove that the assertion of Theorem 7.1.1 holds for the local minimax introduced in Remark 7.1.3.

7.1.17 Prove the assertion of Remark 7.1.4. Hint: Follow along the lines of the proof of Theorem 7.1.1 and use the following technical results: $\sum_{j=0}^{J_\gamma}[\int_0^t \varphi_j(x)dx]^2 = t - 2\pi^{-2}J_\gamma^{-1}(1+o_\sigma(1))$, which may be proved using Parseval identity for the function $I_{\{x\in[0,t]\}}$ (see the line below (7.2.3)); Check that $E\{(E\{\Theta_j|Y_j\}-\Theta_j)^2\} = \sigma^2(1+o_\sigma(1)\ln^{-1}(\sigma^{-1}))$.
7.1.18 Can a second-order efficient estimator be suggested under MISE criteria?
7.1.19 Check (7.1.34)–(7.1.36).
7.1.20 Verify the assertion of the last paragraph in Remark 7.1.5.
7.1.21 Let us compare the optimal smoothing and hard thresholding. Suppose that $Y = \theta + \sigma\xi$ where ξ is a standard normal random variable and θ is an estimated parameter. (a) Show that $\lambda^* = \theta^2/(\theta^2+\sigma^2)$ minimizes $\mathrm{MSE}(\lambda) := E\{(\lambda Y - \theta)^2\}$ over all real λ. (b) Show that $\lambda' = I_{\theta^2>\sigma^2}$ minimizes the $\mathrm{MSE}(\lambda)$ over all $\lambda \in \{0,1\}$. (c) Verify the relation $\frac{1}{2} \le \mathrm{MSE}(\lambda^*)/\mathrm{MSE}(\lambda') \le 1$.
7.2.1 Show that (7.2.1) is an unbiased estimate of $f_k(t)$, i.e., $E\{y_k(t)\} = f_k(t)$.
7.2.2 Let Z_j, $j = 0,1,\ldots$, be iid normal $N(0,\sigma^2)$. Show that for each $t \in [0,1]$ a random variable $B_{2k}(t) := \sigma^{-1}\sum_{j=0}^{2k} Z_j \int_0^t \varphi_j(u)du$ is normal with zero mean and variance $\sum_{j=0}^{2k}(\int_0^t \varphi_j(u)du)^2$.
7.2.3 One of the classical definitions of a standard Brownian motion is as follows. A standard Brownian motion starting at level zero and defined on $[0,T]$ is a stochastic process $\{B(t),\ 0 \le t \le T\}$ satisfying the conditions: (a) $B(0) = 0$; (b) $B(t_2) - B(t_1)$, $B(t_3) - B(t_2), \ldots, B(t_n) - B(t_{n-1})$ are independent for every integer $n \ge 3$ and every $0 \le t_1 < t_2 < \cdots < t_n \le T$; (c) The random variable $B(u) - B(v)$ is normal $N(0, u-v)$ for $0 \le v < v \le T$. Then, let $B_1(t)$ and $B_2(t)$ be two independent standard Brownian motions. Describe properties of a linear combination of these processes.
7.2.4 Let a standard Brownian motion be defined according to Exercise 7.2.3. Then existence of this process may be established by the following experiment. Let Z'_l, $l = 1,2,\ldots$, be iid realizations of a standard normal random variable and let $\{g_j,\ j = 1,2,\ldots\}$ be elements of an orthonormal basis in $L_2[0,T]$. Set $B_k(t) := \sum_{j=1}^k Z'_j \int_0^t g_j(u)du$. Show that $B_k(t)$ converges almost surely to a standard Brownian motion on $[0,T]$ as $k \to \infty$. Hint: See the book by Gihman and Skorohod (1974).
7.2.5 Consider the communication system (7.1.1) shown in Figure 7.1. Assume that the noise is normal and coefficients $\{X_j\}$ of an input signal $f(t) = \sum_{j=1}^k X_j g_j(t)$ are sent n times via this system to improve the reliability of the transmission. Show that such a method makes this system equivalent to the system (7.1.1) with the smaller noise level $\sigma^* = \sigma/n^{1/2}$.
7.2.6 Write down all the steps of the universal estimate for filtering a signal from a white noise.
7.2.7 Use the principle of equivalence and formulate Theorems 7.1.1 and 7.1.2 for a density model.

7.2.8 Use the principle of equivalence and formulate Theorems 7.1.1 and 7.1.2 for a regression model.
7.2.9 Use the principle of equivalence and formulate Theorems 7.1.1 and 7.1.2 for a density model with right-censored data. Hint: Recall that $\sigma^2 = d/n$ and use the coefficient of difficulty introduced in Section 3.4.
7.2.10 Prove the assertion of Remark 7.2.1 for estimation of a density $f \in A_{\gamma,Q} \cap \{\psi : |\psi(t) - f^*(t)| < n^{-1/3},\ f^*(t) > C > 0,\ 0 \le t \le 1\}$.
7.3.1 The proof of Theorem 7.3.1 has outlined how to establish the fact that the estimator (7.3.5) is both globally and pointwise rate optimal. Write down a step-by-step proof.
7.3.2 Prove that all the conditions of Theorem 7.3.1 hold for the case of a filtering model and $\hat\theta_j = \int_0^1 \varphi_j(t)dY(t)$.
7.3.3 Explain all the steps in the lines (7.3.14).
7.3.4 Prove Bernstein's inequality (7.3.15). Hint: Begin with

$$(\sum_{j=0}^{\infty} |\theta_j|)^2 = (\sum_{j=0}^{\infty} (1+j)^{-\beta}(1+j)^{\beta}|\theta_j|)^2 \le \sum_{j=0}^{\infty}(1+j)^{-2\beta} \sum_{j=0}^{\infty}(1+j)^{2\beta}|\theta_j|^2.$$

The assumption $\beta := r + \alpha > 0.5$ implies that $\sum_{j=0}^{\infty}(1+j)^{-2\beta}$ converges. The inequality $\sum_{j=0}^{\infty}(1+j)^{2\beta}\theta_j^2 < C < \infty$ is verified by a direct calculation. (Several different proofs may be found in Bary 1964, Section 2.3.)
7.3.5 Finish the outlined proof of (7.3.9) for the density model.
7.3.6 Verify that conditions of Theorem 7.3.1 hold for the estimate (7.3.16).
7.3.7 Consider the model of the spectral density estimation discussed in Section 5.2, and suggest an estimator $\hat\theta_j$ of the correlation coefficients $\theta_j = E\{X_t X_{t+j}\}$ that satisfies the conditions of Theorem 7.3.1.
7.4.1 Verify (7.4.3).
7.4.2 Verify (7.4.6).
7.4.3 Establish (7.4.8). Hint: Use (2.4.19) and find the largest J such that $\theta_j^2 = c\ln(n)/n$, $0 \le j \le J$, and the corresponding f still belongs to $W_{\beta,Q}$. Then calculate the value of $Jc\ln(n)/n$.
7.4.4 Verify (7.4.8) for a Hölder space and a wavelet basis. Hint: Use the characterization (2.5.3).
7.4.5 Let $Y = \theta + \sigma Z'$, where Z' is a standard normal random variable. Show that $Y^2 - \sigma^2$ is an unbiased estimate of θ_j^2.
7.4.6 Verify (7.4.9).
7.4.7 Explain why in (7.4.10) the maximal J_n^* can be chosen as $\lfloor n/\ln(n) \rfloor$.
7.4.8 Explain how (7.4.16) is obtained.
7.4.9 Verify (7.4.17).
7.4.10 Check the optimality of the block shrinkage (7.4.24).
7.4.11 Prove (7.4.25). Hint: $E\{(Z')^{2k}\} = (2k-1)(2k-3)\cdots 1$.
7.4.12 Show that MSE of the hard-threshold estimator (7.4.45) attains the adaptive minimax rate.
7.4.13 Repeat Figure 7.5 with different coefficients and find optimal ones.

7.4.14 Consider the problem of adaptive estimation of $f(t_0)$ from $Lip_{r,\alpha,L}$ for the filtering model (7.3.1). It has been shown in Section 7.3 that if the parameter of smoothness $\beta = r + \alpha$ is given, then a plug-in de la Vallée Poussin sum $\hat{V}_J(t_0)$ satisfies (7.3.11). Let $\beta > 0.5$. Then the data-driven cutoff $\hat{J}$ is defined as follows. Set J_m^* to be the integer part of $(\ln(n))^2 d^m$, where $d > 2$ is a fixed constant and $m = 0, 1, \ldots; K$ to be the maximum integer satisfying the inequality $J_K^* < n^{1/2}(\ln(n))^{-3}$; β_m to be a solution of equation $\lfloor n/\ln(n) \rfloor^{1/(2\beta+1)} = J_m^*$; J_m to be equal to J_m^* that is closest to $J_m^*[\ln(n)]^{1/(2\beta_m+1)}$; $\tilde{I}(i,j) = \tilde{f}_n(j,0) - \tilde{f}_n(i,0)$; $\tilde{k} = \min\{l : |\tilde{I}(J_l^*, J_m^*)|^2 \le 6\ln(n) J_m^* n^{-1}, l \le m \le K; 0 \le l \le K\}$. Then we set $\hat{J} = J_{\tilde{k}}$ and use $\hat{V}_{\hat{J}}(t_0)$ as an adaptive estimate of $f(t_0)$. Show that

$$\sup_{f \in Lip_{r,\alpha,L}} \mathrm{MSE}(\hat{V}_{\hat{J}}(t_0), f(t_0)) \le C(n/\ln(n))^{-2\beta/(2\beta+1)}. \tag{7.9.6}$$

Hint: Solution and discussion may be found in Efromovich and Low (1994).

7.4.15 Check all the steps of the proof of Theorem 7.4.1.

7.5.1 Verify that a function (7.5.1) is Hölder H_β whenever $|X_{J,s^m}| \le cn^{-1/2}$ for a sufficiently small c.

7.5.2 Prove the inequality (7.5.3).

7.5.3 Explain (7.5.4).

7.5.4 Verify (7.5.10).

7.6.1 Give examples of statistical models where $\hat{\theta}_j^2 - n^{-1}$ is an unbiased estimate of the squared Fourier coefficient θ_j^2.

7.6.2 Explain the steps in (7.6.5).

7.6.3 Verify (7.6.6)–(7.6.7).

7.6.4 Let $\hat{\theta}_j = \theta_j + n^{-1/2} Z_j'$ where Z_j' are iid standard normal. Check that

$$E\Big\{\Big[\sum_{j=m}^{M} j^{2k}(\hat{\theta}_j^2 - n^{-1} - \theta_j^2)\Big]^2\Big\} = n^{-1}\Big[4\sum_{j=m}^{M} j^{4k}\theta_j^2 + 2n^{-1}\sum_{j=m}^{M} j^{4k}\Big]. \tag{7.9.7}$$

7.6.5 Using the result of the previous exercise, check the validity of (7.6.12) for the case $\beta > 2s + 0.25$.

7.6.6 Suggest an estimate of $F_0(f)$ whose mean squared error converges as $n^{-8\beta/(4\beta+1)}$ for $f \in W_{\beta,Q}$ where $0 < \beta \le 0.25$.

7.6.7 Verify (7.6.14).

7.6.8 Consider the heteroscedastic nonparametric random design regression discussed in Section 4.2. Suggest a data-driven estimator of the integral of the squared regression function.

7.7.1 Consider a monotone estimate, that is, an estimate that can be only monotone. Show that such an estimate cannot fit a nonmonotone density in the sense that the MISE does not decay.

7.7.2 Calculate the Fisher information for the model (i) of a location parameter. Is it always constant?

7.7.3 Consider a binomial regression and calculate the optimal design of the experiment. Hint: Use (7.7.6).

7.7.4 Consider the Exercise 7.7.3 for a scale regression.
7.8.1 Show that for $Y = \theta + \sigma\xi$, with ξ being standard normal, the probability for the parameter θ to be covered by the interval (7.8.1) is $1 - \alpha$.
7.8.2 Find $\beta(\theta) := P(Y \in R(\theta_0, \alpha)|\theta)$, where a rejection region $R(\theta_0, \alpha)$ is defined in (7.8.2). The $\beta(\theta)$ is called the *power function* of the test.
7.8.3 Consider a class of tests such that the probability of a rejection region given a null hypothesis is at most the level of significance α and the probability of a rejection region given an alternative hypothesis (*power*) is at least α. Such tests are called *unbiased*. If additionally under any alternative hypothesis an unbiased test maximizes the power (minimizes the second type error) over all unbiased tests, then this test is called *uniformly most powerful unbiased* (UMPU). The corresponding inverted confidence interval is called *uniformly most accurate unbiased*. Indicate such tests and confidence intervals among those considered in the section.
7.8.4 Verify that the probability that $f(t)$ is covered by the band (7.8.11) for all $0 \le t \le 1$ is at least $1 - \alpha + o_n(1)$.
7.8.5 Suggest a test dual to the confidence band (7.8.11). Hint: Invert (7.8.11).

7.10 Notes

First, let us make several general remarks.

• There is a deep connection between statistics and communication (information) theory. As an example, let $Y_1, \dots, Y_n$ be random variables generated according to a density $f(y^n|\theta)$, $y^n := (y_1, \dots, y_n)$, where θ is a realization of a random variable Θ with density g. This problem may be treated as a classical statistical (Bayesian) one where based on n observations the statistician should estimate θ. It also may be considered as a classical problem of communication (information) theory where a signal θ is sent n times via a channel whose outcomes are distributed according to the density $f(y^n|\theta)$. As we have discussed in Section 7.7 (see also Lehmann and Casella 1998, Section 2.5), for the iid case the Fisher information $I(\theta) = \int (\partial f(y|\theta)/\partial\theta)^2 f^{-1}(y|\theta) dy$ is the quantity that describes the statistical setting in terms of an optimal estimation. For communication theory a similar quantity is the *Shannon information*

$$S(\Theta, Y^n) := \int g(\theta) f(y^n|\theta) \ln\left(\frac{f(y^n|\theta)}{\int g(u) f(y^n|u) du}\right) dy^n d\theta. \qquad (7.10.1)$$

This quantity describes the information in the output signals $Y_1, \dots, Y_n$ about the input signal Θ. It was established in the 1970s that there is a precise asymptotic relationship between these two informations. For instance,

for the iid case under mild assumptions,

$$S(\Theta, Y^n) = \frac{1}{2}\ln\left(\frac{n}{2\pi e}\right) + \int g(\theta)\log\left(\frac{I^{1/2}(\theta)}{g(\theta)}\right)d\theta + o_n(1). \qquad (7.10.2)$$

Similar assertions are established for the multiparameter case, dependent observations, sequential samplings, etc. Pinsker (1972) initiated this research by proving (7.10.2) for the case of additive channels discussed in Section 7.1 and by establishing beautiful results about connections between the information theory and statistics. These results were further developed in a series of publications; see, for instance, Ibragimov and Khasminskii (1973), Efromovich (1980a), and Clarke and Barron (1990).

The exact asymptotic relationship between Shannon and Fisher informations not only explains why the filtering model was used in this chapter as the basic one for the asymptotic analysis, but it also sheds light on the equivalence principle discussed in Section 7.2.

• In this book only fixed sample sizes have been considered. In many cases, especially if a given precision of estimation is required, sequential plans of sampling may be needed. These are the plans where based on previous observations, the data analyst may either stop observations and take an action (estimate an underlying function or test a hypothesis) or continue the sampling and ask about the next observation. This is the setting where adaptive estimators shine, and the theory becomes extremely interesting (and mathematically more involved). The sequential setting is discussed in Prakasa Rao (1983), Efromovich and Pinsker (1989), and Efromovich (1980b, 1989, 1995b, 1999c), where further references may be found.

• The recent books by Härdle et al. (1998) and Nemirovskii (1999) cover some modern topics in series estimation. Ibragimov and Khasminskii (1981, Chapter 7), Devroye and Györfi (1985, Chapter 12) and Eubank (1988, Chapter 3) also present the asymptotic results for series estimates.

7.1 The first sharp minimax result is due to Pinsker (1980). It was established for a filtering model and Sobolev spaces, and a series approach was used. This series approach was used later for obtaining sharp minimax results for density, spectral density, and different regression models, including heteroscedastic models and generalized models. See the discussion in Efromovich and Pinsker (1982, 1984, 1996b) and Efromovich (1986, 1996a).

Ibragimov and Khasminskii (1984) introduced the function $\mu(c)$ used in Lemma 7.1.1. Donoho, Liu, and MacGibbon (1990) give a historical overview of exploring this function and the lower bound.

Results on sharp estimation of analytic densities supported on the real line may be found in Golubev and Levit (1996) and Golubev, Levit, and Tsybakov (1996), where second-order efficiency is also discussed. Sharp results are also known for different loss functions, for instance, the case of the sup-norm is discussed in Korostelev (1993) and Donoho (1994). The research on sharp local minimax was initiated by Golubev (1991).

There is a deep connection between the parametric, semiparametric, and nonparametric approaches. See the article by Koshevnik and Levit (1976) and the book by Bickel et al. (1993).

Bayesian approach is discussed in Berger (1985); see also Zhao (1993).

7.2 The principle of equivalence is formulated in the articles by Brown and Low (1996a) and Nussbaum (1996). Limits of the equivalence are discussed in Efromovich and Samarov (1996), Brown and Zhang (1998), and Efromovich (1999b). A discussion of Brownian motion and related filtering models may be found in the books by Ibragimov and Khasminskii (1981) and Mallat (1998).

7.3 The first results about rate optimal series estimation are due to Chentsov (1962); see also the book by Chentsov (1980). Rate optimal series estimates for different loss functions are discussed in the book by Ibragimov and Khasminskii (1981, Section 7.4). See also Stone (1980) and Samarov (1992). Influence of Kolmogorov's ideas on optimal nonparametric curve estimation is discussed by Ibragimov and Khasminskii (1990). Rate optimal estimation of linear functionals is discussed, for instance, in Ibragimov and Khasminskii (1987) and Donoho and Liu (1991).

7.4 Donoho and Johnstone (1994) is a good reference to read about the universal threshold procedure. The books on application of wavelets, mentioned in the notes to Chapter 4, also consider this basic method. Polyak and Tsybakov (1990) discuss optimality of the empirical risk minimization procedure for a projection series estimator. This article also contains further references. The penalization method is discussed in Barron, Birgé, and Massart (1999). The cross-validation technique is analyzed in the books by Eubank (1988) and Wahba (1990). DeVore and Temlyakov (1995) discuss the mathematical results on linear and nonlinear approximations. Important asymptotic results are due to Rubin and Vitale (1980).

The block shrinkage estimator is introduced and discussed in detail in Efromovich and Pinsker (1984, 1996) and Efromovich (1985, 1986, 1996c, 1997c, 1998c, 1999a). In those articles further references may be found.

Rate optimality of SureShrink over Besov spaces is established in Donoho and Johnstone (1995), where results of Monte Carlo simulations are presented as well; see also Goldenshluger and Nemirovski (1997). The block threshold method is suggested and explored in Hall, Kerkyacharian, and Picard (1998). See also Efromovich (1995a, 1996c) and Cai (1999).

The first result on the loss of a logarithmic factor for adaptive estimation of a function at a point and the first adaptive kernel estimator based on Lepskii procedure is due to Lepskii (1990), see also Lepskii (1992). Juditsky (1997) used this procedure to find an adaptive hard threshold wavelet estimator for the case of the L_p-loss function. The research on exploring subsets of Lipschitz spaces that are free from the logarithmic penalty was initiated in Efromovich and Low (1994, 1996a,b). In those articles the two-nets algorithm of bias–variance tradeoff is suggested, and its property to control both the penalty-free subset and data compression is analyzed.

In this section we discussed the cases of analytic and Lipschitz spaces. What will happen if a function belongs to the union of these spaces, that is, may it be smooth or supersmooth? Surprisingly, all the results remain valid, and no additional penalties should be paid; see Efromovich (1998a). In that article the loss of a sharp constant for adaptive MSE convergence and the case of analytic functions is established.

Rate-optimal, sharp, adaptive, versatile, and robust series estimators for different deconvolution models are discussed in Ermakov (1992), Korostelev and Tsybakov (1993, Chapter 9), Efromovich (1994c, 1997a,b), Hengartner (1997), and Efromovich and Ganzburg (1999), among others.

7.5 Sharp minimax results, including an adaptive estimator for spherical data, may be found in Efromovich (1994b).

7.6 Efromovich (1994a), Efromovich and Low (1996a,b), and Efromovich and Samarov(1996, 1999) discuss adaptive estimation of quadratic functionals.

7.7 Pinsker (1980) pioneered the race for constants. Proofs of the results may be found in Efromovich (1996a, 1997d). Nemirovskii, Polyak, and Tsybakov (1985) pioneered the discussion about linear versus nonlinear estimation.

7.8 Hart (1997) gives a book-length treatment of series approaches for hypothesis testing and related topics; see also Ingster (1993).

8

Nonseries Methods

This chapter reviews the main nonseries methods of nonparametric curve estimation. Whenever it is worthwhile, a method is explained for all the given statistical models.

8.1 The Histogram

The most widely used probability density estimator is the histogram. It is also the only density estimator that is studied in all undergraduate statistical classes and is supported by all statistical software.

Given an *origin* x_0 and a *bin width* h, we define the *bins* of a histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$, $m = 0, \pm 1, \pm 2, \ldots$.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. Then the *histogram* is defined by the formula

$$\hat{f}_n(x) := (1/nh)[\text{number of observations in the same bin as } x]. \quad (8.1.1)$$

To construct a histogram we have to choose both an origin and a bin width; the choice of a bin width primarily controls the amount of smoothing inherent in the procedure.

Three histograms with different width of bins for the same data simulated according to the Bivariate density (recall Figure 2.1) are shown in Figure 8.1. Here the origin $x_0 = 0$. The left histogram, based on only 5 bins over the interval $[0, 1]$, oversmooths the data, and as a result only one mode is shown. The histogram in the middle, based on 9 bins, correctly

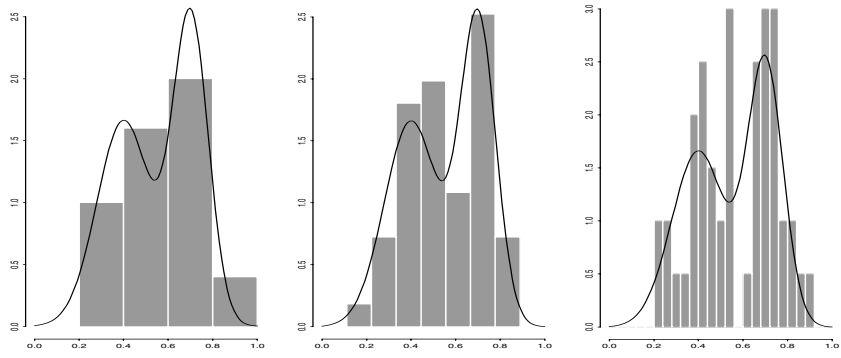

FIGURE 8.1. Histograms with different numbers of bins. A sample of size $n = 50$ is generated from the Bivariate corner density shown by the solid line. {The arguments n, *cdensity*, and *set.nb* allow one to choose the sample size, an underlying corner density from our set of 8 densities shown in Figure 2.1, and numbers of bins for the 3 diagrams, respectively.} [*n=50, cdensity=3, set.nb=c(5,9,25)*]

shows the number of modes. Finally, the right histogram, based on 25 bins, undersmooths the underlying density, but it nicely shows the data.

As we see, undersmoothing produces a wiggly picture with many artificial and confusing modes, while oversmoothing hides modes and obscures the fine structure. The patterns in Figure 8.1 reflect the most important issue for any nonparametric estimator, namely, how to smooth data. Some software, including S–PLUS, use approximately $\log_2(n)+1$ number of bins, which is the Sturge's formula motivated by a normal underlying density.

8.2 The Naive Density Estimator

From the definition of a probability density, if X has density $f(x)$, then

$$f(x) = \lim_{h \to 0} (2h)^{-1} P(x - h < X < x + h). \tag{8.2.1}$$

The *naive estimate*, based on n iid observations $X_1, \ldots, X_n$ of X, straightforwardly mimics this equality,

$$\hat{f}_n(x) := (2hn)^{-1}[\text{number of observations falling in } (x-h, x+h)]. \tag{8.2.2}$$

To express the estimator more transparently, define the *weight function*

$$w(x) := \frac{1}{2} I_{\{|x| \le 1\}}. \tag{8.2.3}$$

Then the naive estimator may be written as

$$\hat{f}_n(x) = (nh)^{-1} \sum_{l=1}^{n} w((x - X_l)/h). \tag{8.2.4}$$

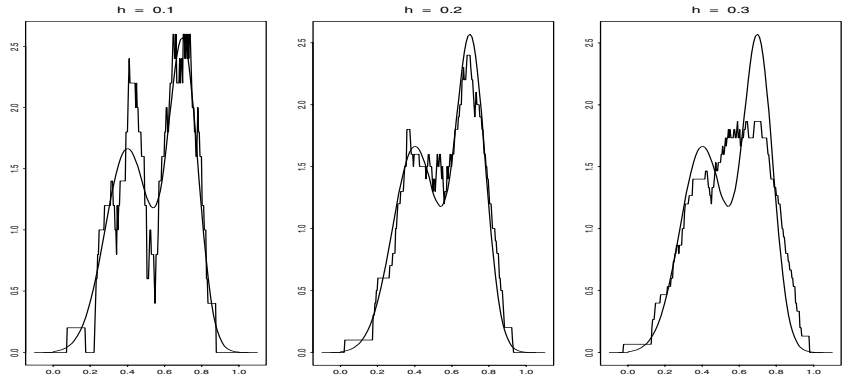

FIGURE 8.2. Naive density estimates of the Bivariate density with different widths h obtained for the same data set of size $n = 50$. {The choice of widths is controlled by the argument *set.h*, and the choice of an underlying corner density by *cdensity.*} [*n=50, cdensity=3, set.h=c(.1,.2,.3)*]

Figure 8.2 shows three naive estimates with different widths h for the case of 50 observations generated from the Bivariate density. As we see, the naive estimates are not wholly satisfactory either esthetically or for use as an estimate for presentation. Moreover, the ragged character of the plots can give a misleading impression about artificial modes. This behavior of the estimator follows from the definition that an estimate is step-wise constant with jumps at the points $X_l \pm h$.

A principal difference between the naive estimator and the histogram is that there is no origin x_0, and the center of the bin is an observation. We shall see in the next section that the naive estimator is a particular case of a large family of kernel estimators.

8.3 Kernel Estimator

In this section we discuss one of the main methods of modern nonparametric curve estimation theory, which is called kernel smoothing. Kernel smoothing, as with series estimation, can be used for any statistical model, so below we discuss its applications for probability density estimation, nonparametric regression, and spectral density estimation. Kernel estimators are certainly the most mathematically studied nonparametric method, and the fundamentals of this theory will be given in Section 8.9.

• **Probability Density Estimation.** The naive estimator introduced in the previous section was the first example of kernel estimation. Recall that the main drawback of the naive estimator was its stepwise nature. But it is easy to generalize the naive estimator to overcome this drawback. Namely, it suffices to replace in (8.2.4) the rectangular weight function

(8.2.3) by a smooth weight function. Kernel estimation theory refers to a weight function as a *kernel function* (or simply *kernel*) and denotes it by $K(x)$. By assumption the kernel function (as with the rectangle weight function) is integrated to unity, that is,

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{8.3.1}$$

Due to the last equation, any probability density function (for instance, a normal density) can be used as the kernel function. Then, as with the naive estimator, a *kernel density estimator* is defined by

$$\hat{f}_n(x) := (nh)^{-1} \sum_{l=1}^{n} K((x - X_l)/h), \tag{8.3.2}$$

where h is referred to as either the *bandwidth*, or *window width*, or *smoothing parameter*. To analyze a kernel estimator it is useful to keep in mind that if $K(x)$ is the density of a random variable Z, then

$$K_h(x) := h^{-1}K(x/h)$$

is the density of the scaled random variable hZ, that is, h is a scaling parameter. In particular, if $K(x)$ is a standard normal density, then h plays the role of the standard deviation. This interpretation of the bandwidth is helpful in choosing reasonable values for h. Note that the choice of a nonnegative kernel implies that an estimate is also nonnegative.

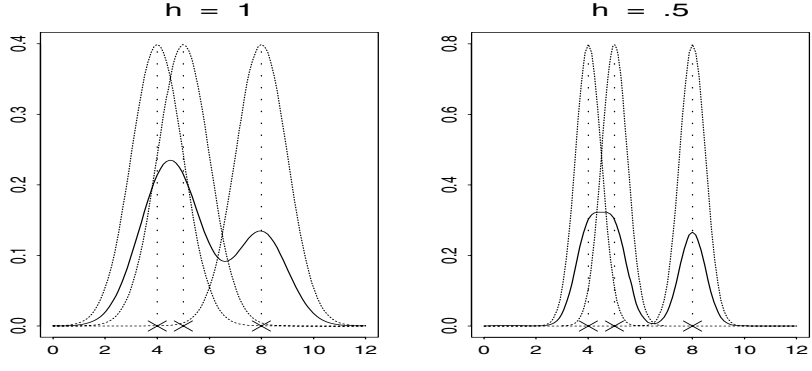

FIGURE 8.3. How the kernel density estimator works. In the left diagram 3 particular observations are shown by crosses. Standard Gaussian kernel functions (shown by dotted lines) are centered at each observation, and then the kernel density estimate (solid line) is the normed sum of these kernel functions. The right diagram shows a similar procedure only with the bandwidth $h = 0.5$. {The choice of 3 observations from the interval $[0, 12]$ is controlled by the argument *set.X*, and the choice of 2 bandwidths is controlled by the argument *set.h.*} [*set.h=c(1,.5), set.X=c(4,5,8)*]

Figure 8.3 illustrates how the kernel estimator (8.3.2) works for a particular case of 3 observations $X_1 = 4$, $X_2 = 5$, $X_3 = 8$, a standard Gaussian kernel (that is, $K(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is the standard normal density), bandwidths $h = 1$ and $h = 0.5$. The kernel estimator is constructed by centering a scaled kernel at each observation; then the value of a kernel estimate at a point x is the average of the 3 kernel ordinates at that point. We see that the kernel estimate spreads each "bump" with weight $1/n$ and therefore the combined contributions from each data point are larger in regions where there are many observations. Clearly, in these regions it is expected that the underlying density has a relatively large value. The opposite occurs in regions with relatively few observations.

A comparison of the estimates in Figure 8.3 to those in Figure 8.2 shows that the kernel estimate inherits all the continuity and differentiability properties of the kernel function. Also note that the effect of the bandwidth is very important. Figure 8.3 shows that changing the bandwidth from 1 to 0.5 dramatically affects the shape of the kernel estimate and transforms it from a bimodal shape to a strata shape. Second, the support of the kernel estimate shrinks from approximately $[0, 12]$ to $[2, 10]$. Thus, the bandwidth may dramatically change the shape of a kernel estimate.

Let us consider a numerical example for the case of 100 observations drawn from the Strata density; see Figure 8.4. Here the S–PLUS function **density** is used with a Gaussian kernel (note that this is not a standard

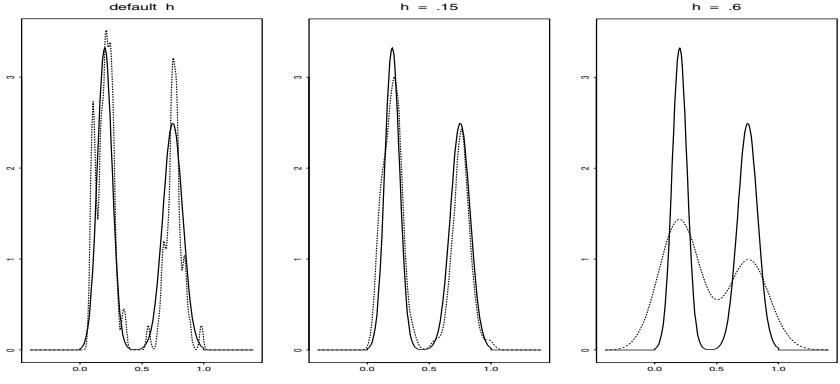

FIGURE 8.4. The effect of bandwidth on kernel density estimation. Each diagram corresponds to a specific bandwidth shown in the title. Kernel density estimators, based on $n = 100$ iid realizations from the Strata density, are shown by dotted lines. The solid line shows the underlying Strata. {The underlying density may be changed by the argument *cdensity*. The choice of the kernel is controlled by the argument *kernel*, the default is *kernel="gaussian"*, and the possible alternatives are *"cosine"*, *"triangular"*, and *"rectangular"*. The bandwidths for the second and third diagrams are controlled by the argument *set.h*.} [*cdensity=4, n=100, set.h=c(.15,.6), kernel="gaussian"*]

normal density). The default h is equal to the range of data divided by $2(1+\log_2 n)$; this rule is again motivated by Sturge's formula mentioned in Section 8.1. We see that this bandwidth is too small, and thus the spurious fine structure becomes visible. Furthermore, the small bandwidth creates the illusion of narrower support of an underlying density. On the other hand, the bandwidth $h = 0.6$ is too large for those particular data, and the strata nature of the underlying density is obscured. Also note that in this case the estimate gives a false impression by overestimating the support of the distribution. The graph in the middle indicates that the bandwidth $h = 0.15$ is just right for this particular data set.

What we see in these three graphs is rather typical for nonparametric estimates. The left graph is undersmoothed, the right graph is oversmoothed, and the middle graph looks good. This is the reason why the nonparametric estimation is often referred to as a smoothing technique.

• **Nonparametric Regression.** The kernel method may be used for other statistical models as well. Consider, for example, the following fixed design homoscedastic regression model,

$$Y_l := f(l/n) + \sigma\varepsilon_l, \quad l = 1, 2, \ldots, n, \tag{8.3.3}$$

where the errors ε_l are independent with zero mean and unit variance. Then a kernel estimator is defined by the formula

$$\hat{f}_n(x) := (nh)^{-1} \sum_{l=1}^{n} Y_l K((x - l/n)/h). \tag{8.3.4}$$

Note that this estimator simply performs moving averaging, or in other words, local averaging of responses. An example of such averaging for the case of the rectangular kernel (8.2.3) and the bandwidth $h = 0.3$ is shown in Figure 8.5. The data set, shown by crosses, is simulated by adding to the corner function Steps normal $N(0, (0.5)^2)$ errors.

While this particular kernel estimate for only 10 observations looks not too bad (but not too good either), it is not difficult to realize that the rectangular kernel may lead to extremely erratic and confusing estimation. {To see this using Figure 8.5, try, for instance, several simulations with the arguments $n = 20$ and $h = 0.08$.} Another important conclusion is that even for this rather artificial example with only 10 observations the kernel estimate is relatively good for the interior region but it is much worse near the boundaries, where it is about a half of the values of the underlying regression curve. This is especially apparent for the right boundary, where despite the presence of the large response, the kernel estimate goes downward.

To understand why the estimator exhibits such different behavior for interior and boundary points, it is worthwhile to explore this estimator mathematically. This is especially easy to do for the case of the kernel being the rectangular weight function (8.2.3). First, we plug (8.3.3) into

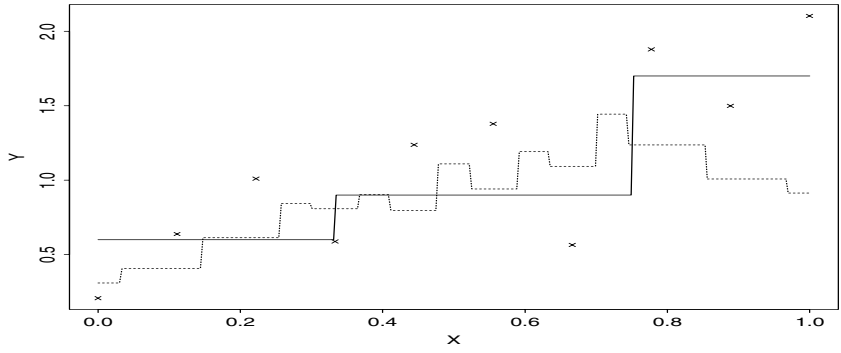

FIGURE 8.5. The performance of the kernel estimator (8.3.4) with the rectangular kernel function for regression model (8.3.3). The scatter plot is shown by crosses. The underlying regression function is the Steps, and it is shown by the solid line, and the estimate by the dotted line. {The argument *regrfun* allows one to change an underlying regression function.} [*n=10, h=.3, regrfun=8, sigma=.5*]

(8.3.4) and get

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{l=1}^{n} f(l/n)w((x-l/n)/h) + \frac{\sigma}{nh}\sum_{l=1}^{n}\varepsilon_l w((x-l/n)/h), \quad (8.3.5)$$

where $w(x)$ is the "box" (8.2.3). Let $h = d_n/n$, where d_n is an integer-valued sequence in n. Assume that x is an "interior" point of the support $[0,1]$, that is, $d_n/n < x < 1 - d_n/n$. Then (8.3.5) may be written as

$$\hat{f}_n(x) = \frac{1}{2d_n}\sum_{\{l:\, -d_n/n \le x-l/n \le d_n/n\}} f(l/n) + \frac{\sigma}{2d_n}\sum_{\{l:\, -d_n/n \le x-l/n \le d_n/n\}} \xi_l.$$

This expression allows us to understand all the main features of kernel estimators for the interior points. For example, consider the mean squared error for a point $x \ne l/n$, $l = 1, 2, \ldots$. Write

$$MSE := E\{[\hat{f}_n(x) - f(x)]^2\} \quad (8.3.6)$$

$$= E\{[(E\{\hat{f}_n(x)\} - f(x)) + (\hat{f}_n(x) - E\{\hat{f}_n(x)\})]^2\}$$

$$= (E\{\hat{f}_n(x)\} - f(x))^2 + E\{(\hat{f}_n(x) - E\{\hat{f}_n(x)\})^2\} =: \text{SBIAS} + \text{VAR} \quad (8.3.7)$$

$$= \Big[(2d_n)^{-1}\sum_{\{l:\, -d_n/n \le x-l/n \le d_n/n\}} \{f(l/n) - f(x)\}\Big]^2 + (2d_n)^{-1}\sigma^2. \quad (8.3.8)$$

In the line (8.3.7) we used the relation $E\{(c+Z)^2\} = c^2 + E\{Z^2\}$, which holds for any constant c and random variable Z with zero mean and finite

variance. Also, in this line the familiar notation for the squared bias and the variance terms of MSE were introduced.

Thus (8.3.6)–(8.3.8) show that optimal estimation is based on a tradeoff between the squared bias term and the variance term. The bandwidth h plays a crucial role in this tradeoff. Namely, the bias is decreased by making the bandwidth smaller, but the variance is decreased by making the bandwidth larger. (This resembles how a cutoff affects the MSE of an orthogonal series estimate, and it is possible to show that an optimal bandwidth is, roughly speaking, inversely proportional to an optimal cutoff. This remark is helpful to understand the dynamics of kernel estimation.)

Now let us explore more precisely how the bandwidth affects the MSE. Consider the case where an underlying regression function f is Lipschitz $Lip_\alpha(L)$ of order α, that is, $|f(x+\delta) - f(x)| \le L|\delta|^\alpha$, $0 < \alpha \le 1$. (Note that we do not assume here that f is 1-periodic and to emphasize this use a slightly different notation than in Section 2.4.) Then

$$\begin{aligned} \text{SBIAS} &\le \Big[(L/2d_n) \sum_{\{l:\, -d_n/n \le x - l/n \le d_n/n\}} |x - l/n|^\alpha \Big]^2 \\ &\le (1+\alpha)^{-2} L^2 h^{2\alpha} (1 + 1/(hn))^{2\alpha}. \end{aligned} \tag{8.3.9}$$

Here we used the definition $h = d_n/n$. Using this inequality on the right-hand side of (8.3.8), we get

$$\text{MSE} \le (1+\alpha)^{-2} L^2 h^{2\alpha} (1 + 1/(hn))^{2\alpha} + \sigma^2/(2hn). \tag{8.3.10}$$

Let us find the optimal bandwidth h_n^* that minimizes the right-hand side of (8.3.10) for sufficiently large n. Exercise 8.3.13 gives us the solution,

$$h_n^* = [(1+\alpha)^2 \sigma^2 / 4\alpha L^2]^{1/(2\alpha+1)} n^{-1/(2\alpha+1)} (1 + o_n(1)). \tag{8.3.11}$$

Recall that $o_n(1)$ denotes a generic decaying sequence in n as $n \to \infty$, i.e., $o_n(1) \to 0$ as $n \to \infty$. Substituting the optimal h_n^* into the right-hand side of (8.3.10) gives us for $f \in Lip_\alpha(L)$,

$$\text{MSE} \le \Big[\frac{L\sigma^{2\alpha}}{(1+\alpha)(4\alpha)^\alpha}\Big]^{2/(2\alpha+1)} (1+2\alpha) n^{-2\alpha/(2\alpha+1)} (1 + o_n(1)). \tag{8.3.12}$$

Thus, the smoother the underlying function (i.e., the larger α), the larger the optimal bandwidth and the smaller the corresponding MSE. This result together with the conclusion of Section 7.3, that $\sup_{f \in Lip_\alpha(L)} \text{MSE}$ cannot decrease faster than $n^{-2\alpha/(2\alpha+1)}$, shows that this kernel estimator is asymptotically rate optimal for an interior point x and for Lipschitz functions of a known order $0 < \alpha \le 1$.

The situation changes drastically for boundary points. Consider the case of a continuous regression function. Then it follows from (8.3.5) that when x is 0 or 1 we have $E\{\hat f(x)\} \to f(x)/2$ as $h \to 0$, and this is what we have seen in Figure 8.5. This implies that the kernel estimator is not even consistent at the boundary points unless $f(0) = f(1) = 0$. An intuitively simple way

of improving the estimator is clear from the first term on the right-hand side of (8.3.5). Namely, instead of the denominator nh one should use a denominator that is equal to the number of nonzero summands in that sum. Another way to solve the problem is to use special boundary kernels.

Apart from thinking about the kernel estimate (8.3.4) as a moving average, there is another way of looking at this estimate that is also useful for understanding how to generalize this estimate for the case of randomly or unequally spaced predictors.

Let us assume that the kernel function $K(x)$ is unimodal, symmetric about zero, integrated to unity, and it has vanishing tails. All the examples of kernel functions considered so far satisfy this assumption. Also assume that $f(x)$ is continuous in x near some point x_0. Then for sufficiently small h we have

$$f(x_0) \approx \int_{-\infty}^{\infty} f(x) K_h(x - x_0) dx, \tag{8.3.13}$$

where $K_h(x) = h^{-1}K(x/h)$. The integral in (8.3.13) is called the *convolution* integral, and the idea of such an approximation is called the approximation in h by an *integral operator with kernel* K.

Let us verify (8.3.13) for the case of the rectangular kernel (8.2.3). Write

$$h^{-1} \int_{-\infty}^{\infty} f(x) w((x - x_0)/h) dx = (2h)^{-1} \int_{x_0 - h}^{x_0 + h} f(x) dx \approx f(x_0),$$

where the last relation holds because due to the continuity of $f(x)$ near point x_0 we have

$$\max_{|t| \le h} |f(x_0 + t) - f(x_0)| \to 0 \quad \text{as} \quad h \to 0.$$

Figure 8.6 illustrates graphically both the convolution formula and the approximation by the integral operator. Here the kernel is a standard normal density and $h = 0.05$. The solid line shows a function $f(x)$ that is approximated at point $x_0 = 0.6$. The dotted line shows the kernel $K_h(x - x_0)$ centered at the point x_0. Then the dashed line shows the product $f(x)K_h(x - x_0)$. Note that this product is asymmetric about x_0, since the function $f(x)$ is increasing in x near x_0. Then, according to (8.3.13), the value $f(x_0)$, shown by the cross, is approximately equal to the integral of the dashed line. To assess this integral, the dot–dash line shows the function $\psi(x) = \int_{-\infty}^{x} f(u) K_h(u - x_0) du$, which asymptotically (as $x \to \infty$) is equal to the integral (8.3.13). We see that to the right of $x = 0.8$ the function $\psi(x)$ flattens out, so $\psi(1) \approx \psi(\infty)$. The value $\psi(1)$ is shown by the crossed rectangle. The last step is to check that $f(x_0) \approx \psi(1)$, and the validity of this relation is apparently supported by the graph.

Another useful interpretation of the convolution formula, based on some probabilistic ideas, is discussed in Exercise 8.3.15.

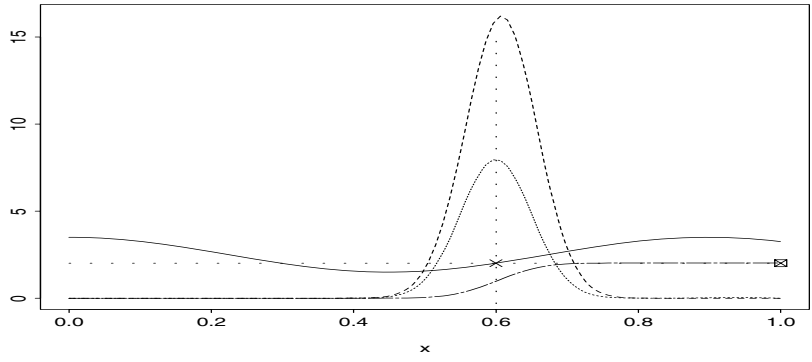

FIGURE 8.6. Illustration of the convolution formula (8.3.13).

Now we are in a position to use the approximation by the kernel operator for the case of random design regression. Assume that a data set (Y_l, X_l), $l = 1, 2, \dots, n$, is generated by the model $Y := f(X) + \varepsilon$ where the predictor X is a random variable with density $g(x)$ and ε is a zero-mean error independent of the predictor. Then using (8.3.13) we write,

$$f(x_0) \approx \int_{-\infty}^{\infty} f(x) K_h(x - x_0) dx = E\{f(X) K_h(X - x_0)/g(X)\}$$
$$= E\{Y K_h(X - x_0)/g(X)\}. \tag{8.3.14}$$

Thus, the estimated function is equal to the expectation of the product $Y K_h(X - x_0)/g(X)$. Using the familiar sample mean estimator we get the following kernel estimator:

$$\tilde{f}(x_0) := n^{-1} \sum_{l=1}^{n} Y_l K_h(X_l - x_0)/g(X_l). \tag{8.3.15}$$

Note that this estimator coincides with (8.3.4) if $g(x) = 1$, $0 \le x \le 1$ (the design density is uniform on $[0, 1]$). It is also a simple exercise to repeat these calculations and see that the estimator is absolutely natural for the case of fixed design predictors with design density $g(x)$; see (4.2.2).

In many practical applications the design density $g(x)$ is unknown and should be estimated based on the data. The first and absolutely natural idea is to plug in the kernel density estimate (8.3.2). Such substitution implies the *Nadaraya–Watson* kernel estimator

$$\hat{f}_n(x) := \frac{\sum_{l=1}^{n} Y_l K_h(x - X_l)}{\sum_{l=1}^{n} K_h(x - X_l)}. \tag{8.3.16}$$

Another idea of estimating $g(x)$, discussed in detail in Section 4.2, is based on the fact that spacings between ordered predictors are inversely proportional to the underlying density. Using this idea in (8.3.15) leads to

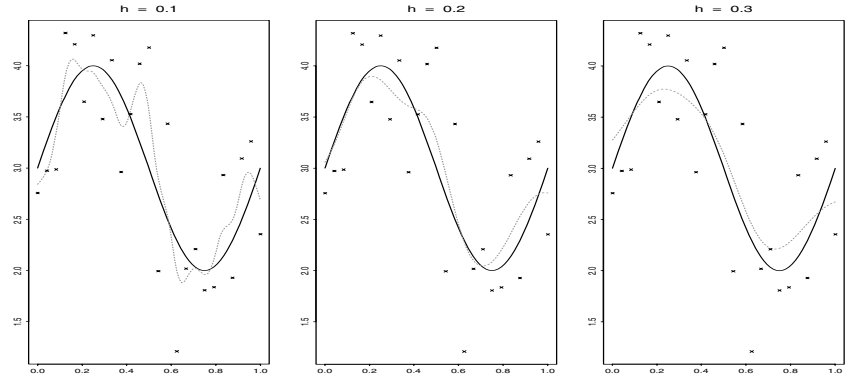

FIGURE 8.7. Nadaraya–Watson kernel estimates with 3 different bandwidths for equidistant regression model with normal $N(0, \sigma^2)$ additive errors. A scatter plot for $n = 25$ observations is shown by crosses. Estimates are shown by the dotted lines, underlying regression function by the solid line. {The default kernel is *"normal"*, the possible alternatives are *"box"* and *"triangle"*. The standard deviation σ is controlled by the argument *sigma*. The argument *set.h* allows one to choose 3 bandwidths.} [*n=25, sigma=.5, kernel= "normal", set.h = c(.1,.2,.3)*]

either the *Priestly–Chao* kernel estimator

$$\hat{f}_n(x) := \sum_{l=1}^{n} Y_{(l)}(X_{(l)} - X_{(l-1)})K_h(X_{(l)} - x) \tag{8.3.17}$$

or the *Gasser–Müller* kernel estimator

$$\hat{f}_n(x) := \sum_{l=1}^{n} Y_{(l)} \int_{s_{l-1}}^{s_l} K_h(u - x)du, \tag{8.3.18}$$

with $s_{l-1} := (X_{(l)} + X_{(l-1)})/2$, $(Y_{(l)}, X_{(l)})$ being sorted according to ordered predictors $X_{(0)} \le X_{(1)} \le \cdots \le X_{(n+1)}$, $X_{(0)} = -\infty$, $X_{(n+1)} = \infty$.

These are the three main types of classical kernel estimators.

The Nadaraya–Watson estimator is supported by the S–PLUS function **ksmooth**. Figure 8.7 illustrates the performance of this estimator based on 25 simulated data generated by the underlying regression function $2.5+\cos(7x)$ and normal $N(0, (0.5)^2)$ additive error. Three kernel estimates with different bandwidths are shown. The left one, with the smallest bandwidth, apparently undersmooths the data, and as a result, the estimate is too wiggly and too sensitive to "outliers." The second estimate, corresponding to $h = 0.2$, looks better. However, look at the following peculiar feature of the estimate: It is too low and too high at the peak and valley, respectively. This is a typical performance of a kernel estimator whose bias is smallest where the underlying regression function is almost linear. The right diagram shows that the bandwidth $h = 0.3$ is too large and the data are oversmoothed.

• **Spectral Density Estimation.** The notion of the spectral density and its importance in time series analysis was discussed in Section 5.2. Recall that if $\{X_t,\ t = \ldots, -1, 0, 1, \ldots\}$ is a second-order stationary time series with mean 0 and autocovariance function $\gamma(j) := E\{X_{t+j}X_t\}$, then under mild assumptions (for instance, the condition $\sum_{j=0}^{\infty} |\gamma(j)| < \infty$ is sufficient), the spectral density function is defined as

$$f(\lambda) := (2\pi)^{-1}\gamma(0) + \pi^{-1}\sum_{j=1}^{\infty}\gamma(j)\cos(j\lambda), \qquad -\pi < \lambda \le \pi. \tag{8.3.19}$$

Here the frequency λ is in units radians/time.

Let a finite sample $X_1, \ldots, X_n$ be given; then the familiar sample autocovariance estimator is defined by $\hat\gamma(j) := n^{-1}\sum_{l=1}^{n-j} X_{l+j}X_l$. Then, a natural step to estimate the spectral density is to plug the sample autocovariance function into the right-hand side of (8.3.19). At the so-called Fourier frequencies $\lambda_k := 2\pi k/n$, the resulting estimator (up to the factor $1/2\pi$) is called a periodogram,

$$I(\lambda_k) := \hat\gamma(0) + 2\sum_{j=1}^{n}\hat\gamma(j)\cos(j\lambda_k). \tag{8.3.20}$$

The underlying idea of kernel spectral density estimation is based on the remarkable theoretical result that under mild assumptions, for sufficiently large n the periodogram can be approximately written as

$$(2\pi)^{-1}I(\lambda_k) \approx f(\lambda_k) + f(\lambda_k)\xi_k, \qquad \lambda_k \in (0, \pi), \tag{8.3.21}$$

where ξ_k are zero-mean random variables with bounded moments.

Thus, at least for large n the problem of spectral density estimation resembles an equidistant nonparametric regression model where values of the periodogram at Fourier frequencies play the role of responses. Then a kernel estimation (smoothing) may be used straightforwardly.

8.4 Local Polynomial Regression

Let us begin with recalling the underlying idea of a linear least-squares regression. It is assumed that pairs (Y_l, X_l) of observations satisfy the linear model

$$Y_l := \beta_0 + \beta_1 X_l + \varepsilon'_l, \quad l = 1, 2, \ldots, n, \tag{8.4.1}$$

where the errors ε'_l are independent random variables with zero mean and finite variances. Using the least-squares criterion, the estimated y-intercept $\hat\beta_0$ and the slope $\hat\beta_1$ are defined as the minimizers of the sum of squared

errors,

$$(\hat{\beta}_0, \hat{\beta}_1) := \text{argmin}_{(\beta_0,\beta_1)} \sum_{l=1}^{n} (Y_l - \beta_0 - \beta_1 X_l)^2. \tag{8.4.2}$$

Then the linear least-squares regression is defined as $\hat{f}(x) := \hat{\beta}_0 + \hat{\beta}_1 x$.

Now let is look at the scatter plot shown by crosses in Figure 8.8. It is absolutely clear that a straight line cannot satisfactorily fit these data. But does this mean that the least-squares idea has no application for such a data set? The answer is "no."

For instance, one may try to fit the data locally by a straight line. Indeed, if $f(x)$ is sufficiently smooth, then Taylor's expansion implies

$$f(x) \approx f(x_0) + (x - x_0) f'(x_0) \tag{8.4.3}$$

for all x in a small neighborhood of x_0. Thus, a straight line can fit a data set locally. In this case the classical linear model (8.4.1) becomes

$$Y_l := \beta_0(x) + \beta_1(x) X_l + \varepsilon''. \tag{8.4.4}$$

Define the functions $\hat{\beta}_0(x)$ and $\hat{\beta}_1(x)$, which are the minimizers for a sum of locally weighted squared errors,

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) := \text{argmin}_{(\beta_0(x),\beta_1(x))} \sum_{l=1}^{n} (Y_l - \beta_0(x) - \beta_1(x) X_l)^2 K_h(x - X_l). \tag{8.4.5}$$

Here, as in the previous section, $K(x)$ is the kernel (kernel function), $K_h(x) := h^{-1} K(x/h)$, and h is the bandwidth. Then the estimator

$$\hat{f}(x) := \hat{\beta}_0(x) + \hat{\beta}_1(x) x \tag{8.4.6}$$

is called a *local linear regression smoother* or *local linear fit*. Note that in (8.4.5) every observation (the pair of predictor and response) affects the choice of the local y-intercept and slope with weight equal to the height of the function K_h at the point equal to the distance between the predictor and the point x. Thus, the farther the predictor from x, the smaller the effect of the response on the estimate. As a result, the bandwidth h dramatically affects that influence. Another useful point of view in (8.4.5) is to consider a local linear fit as a weighted least-squares regression at the point x.

The idea of local linear regression is illustrated by Figure 8.8. The scatter plot for 50 equidistant predictors is shown by crosses, and the observations are generated by a regression function (the dashed line) plus normal $N(0, (0.5)^2)$ errors. Let us consider the local linear estimation for the point $x = 0.3$ using the rectangular kernel (8.2.3) and $h = 0.15$. The dotted line shows $K_{0.15}(x - 0.3)$ centered at the point $x = 0.3$. Then the formula (8.4.5) implies that for this rectangular kernel all weights are the same within the support of this kernel, and thus one needs to find an ordinary

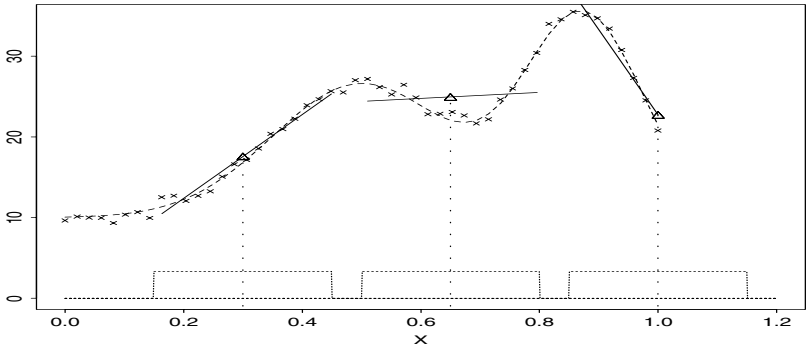

FIGURE 8.8. Idea of a local linear fit. The dashed line is an underlying regression function. Simulated data are shown by crosses. The dotted line shows rectangular kernels with $h = 0.15$. Local linear fits for points $x = 0.3$, $x = 0.65$, and $x = 1$ are shown by triangles, and the corresponding local linear least-squares regression lines are shown by the solid lines.

least-squares regression based only on observations with predictors X_l such that $0.3 - h \le X_l \le 0.3 + h$, that is, with predictors that belong to the support of this rectangular kernel. This ordinary regression is shown by the solid line, and its value at the point $x = 0.3$, shown by the triangle with the X-coordinate $x = 0.3$, is the value of the local linear regression at $x = 0.3$. Note that this triangle fits the underlying curve very nicely. The situation is not so rosy for the case $x = 0.65$, where a similar approach gives a poor fit. Clearly, the reason is that the valley requires an essentially smaller value of bandwidth. The third point is $x = 1$, which nicely illustrates why the idea of local linear regression is so appealing for boundary points.

Figure 8.9 illustrates how the kernel function and bandwidth affect the local linear estimation. The data set is simulated as in Figure 8.8. The rectangular kernel is (8.2.3), and the Gaussian kernel is a standard normal density. We see that as with the classical kernel estimators, discussed in the previous section, local linear estimates are affected by the smoothness of the kernel. However, this effect is not so drastic as a wrongly chosen bandwidth. In short, for samples of small sizes bandwidth is the main factor to look for, and typically all smooth kernels will do a similar job. Now let us turn our attention to the boundary points. For the points near the right edge, the local linear estimation is almost perfect whenever the bandwidth is not too large (in this case oversmoothing occurs). On the other hand, the situation is more complicated with the left edge, where the estimates do not flatten out. This tiny left tail is a challenging problem for a local linear estimator because the underlying curve sharply changes, and thus a smaller bandwidth is needed. On the other hand, a smaller bandwidth increases the

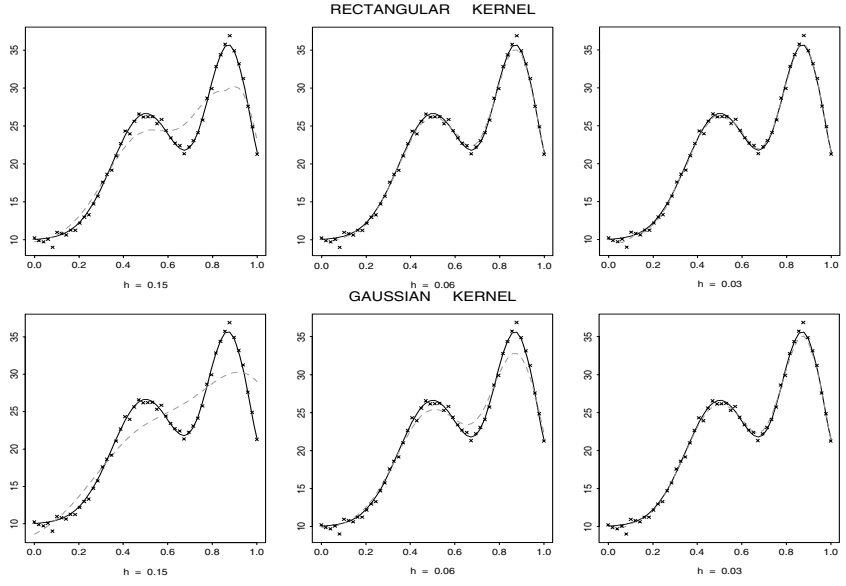

FIGURE 8.9. The effect of the kernel and the bandwidth on the local linear regression. A scatter plot of 50 observations of an equidistant regression with iid normal $N(0, \sigma^2)$, $\sigma = 0.5$, additive errors (the same for all the 6 diagrams) is shown by crosses. The underlying regression function is shown by the solid line and the local linear regressions by the dashed lines. The estimates in the top row of the diagrams are obtained using the rectangular kernel, the estimates in the bottom row by the Gaussian kernel. The bandwidths are shown in the subtitles. {The argument *set.h* controls bandwidths.} [*n=50, sigma=.5, set.h = c(.15,.06,.03)*]

effect of particular errors, and we clearly see how a single relatively large error (which implies the small response at $x = 0.08$) affects the estimates.

Of course, a local linear fit is not the only possible one. *Local constant* fit is another possibility, and in general, a *local polynomial fit* is an alternative. Let p be the degree of the polynomial being fit. At a point x_0 the estimator $\hat{f}(x, p, h)$ is obtained by fitting the polynomial

$$\beta_0 + \beta_1(x - x_0) + \cdots + \beta_p(x - x_0)^p \tag{8.4.7}$$

to a data set $\{(Y_l, X_l), l = 1, 2, \ldots, n\}$ using weighted squares with kernel weights $h^{-1}K((X_l - x_0)/h)$. The value of

$$\hat{f}(x, p, h) := \sum_{j=0}^{p} \hat{\beta}_j(x) x^j, \tag{8.4.8}$$

is the height of the polynomial fit at the point x where the vector-function $(\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))$ minimizes

$$\sum_{l=1}^{n} [Y_l - (\beta_0(x) + \beta_1(x)(X_l - x) + \cdots + \beta_p(x)(X_l - x)^p)]^2 K_h(X_l - x). \quad (8.4.9)$$

The estimate (8.4.8) is called a *local polynomial regression of order p.*

Simple explicit formulae exist for the case of a local constant regression ($p = 0$) where the estimator coincides with the Nadaraya–Watson kernel estimator,

$$\hat{f}(x, 0, h) = \frac{\sum_{l=1}^{n} Y_l K_h(x_l - x)}{\sum_{l=1}^{n} K_h(x_l - x)}. \quad (8.4.10)$$

The local linear estimator ($p = 1$) is

$$\hat{f}(x, 1, h) = \frac{n^{-1} \sum_{l=1}^{n} [\hat{s}_2(x, h) - \hat{s}_1(x, h)(X_l - x)] K_h(X_l - x) Y_l}{\hat{s}_2(x, h)\hat{s}_0(x, h) - [\hat{s}_1(x, h)]^2}, \quad (8.4.11)$$

where

$$\hat{s}_r(x, h) := n^{-1} \sum_{l=1}^{n} (X_l - x)^r K_h(X_l - x). \quad (8.4.12)$$

The local polynomial smoothers inherit the drawbacks of classical polynomial regressions, and one of the main ones is their high sensitivity to extreme observations, i.e., to *outliers* in response variables. Thus it is preferable to have a *robust* method that is resistant to outliers.

Locally weighted scatter plot smoothing (LOWESS) is a procedure that makes the locally weighted least squares method more robust to outliers. This procedure is supported by the S–PLUS function **lowess**. There are several steps in this procedure. First, a local polynomial fit is calculated. Second, residuals are found. Third, weights are assigned to each residual: Large (respectively small) residuals receive small (respectively large) weights. Fourth, a local polynomial fit is calculated one more time, but now by assigning to each observation a new weight that is the product of the weight at the initial fit and the weight assigned to its residual from that initial fit. Thus the observations showing large residuals in the initial fit (and which are possibly outliers) are downweighted in the second fit. The above process is repeated several times, resulting in the estimator LOWESS.

8.5 The Nearest Neighbor Method

The nearest neighbor method is a procedure typically used for density estimation. The underlying idea of the method is based on the definition of the probability density,

$$f(x) = \lim_{h \to 0} (2h)^{-1} P(x - h < X < x + h),$$

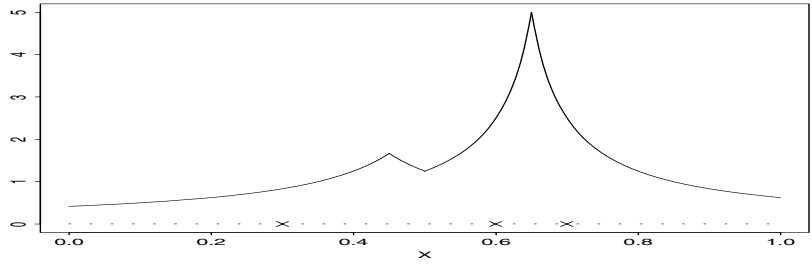

FIGURE 8.10. Nearest neighbor density estimate with $k = 2$ for the case of 3 observations. The estimate is shown by the solid line over the unit interval; the observations are shown by crosses.

and then noting that we expect $k = n(2h)f(x)$ observations falling in a box of width $2h$ and centered at the point of interest x.

Recall that the naive estimator, discussed in Section 8.2, is based on using a fixed bandwidth h, calculating the number $\hat{k}$ of observations such that they belong to the interval $[x - h, x + h]$, and then setting

$$\hat{f}_n(x) := \frac{\hat{k}}{n2h}. \tag{8.5.1}$$

In contrast, the nearest neighbor method is based on a fixed number of points k that determines the width of a box in a search. Thus, we calculate the Euclidean distance $\hat{h}$ from the point of interest x to the distant kth observation and define the *kth nearest neighbor density estimate* by

$$\tilde{f}_n(x) := \frac{k}{n2\hat{h}}. \tag{8.5.2}$$

The similarity between (8.5.1) and (8.5.2) is striking.

Note that for x less than the smallest data point $X_{(1)}$ we have $\hat{h}(x) = X_{(k)} - x$. Here $X_{(k)}$ denotes the kth ordered observation. Thus the estimate (8.5.2) is proportional to $|x|^{-1}$ as $x \to -\infty$. We observe the same behavior for the right tail of the estimate. Thus, tails are estimated extremely badly and need to be discarded from consideration.

Figure 8.10 illustrates this estimator for the case $k = 2$ and three data points 0.3, 0.6, and 0.7. The graph nicely shows the underlying idea of this estimator, namely, that density is inversely proportional to the size of the box needed to contain a fixed number k of observations. The drawback of the nearest neighbor method is that the derivative of a nearest neighbor estimate is discontinuous. As a result, the estimate can give a wrong impression. Also, this estimate is not integrable due to its heavy tails.

On the other hand, the underlying idea of this estimate that the probability density is inversely proportional to the distance between ordered observations (spacing) is very attractive. Recall that it has been intensively used, for instance, in Sections 4.2 and 8.3.

The idea of the nearest neighbor method can be used in a kernel estimator where the bandwidth is chosen to be $\hat{h}$. Such a kernel estimator is called a *kth neighbor kernel estimate*,

$$\tilde{f}_n(x) := \big(n\hat{h}(x)\big)^{-1} \sum_{l=1}^{n} K\big((x - X_l)/\hat{h}(x)\big), \tag{8.5.3}$$

and this is a kernel estimate with a data-driven bandwidth. However, this is not an entirely data-driven method, because a choice of k should be made. Note that this generalized estimate becomes the ordinary kth nearest neighbor estimate when the kernel function is rectangular.

An appealing feature of the nearest neighbor method is its simplicity in expanding to a multivariate setting. To define a nearest neighbor estimate in s-dimensional space, let $d_k(\mathbf{x})$ be the Euclidean distance from $\mathbf{x}$ to the kth nearest data point, and let $V_k(\mathbf{x})$ be the (s-dimensional) volume of the s-dimensional sphere of radius $d_k(\mathbf{x})$. Thus $V_k(\mathbf{x}) = c_s[d_k(\mathbf{x})]^d$, where c_s is the volume of the s-dimensional sphere with unit radius, that is, $c_1 = 2$, $c_2 = \pi$, $c_3 = 4\pi/3$, etc. Then, the nearest neighbor method is defined by

$$\tilde{f}_n(\mathbf{x}) := \frac{k}{nV_k(\mathbf{x})}. \tag{8.5.4}$$

Note that if we set the kernel function $K(\mathbf{x}) := 1/c_k$ within the sphere of unit radius and $K(\mathbf{x}) := 0$ otherwise, then the nearest neighbor method is identical to a kernel smoothing. This connection between the kernel and nearest neighbor method demonstrates that a study of the nearest neighbor method can be based on the theory of kernel estimation.

8.6 The Maximum Likelihood Method

We begin with recalling the idea of maximum likelihood estimation of a parameter. A maximum likelihood estimate (MLE) of a parameter is defined as the value of the parameter that maximizes the probability (likelihood) of the observed data. Let f_θ be an underlying density given the parameter θ and let n iid observations $X_1, X_2, \ldots, X_n$ be given. Then a *maximum likelihood estimate* $\hat{\theta}$ is defined as the maximizer of the *likelihood* $\prod_{l=1}^{n} f_\theta(X_l)$, that is, $\prod_{l=1}^{n} f_{\hat{\theta}}(X_l) = \max_\theta \prod_{l=1}^{n} f_\theta(X_l)$.

For example, let X be normal with unknown mean θ and variance σ^2. Then the maximum likelihood estimate, based on only one observation X_1, is $\hat{\theta} = X_1$, because this value of the parameter θ maximizes the likelihood $(2\pi\sigma^2)^{-1/2} \exp\{-(X_1 - \theta)^2/2\sigma^2\}$.

Similarly, for the case of a density estimation based on n iid realizations $X_1, \ldots, X_n$, we can define the *likelihood* of a density $f(x)$ as

$$L(f|X_1, \ldots, X_n) := \prod_{l=1}^{n} f(X_l). \tag{8.6.1}$$

Unfortunately, if we shall try to maximize this likelihood over all possible densities, the likelihood can be made arbitrarily large. Indeed, just think about a density that is a mixture of n normal densities with means X_l and variance σ^2. Such a density has an arbitrarily large likelihood as the variance becomes smaller, but due to its "bumpy" structure it is clearly far from any smooth underlying density.

There are several possible cures for this problem. Before exploring them, let us consider a problem that became the glory of the maximum likelihood method and that was elegantly solved in the 1950s by Grenander.

Assume that it is known that an underlying density is bounded and monotone (nonincreasing) on its support $[0, \infty)$. Then Grenander explored the maximum likelihood estimate for this particular class of densities, that is, when the maximum of (8.6.1) was taken only over the bona fide (monotone) densities. His first assertion was that the MLE is a step function with breakpoints (jumps) at the order statistics $X_{(l)}$, $l = 1, 2, \ldots, n$. And the next step was a closed procedure for finding the MLE of an underlying density, namely, the MLE is the density whose distribution function is the smallest concave majorant of the empirical distribution function.

All the steps of Grenander's estimate are illustrated in Figure 8.11.

Note that the success of the MLE for the case of monotone densities has been due to the restriction of a class of considered densities. Thus, it is not surprising that Grenander suggested to restrict the class of densities for the general setting as well and to choose f only from a given sequence

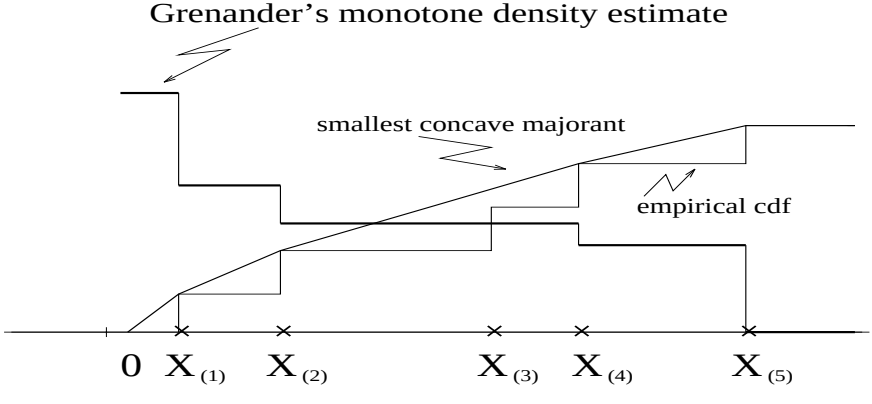

FIGURE 8.11. Steps for finding the MLE estimate of a bounded nonincreasing density supported on $[0, \infty)$. The first step is to arrange the data (here 5 observations shown by crosses) in ascending order from the smallest to the largest, i.e., find order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(5)}$. The second step is to draw the empirical cumulative distribution function, which is a step function with jumps equal to the inverse sample size (here 0.2) at the order statistics. The third step is to find the smallest concave majorant of the empirical cdf. Finally, the slopes of this majorant give us the stepwise MLE, which is Grenander's estimate.

of classes $\mathcal{S}_n$. Such a sequence is called a *sieve*, and the estimation method is called the *method of sieves*.

A particular example of such an estimate is the histogram. To see this, choose $\mathcal{S}_n := \{f : f \text{ is constant on the interval } [(j-1)h_n, jh_n),\ j \text{ is integer}, h_n \text{ decays as } n \to \infty\}$. Then it is not difficult to see that in this case the maximum likelihood sieve estimator is

$$\hat{f}_n(x) := (h_n n)^{-1} \sum_{l=1}^{n} I_{\{X_l \in [(j-1)h_n, jh_n)\}}, \quad x \in [(j-1)h_n, jh_n). \quad (8.6.2)$$

Indeed, we are looking for parameters c_j (heights of a maximum likelihood histogram) such that $h_n \sum_{j=-\infty}^{\infty} c_j = 1$ and the likelihood $\prod_{j=-\infty}^{\infty} c_j^{\nu_j}$ (or the logarithm of this product $\sum_{j=-\infty}^{\infty} \nu_j \ln(c_j)$, which is called the *log-likelihood*) takes on a maximum value. Here $\nu_j = \sum_{l=1}^{n} I_{\{X_l \in [(j-1)h_n, jh_n)\}}$, and it is assumed that $0^0 := 1$ and $0 \times \infty := 0$. According to the method of Lagrange multipliers, optimal parameters should maximize

$$\sum_{j=-\infty}^{\infty} \nu_j \ln(c_j) - \mu h_n \sum_{j=-\infty}^{\infty} c_j, \quad (8.6.3)$$

where μ is a Lagrange multiplier (a real number).

Recall that the *method of Lagrange multipliers* states that all local extrema of the function $f(x_1, \ldots, x_s)$, subject to the constraint $g(x_1, \ldots, x_s) = 0$, will be found among those points $(x_1, \ldots, x_s)$ for which there exists a real number μ such that

$$\partial F(x_1, \ldots, x_s, \mu)/\partial x_l = 0, l = 1, \ldots, s,\ \partial F(x_1, \ldots, x_s, \mu)/\partial \mu = 0, \quad (8.6.4)$$

where

$$F(x_1, \ldots, x_s, \mu) := f(x_1, \ldots, x_s) - \mu g(x_1, \ldots, x_s), \quad (8.6.5)$$

assuming that all the indicated partial derivatives exist. In short, given a constraint one should find relative extrema of a weighted sum of the aim function and the constraint function. (Exercise 8.6.4 gives an elegant example of using this method.)

Via differentiation of (8.6.3) we see that the optimal parameters are $c_j^* = \nu_j/\mu h_n$, and then $\mu = n$ because the constraint $h_n \sum_{j=-\infty}^{\infty} c_j^* = 1$ should hold and $\sum_{j=-\infty}^{\infty} \nu_j = n$ by definition of ν_j. This implies (8.6.2).

Another particular example of a sieve is $\mathcal{S}_n = \{f : \ f \in D, \ \ R(f) \le C\}$. Such an approach leads to a *maximum penalized likelihood* method. Here D is a given class of densities, $R(f)$ is a *penalty function*, and C is a constant.

The Lagrange multipliers method implies choosing the f in D that maximizes the *penalized log-likelihood*,

$$l_\mu(f) := \sum_{l=1}^{n} \ln(f(X_l)) - \mu R(f), \quad (8.6.6)$$

where $\mu \geq 0$ is the Lagrange multiplier or so-called smoothing parameter. The probability density that maximizes this likelihood is called a *maximum penalized likelihood density estimate*. Several examples of the penalty function are $R(f) := \int_{-\infty}^{\infty}(a[df(x)/dx]^2 + b[d^2f(x)/d^2x]^2)dx$ and $R(f) := \int_{-\infty}^{\infty}([d^2f(x)/d^2x]^2/f(x))dx$.

Note that while (8.6.6) is motivated by the sieve approach, it may be looked at as a penalization approach. In this case μ should be chosen rather than the constant C in the constraint $R(f) \leq C$. Since the choice of μ is more intuitively clear, it is customary to discuss this problem via the parameter μ rather than C.

The parameter μ in (8.6.6) (or respectively the constant C) becomes similar to the bandwidth or the cutoff for kernel and series estimates, respectively. It controls the amount of smoothing, and increasing μ leads to smoother estimates and vice versa. This is why μ is called a *smoothing parameter*.

As with parametric maximum likelihood estimation, the main problem in the implementation of a nonparametric maximum likelihood method is computational: How does one find that maximum? Curiously enough, some recommended computational approaches are based on an orthogonal series approach. Namely, one writes $\hat{f}(x) := \sum_{j=0}^{J}\theta_j\varphi_j(x)$, where φ_j are elements of an orthogonal basis, and then applies the penalized maximum likelihood method to $J+1$ parameters $(\theta_0, \ldots, \theta_J)$. As a result, such a maximum likelihood method is converted into the series approach discussed in the previous chapters.

8.7 Spline Approximation

We begin with fundamentals.

• **Review of Interpolation Theory.** Assume that a table of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is given, and assume that the x_l's form an increasing sequence of distinct points. The table represents n points in the Cartesian plane, and we would like to connect these points by a smooth curve. Thus we seek to determine a curve that is defined for all x and that takes on the corresponding values y_l for each of the distinct x_l's in this table. Such a curve is said to *interpolate* the table, and the points x_l are called *nodes*.

The first and absolutely natural idea of finding an interpolating curve is to use a polynomial function in x, that is, $p_n(x) := \sum_{i=0}^{n-1} a_i x^i$. Such a function is called an *interpolating* polynomial of degree n. Note that at each of n nodes this polynomial satisfies $p_n(x_l) = y_l$. Then a natural expectation is that if the table represents points of an underlying function $f(x)$ (that is, $f(x_l) = y_l$), then the function $f(x)$ will be well approximated by the interpolating polynomial $p_n(x)$ at all intermediate points. Moreover,

it might be expected that as the number of nodes increases, this agreement will become better and better.

However, in the history of mathematics a severe shock occurred when it was realized that this expectation was ill-founded. A counterexample is provided by the Runge function $f(x) := 1/(1 + x^2)$, $x \in [-5, 5]$, with $n + 1$ equidistant nodes including the endpoints. (Note that the Runge function is proportional to a truncated Cauchy density, which is a familiar "troublemaker" in statistics.) It has been proven that

$$\lim_{n\to\infty} \max_{x\in[-5,5]} |p_n(x) - f(x)| = \infty.$$

In short, a polynomial approximation of this extremely smooth function is wiggly, and the maximal error of the approximation at nonnodal points increases beyond all bounds!

The moral is that a polynomial interpolation of a high degree with many nodes is a risky operation, since the resulting polynomials may be very unsatisfactory as representations of the underlying functions. (It is of interest to note that this conclusion resembles the conclusion of the theory of orthogonal series estimation that recommends against using large cutoffs.)

To avoid this phenomenon, it can be worthwhile to relax the assumption that $f(x)$ should be globally (for all x) interpolated by a polynomial, and instead use a piecewise local polynomial interpolation. Such an interpolation is called a *spline function* or simply a *spline*.

A spline function $S(x)$ is a function that consists of polynomial pieces joined together with certain smoothness conditions.

A simple example is a *polygonal* function (or first-degree spline) whose pieces are linear polynomials joined together. See Figure 8.12, where an example of such a spline is shown by the dotted line.

An x-coordinate at which a spline function changes its character is called a *knot*. In Figure 8.12 the knots are 0, 0.2, 0.4, 0.6, 0.8, and 1. Between knots x_j and x_{j+1} we define a first-degree spline as $S(x) = a_j x + b_j =: S_j(x)$. This spline is piecewise linear. Usually, $S(x)$ is defined as $S_1(x)$ for $x < x_1$ and as $S_{n-1}(x)$ for $x > x_n$, where x_1 and x_n are the boundary knots.

A *second-degree spline* is a piecewise quadratic polynomial such that $S(x)$ and its derivative $S^{(1)}(x)$ are continuous.

The negative side of linear and quadratic splines is that the slope changes abruptly for a linear spline, as it does for the second derivative of a quadratic spline. This makes the curve not pleasing to the eye.

A cubic (third-degree) spline is such that $S(x)$ is a piecewise cubic polynomial with continuous first $S^{(1)}(x)$ and second $S^{(2)}(x)$ derivatives. This is the spline that is most often used in applications, and the reason why is clear from Figure 8.12.

Let us calculate the number of parameters (degrees of freedom) of a cubic spline based on n knots and the number of restrictions (side conditions). Since between knots x_j and x_{j+1} a cubic spline is a cubic polynomial, that

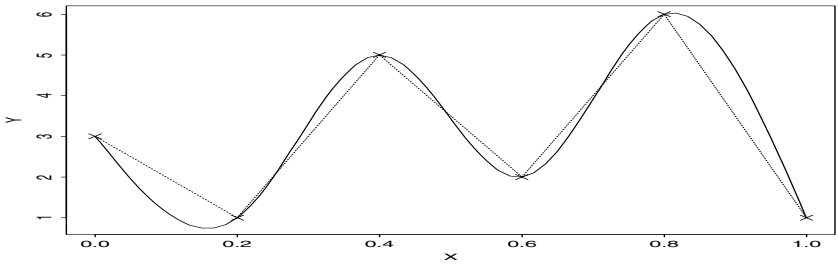

FIGURE 8.12. First-degree spline function (dotted line) and natural cubic spline (solid line) for the case of 6 equally spaced knots.

is, $S_j = a_j + b_j x + c_j x^2 + d_j x^3$, there are $4(n-1)$ variables (check this with Figure 8.12) and $2 + 2(n-2) + 2(n-2) = 4(n-1) - 2$ constraints. These constraints are due to the necessity for a cubic spline to be equal to given values at every knot, and its first and second derivatives should be continuous at every interior point. Since we are two restrictions short, let us add two constraints, $S^{(2)}(x_1) = S^{(2)}(x_n) = 0$, and then refer to such a spline as the *natural cubic spline*.

As we mentioned in the beginning of this section, a global polynomial interpolation may lead to undesired oscillations. In contrast, natural cubic spline interpolation nicely matches the smoothness of an underlying function. This follows from the following famous result of spline theory.

Theorem 8.7.1. *If S is the natural cubic spline function that interpolates a twice differentiable function f at knots $x_1 < x_2 < \cdots < x_n$, then*

$$\int_{x_1}^{x_n} [S^{(2)}(x)]^2 dx \le \int_{x_1}^{x_n} [f^{(2)}(x)]^2 dx. \tag{8.7.1}$$

Proof. Let $g(x) := f(x) - S(x)$. Then at the knots we have $g(x_l) = 0$ because S exactly fits f at the knots. Also, $f^{(2)}(x) = S^{(2)}(x) + g^{(2)}(x)$. Thus,

$$\int [f^{(2)}(x)]^2 dx = \int [S^{(2)}(x)]^2 dx + \int [g^{(2)}(x)]^2 dx + 2 \int g^{(2)}(x) S^{(2)}(x) dx.$$

We see that the equality $\int_{x_1}^{x_2} g^{(2)}(x) S^{(2)}(x) dx = 0$ implies (8.7.1), so it suffices to verify this equality. We shall do this by applying the technique of integration by parts (recall (2.2.3)) to the integral in question. Write

$$\int_{x_1}^{x_n} g^{(2)}(x) S^{(2)}(x) dx = \sum_{j=1}^{n-1} \Big[S^{(2)}(x) g^{(1)}(x) \Big|_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} S^{(3)}(x) g^{(1)}(x) dx \Big]$$

$$= \left[S^{(2)}(x_n)g^{(1)}(x_n) - S^{(2)}(x_1)g^{(1)}(x_1)\right] - \sum_{j=1}^{n-1} \int_{x_j}^{x_{j+1}} S^{(3)}(x)g^{(1)}(x)dx.$$

The first term is equal to zero because our spline is a natural cubic spline, i.e., $S^{(2)}(x_1) = S^{(2)}(x_n) = 0$. To calculate the second term we note that a cubic spline between two knots can be written as $a_j + b_j x + c_j x^2 + d_j x^3$, and then

$$\int_{x_j}^{x_{j+1}} S^{(3)}(x)g^{(1)}(x)dx = 6d_j \int_{x_j}^{x_{j+1}} g^{(1)}(x)dx = 0$$

because $g(x_j) = f(x_j) - S(x_j) = 0$. The theorem is proved.

While spline functions are an appealing tool for interpolating smooth functions, finding them numerically is not a simple task. Thus, special spline functions have been developed that are well adapted to numerical tasks. The first example is the *B-spline*, which forms a "basis" for the set of all splines.

Suppose we have an infinite set of knots $\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots$. Then the jth B-spline of zero-degree is defined by $B_j^0(x) = 1$ if $x_j \le x < x_{j+1}$ and $B_j^0(x) = 0$ otherwise. In short, $B_j^0(x)$ is a "box" of unit height (a rectangular kernel) placed on the interval $[x_j, x_{j+1})$.

With the function B_j^0 as a starting point we now generate all the higher-degree B-splines by a simple recursive formula:

$$B_j^k(x) = \frac{(x - x_j)B_j^{k-1}(x)}{x_{j+k} - x_j} + \frac{(x_{j+k+1} - x)B_{j+1}^{k-1}(x)}{x_{j+k+1} - x_{j+1}}, \quad k \ge 1. \tag{8.7.2}$$

Then a kth-degree B-spline is defined as

$$S^k(x) := \sum_{j=-\infty}^{\infty} \theta_j^k B_{j-k}^k(x). \tag{8.7.3}$$

A basic question is how to determine the coefficients θ_j^k in this expansion. Note that since B-splines of positive degree are not orthogonal, there are no simple formulae similar to those we have for an orthogonal series expansion. Nevertheless, direct calculations show that the zero- and first-degree interpolating B-splines are extremely simple,

$$S^0(x) = \sum_{j=-\infty}^{\infty} y_j B_j^0(x), \qquad S^1(x) = \sum_{j=-\infty}^{\infty} y_j B_{j-1}^1(x).$$

For higher-degree splines, some arbitrariness exists in choosing these coefficients. We shall not pursue this issue further, since in statistical applications one is interested in an approximation rather than an interpolation, and then a least-squares approach can be used. But it is worthwhile to note that as in the zero- and first-degree cases, interpolating splines are linear in y_j.

There exist several other useful spline bases, including the Kimeldorf–Wahba and the Demmler–Reinsch bases. These two bases are particularly useful as theoretical tools, and the latter resembles a trigonometric basis.

Finally, let us recall the following illuminating result, which has motivated the smoothing spline estimator discussed below.

Assume that an m-fold differentiable function $f(x)$ is supported on an interval $[a, b]$ and that it satisfies the following restrictions: (i) $f(x_l) = y_l$, $l = 1, 2, \ldots, n$; (ii) the $(m-1)$th derivative $f^{(m-1)}(x)$ is continuous in x. Then the problem is to find among all such functions the function that has the minimal integral of its squared mth derivative, that is, the function with the minimal value of $\int_a^b (f^{(m)}(x))^2 dx$.

It has been shown that the solution of this problem is unique and the function in question is a polynomial spline satisfying the restriction (i) with x_l being knots in addition to satisfying the following three side conditions: (a) f is a polynomial of degree not larger than $m-1$ when $x \in [a, x_1]$ and $x \in [x_n, b]$; (b) f is a polynomial of degree not larger than $2m-1$ for the interior points $x \in [x_l, x_{l+1}], l = 1, 2, \ldots, n$; (c) $f(x)$ has $2m-2$ continuous derivatives on the real line.

In short, the minimizer f^* is a spline with polynomial pieces joined at the knots x_l so that f^* has $2m-2$ continuous derivatives. Note that in many applications the assumption $m = 2$ is quite reasonable, and in this case the solution is a natural cubic spline. This case has also the following nice physical interpretation. Assume that $f(x)$ is an interpolating curve created by a metal strip. Then the integral $\int_a^b (f^{(2)}(x))^2 dx$ is proportional to the potential energy of the strip. Thus, since a strip in equilibrium should have minimal potential energy, the equilibrium curve of such a strip is a natural cubic spline.

• **Spline Smoothing for Nonparametric Regression.** Consider the homoscedastic regression model

$$Y_l := f(X_l) + \varepsilon_l, \quad l = 1, 2, \ldots, n, \tag{8.7.4}$$

where ε_l are independent and identically distributed zero-mean errors.

One of the possible methods of using splines to approximate an underlying regression function is to use a spline basis, for instance, a cubic B-spline basis. In this case one chooses a fixed knot sequence $-\infty < t_1 < t_2 < \cdots < t_J < \infty$, which may differ from predictors, and then calculates elements of a corresponding cubic spline basis. It is possible to show that only $J+4$ elements of this basis are needed. With some abuse of notation, denote these elements by $B_j(x)$ and write a corresponding polynomial spline as

$$S(x) = \sum_{j=1}^{J+4} \theta_j B_j(x). \tag{8.7.5}$$

Then the coefficients θ_j can be calculated, for instance, as parameters minimizing the sum of squared errors

$$\sum_{l=1}^{n}\Big[Y_l - \sum_{j=1}^{J+4}\theta_j B_j(X_l)\Big]^2. \tag{8.7.6}$$

Denote by $\hat{\theta}_j$ the least-squares estimates and then define a polynomial spline estimator by the formula

$$\hat{f}_n(x) := \sum_{j=1}^{J+4}\hat{\theta}_j B_j(x). \tag{8.7.7}$$

Note that this estimator is similar to a series estimator, and the number of knots J defines the roughness of this estimate.

Another approach is based on the idea of finding a smooth curve that minimizes the penalized sum of squared errors

$$n^{-1}\sum_{j=1}^{n}(Y_j - f(X_j))^2 + \mu\int_a^b [f^{(m)}(x)]^2 dx \tag{8.7.8}$$

for some nonnegative μ. Then, as in the earlier interpolation approach, the solution of this minimization problem is a spline, and it is called a *smoothing spline* estimator.

In particular, for the case of $m = 2$ the minimizer of (8.7.8) is a natural cubic spline. Note that μ plays the role of a smoothing parameter. Indeed, the first sum in (8.7.8) penalizes the lack of fidelity of the spline approximation to the data. The second term is responsible for the smoothness of the spline approximation. To see this, let us consider the extreme

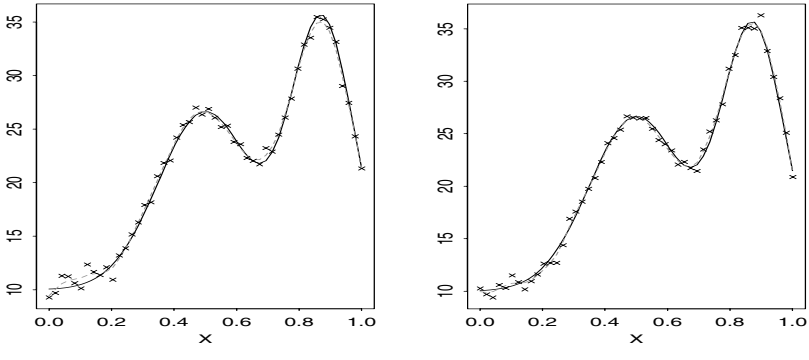

FIGURE 8.13. Smoothing spline estimates for two simulated data sets of nonparametric equidistant regression with iid $N(0, \sigma^2)$ errors, $\sigma = 0.5$. Estimates and the underlying regression function are shown by dashed and solid lines. Scatter plots are shown by crosses. The sample size is $n = 50$. [*n=50, sigma=.5*]

cases $\mu = 0$ and $\mu = \infty$. The former case leads to an interpolation, that is, $\hat{f}(X_l) = Y_l$, $l = 1, \ldots, n$. The latter leads to a linear regression because it implies $f^{(2)}(x) \equiv 0$.

Thus, μ is referred to as a smoothing parameter, and it controls the shape of the smoothing spline estimator, which can be changed from the most complicated and "wiggly" interpolation model to the simplest and smoothest linear model.

In other words, (8.7.8) represents a tradeoff between the fidelity of fit to the data, as represented by the residual sum of squares, and the smoothness of the solution, as represented by the integral of the squared mth derivative.

The smoothing spline estimator for the case $m = 2$ is supported by the S–PLUS function **smooth.spline**, which chooses the smoothing parameter μ by a cross-validation procedure (Section 8.10 explains the idea of this procedure). Estimates for two particular simulated data sets are shown by dashed lines in Figure 8.13. Observations are simulated according to (8.7.4) with normal $N(0, (0.5)^2)$ errors. As we see, the smoothing spline estimator gives a good approximation.

8.8 Neural Networks

The exciting idea of neural networks stems from attempts to model the human brain. Note that all the previously discussed nonparametric methods have been motivated by either mathematical results or numerical methods or fast computers. However, the human brain has many abilities, such as understanding fuzzy notions or making inferences based on past experience and relating them to situations that have never been encountered before. Such abilities would also be desirable in nonparametric estimation. This explains in part the motivation to understand and model the human brain.

The basic computational unit of the brain is the neuron. A human brain has approximately 10^{11} neurons, which act in parallel and which are highly interconnected. Figure 8.14 shows a simplified model of the neuron. The neuron receives a weighted sum of inputs and then, using its *activation* function $\mathrm{s}(u)$, calculates an output $\mathrm{Y} = \mathrm{s}(\sum_{j=1}^{n} \mathrm{w}_j \mathrm{X}_j)$. For instance, the McCulloch–Pitts model assumes that $\mathrm{s}(u)$ is a step (threshold) function such that $\mathrm{s}(u) = 0$ if the net input u is smaller than a unit threshold level ($|u| < 1$) and $\mathrm{s}(u) = 1$ otherwise. This implies that this neuron fires ($\mathrm{Y} = 1$) or not ($\mathrm{Y} = 0$) depending on the value of the weighted sum of inputs. Although this particular model is simple, it has been demonstrated that computationally it is equivalent to a digital computer. In other words, a set of interconnected McCulloch–Pitts neurons can perform as a conventional digital computer.

Moreover, it is easy to see that this neuron can solve classical statistical problems like regression analysis or hypothesis testing. As an example, con-

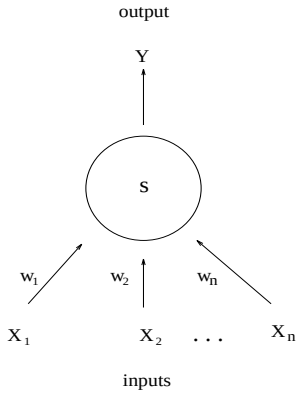

FIGURE 8.14. Model of a neuron. The inputs X_j are multiplied by weights w_j. $\mathrm{Y} = \mathrm{s}(\sum_{j=1}^{n} \mathrm{w}_j \mathrm{X}_j)$ is an output.

sider the problem of nonparametric equidistant regression where responses $X_l = f(l/n) + \sigma\varepsilon_l$, $l = 1, 2, \ldots, n$. Then all the earlier methods, including least-squares linear regression, kernel smoothing, and spline estimation, can be written as a weighted sum of the responses, $\hat{f}_n(x) = \sum_{l=1}^{n} w_l(x) X_l$. Thus $\hat{f}_n(x)$ can be computed by the neuron shown in Figure 8.14 with an identity activation function $\mathrm{s}(u) = u$. The only issue is that appropriate weights should be chosen.

Another appealing example is hypothesis testing. Consider the classical one-tailed test $H_0 : \theta \le 0$ versus $H_a : \theta > 0$ with the level of significance α. The data are iid normal random variables X_l with mean θ and variance σ^2. The well-known solution of the problem is to reject the null hypothesis H_0 if $\sum_{l=1}^{n} n^{-1} X_l > (\sigma/n^{1/2}) z_\alpha$, where z_α satisfies $P(\xi_0 > z_\alpha) = \alpha$ with ξ_0 being a standard normal random variable. Thus, in this case the McCulloch–Pitts neuron solves this hypothesis testing problem with identical weights $\mathrm{w}_j = 1/(\sigma z_\alpha n^{1/2})$. Note that the neuron fires ($\mathrm{Y} = 1$) if the null hypothesis is rejected.

Thus as soon as the solution to a problem can be written as a linear combination of input variables that is transformed by a function, this solution can be obtained by a neuron.

More complicated problems can be solved by a neural network, which is a set of neurons that are highly interconnected. A generic example is given in Figure 8.15. This network is constructed with layers of units, and thus it is called a multilayer network. A layer of units is composed of units that perform similar tasks. A feed-forward network is one where units in one layer pass data only onto units in the next upper layer. The zero (input) layer consists of the input units, which are simply inputs X_j. An ith unit of the input layer is connected to a jth unit of the first hidden layer according to a weight w^0_{ji}. This weight is interpreted as the strength of the connection from this ith unit on the zero (input) layer to the jth unit on the first

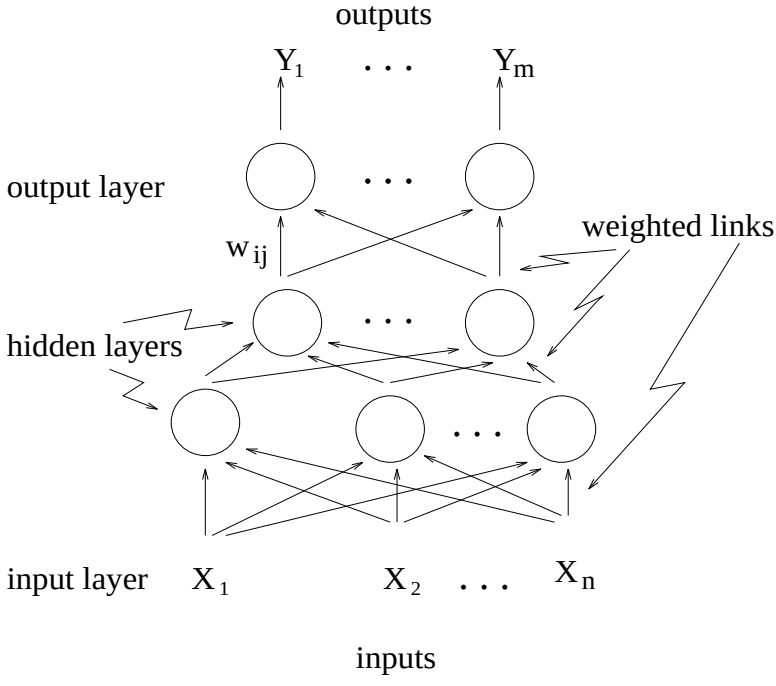

FIGURE 8.15. A generic multilayer feed-forward neural network.

(hidden) layer. Then the output of this jth unit is $\mathrm{Y}_j = \mathrm{s}_1(\sum_i \mathrm{w}^0_{ji}\mathrm{X}_i)$. Units on other layers perform similarly. When counting layers it is common practice not to count the input layer, because it does not perform any computation, but simply passes data onto the first layer. So the network shown is termed a three-layer one.

The main problem in using neural networks is to compute optimal weights. Typically this is done by calculating the gradient of a risk function, and this requires smooth activation functions. Two common choices are the logistic function $\mathrm{s}(u) := 1/(1 + e^{-u})$ or the hyperbolic tangent $\mathrm{s}(u) := (e^u - e^{-u})/(e^u + e^{-u})$. These activation functions are smooth approximations of the step function, and they are referred to as *sigmoid* functions due to their s-shape.

One of the reasons why multilayer neural networks, and two-layer networks in particular, are interesting is that it has been shown that a two-layer feed-forward neural network with a significant number of hidden units can approximate a continuous function to any degree of accuracy. Thus, such a neural network becomes a powerful modeling tool.

The main practical question in operating a neural network is how to adapt it to a particular problem. The human brain, which is the prototype, can *learn* or be *trained* with or without the help of a supervisor. *Supervised* learning (learning with a teacher) occurs when there is a known target

value associated with each input in the training set. The brain compares its outcome with the target shown by a teacher, and the difference is used for adaptation. *Unsupervised* learning is needed when the training data lack a target output; for instance, the brain can learn that a gray sky is more likely to produce rain than a blue one.

Similar procedures of supervised and unsupervised learning can be developed for a neural network. For a given structure of a network, mathematically it means an adaptive choice of weights $\{w_{ij}^k\}$. A widely used method is a back-propagation algorithm that minimizes a discrepancy between outputs and some target values by calculating the gradient of the risk according to the Lagrange multipliers technique. Typically there are many local minima, so heuristics and experience in using this method are required. There is no surprise that a commonly used criterion is that of least squares, which, to ensure a smooth curve, may be penalized by adding a function based on the second derivative of the modeled curve. Recall that a similar approach was used for splines.

8.9 Asymptotic Analysis of Kernel Estimates

In this section we consider a problem of estimation of the density f of a random variable X based on its n iid realizations $X_1, \dots, X_n$ when n is large. We are interested in optimal kernel estimation, namely, in understanding what kind of kernels and bandwidths delivers an optimal estimation for large samples (as $n \to \infty$).

In Section 8.3 we introduced the kernel density estimator,

$$\hat{f}(x) := (nh)^{-1} \sum_{l=1}^{n} K((x - X_l)/h). \tag{8.9.1}$$

Recall that K is called the kernel and h the bandwidth.

We would like to study the pointwise mean squared error of the kernel estimate,

$$\text{MSE} := E\{(\hat{f}(x) - f(x))^2\}, \tag{8.9.2}$$

and the global mean squared integrated error of the kernel estimate,

$$\text{MISE} := E\Big\{ \int (\hat{f}(x) - f(x))^2 dx \Big\}. \tag{8.9.3}$$

In this section the integral is taken over the real line.

Consider the case where an underlying density f belongs to a Lipschitz space $Lip_{r,\alpha}(L)$, that is, f is bounded, r-times differentiable, and its rth derivative $f^{(r)}$ is Lipschitz of order α, $0 < \alpha \le 1$. In short,

$$|f(x)| \le C, \quad |f^{(r)}(x_1) - f^{(r)}(x_2)| \le L|x_1 - x_2|^\alpha. \tag{8.9.4}$$

Then, as we know from Chapter 7, f may be such that regardless of the procedure of estimation, both MSE and MISE decay at a rate not faster than $n^{-2\beta/(2\beta+1)}$, where $\beta := r + \alpha$.

Our aim is to explore kernel estimates that attain this rate.

To do this, we begin by introducing the class of kernels that will be considered. Define a class $S_{r,\alpha}$ of kernels $K(t)$ that are symmetric about zero (i.e., even), bounded, and such that

$$\int K(t)dt = 1, \tag{8.9.5}$$

$$\int t^s K(t)dt = 0, \quad s = 1, \dots, \max(1, r), \tag{8.9.6}$$

$$\int |t^{r+\alpha} K(t)|dt < \infty. \tag{8.9.7}$$

In words, this is the class of kernels that are even, bounded, integrated to unity, all their moments up to the rth are zero, and the product $t^\beta K(t)$ is absolutely integrable. Note that in (8.9.6) only even s impose restrictions (since the kernel is even), so for $r \le 1$ this class includes all the previously discussed density kernels. On the other hand, if $r \ge 2$, then nonnegative kernels do not satisfy (8.9.6) and thus do not belong to $S_{r,\alpha}$. As we shall see, a kernel taking negative values is required for asymptotically optimal estimation of densities with smooth second derivative.

To make the following mathematical manipulation easier to understand, let us recall some elementary relations related to the variance of a random variable Z,

$$\mathrm{Var}(Z) = E\{(Z - E\{Z\})^2\} = E\{Z^2\} - (E\{Z\})^2 \le E\{Z^2\}, \tag{8.9.8}$$

and if $Z_1, \dots Z_n$ are iid realizations of the Z, then

$$\mathrm{Var}(n^{-1}\sum_{l=1}^{n} Z_l) = n^{-1}\mathrm{Var}(Z). \tag{8.9.9}$$

Now we are in a position to study the MSE of the kernel estimate (8.9.1). It is convenient to use the notation $K_h(x) := h^{-1}K(x/h)$ because it allows us to rewrite this estimate as a sample mean estimate,

$$\hat{f}(x; h) = n^{-1}\sum_{l=1}^{n} K_h(x - X_l). \tag{8.9.10}$$

Using (8.9.8), the MSE may be written as a sum of the variance and the squared bias terms,

$$\mathrm{MSE} = \mathrm{Var}(\hat{f}(x)) + [E\{\hat{f}(x)\} - f(x)]^2 =: \mathrm{VAR} + \mathrm{SBIAS}. \tag{8.9.11}$$

Using (8.9.8)–(8.9.9) and the fact that f is the density of X, we may write for the variance term,

$$\mathrm{VAR} = n^{-1}\mathrm{Var}(K_h(x - X)) = n^{-1}(E\{[K_h(x - X)]^2\} - [E\{K_h(x - X)\}]^2)$$

$$= n^{-1}\Big(\int [K_h(x-u)]^2 f(u)du - \Big[\int K_h(x-u)f(u)du \Big]^2 \Big). \quad (8.9.12)$$

The squared bias term we rewrite using (8.9.5):

$$\text{SBIAS} = \Big[\int K_h(x-u)(f(u)-f(x)du \Big]^2 . \quad (8.9.13)$$

Note that both these terms have a similar structure, so to continue the study we need the following simple relation for $m = 1, 2$:

$$\int [K_h(x-u)]^m f(u)du = \int h^{-m}[K((x-u)/h)]^m f(u)du$$
$$= h^{-m+1} \int [K(t)]^m f(x+ht)dt. \quad (8.9.14)$$

Here the first equality is due to the definition of K_h, and the second is due to the change of variable $t = (u-x)/h$ and the symmetry of $K(t)$.

Using (8.9.14) in (8.9.12), we get for the variance term,

$$\text{VAR} = (nh)^{-1} \int [K(t)]^2 f(x+ht)dt - n^{-1}\Big(\int K(t)f(x+ht)dt \Big)^2. \quad (8.9.15)$$

Now is the moment to use the assumption about smoothness of the underlying density f. We replace $f(x+ht)$ by $f(x) + (f(x+ht) - f(x))$ and then note that the following relation holds for densities from $Lip_{r,\alpha}(L)$:

$$\int |K(t)|^m |f(x+ht) - f(x)|dt = o_h(1), \quad m = 1, 2. \quad (8.9.16)$$

Here $o_h(1)$ denotes a generic sequence in h such that $o_h(1) \to 0$ as $h \to \infty$.

Let us check (8.9.16) for the case of the least smooth density with $r = 0$. Using (8.9.4), (8.9.7), and boundness of the kernel we may write for $m = 1, 2$,

$$\int |K(t)|^m |f(x+ht) - f(x)|dt \le L|h|^\alpha \int |K(t)|^m |t|^\alpha dt = o_h(1). \quad (8.9.17)$$

The case of differentiable densities with $r \ge 1$ is left as Exercise 8.9.8.

Using (8.9.16) in (8.9.15) after the replacement of $f(x+th)$ we get

$$\text{VAR} = (nh)^{-1} f(x) \int [K(t)]^2 dt + o_h(1)(nh)^{-1}. \quad (8.9.18)$$

Now we consider the SBIAS at (8.9.13). If $r = 0$, then using (8.9.14) and then (8.9.4) gives us

$$\text{SBIAS} = \Big[\int K(t)(f(x+ht) - f(x))dt \Big]^2 \le h^{2\alpha} \Big[L \int |t^\alpha K(t)|dt \Big]^2. \quad (8.9.19)$$

The case $r > 0$ is considered similarly with the help of the *Taylor expansion*

$$f(x+ht) = f(x) + \sum_{j=1}^{r-1} \frac{(ht)^j}{j!} f^{(j)}(x) + \frac{(ht)^r}{r!} f^{(r)}(y_{ht}), \tag{8.9.20}$$

where y_{ht} is a point between $x + ht$ and x. Write

$$\text{SBIAS} = \Big[\int K(t) \Big(\sum_{j=1}^{r-1} \frac{(ht)^j}{j!} f^{(j)}(x) + \frac{(ht)^r}{r!} f^{(r)}(y_{ht}) \Big) dt \Big]^2 \tag{8.9.21}$$

$$= \Big[\int K(t) \frac{(ht)^r}{r!} \big(f^{(r)}(y_{ht}) - f^{(r)}(x) + f^{(r)}(x) \big) dt \Big]^2 \tag{8.9.22}$$

$$\le h^{2\beta} \Big[\frac{L}{r!} \int |t^\beta K(t)| dt \Big]^2. \tag{8.9.23}$$

Here the equality (9.8.22) is obtained using (8.9.6), and the inequality (9.8.23) using (8.9.4) and (8.9.6).

The comparison of (8.9.23) to (8.9.19) shows that (8.9.23) holds for all r. Thus, plugging the obtained upper bounds for the variance and the squared biased terms into (8.9.11) gives us

$$\text{MSE} \le (nh)^{-1} f(x) \int [K(t)]^2 dt + h^{2\beta} \Big[\frac{L}{r!} \int |t^\beta K(t)| dt \Big]^2 + o_h(1)(nh)^{-1}. \tag{8.9.24}$$

Using a bandwidth h_n^* that is proportional to $n^{-1/(2\beta+1)}$, we obtain that $\text{MSE} \le C n^{-2\beta/(2\beta+1)}$ (recall that C is a generic finite constant). Thus, a kernel estimator is pointwise rate optimal.

Let us make a remark about the necessity of condition (8.9.6) which excludes nonnegative density kernels if $r \ge 2$. Let f belong to a Lipschitz class with $r \ge 2$. Then, according to (8.9.21)–(8.9.22), the SBIAS always has the terms $[h^j \int t^j K(t) dt f^{(j)}(x)/j!]^2$, $j = 1, \dots, r$. These terms should be of the optimal order $h^{2(r+\alpha)}$, $0 < \alpha \le 1$, for all small h and all x. This can be the case only if all these terms are 0. Thus, (8.9.6) is a necessary condition for a kernel estimate to be rate optimal.

Let us formulate this result as a mathematical proposition.

Theorem 8.9.1. *A kernel estimate with the kernel from the class $S_{r,\alpha}$ and the bandwidth proportional to $n^{-1/(2\beta+1)}$ is rate optimal over a Lipschitz class $Lip_{r,\alpha}(L)$, $0 < \alpha \le 1$, that is,*

$$\sup_{f \in Lip_{r,\alpha}(L)} \text{MSE} \le C n^{-2\beta/(2\beta+1)}, \tag{8.9.25}$$

where $\beta = r + \alpha$ and C is a finite constant. Also, the condition (8.9.6) is necessary for the rate optimal estimation.

So far we have considered the MSE. According to (8.9.3), MISE is the integrated MSE over the real line. Thus, we can directly integrate the right

hand side of (8.9.25) only if a global integrated risk is considered over a finite interval. To consider the MISE over the real line, one may assume that $L = L(x)$ in the definition of the Lipschitz class and then add some properties of $L(x)$ that will allow the integration. We leave this as Exercise 8.9.12.

The relation (8.9.24) gives us an upper bound in terms of the class of densities, while in many cases it would be better to have an upper bound expressed in terms of an underlying density. For instance, this may be beneficial for adaptive estimation, as we shall see in Section 8.10.

Let us explore this problem for a case of a kernel being bounded even density with finite fourth moment; such a kernel is widely used by data analysts. We shall restrict this case to densities $f \in Lip_{2,\alpha}(L)$ where α may be as small as desired and L may be as large as desired. In other words, we assume that an underlying density is twice differentiable and its second derivative is Lipschitz of any order, which is unknown. The reason why we consider such densities is clear from Theorem 8.9.1, which states that twice-differentiable densities are the boundary case for nonnegative kernels to be optimal.

Our aim is to find an expression for the MSE and optimal bandwidth via an underlying density f (but not via β and L as in (8.9.24)).

The relation (8.9.18) for the variance term is written via f. Thus we need to get an expression for the SBIAS. Using the Taylor expansion (8.9.20) with $r = 2$, (8.9.21), and the assumption $\int tK(t)dt = 0$, we get

$$\text{SBIAS} = \frac{h^4}{4}\Big[\int t^2 K(t) f^{(2)}(y_{ht})dt\Big]^2. \tag{8.9.26}$$

Combining this result with (8.9.18) we obtain that

$$\text{MSE} = \frac{f(x)}{nh}\int [K(t)]^2 dt(1+o_h(1)) + \frac{h^4}{4}\Big[\int t^2 K(t) f^{(2)}(y_{ht})dt\Big]^2. \tag{8.9.27}$$

Then as in (8.9.17) we get that

$$\text{MSE} = \Big(\frac{f(x)}{nh}\int [K(t)]^2 dt + \frac{h^4}{4}\Big[f^{(2)}(x)\int t^2 K(t)dt\Big]^2\Big)(1+o_h(1)). \tag{8.9.28}$$

This is the kind of expression for the MSE that we wanted to get because it is based solely on the kernel, the bandwidth, and the underlying density.

The optimal bandwidth $h^*(x)$, which minimizes the right-hand side of (8.9.28), is

$$h^*(x) := \frac{\big[f(x)\int (K(t))^2 dt\big]^{1/5}}{\big[f^{(2)}(x)\int t^2 K(t)dt\big]^{2/5}} n^{-1/5}. \tag{8.9.29}$$

Here we ignored the factor $1+o_h(1)$ because it is close to 1 for large n, and then we used the elementary fact that $cn^{-1}y^{-1} + y^4/4$ takes on its minimal value at $y^* = (cn^{-1})^{1/5}$.

The bandwidth (8.9.29) performs optimal variance–bias tradeoff. Note that the larger the second derivative, the smaller the optimal bandwidth, and vice versa. This is a rather natural property because the value of the second derivative tells us how far an underlying curve is from a straight line where all observations should be used with the same weights.

Substituting this optimal bandwidth into (8.9.28), we get

$$\text{MSE} = (5/4)C_K[f^{(2)}(x)]^{2/5}[f(x)]^{4/5}n^{-4/5}(1+o_n(1)), \tag{8.9.30}$$

where

$$C_K := \Big[\int t^2K(t)dt\Big]^{2/5}\Big[\int (K(t))^2dt\Big]^{4/5}. \tag{8.9.31}$$

Now we are in a position to explore an optimal kernel function $K(x)$. Under the previous assumptions, the optimal kernel must minimize C_K. There is no unique minimizer because C_K is the scale invariant. Thus, if we add an additional restriction on the second moment of the kernel, $\int t^2K(t)dt = \sigma^2$, then the minimizer is $K_e(t,\sigma) := (3/(4\sqrt{5\sigma^2}))[1-t^2/(5\sigma^2)]I_{\{t^2\le 5\sigma^2\}}$. This kernel is called the *Epanechnikov kernel*. It is customary to choose a kernel supported on $[-1,1]$, which implies $\sigma^2 = 0.2$.

It is useful to note that C_K is not too sensitive to changing the kernel. For instance, if instead of the Epanechnikov kernel a standard normal density is used (this choice allows us to analyze any derivative, and it makes the estimate extremely smooth), then C_K increases 1.05 times. If the uniform on $[-1,1]$ density is used (this choice leads to an estimate that looks like a histogram), then the increase is 1.08 times. Therefore, many statisticians believe that the choice of a kernel is not too important. Unfortunately, this is not the case asymptotically, because as we know from our results on series estimators, the rate of convergence may be much faster than $n^{-4/5}$, while positive kernels can give us only the rate $n^{-4/5}$. To improve the convergence one must consider kernels that take on negative values.

Let us formulate these results as a mathematical proposition.

Theorem 8.9.2 *Let a kernel K be a bounded even density with a finite fourth moment. Let an underlying density be from $Lip_{2,\alpha}(L)$, $\alpha > 0$. Then the mean squared error of the kernel estimate (8.9.1) with the bandwidth (8.9.29) satisfies (8.9.30). Also, among all such kernels the asymptotically optimal one is the Epanechnikov kernel.*

Now let us briefly consider a global approach for the case of nonnegative kernels. Under mild assumptions it is possible to show that the factor $1 + o_h(1)$ in (8.9.28) tends to 1 uniformly over all x. Then the integration of both sides of (8.9.28) yields

$$\text{MISE} = \Big(\frac{1}{nh}\int (K(t))^2dt + \frac{h^4}{4}\int [f^{(2)}(x)]^2dx\Big[\int t^2K(t)dt\Big]^2\Big)(1+o_h(1)). \tag{8.9.32}$$

Then the optimal global bandwidth is

$$h^* = \left[\frac{\int (K(t))^2 dt}{\int (f^{(2)}(x))^2 dx \big(\int t^2 K(t) dt \big)^2} \right]^{1/5} n^{-1/5}. \tag{8.9.33}$$

This is the formula that is used by the plug-in adaptive technique discussed in the next section because the only characteristic of f used by this bandwidth is the quadratic functional of its second derivative. As we know from Section 7.6, this functional may be easily estimated.

Finally, we plug this bandwidth into (8.9.32) and get the desired formula for the optimal MISE,

$$\begin{aligned} \text{MISE} = \frac{5}{4} \Big[\int t^2 K(t) dt \Big]^{2/5} \Big[\int (K(t))^2 dt \Big]^{4/5} \\ \times \Big[\int (f^{(2)}(x))^2 dx \Big]^{1/5} n^{-4/5} (1 + o_n(1)). \end{aligned} \tag{8.9.34}$$

8.10 Data-Driven Choice of Smoothing Parameters

In this section we consider several methods of data-driven choice of smoothing parameters that are used for small sample sizes.

• **Reference Method.** Were an underlying estimated function (probability density, regression function, etc.) known, then for all the estimators discussed a correct optimal smoothing parameter could be calculated. The idea of the reference method is to pretend that an underlying function is a particular one, choose it as a *reference*, and then use the corresponding optimal smoothing parameter.

As an exercise, let us find an optimal global bandwidth (8.9.33) using the reference method. Let the reference density $\phi_\sigma(x)$ be normal $\phi_\sigma(x) = \sigma^{-1}\phi_1(x/\sigma)$, where $\phi_1(x)$ is the standard normal density. We may use the chain rule to find the first derivative $\phi_\sigma^{(1)}(x) = \sigma^{-2}\phi_1^{(1)}(x/\sigma)$ and then the second derivative $\phi_\sigma^{(2)}(x) = \sigma^{-3}\phi_1^{(2)}(x/\sigma)$. Then, using the change of variable $u = x/\sigma$ we calculate

$$\begin{aligned} \int [\phi_\sigma^{(2)}(x)]^2 dx &= \sigma^{-6} \int [\phi_1^{(2)}(x/\sigma)]^2 dx = \sigma^{-5} \int [\phi_1^{(2)}(u)]^2 du \\ &= [3/(64\pi)^{1/2}] \sigma^{-5} \approx 0.2\sigma^{-5}. \end{aligned} \tag{8.10.1}$$

Here the integrals are over the real line.

Then we plug this result in the (8.9.33) and get the reference method bandwidth for a given σ,

$$h_n := \left[\frac{(64\pi)^{1/2} \int [K(t)]^2 dt}{3 \big[\int t^2 K(t) dt \big]^2} \right]^{1/5} \sigma n^{-1/5}. \tag{8.10.2}$$

Finally, any reasonable estimate of σ may be plugged in.

An advantage of the reference method is its simplicity, and it may be very good if luckily an underlying function resembles the reference. A disadvantage is that the estimate may be bad if that luck fails.

• **Cross-Validation.** This is the adaptation technique when part of a sample is used to obtain information about another part. We have discussed this technique for the example of regression in Section 7.4, so let us here explain it for the probability density model.

Suppose that we would like to find an optimal smoothing parameter h that minimizes the MISE of an estimate $\hat{f}(x;h)$ based on n iid observations from this density. Write

$$\text{MISE}(\hat{f}(x;h),f) = E\Big\{\int [\hat{f}(x;h)]^2dx - 2\int \hat{f}(x;h)f(x)dx\Big\} + \int [f(x)]^2dx.$$

The last term does not depend on the smoothing parameter h, so we should minimize the expectation. Note that we cannot do this directly because the underlying density f is unknown. Thus, let us assume that an extra X_{n+1}th observation is given. Using the independence of the observations we may write

$$E\Big\{\int \hat{f}(x;h)f(x)dx\Big\} = E\big\{\hat{f}(X_{n+1};h)\big\}, \tag{8.10.3}$$

and then estimate this expectation via a sample mean.

We do not have extra observations, but we can consider a sample with one deleted observation. This leads to the so-called leave-one-out estimator $n^{-1}\sum_{l=1}^n \hat{f}_{-l}(X_l;h)$ of $E\{\int \hat{f}(x;h)f(x)dx\}$, where $\hat{f}_{-l}(X_l;h)$ is the estimator based on the sample with deleted lth observation.

Thus the least-squares leave-one-out cross-validation implies the choice of h^* that minimizes

$$\text{LSCV}(h) := \int [\hat{f}(x;h)]^2dx - 2n^{-1}\sum_{l=1}^n \hat{f}_{-l}(X_l;h). \tag{8.10.4}$$

Clearly, a portion of observations may be deleted as well and then the corresponding sample mean used.

• **Plug-in Method.** This is another popular method, which is typically motivated by asymptotic results that claim a formula $h^*(f)$ for a smoothing parameter when the underlying function f is supposed to be known. Then an estimate of f, or if this is the case, estimates of some functionals of f, are plugged in.

As an example, consider the formula (8.9.33) for the asymptotically optimal global bandwidth. This bandwidth depends on the unknown quadratic functional $F_2(f) = \int [f^{(2)}(x)]^2dx$, which is the integral of the squared second derivative. Thus, we may estimate this functional (see Section 7.6) and then plug an estimate in.

8.11 Practical Seminar

The aim of this seminar is to gain experience in using the Nadaraya–Watson kernel regression estimator for the analysis of real data sets. This estimator was discussed in Section 8.3, and Figure 8.7 illustrated its performance for a simulated data set.

Recall that the Nadaraya–Watson estimator is supported by the S–PLUS function **ksmooth** with two arguments: *kernel* and *bandwidth*. Four kernels may be chosen: "box", "triangle", "parzen" (which is a box kernel convolved with a triangle kernel), and "normal". Recall that all these kernels are nonnegative and that the estimate inherits the smoothness of the kernel. However, among these two arguments, bandwidth is the critical one.

Now, after this short review, let us consider a particular data set and apply our method of a "running" argument to assess the performance of this kernel estimator.

The particular data set is **saving.x**, which is a matrix with 5 columns (variables) describing averaged statistics over 1960–1970 (to remove business cycles or other short-term fluctuations) for 50 countries. Just for the record, the variables (columns) are (1) Percentage of population younger than 15 years old; (2) Percentage of population older than 75; (3) Income, which is per capita disposable income in U.S. dollars; (4) Growth, which is the percent rate of change in per capita disposable income; (5) Savings rate, which is aggregated personal savings divided by disposable income.

Here we consider the regression of Percentage of population older than 75 (the response Y) on the Income (the predictor X). In short, we would like to know how the nation's prosperity (measured in units of Income) affects the percentage of the nation's elderly population.

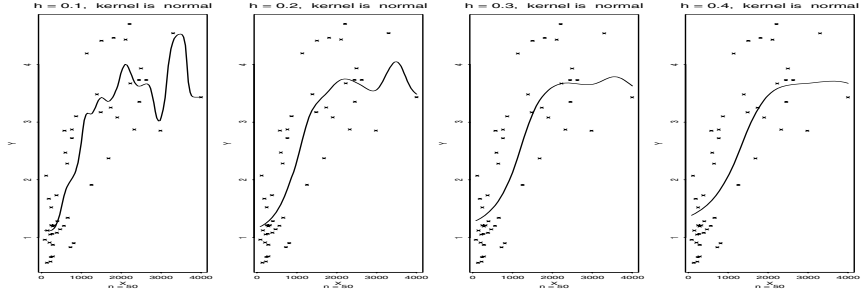

FIGURE 8.16. Nadaraya–Watson kernel estimates with different bandwidths for a real data set with the predictors X and responses Y. Here the particular data set is **saving.x** with the predictor being Income and the response being Percentage (of population older than 75). The sample size is 50, and it is shown in the subtitles. Titles indicate the bandwidth h and the kernel. {A data set is controlled by the arguments X and Y, which should be columns.} [*X=saving.x[,3], Y=saving.x[,2], kernel= "normal", set.h = c(.1,.2,.3,.4)*]

Figure 8.16 shows us how the Nadaraya–Watson estimator with normal kernel highlights the relationship between the Income and the Percentage. As we see, all four estimates give us a correct overall description of the data at hand. The estimate with the smallest bandwidth $h = 0.1$ tries to fit all the tiny details and to inform us about them. Note that the 3 countries with the largest Incomes are Canada ($X = 2983$), Sweden ($X = 3299$), and United States ($X = 4002$). Thus, while the right tail of the estimate looks strange, the reality is that in the 1960s both Canada and the United States enjoyed unprecedented prosperity together with the so-called period of baby boomers (the percentages of youngsters with age at most 15 are 32 and 29, respectively; see more in Section 4.12). Sweden also prospered, but with a larger percentage of senior citizens (in part because in the 1960s their youngsters constituted only 23 percent of the total population). These three countries are almost solely responsible for approximately a quarter of the range of Income and for the relationship between Income and Percentage for superrich nations. And this is the message of the right tail of the kernel estimate with the smallest bandwidth 0.1.

The larger the bandwidth, the smoother the estimate and the smaller the number of tiny details we can see. Note that when you analyze a real data set like the one considered, it is not a clear-cut issue to say that the estimate oversmooths data and that the estimate undersmooths it. After all, all depends on what you would like to see. All these estimates are almost identical for low and moderate Incomes (only the first one is a bit "undersmoothed"), so the only important difference between them is in the message about Percentages for the largest Incomes. The estimate with $h = 0.1$ may look undersmoothed for the largest Incomes, but at the same time, it sends us the strongest (and absolutely correct) message that for rich countries the relationship between a nation's prosperity and Percentage of elderly population is very complicated. Indeed, the percentage of the elderly population in the richest countries was very sensitive to such parameters as participation in wars and current birth rates, among others. On the

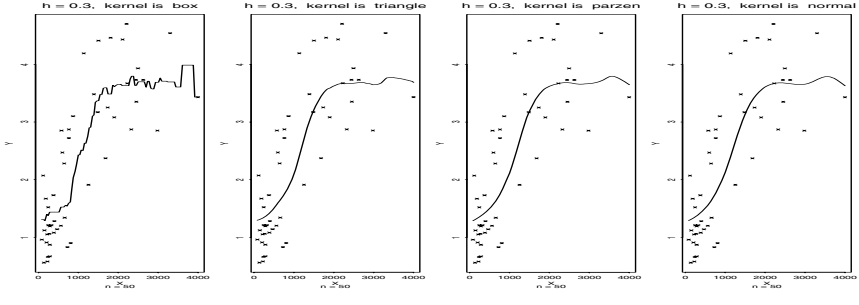

FIGURE 8.17. Nadaraya–Watson kernel estimates with 4 different kernels supported by S–PLUS. The data set is the same as in Figure 8.16. [*X=saving.x[,3], Y=saving.x[,2], h=.3, set.kernel= c("box", "triangle", "parzen", "normal")*]

other hand, the smoothest estimate with $h = 0.4$ looks right in terms of a hypothetical relationship between a nation's prosperity, defined as income per capita, and the percentage of senior citizens. After all, we all want to believe that this is the case, and probably this should be the case.

Figure 8.17 shows us how a choice of the kernel affects the Nadaraya–Watson estimate. We see that the estimate inherits the smoothness of the underlying kernel. The "box" kernel does not look right here, and the "triangle" is just a bit rough. The two others are almost identical, but the normal kernel implies the smoother estimate.

8.12 Exercises

8.1.1 Can the histogram (8.1.1) be referred to as a sample mean estimate?
8.1.2 The classical definition of the histogram is based on using bins of identical widths. Some statisticians argue in favor of using variable bins. Discuss possible pros and cons of a histogram with variable bins.
8.1.3 Let f be a continuous density supported on [0,1]. Suggest a histogram estimate $\hat{f}_n$ such that $E\{\int_0^1 (\hat{f}_n(x) - f(x))^2 dx\} \to 0$ as $n \to \infty$.
8.1.4 Consider sample sizes 50 and 100 and corner densities the Strata and the Monotone. Using Figure 8.1, for each pair of a sample size and a density, find the optimal width of bins. Then explain the result.
8.1.5 Give an example of a data set where two reasonable choices of the origin may crucially affect the histograms.
8.2.1 Is the naive estimate a sample mean estimate? Is it a pointwise unbiased estimate?
8.2.2 Let an underlying density $f(x)$ be positive and continuous at the point x_0. Find the variance of $\hat{f}_n(x_0)$ for the case of small h. Hint: Recall the binomial random variable.
8.2.3 Repeat Exercise 8.1.4 with the naive estimate and Figure 8.2.
8.2.4 What are the main differences between the histogram and naive density estimate? When would you prefer to use each of them?
8.2.5 Is a naive estimate differentiable (smooth)? Suggest a modification that would lead to a twice differentiable (smooth) estimate.
8.3.1 Is the kernel estimate (8.3.2) an unbiased estimate of an underlying density at a point x_0? Does the location of this point (interior or boundary) or the value of the bandwidth affect the answer?
8.3.2 Let (8.3.1) hold and let the kernel be nonnegative. Is the kernel estimate (8.3.2) integrated to one and nonnegative?
8.3.3 Use Figure 8.4 with sample sizes 50, 100, and 200, and find corresponding optimal bandwidths. Then repeat the same experiment for the Uniform density. Draw a conclusion on how the sample size and the underlying density affect the choice of optimal bandwidth.

8.3.4 Using Figure 8.5, answer the following question. How does the standard deviation σ of the noise term affect the optimal bandwidth?
8.3.5 Increase the sample size n in Figure 8.5 and find a qualitative relationship between the sample size and optimal bandwidth.
8.3.6 Let $\sigma = 0$ (no error term) and thus $Y_l = f(l/n)$. Explain the performance of the kernel estimate (8.3.4) for this particular case. Also, what bandwidth would you recommend to use for this case?
8.3.7 Explain lines (8.3.6)–(8.3.8).
8.3.8 As in (8.3.6)–(8.3.8), consider the case $x = l/n$.
8.3.9 Use (8.3.6)–(8.3.8) and answer the following question. Let the bandwidth h be increased. How does this affect the squared bias and the variance terms in the MSE?
8.3.10 Give an example of a regression function that is (a) Lipschitz of order 0.5 but not differentiable; (b) not Lipschitz.
8.3.11 Explain all the steps in (8.3.9). Hint: Think about how to bound from above the series $\sum_{j=1}^{k} j^{\alpha}$.
8.3.12 Explain how the inequality (8.3.10) was obtained.
8.3.13 For large n find h that minimizes (8.3.10). Hint: If a function $g(h)$ is differentiable, then its derivative is equal to zero at the local extrema (minima and maxima). If this function is twice differentiable, then the second derivative is negative at local maxima and positive at local minima.
8.3.14 Can the inequality in (8.3.12) be replaced by equality if the factor $1 + o_n(1)$ is replaced by $O_n(1)$? The $O_n(1)$ is a generic bounded sequence in n. Also, can we do this if additionally $\sup_{f \in Lip_{\alpha}(L)}$ MSE is used in place of MSE on the left-hand side of (8.3.12)?
8.3.15 Let $Z := X + hY$, where X and Y are independent random variables with densities $p^X(u)$ and $p^Y(u)$. Show that the density $p^Z(u)$ of the sum may be written as the convolution integral $p^Z(u) = \int_0^{\infty} p^X(x) h^{-1} p^Y((u - x)/h) dx$. Then, use this result to explain (8.3.13).
8.3.16 Explain all steps in obtaining (8.3.14).
8.3.17 Consider the case of a fixed-design regression with design density $g(x)$. Then, as in (8.3.14)–(8.3.15), suggest a kernel estimator.
8.3.18 Use definition (4.2.2) of the design density for a fixed design regression to explain why both (8.3.17) and (8.3.18) are good data-driven estimates of the right-hand side of (8.3.14). Then, discuss the same question for a random-design regression. Hint: Recall the fact, discussed in Section 4.2, that under mild assumptions the difference $X_{(l+s)} - X_{(l-s)}$ between ordered predictors is inversely proportional to $nh(X_{(l)})/(2s)$.
8.3.19 Write down a kernel estimate for the spectral density that smooths the periodogram at the Fourier frequencies.
8.4.1 Explain how the kernel function and the bandwidth affect a local linear fit.
8.4.2 Find the mean and variance of a local linear estimate at an interior point.

8.4.3 Explain why a local linear fit performs well at boundary points, while a kernel estimate does not.
8.4.4 Using Figure 8.9, explain how the sample size and variance of errors affect an optimal bandwidth.
8.4.5 Suppose that a variable bandwidth $h(x)$ is available. Using Figure 8.9, suggest a reasonable variable bandwidth.
8.4.6 What order p of a local polynomial regression would you suggest for the regression function in Figure 8.9?
8.5.1 Explain the idea of the kth nearest neighbor method.
8.5.2 Describe the shape of tails of a nearest neighbor estimate.
8.5.3 Let n be increased. How should k be changed?
8.5.4 Consider the case of 2 observations and draw a sketch of the nearest neighbor estimate with $k = 1$.
8.5.5 Consider the case of 4 observations of a pair of random variables. Draw a sketch of the nearest neighbor estimate of an underlying bivariate density with $k = 2$.
8.5.6 What can be said about the tails of the nearest neighbor estimate for a bivariate density?
8.5.7 Explain a possible difference in performances of kth neighbor kernel estimators with Gaussian and rectangular kernels.
8.5.8 Write down the multivariate nearest neighbor estimator as a kernel estimator.
8.6.1 Let $X_1, \dots, X_n$ be iid normal $N(\theta, \sigma^2)$. Then (a) given σ^2, find the MLE of θ; (b) given θ, find the MLE of σ^2; (c) find the MLE of the pair (θ, σ^2).
8.6.2 Let a random variable X have the binomial distribution $B(p, n)$. Find the MLE of the probability of success p.
8.6.3 Explain the maximum likelihood sieve estimator (8.6.2).
8.6.4 Using the Lagrange multipliers method, find the pair (x, y) such that their product xy takes on the maximum possible value given $x^2 + y^2 = C$. Also, give a geometric interpretation of this problem.
8.7.1 Explain the problem of interpolation. Suggest several possible solutions, and then discuss their pluses and minuses.
8.7.2 What kind of a cubic spline is called a natural cubic spline? What is the reason for introducing a natural cubic spline?
8.7.3 Explain possible practical implications of Theorem 8.7.1.
8.7.4 What is the place in the proof of Theorem 8.7.1 where we use the fact that the cubic spline is natural?
8.7.5 Explain the least-squares spline estimator (8.7.7).
8.7.6 What are the estimates that minimize (8.7.8) with the smoothing parameters $\mu = 0$ and $\mu = \infty$?
8.8.1 Explain how a kernel density estimator may be constructed using a neural network.
8.8.2 Consider the previous exercise for the case of a regression.
8.8.3 Can a neural network be used for orthogonal series estimation?

8.8.4 Explain how a neural network solves hypothesis testing problems.
8.8.5 Suggest a problem where using a 2-layer neural network may be beneficial.
8.9.1 Does a twice-differentiable function belong to $Lip_{1,1}(L)$?
8.9.2 What are the boundary values of r and α such that a symmetric, bounded density with finite second moment still belongs to the class $S_{r,\alpha}$?
8.9.3 Let $K(x) := \frac{3}{4}(1-x^2)I_{|x|\le 1}$. What class $S_{r,\alpha}$ is this kernel from?
8.9.4 Consider kernels $K(x) := a + bx + cx^2 + dx^3 + ex^4$ supported on $[-1, 1]$. Find the parameters that imply $K \in S_{3,\alpha}$, $0 < \alpha \le 1$.
8.9.5 A kernel that belongs to all $S_{r,\alpha}$, $r \ge 0$, is called a *superkernel*. Give an example and explain why such kernels may be useful. Hint: Look at $K(x) := (1/2\pi)\int \cos(tx)[1 - \exp\{-1/t^2\}]dt$, and recall Theorem 8.9.1.
8.9.6 A kernel $K(x) := C\exp\{-1/(1-x^2)\}I_{|x|\le 1}$ is called a *mollifier*. What class $S_{r,\alpha}$ does this kernel belong to? What are the properties of this kernel? Hint: Think about how smooth this kernel is and look at its support. Also, we discussed this kernel before.
8.9.7 Explain all the steps in (8.9.14).
8.9.8 Establish (8.9.16) for $r > 0$.
8.9.9 Is the assumption (8.9.5) used to obtain lines (8.9.21)–(8.9.23)?
8.9.10 MISE is simply the integrated MSE. Can we get an upper bound for MISE by formal integration of the right-hand side of (8.9.24)?
8.9.11 Consider the case $f \in Lip_{0,\alpha}(L)$. Find a nonasymptotic upper bound for the MSE, that is, a bound without $o(1)$.
8.9.12 Solve the problem of finding an upper bound for the MISE formulated after Theorem 8.9.1.
8.9.13 Check that the Epanechnikov kernel minimizes (8.9.31).
8.9.14 Prove (8.9.32) using any necessary assumption about f.
8.9.15 Is the Epanechnikov kernel also optimal for the MISE criteria?
8.9.16 Establish a result, similar to Theorem 8.9.1, for a regression model.
8.10.1 The double-exponential (Laplacian) density is $f_\lambda(x) := \frac{1}{2}\lambda e^{-\lambda|x|}$, $-\infty < x < \infty$. Use the reference method and this density to find the data-driven bandwidth.
8.10.2 Let the data-driven kernel estimator of the previous exercise be used to estimate a normal density. Compare its MISE with the MISE of a similar kernel estimator with the correct reference density.
8.10.3 Suggest a leave-one-out cross-validation procedure for a kernel (a) density estimator and (b) spectral density estimator.
8.10.4 Consider the smoothing spline estimator (8.7.8) and suggest an adaptive method for choosing the smoothing parameter.
8.10.5 Explain a leave-m-out cross-validation procedure for regression and density models. Are there situations where using large m may be attractive?
8.11.1 Analyze the relationship between Percentage of youngsters (X = *saving.x[,1]*) and Percentage of senior citizens (Y = *saving.x[,2]*). Then use the result to continue the discussion of Section 8.11.

8.11.2 Choose several data sets and analyze the Nadaraya–Watson kernel estimator using different bandwidths and kernels. Then explain which of the two arguments is more critical.

8.13 Notes

There is a wide choice of books about nonseries estimation methods. Comprehensive bibliographic notes may be found in the books by Fan and Gijbels (1996) and Simonoff (1996). A book-length discussion on classification and regression trees is given in Breiman et al. (1984).

8.1-8.2 The book by Silverman (1986) presents a simple introduction to these basic methods.

8.3 The literature on kernel estimation is extremely vast. A combination of books by Simonoff (1996) and Wand and Jones (1995) may be used for further reading. The first asymptotic results are due to Akaike (1954), Rosenblatt (1956), and Parzen (1962). Among recent ones, Lepskii, Mammen, and Spokoiny (1997) established rate optimality of kernel estimators with variable bandwidth over Besov spaces. Thus the kernel estimates may be an alternative to wavelet estimates for estimation of spatially inhomogeneous functions.

8.4 A book-length treatment of local polynomial regression is given by Fan and Gijbels (1996). An interesting discussion may be also found in Korostelev and Tsybakov (1993).

8.5 Simple rules survive! Since its conception in the 1950s, the nearest neighbor method still attracts the attention of many followers. Probably the simplest further reading is Silverman (1986, Section 5.2). An application to pattern recognition may be found in Devroye, Györfi, and Lugosi (1996).

8.6 The textbook by Devroye (1987) and the book by Grenander (1981) may be recommended for a mathematically mature reader.

8.7 The books by Eubank (1988), Green and Silverman (1994), and Wahba (1990) give a comprehensive account on both theory and applications of spline techniques. Sharp minimax results are discussed in Nussbaum (1985), Speckman (1985), and Golubev and Nussbaum (1992), among others.

8.8 The books by Ripley (1996) and Venables and Ripley (1997) are a good combination for the reader who would like to combine theory with S–PLUS applications.

8.9 In addition to the books mentioned for Section 8.3, the textbook by Devroye (1987) may be recommended.

8.10 All the above-mentioned books discuss data-driven estimators. See also Lepskii (1990), where the first kernel estimator, which attains the optimal adaptive rate of MSE convergence for Lipschitz functions, was suggested.

Appendix A. Fundamentals of Probability and Statistics

Statistics of nonparametric curve estimation is founded on parametric statistics, which, in turn, depends on the theory of probability. It will be sufficient for our purposes to present here only the basic definitions, concepts, and machinery of probability theory and parametric statistics in a form useful for nonparametric statistics. The reader interested in a more detailed and comprehensive account of these theories may refer to books by Ross (1997) and Casella and Berger (1990).

• **Probability Theory.** Many sets of data that are of a practical interest are generated by a *random experiment* which is an act or process that leads to a single outcome that cannot be predicted with certainty in advance. For instance, one may be interested in the number of heads (H) and tails (T) generated by flipping two coins or in a daily stock price. The outcome of an experiment will not be known in advance, but we can always suppose that the set of all possible outcomes is known. For the first example it is a set of four outcomes ({HH}, {HT}, {TH}, {TT}), for the second an interval of possible prices. We shall refer to such a set as the *sample space* and denote it by Ω.

Let us begin our discussion with the case of the *discrete* sample space $\Omega = \{w_1, w_2, \ldots\}$ with a finite or countably infinite number of single outcomes $w_l, l = 1, 2, \ldots$, also called *elementary* or *simple events*. By assumption a simple event cannot be decomposed into simpler outcomes of the experiment. A particular collection of possible outcomes of an experiment is called an *event*. In other words, an event is any subset of Ω including Ω itself. For instance, the event "at least one head" in the experiment of flipping two coins is the collection of 3 single outcomes ({HH}, {HT}, {TH}).

An event of interest can often be viewed as a composition of two or more events. Union and intersection are two typical ways for forming such an event. The *union* of two events A and B is the event that occurs if either A or B or both occur on a single performance of the experiment, and it is denoted by the symbol $A \cup B$. In other words, the union consists of all outcomes that are either in A or in B or in both A and B. The *intersection* of two events A and B is the event that occurs if both A and B occur on a single performance of the experiment, and it is denoted by $A \cap B$. In other words, the intersection consists of all outcomes that are in both A and B. Note that because the intersection of any two simple events is empty, it is natural to introduce such an *empty* (or so-called *null* or *nonoccurring*) event and denote it by $\emptyset$. Also, a very useful concept is the complementary event: The event A^c is the *complement* of the event A if it occurs when A does not, in other words, the complement of A consists of all outcomes in the sample space Ω that are not in A. In particular, $\Omega^c = \emptyset$.

The theory of probability assigns a *probability* (likelihood) to events. For the case of a discrete sample space, there are five steps for calculating the probability of an event: (1) Define the experiment. (2) List the simple events. (3) Assign probabilities to the simple events in such a way that they are numbers between 0 and 1 and their total is 1. (4) Determine the collection of simple events contained in the event of interest. (5) Sum probabilities of the simple events from that collection to obtain the probability of the event of interest.

We shall denote the probability of an event A by $P(A)$. Note that we may write $P(A) = \sum_{l:\, w_l \in A} P(w_l)$, $P(A) + P(A^c) = 1$, $P(\Omega) = 1$, and $P(\emptyset) = 0$.

Example A.1 Consider the experiment of tossing two coins assuming that they are balanced (fair). Let A be the event "two heads," let B be the event "no heads," and let C be the event "at least one tail." Find the probabilities of the following events: (i) $A \cup B$, (ii) $A \cap B$, (iii) $B \cap C$, (iv) A^c.
Solution: Recall that the sample space of this random experiment consists of four simple events; using our notation we may write that $w_1 = \{H, H\}$, $w_2 = \{H, T\}$, $w_3 = \{T, H\}$, and $w_4 = \{T, T\}$. It is given that the coins are balanced (fair), and thus all these simple events occur with the same likelihood, i.e., $P(w_1) = P(w_2) = P(w_3) = P(w_4) = \frac{1}{4}$. Note that all these probabilities are between 0 and 1 with the total 1; thus the requirement of step 3 for calculating probabilities of events is satisfied. Then we do steps 4 and 5 separately for each of the events of interest. (i) Note that $A = \{H, H\}$ and $B = \{T, T\}$, so their union is $A \cup B = (w_1, w_4)$, which, according to step 5 for calculating probabilities of events, implies $P(A \cup B) = P(w_1) + P(w_4) = \frac{1}{2}$. (ii) The intersection of A and B is the empty event; thus $P(A \cap B) = P(\emptyset) = 0$. (iii) Note that $C = (\{H, T\}, \{T, H\}, \{T, T\})$. Thus $B \cap C = \{T, T\}$ and $P(B \cap C) = P(w_4) = \frac{1}{4}$. (iv) Note that the complement of A is the event C. Thus $P(A^c) = P(C) = P(w_2) + P(w_3) + P(w_4) = \frac{3}{4}$.

Another way to find this probability is to use the formula $P(A^c) = 1 - P(A) = 1 - P(w_1) = 1 - \frac{1}{4} = \frac{3}{4}$.

Example A.2 A ball is "randomly drawn" from a bowl containing two white and three red balls. What is the probability that the ball is red?
Solution: The phrase "randomly drawn" means that we at random (with the same likelihood) draw a ball. Here the sample space consists of 5 outcomes (two white and three red balls), and all outcomes have the same probability $\frac{1}{5}$. Because there are 3 red balls in the bowl, the probability of the event that the first ball is red is equal to $\frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5}$.

In many cases it is desirable to study two events simultaneously. The events A and B are called *independent* if $P(A \cap B) = P(A)P(B)$. Otherwise, we refer to such events as *dependent*.

Example A.3 Consider the experiment of Example A.1. Define which of the following events are independent: (i) The first flip is a head and both flips are heads; (ii) The first flip is a head and the second is a tail.
Solution: (i) The intersection of these events is the elementary event $\{H, H\}$, and its probability is equal to $\frac{1}{4}$. The probability of the first event is $\frac{1}{2}$ and that of the second is $\frac{1}{4}$, and because $\frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{4}$ we conclude that these events are dependent. This conclusion supports our intuition about dependency, because if the second event occurs (we know that both flips are heads) then this implies that the first flip is a head. (ii) The intersection of these two events is the elementary event $\{HT\}$ and its probability is $\frac{1}{4}$. The probability that the first flip is a head is equal to $P(\{HH\}) + P(\{HT\}) = \frac{1}{2}$ and the probability that the second flip is a tail is equal to $P(\{HT\}) + P(\{TT\}) = \frac{1}{2}$ (recall that the probability of any event is to be calculated via the sum of probabilities of elementary events that imply the event). Thus we get $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$, and therefore these two events are independent.

To finish our discussion of the events and their probabilities, two useful remarks are due. First, if we set $\cup_{l=1}^{n} A_l = A_1 \cup A_2 \cup \cdots \cup A_n$ and similarly $\cap_{l=1}^{n} A_l = A_1 \cap A_2 \cap \cdots \cap A_n$, then DeMorgan's laws $(\cup_{l=1}^{n} A_l)^c = \cap_{l=1}^{n} A_l^c$ and $(\cap_{l=1}^{n} A_l)^c = \cup_{l=1}^{n} A_l^c$ make the relationship between the basic operations of forming unions, intersections, and complements very simple.

Second, the five steps formulated above of finding probabilities of events are based on the following *three axioms of probability.*
Axiom 1 The probability of any event should be between 0 and 1, that is, $0 \leq P(A) \leq 1$.
Axiom 2 The probability that an outcome of a random experiment belongs to the sample space Ω is equal to 1, that is, $P(\Omega) = 1$.
Axiom 3 For any countable sequence of mutually exclusive events $A_1, A_2, \ldots$ (that is, the events such that $A_i \cap A_j = \emptyset$ when $i \neq j$, or, in words, events with no common outcomes) the following relation holds, $P(\cup_{l=1}^{\infty} A_l) = \sum_{l=1}^{\infty} P(A_l)$.

Discrete Random Variables. In many cases it is convenient to define quantities of interest as numbers. These quantities of interest, or more formally, real-valued functions defined on a sample space, are known as *random variables*. A random variable that can take on at most a countable number of possible values is said to be a *discrete random variable*. For a discrete random variable X we define the *probability mass function* $f(a)$ of X as $f(a) := P(X = a)$. Recall that := or =: means "by definition."

Example A.4 An experiment consists of tossing 2 fair coins. If we let X denote the number of heads, find its probability mass function $f(k) := P(X = k)$, i.e., the probability that the number of heads is equal to k.
Solution: It is clear that $f(k) = 0$ for $k < 0$ and $k > 2$ because there are no elementary events that lead to such events (in other words, these events are empty events). Thus, we should calculate the probability mass function for $k = 0, 1$, and 2. We get $f(0) = P(X = 0) = P(\{TT\}) = \frac{1}{4}$, $f(1) = P(X = 1) = P(\{HT\}) + P(\{TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, and $f(2) = P(X = 2) = P(\{HH\}) = \frac{1}{4}$. Thus the answer is: $f(k) = 0$ for $k < 0$, $f(0) = \frac{1}{4}$, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$, and $f(k) = 0$ for $k > 2$.

The probability mass function $f(x)$ gives us a complete description of a discrete random variable. The other complete characteristic is the *cumulative distribution function* (*cdf*) $F(x) := P(X \le x)$. To stress that the probability mass function and the cumulative distribution function describe a particular random variable X, we may write $f^X(x)$ and $F^X(x)$.

Example A.5 Find the cumulative distribution function of the random variable in Example A.4.
Solution: By definition, $F(x) = P(X \le x) = \sum_{-\infty < k \le x} f(k)$, and therefore $F(x) = 0$ for $x < 0$, $F(x) = \frac{1}{4}$ for $0 \le x < 1$, $F(x) = \frac{3}{4}$ for $1 \le x < 2$, and $F(x) = 1$ for $x \ge 2$. Thus, the cumulative distribution function is a step function with jumps equal to $P(X = k)$ at points $k = 0$, $k = 1$, and $k = 2$.

In many situations it is of interest to study more than one random variable associated with the same experiment. In order to deal with such cases, we begin with the case of two discrete random variables X and Y.

The *joint cumulative distribution function* of X and Y is defined by $F^{XY}(x, y) := P((X \le x) \cap (Y \le y))$, and the corresponding *joint probability mass function* by $f^{XY}(x, y) := P((X = x) \cap (Y = y))$.

The joint cumulative distribution as well as the probability mass function completely define two random variables. In particular, the *marginal* cumulative distribution and marginal probability mass function of X are defined by $F^X(x) := F^{XY}(x, \infty)$ and $f^X(x) := \sum_y f^{XY}(x, y)$, respectively. Here $\sum_y f^{XY}(x, y)$ means the summation over all values y of Y. Note that it suffices to sum only over values y such that $f^Y(y) > 0$ because $\sum_y f^{XY}(x, y) = \sum_{y: f^Y(y) > 0} f^{XY}(x, y)$.

Example A.6 Consider again Example A.4 and let Y denote the number of tails. Find the joint probability mass function of X and Y.
Solution: Note that the only possible values (x, y) of X and Y are such that they are nonnegative and $x + y = 2$ (the total number of heads and tails should be 2). Thus we conclude that $f^{XY}(0, 2) = P(\{TT\}) = \frac{1}{4}$, $f^{XY}(1, 1) = P(\{HT\} \cup \{TH\}) = \frac{1}{2}$, $f^{XY}(2, 0) = P(\{HH\}) = \frac{1}{4}$, and $f^{X,Y}(x, y) = 0$ for all other values of (x, y).

Example A.7 Let X and Y be discrete random variables with the joint probability mass function $f^{XY}(x, y)$. Find the probability mass function for the sum $Z = X + Y$.
Solution: Because $f^{X+Y}(z) := P(X + Y = z)$, using the third axiom of probability yields that $P(X + Y = z) = \sum_{x: f^X(x)>0} P((X = x) \cap (Y = z - x)) = \sum_{x: f^X(x)>0} f^{XY}(x, z - x)$. This gives us the answer,

$$f^{X+Y}(z) = \sum_{x: f^X(x)>0} f^{XY}(x, z - x). \tag{A.1}$$

As in the definition of independent events, two discrete random variables X and Y are called *independent* if $f^{XY}(x, y) = f^X(x) f^Y(y)$ (in terms of distribution functions if $F^{XY}(x, y) = F^X(x) F^Y(y)$).

As an example, note that for the case of independent X and Y the formula (A.1) is simplified,

$$f^{X+Y}(z) = \sum_{x: f^X(x)>0} f^X(x) f^Y(z - x). \tag{A.2}$$

Let A and B be two events and $P(B) > 0$. Then the *conditional probability* of A given B is defined by $P(A|B) := P(A \cap B)/P(B)$. Similarly, if X and Y are two discrete random variables, we define the *conditional probability mass function* of X given $Y = y$ by $f^{X|Y}(x|y) := f^{XY}(x, y)/f^Y(y)$, and we define the *conditional cumulative distribution function* by $F^{X|Y}(x|y) := P(X \le x|Y = y) = \sum_{u:\, u \le x} f^{X|Y}(u|y)$. Note that if X is independent of Y, then the conditional probability mass function and the conditional distribution function are equal to the unconditional ones.

Example A.8 In Example A.7, find the conditional probability mass function of Y given $Z = z$.
Solution: Using (A.1) we write,
$f^{Y|Z}(y|z) = f^{YZ}(y, z)/f^Z(z) = f^{XY}(z - y, y)/\sum_{x:\, f_X(x)>0} f^{XY}(x, z - x)$.

Besides the probability mass function and cumulative distribution function, which completely describe random variables, several other characteristics are customarily used and give some partial descriptions of random variables.

Four of the most important such characteristics are: (i) The *expectation* (the *expected value* or the *mean*) of X, denoted by $E\{X\}$ and defined by

$$E\{X\} := \sum_{x: f^X(x)>0} x f^X(x) =: \int_{-\infty}^{\infty} x dF^X(x). \tag{A.3}$$

(ii) The *variance* of X, denoted by $\mathrm{Var}(X)$ and defined by

$$\mathrm{Var}(X) := E\{[X - E\{X\}]^2\} = \sum_{x: f^X(x)>0} [x - E\{X\}]^2 f^X(x). \tag{A.4}$$

(iii) The *kth moment* of X is defined by $\mu_k(X) := E\{X^k\}$, $k = 1, 2, \ldots$. Clearly, the first moment is the mean, while the second moment is equal to the variance plus the squared mean, i.e.,

$$E\{X^2\} = \mathrm{Var}(X) + [E\{X\}]^2. \tag{A.5}$$

(iv) The *covariance* of two random variables X and Y, denoted by $\mathrm{Cov}(X, Y)$ and defined by

$$\mathrm{Cov}(X, Y) := E\{(X - E\{X\})(Y - E\{Y\})\}. \tag{A.6}$$

The *standard deviation*, which is equal to the square root of the variance, is another useful characteristic. Note that the standard deviation has the same units as the variable, while the variance is measured in squared units. Also, the *correlation* of two random variables X and Y, denoted by $\rho(X, Y)$ and defined by $\rho(X, Y) := \mathrm{Cov}(X, Y)/[\mathrm{Var}(X)\mathrm{Var}(Y)]^{1/2}$, is often used.

Example A.9 In Example A.4 calculate the mean, the variance, and the standard deviation of X.
Solution: We begin by calculating the mean. Write, $E\{X\} = \sum_{k=0}^{2} k f(k) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 2 \cdot \frac{1}{4} = 1$. This is a rather obvious outcome because on average we should get one head after tossing two fair coins. Then we may calculate the variance, $\mathrm{Var}(X) = \sum_{k=0}^{2} (k-1)^2 f(k) = (-1)^2 \frac{1}{4} + (0)^2 \frac{2}{4} + (1)^2 \frac{1}{4} = \frac{1}{2}$. Finally, the standard deviation is equal to $2^{-1/2}$.

Suppose that we are given a random variable X along with its probability mass function and we want to calculate not the expected value of X but the expectation of some function $g(X)$, for instance, X^4. The straightforward way to do this is as follows: define $Y := g(X)$, find the probability mass function $f^Y(y)$, and then calculate $\sum_{y: f^Y(y)>0} y f^Y(y)$. However, it turns out that the expectation of $g(X)$ may be calculated much simpler by the formula

$$E\{g(X)\} = \sum_{x:\ f^X(x)>0} g(x) f^X(x). \tag{A.7}$$

Similarly, for any two random variables X and Y and a bivariate function $g(x, y)$,

$$E\{g(X,Y)\} = \sum_{(x,y):\, f^{XY}(x,y)>0} g(x,y) f^{XY}(x,y). \tag{A.8}$$

Below, whenever it is not confusing, we may write $\sum_x g(x)f(x)$ or simply $\sum g(x)f(x)$ in place of $\sum_{x:\, f(x)>0} g(x)f(x)$.

Example A.10 Prove that if X and Y are independent then for any constants a, b, c, and d,

$$E\{(aX+b)(cY+d)\} = (aE\{X\}+b)(cE\{Y\}+d).$$

Solution: Using (A.8) we write, $E\{(aX+b)(cY+d)\} = \sum_{(x,y)}(ax+b)(cy+d)f^{XY}(x,y)$. Then, because X and Y are independent, $\sum_{(x,y)}(ax+b)(cy+d)f^{XY}(x,y) = [\sum_x (ax+b)f^X(x)][\sum_y (cy+d)f^Y(y)]$. This yields the result.

Example A.11 Prove that for any two random variables X and Y the correlation $\rho(X,Y)$ (i) takes on values between -1 and 1; (ii) is equal to 0 whenever X and Y are independent; (iii) is equal to 1 if $X = Y$ and equal to -1 if $X = -Y$.

Solution: (i) The proof is based on the famous *Cauchy–Schwarz* inequality

$$|E\{Z_1 Z_2\}|^2 \le E\{Z_1^2\}E\{Z_2^2\}, \tag{A.9}$$

which holds for any two random variables Z_1 and Z_2. It is proved for a general setting in Section 2.3; see the paragraph below line (2.3.4). Set $Z_1 = X - E\{X\}$, $Z_2 = Y - E\{Y\}$ and then using the Cauchy–Schwarz inequality we get $|\mathrm{Cov}(X,Y)| \le [\mathrm{Var}(X)\mathrm{Var}(Y)]^{1/2}$. The last inequality implies the desired result. (ii) Because X and Y are independent, the assertion follows from Example A.10. (iii) Note that $E\{(X-E\{X\})(X-E\{X\})\} = \mathrm{Var}(X)$, which together with Example A.10 implies the assertion.

Example A.11 shows that the correlation is a very convenient numerical characteristic of dependency between two random variables.

There are several other useful formulae that are proved similarly to Example A.10. Let X and Y be random variables and a and b be constants. Then,

$$E\{aX+bY\} = aE\{X\} + bE\{Y\}, \tag{A.10}$$

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(XY). \tag{A.11}$$

An important corollary of (A.11) and Example A.11(ii) is that if X and Y are independent, then

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y). \tag{A.12}$$

Let us define the *indicator* of an event B as a function $I_{\{B\}}$ such that $I_{\{B\}} = 1$ if B occurs and $I_{\{B\}} = 0$ if B fails to occur. For instance,

$I_{\{3>2\}} = 1$, while $I_{\{2=3\}} = 0$. This function allows us to write the probability $P(X \in A)$ of the event $X \in A$ as the expectation of $I_{\{X\in A\}}$, that is, $P(X \in A) = E\{I_{\{X\in A\}}\}$. Indeed, write $E\{I_{\{X\in A\}}\} = \sum_x I_{\{x\in A\}} f^X(x) = \sum_{x:\, x\in A} f^X(x) = P(X \in A)$, or even simpler $E\{I_{\{X\in A\}}\} = P(X \in A)1 + P(X \notin A)0 = P(X \in A)$.

To study numerical characteristics of a random variable X given $Y = y$, define the *conditional expectation* of X given $Y = y$,

$$E\{X|Y = y\} := \sum_x x f^{X|Y}(x|y). \tag{A.13}$$

The conditional expectation plays a central role in both probability and statistics due to the formula

$$E\{X\} = E\{E\{X|Y\}\}, \tag{A.14}$$

which allows one to calculate the expectation of X via the expectation of its conditional expectation given a random variable Y. Here $E\{X|Y\} := \varphi(Y)$ where $\varphi(y) := E\{X|Y = y\}$. To prove (A.14) write,

$$\begin{aligned} E\{E\{X|Y\}\} &= \sum_{y:\, f^Y(y)>0} \Big[\sum_x x f^{X|Y}(x|y)\Big] f^Y(y) \\ &= \sum_x \sum_{y:\, f^Y(y)>0} x\big[f^{XY}(x,y)/f^Y(y)\big] f^Y(y) \\ &= \sum_x \sum_y x f^{XY}(x,y) = \sum_x x\Big[\sum_y f^{XY}(x,y)\Big] \\ &= \sum_x x f^X(x) = E\{X\}. \end{aligned}$$

Example A.12 A miner is trapped in a mine containing two doors. The first door leads to a tunnel that will take him to safety after 2 hours; the second door leads to a tunnel that will return him to the mine after 1 hour of travel. Assume that the miner is equally likely to choose either door each time. Find the expected length of time until he reaches safety.

Solution: Let X denote the amount of time (in hours) until the miner reaches safety, and let Y denote the door he initially chooses. Then $E\{X\} = E\{E\{X|Y\}\} = E\{X|Y = 1\}P(Y = 1) + E\{X|Y = 2\}P(Y = 2) = (1/2)[E\{X|Y = 1\} + E\{X|Y = 2\}]$. Clearly, $E\{X|Y = 1\} = 2$, while $E\{X|Y = 2\} = 1 + E\{X\}$ because after returning to the mine everything begins again. Combining the results we get $E\{X\} = \frac{1}{2}[2+1+E\{X\}]$, which yields $E\{X\} = 3$. Thus on average it takes three hours for the miner to reach safety.

To finish our discussion of discrete random variables, let us introduce two classical discrete random variables.

The Binomial Random Variable. Suppose that the outcome of a trial (random experiment) can be classified as either a "success" or a "failure."

The example of tossing a coin is a particular case. If we let $Z = 1$ when the outcome is a success and $Z = 0$ when it is a failure, then the probability mass function is given by $f^Z(1) = p$ and $f^Z(0) = 1 - p$, where p, $0 \leq p \leq 1$, is the probability that the trial is a "success." Suppose that we independently repeat this experiment n times. Then the random variable X that is equal to the number of successes is called *binomial*, and in this case we write that X is distributed according to $B(n, p)$. The probability mass function of a $B(n, p)$ random variable is given by the formula $f(k) = [n!/k!(n-k)!]p^k(1-p)^{n-1}$, $E\{X\} = np$, and $\mathrm{Var}(X) = np(1-p)$. Here $0 \leq k \leq n$, $k! = k \cdot (k-1) \cdots 1$ for $k > 1$, $1! = 1$, and $0! = 1$. A binomial random variable $B(1, p)$ (that is, the outcome of a single trial, denoted above by Z) is often referred to as a Bernoulli random variable.

The Poisson Random Variable. A random variable X, taking on values $0, 1, 2, \ldots$, is said to be a *Poisson* random variable with parameter $\lambda > 0$ if $P(X = k) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$. Note that the familiar Taylor expansion, $e^\lambda = \sum_{k=0}^{\infty} \lambda^k/k!$ yields that the Poisson random variable is defined correctly. Straightforward calculation shows that $E\{X\} = \mathrm{Var}(X) = \lambda$. This explains why λ is called the *intensity* of the Poisson random variable. The Poisson random variable is closely related to the Binomial random variable, since the cumulative distribution function of a Poisson random variable with the intensity λ is the limit of the cumulative distribution function of $B(n, p)$ when $n \to \infty$ and $np \to \lambda$. The latter also sheds light on the formulae for the expectation and the variance of a Poisson random variable.

Continuous Random Variables. A random variable X is called continuous if there exists a nonnegative function $f^X(x)$ such that $\int_{-\infty}^{\infty} f^X(x)dx = 1$, and the cumulative distribution function $F^X(x) := P(X \leq x)$ of X may be written as

$$F^X(x) = \int_{-\infty}^{x} f^X(u)du. \tag{A.15}$$

The function $f^X(x)$ is called the *probability density function* or simply density of the continuous random variable X. From the definition we get $P(X = x) = P(X \leq x) - P(X < x) = 0$ for any number x, and this represents a major distinction between continuous and discrete random variables. Also, for any two constants $a \leq b$ we get $P(a \leq X \leq b) = \int_a^b f^X(x)dx$. Moreover, $dF^X(x)/dx = f^X(x)$ at the points of continuity of the density, thus a continuous probability density is the derivative of the cumulative distribution function. Both the cumulative distribution function and the probability density give a complete description of the corresponding random variable. Also note that if a function is integrable to 1 and nonnegative, then it is the probability density of a random variable.

Example A.13 Suppose that X is a continuous random variable that describes the amount of time in days that a printer functions before breaking down, with probability density function $f^X(x) = Ce^{-x/50}$ for $x \geq 0$ and $f^X(x) = 0$ otherwise. Find C and the probability that it will function less than 30 days.
Solution: Due to axiom 2,

$$P(X < \infty) = \int_{-\infty}^{\infty} f^X(x)dx = 1. \tag{A.16}$$

Recall the integration formula $\int_a^b Ce^{-x/\lambda}dx = C\lambda[e^{-a/\lambda} - e^{-b/\lambda}]$. This formula together with (A.16) gives $C = 1/50$. Then,
$P(X < 30) = \int_0^{30}(1/50)e^{-x/50}dx = [e^{0/50} - e^{-30/50}]/50 = [1 - e^{-3/5}]/50$.

As in the discrete case, we define the expectation of a continuous random variable as

$$E\{X\} := \int_{-\infty}^{\infty} xf^X(x)dx = \int_{-\infty}^{\infty} xdF^X(x).$$

Recall that notions of variance, standard deviation, covariance, and correlation are defined via the expectation, so there is no need for new definitions. Another group of important characteristics of a distribution of a continuous random variable X are the αth quantiles q_α^X such that $P(X \leq q_\alpha^X) = \alpha$. While moments of X may not exist (the Cauchy random variable with the density $f^X(x|\theta) = 1/(\pi(1 + (x - \theta)^2))$ has no moments), the quantiles always exist. The customarily analyzed quantiles are the first quartile, the median (second quartile), and the third quartile, which correspond to $\alpha = 0.25,\ 0.5,\ 0.75$.

Let us define two specific continuous random variables.

The Uniform Random Variable. A random variable X is said to be uniformly distributed over the interval $[a, b]$ if its probability density function is given by $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and $f(x) = 0$ otherwise. We shall often refer to this random variable by saying that X is $U(a, b)$. Note that f is nonnegative and integrated to unity, so it is indeed a probability density. It is customary to refer to $[a, b]$ as the support of the density. In general, the *support* is a minimal set such that the probability density vanishes (is equal to zero) beyond the set.

The name "uniform" is explained by the fact that for any $a \leq a_1 \leq b_1 \leq b$ we have $P(a_1 \leq X \leq b_1) = \int_{a_1}^{b_1}(b-a)^{-1}dx = (b_1-a_1)/(b-a)$. In words, the probability that a uniform random variable is in any particular subinterval of $[a, b]$ is proportional to the length of the subinterval. A straightforward calculation shows that $E\{X\} = (a + b)/2$ and $\mathrm{Var}(X) = (b - a)^2/12$.

The Normal Random Variable. We say that X is a normal random variable $N(\mu, \sigma^2)$ if its density is given by

$$f(x) = (2\pi\sigma^2)^{-1/2}\, e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty\,. \tag{A.17}$$

Direct calculations show that $E\{X\} = \mu$ and $\mathrm{Var}(X) = \sigma^2$, that is, the normal random variable is completely defined by its mean and variance. There are several properties of a normal random variable that will be important to us. The first one is that the graph of a normal density is a symmetric bell-shaped curve about the mean. The second is that $f(x)$ practically vanishes (becomes very small) whenever $|x - \mu| > 3\sigma$ (the so-called rule of three sigma). The third is that the sum of two independent normal random variables with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) is again a normal random variable $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Example A.14 A random variable X is called *standard normal* if it is $N(0, 1)$. Show that if Y is $N(\mu, \sigma^2)$, then $(Y - \mu)/\sigma$ is a standard normal random variable.
Solution: Write, $P((Y - \mu)/\sigma \le y) = P(Y \le \mu + y\sigma) = (2\pi\sigma^2)^{-1/2} \int_{-\infty}^{\mu+y\sigma} e^{-(u-\mu)^2/2\sigma^2} du$. Then the substitution $v = (u - \mu)/\sigma$ gives the desired $P((Y - \mu)/\sigma \le y) = (2\pi)^{-1/2} \int_{-\infty}^{y} e^{-v^2/2} dv$.

We say that X and Y are *jointly continuous* if there exists a function $f^{XY}(x, y)$ (the *two-dimensional* or *bivariate probability density*) such that $F^{XY}(x, y) := P(X \le x, Y \le y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f^{XY}(u, v) du\, dv$.

If X and Y are jointly continuous, then they are individually continuous, and the *marginal* density of X can be obtained by the integration, $f^X(x) := \int_{-\infty}^{\infty} f^{XY}(x, y) dy$. The marginal density of Y is defined absolutely similarly. For such random variables, necessary and sufficient conditions for their independence is $f^{XY}(x, y) = f^X(x) f^Y(y)$ for all x and y.

Example A.15 The joint density function of independent random variables X and Y is given. Find the probability density of $X + Y$. (This example is the continuous counterpart of Example A.7.)
Solution: Typically, the simplest way to solve such a problem is first to find the cumulative distribution function and then take the derivative. We write $F^{X+Y}(z) = P(X + Y \le z) = \int \int_{x+y \le z} f^{XY}(x, y) dx\, dy = \int_{-\infty}^{\infty} f^X(x) [\int_{y:\, y \le z-x} f_Y(y) dy] dx = \int_{-\infty}^{\infty} f^X(x) F^Y(z - x) dx$. Taking the derivative, we get

$$f^{X+Y}(z) = \int_{-\infty}^{\infty} f^X(x) f^Y(z - x) dx. \tag{A.18}$$

The right side of (A.18) is called the *convolution* of f_X and f_Y on the real line. Thus, the density of the sum of two independent random variables is equal to the convolution of their densities.

The *conditional* probability density function of X given $Y = y$ is defined for all values of y such that $f^Y(y) > 0$ by the formula

$$f^{X|Y}(x|y) := f^{XY}(x, y)/f^Y(y). \tag{A.19}$$

The conditional expectation of $g(X)$ given $Y = y$ is calculated by the formula

$$E\{g(X)|Y = y\} = \int_{-\infty}^{\infty} g(x) f^{X|Y}(x|y) dx. \tag{A.20}$$

Also, as in the discrete case,

$$P(X \in A) = E\{I_{\{X \in A\}}\} = \int_A f^X(x) dx, \tag{A.21}$$

$$\begin{aligned} P((X,Y) \in B) &= E\{P((X,Y) \in B|Y)\} \\ &= \int_{-\infty}^{\infty} P((X,y) \in B|Y = y) f^Y(y) dy. \end{aligned} \tag{A.22}$$

Example A.16 Let X and Y be independent continuous random variable. Find the distribution of $X + Y$ by conditioning on the value of Y.
Solution: Write $P(X + Y \le z) = \int_{-\infty}^{\infty} P(X + Y \le z|Y = y) f^Y(y) dy = \int_{-\infty}^{\infty} P(X \le z - y|Y = y) f^Y(y) dy$. Then using the independence of X and Y we get $P(X + Y \le z) = \int_{-\infty}^{\infty} F^X(z - y) f^Y(y) dy$.

Much of the previous development carries over to the case of more than two random variables. In this case we define an n-dimensional *random vector* or *sequence* $X^n = (X_1, X_2, \ldots, X_n)$ via its joint cumulative distribution function $F^{X^n}(x^n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$. For instance, jointly continuous random variables $X_1, X_2, \ldots, X_n$ are independent if and only if $f^{X^n}(x^n) = f^{X_1}(x_1) f^{X_2}(x_2) \cdots f^{X_n}(x_n)$ for all x^n.

The Multivariate Normal Distribution. We begin with several preliminary notions. Let an n-dimensional random vector $X^n = (X_1, \ldots, X_n)'$ be a column vector each of whose components is a random variable. Here A' denotes the transpose array A (vector or matrix). If $E\{|X_i|\} < \infty$ for each i, then the expectation of X^n is defined as the column vector $E\{X^n\} = (E\{X_1\}, \ldots, E\{X_n\})'$. In the same way the expectation of any array of random variables (e.g, a matrix of random variables) is defined.

For two random vectors X^n and Y^n such that all their entries have a finite second moment, we define the covariance matrix of X^n and Y^n by the matrix $M_{X^n Y^n} = \mathrm{Cov}(X^n, Y^n) = E\{(X^n - E\{X^n\})(Y^n - E\{Y^n\})'\}$. Note that the (l, m) entry of the covariance matrix is the $\mathrm{Cov}(X_l, Y_m)$.

If a^n is an n-component column vector of constants, $B^{n \times m}$ is an $n \times m$ matrix of constants, and X^n is a random vector with elements that have a finite second moment, then the random variable $Y^n = a^n + B^{n \times m} X^m$ has the mean

$$E\{Y^n\} = a^n + B^{n \times m} E\{X^m\}, \tag{A.23}$$

and the covariance matrix

$$M_{Y^n Y^n} = B^{n \times m} M_{X^m X^m} (B^{n \times m})'. \tag{A.24}$$

By definition, a random vector Y^n is said to be multivariate normal (to have a multivariate normal distribution) if and only if there exist a column vector a^n, a matrix $B^{n\times m}$, and a random vector X^m with independent standard normal random components such that $Y^n = a^n + B^{n\times m}X^m$.

The joint probability density of a multivariate normal vector Y^n with expectation μ^n and the $n \times n$ covariance matrix B is

$$f^{Y^n}(y^n) = \frac{1}{(2\pi)^{n/2}(\det B)^{1/2}} \exp\{-(y^n - \mu^n)'B^{-1}(y^n - \mu^n)/2\}. \quad \text{(A.25)}$$

Limit Theorems. Let n *independent and identically distributed* (from now on we use the shorthand notation iid) random variables $X_1, X_2, \ldots, X_n$ be given. In other words, all these random variables have the same distribution as a random variable X, and therefore they may be considered as n independent *realizations* of X. A typical problem of statistics is to estimate the expectation (theoretical mean) $E\{X\}$ based on the realizations. A natural approach is to estimate $E\{X\}$ by the so-called sample mean $\bar{X} = [X_1 + X_2 + \cdots + X_n]/n$.

How close are $\bar{X}$ and $E\{X\}$? A first and rather rough answer is given by the *Chebyshev inequality*, which states that if Y is a random variable with finite mean and finite variance, then for any value $k > 0$,

$$P(|Y - E\{Y\}| \geq k) \leq \mathrm{Var}(Y)/k^2. \quad \text{(A.26)}$$

To study the sample mean with the help of the Chebyshev inequality we calculate

$$E\{\bar{X}\} = E\{X\}, \quad \text{(A.27)}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n, \quad \text{(A.28)}$$

(here (A.10) and (A.12) have been used), and then find that

$$P(|\bar{X} - E\{X\}| \geq k) \leq \mathrm{Var}(X)/nk^2. \quad \text{(A.29)}$$

We conclude that the sample mean becomes closer and closer to the theoretical mean (the expectation) as the sample size n increases.

Example A.17 Suppose that the levels of insulin measured during the day are iid random variables with mean 12 and variance 2. Using the Chebyshev inequality, estimate the probability that the level of insulin takes on values between 10 and 14.

Solution: Let X denote the level of insulin. To use (A.26) set $k = 2$ and write $P(10 < X < 14) = P(|X - E\{X\}| < 2) = 1 - P(|X - E\{X\}| \geq 2)$. Then, using the Chebyshev inequality we get $P(|X - E\{X\}| \geq 2) \leq 2/2^2 = \frac{1}{2}$. Thus we conclude that $P(10 < X < 14) \geq \frac{1}{2}$.

Now we are going to discuss two limit theorems, the first of which is classified as a "law of large numbers" and the second as a "central limit theorem." The former is concerned with the convergence of the sample

mean to the theoretical mean as $n \to \infty$. The latter studies cases where the sample mean may be approximately described by a normal random variable with the expectation equal to the theoretical mean.

Here we give two examples of such results.

The Weak Law of Large Numbers. Let $X_1, X_2, \dots, X_n$ be iid random variables with a finite mean μ. Then for any $\epsilon > 0$

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty. \tag{A.30}$$

The Central Limit Theorem. Let $X_1, X_2, \dots, X_n$ be iid random variables with mean μ and finite variance σ^2, and let ξ be a standard normal random variable. Then for any real x,

$$P\Big(\frac{n^{1/2}(\bar{X} - \mu)}{\sigma} \le x\Big) \to P(\xi \le x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-u^2/2} du \quad \text{as} \quad n \to \infty. \tag{A.31}$$

Note that both the law of large numbers and the central limit theorem tell that $\bar{X}$ should be close to $E\{X\}$ for large n.

• **Parametric Statistics.** Here we discuss basic concepts of parametric statistics that deal with a *sample* of n iid random variables $X^n := (X_1, X_2, \dots X_n)$. We refer to n as the *sample size*. We also often refer to X_l as the lth *realization* (*observation*) of a random variable X with the same distribution. If the realizations are arranged in ascending order (from the smallest to the largest), then they are called *ordered statistics* and denoted by $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, where $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$.

The main assumption of parametric statistics is that the cumulative distribution function $F_\theta^X(x)$ of X is known up to the parameter θ, and $\theta \in S$ where the set S is known. For instance, X may be a normal random variable with unknown nonnegative mean θ and unit variance, in which case $S = [0, \infty)$.

Customarily, three topics are studied: (i) point estimation of a parameter; (ii) confidence interval estimation of a parameter; (iii) hypotheses testing. Let us clarify the topics via an example. Suppose that we know that the temperature at noontime tomorrow has a normal distribution with unknown mean. Then, based on previous measurements of the temperature, (i) point estimation gives us a recipe on how to find an estimate of the mean (how to find a point estimate); (ii) confidence interval estimation gives us a recipe on how to find an interval that covers the mean with at least a given probability; (iii) hypotheses testing gives us a recipe on how to answer questions like "will the mean be above 70°F?" or "will the mean be below 50°F?"

Let us consider methods of these three topics.

Point Estimation of a Parameter. The *estimate* of a parameter θ is a function of observations. We use different diacritics above θ such as $\hat{\theta}$ and $\tilde{\theta}$ to denote estimates or *statistics* based on a given data set. Similarly, if

any functional or function of a parameter is estimated, then the functional or function with a diacritic above denotes an estimate (statistics).

The *mean squared error* $\text{MSE}(\hat{\theta}, \theta) := E_\theta\{(\hat{\theta} - \theta)^2\}$ is traditionally used to measure the goodness of estimating θ by an estimate $\hat{\theta}$. Here $E_\theta\{\cdot\}$ denotes the expectation according to the distribution F_θ, and we use this subscript when it is important to stress that the underlying distribution depends on the parameter.

Example A.18 Let X_1, X_2, and X_3 be iid with a symmetric distribution about θ. We would like to compare the following three estimates of θ: $\hat{\theta}_1 = \bar{X}$, $\hat{\theta}_2 = X_{(2)}$, and $\hat{\theta}_3 = (X_{(1)} + X_{(3)})/2$. For the case where the observations are uniformly distributed on $[0, 1]$ find the best estimate that minimizes the mean squared error.
Solution: Here we should make straightforward calculations. They show that the mean squared error is equal to 1/36, 1/20, and 1/40 for the first, second, and third estimates, respectively. Thus, the third estimate is the best. Note that the answer crucially depends on the assumption about the underlying uniform distribution. For instance, for the case of a normal distribution the first estimate (the sample mean) has the minimal mean squared error.

In the example above we simply suggested several reasonable estimates and then compared their *risks*—here mean squared errors. The theory of point estimation has developed many general methods of finding estimates. Below, we briefly consider several of them.

Plug-In Method. Suppose that θ may be written as $\theta = G(F_\theta)$. Then the *plug-in* estimate is defined as $\hat{\theta} = G(\hat{F})$, where $\hat{F}$ is an estimate of the cumulative distribution function F_θ.

Recall that by definition $F_\theta(x) = P(X \le x|\theta)$, and therefore $F_\theta(x) = E\{I_{\{X \le x\}}|\theta\}$. Thus, $F_\theta(x)$ is the expectation of the random variable $I_{\{X \le x\}}$ and may be estimated by a *sample mean estimate*,

$$\bar{F}_n(x) := n^{-1} \sum_{l=1}^{n} I_{\{X_l \le x\}}. \tag{A.32}$$

The statistic $\bar{F}_n(x)$ plays a central role in statistics and probability, and it is referred to as the *empirical cumulative distribution function*.

Example A.19 Let n iid realizations $X_1, \ldots, X_n$ of a random variable X with a finite unknown mean μ be given. Find a plug-in estimate of μ.
Solution: Recall that $\mu = E\{X\} = \int_{-\infty}^{\infty} x dF_\mu(x)$, where $F_\mu(x)$ denotes the cumulative distribution function of X given the mean μ. Then the plug-in estimate is $\hat{\mu} = \int_{-\infty}^{\infty} x d\bar{F}_n(x)$, and according to (A.3) $\hat{\mu} = n^{-1} \sum_{l=1}^{n} X_l = \bar{X}$. Thus, here the plug-in estimate is the sample mean.

Example A.20 We observe n realizations of a uniform random variable $U(\theta, \theta+1)$. Find the plug-in estimate of θ.
Solution: One of the possible ways to describe θ via the cumulative distribution function $F_\theta(x)$ of X given the parameter θ is to write $\theta = \int_{-\infty}^{\infty} x dF_\theta(x) - 0.5$. In words, we notice that θ is equal to the mean minus a half. Then, Example A.19 implies that the plug-in estimate is $\hat{\theta} = \bar{X} - 0.5$.

The underlying idea of a plug-in estimate is that an empirical cumulative distribution function should approximate an underlying cumulative distribution function. And this is indeed the case. For instance, consider the following measure of closeness between F and $\bar{F}_n$: $\hat{D}_n := \sup_{-\infty<x<\infty} |\bar{F}_n(x) - F(x)|$, known as the *Kolmogorov–Smirnov distance*. Then it can be shown that for iid realizations there exists a finite positive constant C (not depending on F) such that

$$P(\hat{D}_n > d) \le C \exp\big\{-2nd^2\big\}, \quad d > 0. \tag{A.33}$$

This inequality gives us a theoretical justification of the plug-in method.

Maximum Likelihood Method. Let $f_\theta(x)$ be the probability density (or the probability mass function) of X that is known up to the parameter of interest $\theta \in S$. The *maximum likelihood estimate* is a statistic that maximizes the *likelihood function* (joint density at the point X^n) $f_\theta(X^n)$ over $\theta \in S$.

Example A.21 Find the maximum likelihood estimate of the mean μ of a normal random variable based on its n iid realizations.
Solution: The likelihood function is equal to

$$\begin{aligned} f_\mu(X^n) &= (2\pi\sigma^2)^{-n/2} \exp\Big\{-\sum_{l=1}^{n} (X_l - \mu)^2/2\sigma^2\Big\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big\{-\Big[\sum_{l=1}^{n} X_i^2 - 2\mu n\bar{X} + n\mu^2\Big]/2\sigma^2\Big\}, \end{aligned} \tag{A.34}$$

and it is easy to see that it is maximized by $\hat{\mu} = \bar{X}$ because $-2\mu n\bar{X} + n\mu^2 = n(\mu - \bar{X})^2 - n\bar{X}^2$. Thus, for a normal random variable the sample mean is the maximum likelihood estimate. Note that here the maximum likelihood estimate coincides with the plug-in estimate.

Example A.22 In Example A.20 find a maximum likelihood estimate.
Solution: We observe $X_1, \dots, X_n$ that are uniformly distributed over an interval $[\theta, \theta+1]$. To find a maximum likelihood estimate, we should find a convenient expression for the joint density as a function of θ. Note that the likelihood function is equal to 1 if $\theta \le X_{(1)} \le X_{(n)} \le 1 + \theta$ and it is equal to 0 otherwise (here the crucial point is that the likelihood function is explicitly written as a function of θ). Thus any $\hat{\theta}$ such that $X_{(n)} - 1 \le \hat{\theta} \le X_{(1)}$ is a maximum likelihood estimate. This is an example where the maximum likelihood estimate is not uniquely defined. In such

cases we choose any maximum likelihood estimate; in particular, we can set $\hat{\theta} = X_{(1)}$ or $\hat{\theta} = X_{(n)} - 1$.

Example A.23 Consider the following *mixture* of a normal density with parameters (μ, σ^2) and a standard normal density,

$$f_{\mu,\sigma}(x) = \frac{1}{2}(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2} + \frac{1}{2}(2\pi)^{-1/2} e^{-x^2/2}.$$

Here $\theta = (\mu, \sigma^2)$, and one observes n iid realizations of a random variable with the mixture density. Suppose that $\sigma^2 > 0$. Show that the maximum likelihood method fails to estimate θ.

Solution: For iid observations it is always more convenient to deal with the *log-likelihood* function $L_\theta(X^n) = \ln(f_\theta(X^n))$ rather than the likelihood function. Set $\hat{\mu}_0 = X_1$. Then for any given constant C there exists a sufficiently small σ_0^2 such that the log-likelihood will be larger than C. Indeed,

$$\begin{aligned} L_{(\hat{\mu}_0, \sigma_0^2)}(X^n) &> \ln\left(1/2(2\pi\sigma_0^2)^{1/2}\right) + \sum_{l=2}^{n} \ln\left((1/2(2\pi)^{1/2}) e^{-X_i^2/2}\right) \\ &= -\ln(\sigma_0) - \sum_{l=2}^{n} (X_i^2/2) - n \ln\left(2(2\pi)^{1/2}\right) > C. \end{aligned}$$

Thus the maximum likelihood method is in trouble here because the likelihood may be as large as desired when σ_0 decreases. This example shows the limits of the maximum likelihood method.

Unbiased Estimation. An estimate $\hat{\theta}$ is called *unbiased* if $E_\theta\{\hat{\theta}\} = \theta$.

Example A.24 Let us observe n iid realizations of a normal random variable with known mean μ and unknown variance σ^2. Consider the following two estimates of the variance: $\hat{\sigma}_1^2 = n^{-1} \sum_{l=1}^{n} (X_l - \mu)^2$ and $\tilde{\sigma}_2^2 = n^{-1} \sum_{l=1}^{n} (X_l - \bar{X})^2$. Which estimate is unbiased?

Solution: Write for the former estimate $E_\sigma\{\hat{\sigma}_1^2\} = E_\sigma\{(X - \mu)^2\} = \sigma^2$. Thus this estimate is unbiased. Note that $\hat{\sigma}_1^2$ is the sample mean, plug-in estimate, and maximum likelihood estimate simultaneously. The second estimate may be written as $\tilde{\sigma}_2^2 = \hat{\sigma}_1^2 - (\bar{X} - \mu)^2$, and therefore it is biased. Curiously, the bias is always negative. Note that the estimate does not depend on μ, and it is not difficult to modify the estimate and make it unbiased. Namely, the estimate

$$\hat{\sigma}^2 := (n/(n-1))\tilde{\sigma}_2^2 = (n-1)^{-1} \sum_{l=1}^{n} (X_l - \bar{X})^2 \tag{A.35}$$

is unbiased. It is customarily used when the mean is unknown.

Let us rewrite the mean squared error as

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}, \theta) &= E_\theta\{(\hat{\theta} - E\{\hat{\theta}\})^2\} + (E_\theta\{\hat{\theta}\} - \theta)^2 \\ &=: \mathrm{Var}_\theta(\hat{\theta}) + \mathrm{SBIAS}_\theta(\hat{\theta}). \end{aligned} \tag{A.36}$$

The last line is the expansion of the mean squared error into the sum of the variance and the *squared bias* terms. This expansion explains why an unbiased estimation is appealing. On the other hand, for some settings no unbiased estimate exists (see Exercise A.13). Also, even if an unbiased estimate exists, it may be worthwhile to consider a biased one. For instance, consider the case where the unknown mean of a normal random variable is estimated. In this case the sample mean $\bar{X}$ is an unbiased estimate. However, if it is given that the mean is nonnegative (that is, $\theta \in S = [0, \infty)$), then the biased estimate $\tilde{\theta} = \max(0, \bar{X})$ will always be better (has smaller error) than $\bar{X}$. In other words, for this setting the estimate $\bar{X}$ is *inadmissible* because it is *dominated* by $\tilde{\theta}$.

So far we have considered methods based on ideas other than direct minimization of the mean squared error. Let us consider several approaches based on direct minimization of the mean squared error or its functionals. We begin with one curious approach.

Example A.25 Let us observe n iid realizations of a random variable with unknown mean θ and known variance σ^2. As we know, here the sample mean $\bar{X}$ is a natural estimate of θ. Explore the possibility to decrease the mean squared error by a *linear* estimate $\lambda\bar{X}$, $0 \le \lambda \le 1$, where λ (a so-called shrinkage coefficient because it shrinks the sample mean $\bar{X}$ towards origin) does not depend on the data (it is not a statistic) but may depend on both θ and σ^2.

Solution: Consider the mean squared error of the linear estimate $\lambda\bar{X}$,

$$\begin{aligned} &E_\theta\{(\lambda\bar{X} - \theta)^2\} = \lambda^2 E_\theta\{\bar{X}^2\} - 2\lambda\theta^2 + \theta^2 \\ &= E_\theta\{\bar{X}^2\}(\lambda - \theta^2/E_\theta\{\bar{X}^2\})^2 + \theta^2(E_\theta\{\bar{X}^2\} - \theta^2)/E_\theta\{\bar{X}^2\}. \end{aligned}$$

Then, the equality $E_\theta\{\bar{X}^2\} = \sigma^2 n^{-1} + \theta^2$ allows us to write

$$\begin{aligned} &E_\theta\{(\lambda\bar{X} - \theta)^2\} \\ &= (\sigma^2 n^{-1} + \theta^2)[\lambda - \theta^2/(\sigma^2 n^{-1} + \theta^2)]^2 + \theta^2\sigma^2 n^{-1}/(\sigma^2 n^{-1} + \theta^2). \quad \text{(A.37)} \end{aligned}$$

Thus, we conclude that the optimal shrinking coefficient λ^* (that minimizes the mean squared error) is defined by

$$\lambda^* = \frac{\theta^2}{\sigma^2 n^{-1} + \theta^2}. \quad \text{(A.38)}$$

Also, we get the following *lower bound* for the risk,

$$E_\theta\{(\lambda\bar{X} - \theta)^2\} \ge \min_\lambda E_\theta\{(\lambda\bar{X} - \theta)^2\} = \lambda^*\sigma^2 n^{-1}. \quad \text{(A.39)}$$

Of course, λ^* depends on θ, which is unknown. However, several conclusions may be made. Firstly, we see that shrinking may lead to a decrease in the mean squared error. Secondly, the right side of (A.39) gives us a lower bound for the mean squared error over all possible linear estimates. Finally,

one can try to estimate λ^* by a plug-in estimate. Interestingly, all these ideas play a central role in data-driven nonparametric curve estimation.

Bayesian Approach. A Bayes estimate is an estimate of θ that minimizes the averaged mean squared error (or so-called Bayes error), $\int_S E_\theta\{(\hat\theta-\theta)^2\}dG(\theta)$, where G is the *prior* cumulative distribution function with the domain S. As a result, both the Bayes estimate and the Bayes error do not depend on an underlying parameter, but they do depend on the prior distribution. If we let Θ denote a random variable with the prior distribution function G, then the Bayes estimate is calculated by the formula

$$\hat\theta(G)=E\{\Theta|X^n\}. \tag{A.40}$$

Example A.26 Prove that (A.40) is the Bayes estimate.
Solution: First, let us show that if X is a random variable with finite variance, then for any constant c the following inequality holds,

$$E\{(X-E\{X\})^2\}\le E\{(X-c)^2\}, \tag{A.41}$$

with equality if and only if $c=E\{X\}$. Indeed, $E\{(X-c)^2\}=E\{(X-E\{X\}+E\{X\}-c)^2\}=E\{(X-E\{X\})^2\}+(E\{X\}-c)^2$. This implies (A.41). Then according to (A.14) we may write for any estimate $\tilde\theta$ based on observations X^n that $E\{(\tilde\theta-\Theta)^2\}=E\{E\{(\tilde\theta-\Theta)^2|X^n\}\}$. Note that while considering the conditional expectation $E\{(\tilde\theta-\Theta)^2|X^n\}$ we may assume that $\tilde\theta$ is a constant. This together with (A.41) implies the result. In other words,

$$\min_{\tilde\theta} E\{(\tilde\theta-\Theta)^2\}=E\{(E\{\Theta|X^n\}-\Theta)^2\}, \tag{A.42}$$

where the minimum is taken over all possible estimators $\tilde\theta$.

Example A.27 Let $X_1,\dots,X_n$ be iid realizations of a normal $N(\theta,\sigma^2)$ random variable X, and let σ^2 be given. Find a Bayes estimate of θ for the normal $N(\mu,b^2)$ prior distribution.
Solution: The joint density of Θ and X^n is

$$f^{X^n,\Theta}(x^n,\theta)=\frac{1}{(2\pi\sigma^2)^{n/2}(2\pi b^2)^{1/2}}\exp\left\{-\frac{\sum_{l=1}^n(x_l-\theta)^2}{2\sigma^2}-\frac{(\theta-\mu)^2}{2b^2}\right\}.$$

Recall that to obtain the distribution of Θ given X^n, the joint density is divided by the marginal density $f^{X^n}(x^n)$. Write

$$f^{\Theta|X^n}(\theta|x^n)=\psi(x^n)\exp\left\{-\frac{[\theta-(b^2\bar x+\sigma^2n^{-1}\mu)/(b^2+\sigma^2n^{-1})]^2}{2[n^{-1}\sigma^2b^2/(b^2+\sigma^2n^{-1})]}\right\}.$$

Here $\bar x=n^{-1}\sum_{l=1}^n x_l$ and $\psi(x^n):=\psi(x^n,b,\sigma^2,n)$ is a function not involving θ. The main step is to look at the *posterior* density $f^{\Theta|X^n}(\theta|x^n)$ and

realize that as a function in θ it is again a normal density with mean

$$E\{\Theta|X^n = x^n\} = \frac{b^2}{b^2 + \sigma^2 n^{-1}}\,\bar{x} + \frac{\sigma^2 n^{-1}}{b^2 + \sigma^2 n^{-1}}\,\mu \tag{A.43}$$

and variance

$$\mathrm{Var}(\Theta|X^n = x^n) = \frac{b^2\sigma^2 n^{-1}}{b^2 + \sigma^2 n^{-1}}. \tag{A.44}$$

According to (A.40), the estimate (A.43) is the Bayes estimate. Also, (A.44) gives us the Bayes error. Note that the Bayes estimate becomes essentially the estimator $\bar{X}$ for large n and it is close to the prior mean μ for large $\sigma^2 n^{-1}$. This outcome is intuitively reasonable.

Minimax Approach. While the Bayesian approach is based on averaging the mean squared error, the goal of the minimax approach is to select the best possible estimate for worst-case scenario of a parameter in the set S, that is, an estimate $\tilde{\theta}$ is called a *minimax* estimate if

$$\sup_{\theta\in S} E\{(\tilde{\theta} - \theta)^2\} = \inf_{\hat{\theta}} \sup_{\theta\in S} E\{(\hat{\theta} - \theta)^2\}. \tag{A.45}$$

Recall that the *supremum* $\sup_{x\in D}\psi(x)$ is the smallest number a such that $a \geq \psi(x)$ for all $x \in D$. The *infimum* $\inf_{x\in D}\psi(x)$ is the largest number b such that $b \leq \psi(x)$ for all $x \in D$. We use these notions instead of maximum and minimum because the last two may not exist in some settings. For instance, $\max_{x\in(0,1)} x^2$ and $\min_{x\in(0,1)} x^2$ do not exist, whereas $\sup_{x\in(0,1)} x^2 = 1$ and $\inf_{x\in(0,1)} x^2 = 0$.

The minimax approach is more conservative than the Bayesian approach, but at least formally it does not depend on a prior distribution. However, as we shall see from the following assertion, a customarily used method to find a minimax estimate is based on using a Bayesian approach.

Proposition A.1 *If a Bayes estimate has a constant mean squared error (that does not depend on the estimated parameter), then this estimate is minimax. Moreover, if there is a sequence of Bayes estimates whose mean squared errors converge to a constant, then the limit of the Bayes estimates is the minimax estimate.*

Example A.28 For the setting of Example A.27, find a minimax estimate.
Solution: Due to Proposition A.1, we are to find a prior distribution such that the Bayes risk is constant or approximates a constant. If we consider the prior distributions of Example A.27 and then let $b \to \infty$, then the Bayes estimate becomes the familiar sample mean $\bar{X}$ with the risk equal to σ^2/n. This risk is constant (it does not depend on θ), and therefore the sample mean is the minimax estimate of the mean of a normal random variable.

Confidence Interval Estimation. Consider the following problem. Let X be normal $N(\theta, \sigma^2)$; we would like to find an interval (based on X) that "covers" the unknown mean θ with probability at least $1 - \alpha$, $0 \leq \alpha \leq 1$

($1-\alpha$ is referred to as the *confidence coefficient*). Consider the interval

$$(X - z_{\alpha/2}\sigma, X + z_{\alpha/2}\sigma), \tag{A.46}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution (i.e., it is the solution to the equation $P(Z > z_{\alpha/2}) = \alpha/2$, where Z is a standard normal random variable). First, we see that this interval covers θ with the probability $1-\alpha$ because $P_\theta(X - z_{\alpha/2}\sigma < \theta < X + z_{\alpha/2}\sigma) = P_\theta(-z_{\alpha/2} < (X-\theta)/\sigma < z_{\alpha/2}) = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$. Second, it is easy to see from the bell shape of normal density that any interval $(X - z_{\alpha/2}\sigma + c, X + z_{\alpha/2}\sigma + c)$ with $c \neq 0$ covers the mean with probability less than $1-\alpha$. This makes using (A.46) as a $1-\alpha$ confidence interval very appealing.

This simple setting together with the central limit theorem allows us to suggest confidence intervals for more complicated models. Suppose that n iid realizations $X_1, \ldots, X_n$ of a random variable X with unknown mean θ and a finite variance are observed. To find a $1-\alpha$ confidence interval, consider the sample mean $\bar{X}$. By the central limit theorem, for large n the distribution of $\bar{X}$ is approximately normal with mean θ and variance $\sigma_n^2 := \sigma^2/n$. Thus, we may use formula (A.46) and plug-in $\bar{X}$ in place of X and σ_n in place of σ. This yields the $1-\alpha$ confidence interval:

$$(\bar{X} - z_{\alpha/2}\sigma_n, \bar{X} + z_{\alpha/2}\sigma_n). \tag{A.47}$$

If σ is unknown then the estimate (A.35) may be used.

Testing Statistical Hypotheses (Neyman–Pearson Approach). We shall discuss the concepts of this approach via considering the following particular problem. Let a sample $X_1, \ldots, X_n$ of n iid realizations $X_1, \ldots, X_n$, from a normal distribution with unknown mean θ^* and known variance σ^2, be given. Then the problem is to decide whether the mean is equal to a given value θ_0 or not. Thus, the possible probability distributions of the observations are grouped into two aggregates, one of which is called the *null hypothesis* and is denoted by H_0, and the other of which is called the *alternative hypothesis* and is denoted by H_a. In short, we may write that we would like to test H_0: $\theta^* = \theta_0$ versus H_a: $\theta^* \neq \theta_0$.

The particular hypothesis H_0 is called *simple* because the null hypothesis completely specifies the probability distribution; the alternative one is called *composite* because it does not specify the distribution.

According to the Neyman–Pearson paradigm, a decision to reject H_0 in favor of H_a is made only if observations belong to the *rejection (critical) region* R, which completely describes the hypothesis test. The complement R^c of the rejection region is called the *acceptance region*. Then two types of errors may occur:

1. H_0 may be rejected when it is true. Such an error is called a *type I error* (first type error), and its probability is $e_1 = P(X^n \in R | \theta^* = \theta_0)$. The rejection region is chosen in such a way that $e_1 = \alpha$, where the parameter α is preassigned and is called the *significance level* of the test. The level of

significance is customarily chosen between 0 and 0.25.
2. H_0 may be accepted when it is false. Such an error is called a *type II error* (second type error). For our particular example this error is defined for any parameter $\theta \neq \theta_0$, and its probability is $e_2(\theta) = P(X^n \in R^c | \theta^* = \theta)$.

One more useful notion is the *power* of the test defined as $\beta(\theta) := P(X^n \in R | \theta^* = \theta)$ for $\theta \neq \theta_0$. In words, this is the probability that H_0 is rejected when it is false and the underlying mean is equal to θ. Clearly, $e_2(\theta) = 1 - \beta(\theta)$, so the larger the power of a test, the smaller its second type error.

Because the first type error is fixed in advance (it is equal to the level of significance α), the optimal test should maximize the power. Also note that formally we may write $\beta(\theta_0) = e_1$, so the power as a function in $\theta \in (-\infty, \infty)$ is a convenient tool to describe the probabilities of all considered errors.

Now, when we know the setting and the terminology, let us suggest a reasonable test (rejection region). Because the problem is about the mean of iid normal random variables, one can estimate the mean by a sample mean estimate $\bar{X} = n^{-1} \sum_{l=1}^{n} X_l$ and then reject H_0 if the sample mean is far from θ_0. In such a case $\bar{X}$ is called a *test statistic*, and the rejection region is

$$R := \{(X_1, \dots, X_n) : \ |\bar{X} - \theta_0| \geq c\},$$

where c should be such that $e_1 = \alpha$. To fund such c we write

$$e_1 = P(\,(X_1, \dots, X_n) \in R | \theta^* = \theta_0) = P(|\bar{X} - \theta_0| \geq c \,|\theta^* = \theta_0) = \alpha.$$

Under the null hypothesis, $\bar{X}$ is normally distributed with mean θ_0 and variance $\sigma_n^2 = \sigma^2/n$. Thus, $c = z_{\alpha/2}\sigma_n$ gives the solution to the above equation. Recall that $z_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. Thus we obtain that the rejection region is

$$R := \{(X_1, \dots, X_n) : \ |\bar{X} - \theta_0| \geq z_{\alpha/2}\sigma_n\}. \tag{A.48}$$

Note that the level of significance α is the core ingredient of the Neyman–Pearson paradigm; however, its choice is typically subjective. Thus, in many applications it makes more sense to report a statistic that is called the *p-value* (observed level of significance). For a chosen rejection region and a given data set, the p-value is the *smallest value of α for which the null hypothesis will be rejected*. For the test (A.48) it is calculated by the formula

$$p\text{-value} = P\big(|Z| > |\bar{X} - \theta_0|/\sigma_n \,\big|\theta^* = \theta_0\big)\ , \tag{A.49}$$

where Z is a standard normal variable independent of $\bar{X}$.

Let us check that (A.49) is indeed the observed level of significance. Assume that $\alpha \geq \hat{\gamma}$, where $\hat{\gamma}$ is the right part of (A.49). Then

$$z_{\alpha/2} \leq z_{\hat{\gamma}/2} = |\bar{X} - \theta_0|/\sigma_n.$$

This implies $|\bar{X} - \theta_0| \geq z_{\alpha/2}\sigma_n$, which in turn implies the rejection of H_0 according to (A.48). Conversely, if $\alpha < \hat{\gamma}$, then the null hypothesis

is accepted. This completes the proof. As we see, the p-value completely describes a data set for the Neyman–Pearson paradigm.

Finally, according to (A.48) the acceptance region of the test is

$$R^c = \{(X_1, \dots, X_n) : \ |\bar{X} - \theta_0| < z_{\alpha/2}\sigma_n\}. \tag{A.50}$$

As we see, the confidence interval (A.47) is just the inverted acceptance region (A.50). Indeed, according to (A.50), we accept H_0: $\theta^* = \theta_0$ if $\theta_0 \in (\bar{X} - z_{\alpha/2}\sigma_n, \bar{X} + z_{\alpha/2}\sigma_n)$, and according to (A.47) this means that we accept H_0 if θ_0 belongs to the $1 - \alpha$ confidence interval. In other words, the $1 - \alpha$ confidence interval consists precisely of all values of θ_0 for which the null hypothesis is accepted with the level of significance α.

What we have seen is the well-known method of finding confidence interval estimates via inverting hypothesis tests, and vice versa. Also, we have seen that a good estimator may lead to finding an attractive test.

Exercises

A.1 Prove that $P(E \cap F) \geq P(E) + P(F) - 1$.

A.2 Show that if events $E_1, E_2, \dots, E_n$ are independent, then

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = 1 - \prod_{l=1}^{n}(1 - P(E_l)). \tag{A.51}$$

A.3 Prove that if $P(E_i|E_1 \cap \cdots \cap E_{i-1}) > 0$, $i = 1, 2, \dots, n$, then

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1) \cdots P(E_n|E_1 \cap \cdots \cap E_{n-1}).$$

A.4 Express $P(X \geq a)$ via the cumulative distribution function of X.

A.5 Let F be the cumulative distribution function of X. Find the cumulative distribution function of $Y = \alpha X + \beta$, where $\alpha > 0$ and β are constants.

A.6 The joint distribution of two discrete random variables X and Y is $P(X = 0, Y = 0) = .2$, $P(X = 0, Y = 1) = .3$, $P(X = 1, Y = 0) = .3$, $P(X = 1, Y = 1) = .2$. Are these random variables independent?

A.7 The joint probability density function of X and Y is given by $f(x, y) = \exp(-x - y)$, $0 \leq x < \infty$, $0 \leq y < \infty$. Find $P(X < Y)$.

A.8 Show that (i) $E\{(X - a)^2\}$ is minimized at $a = E\{X\}$; (ii) $E\{|X - a|\}$ is minimized at a equal to the median of X.

A.9 Let a, b, c, and d be constants. Show that $\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X, Y)$.

A.10 Show that $\mathrm{Cov}(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, Y_j)$.

A.11 Show that (A.10)–(A.12) hold.

A.12 Let $X_{(1)}, \dots, X_{(n)}$ be ordered iid realizations of a continuous random variable X supported on a finite interval $[a, b]$ and having a continuous density f on $[a, b]$. Check that $P(a \leq X \leq X_{(1)}) = P(X_{(n)} \leq X \leq b) = 1/(n+1)$. Hint: The event $X \leq X_{(1)}$ occurs if and only if all n realizations

are larger than X. Using independence of X and the realizations, get

$$P(a \le X \le X_{(1)}) = \int_a^b P(x \le X_{(1)} | X = x) f(x) dx$$

$$= \int_a^b \prod_{l=1}^{n} P(x \le X_l) f(x) dx = \int_a^b [P(x \le X)]^n f(x) dx$$

Note that $P(x \le X) = 1 - F(x)$ where F is the cumulative distribution function of X. Also, we have $f(x) = dF(x)/dx$, $x \in [a, b]$. Thus, $\int_a^b [P(x \le X)]^n f(x) dx = \int_a^b [1 - F(x)]^n dF(x) = \int_0^1 z^n dz$ where in the last equality we used the substitution $z = 1 - F(x)$, and that $F(a) = 0$ and $F(b) = 1$.

A.13 Let X be distributed according to the binomial distribution $B(p, n)$, $0 < p < 1$. Suggest an unbiased estimate of p and find its mean squared error. Also, show that no unbiased estimate of $1/p$ exists. Hint: Recall that $\sum_{l=0}^{k} a_l x^l = 0$ for all $x \in (a, b)$, $a < b$, if and only if all $a_l = 0$.

A.14 Let $X_1, X_2, \ldots, X_n$ be iid normal $N(\theta, \sigma^2)$. Suggest an unbiased estimate of θ^2. Also, find the mean squared error of this estimate.

A.15 Let X be a random variable, and we are interested in estimating the parameter $\theta = P(a \le X \le b)$. Suggest an unbiased estimate based on n iid realizations of X and find its mean squared error.

A.16 Let X and Y be two random variables. Suggest an unbiased estimate of the parameter $\theta = P(X \le Y)$ and then find its mean squared error. Also, can the independence of these random variables help to solve the problem of estimating θ?

A.17 Let X be a binomial random variable with $n = 100$ trials and the probability p of "success." Suggest a $1 - \alpha$ confidence interval for the p.

A.18 Let X be normal with mean θ^* and variance σ^2. Test the null hypothesis H_0: $\theta^* = \theta_0$ versus H_a: $\theta^* > \theta_0$ at the level of significance α. For the suggested test find the first type error, the second type error for $\theta > \theta_0$, the power function, the p-value (observed level of significance).

Appendix B. Software

The software may be used in the S–PLUS environment under UNIX or under Windows. If you do not have a free access to the S–PLUS then the following information may be useful. The S–PLUS 4.5 Student Edition for PC is sold by Duxbury Press (Web: www.duxbury.com) for the cost of a regular textbook (the wavelets package is extra).

Below we discuss how to install and use the software when the S–PLUS 3.x under UNIX is used (this is the most "complicated" scenario). Consult the file **news.INSTALLATION** at the web site *http://www.math.unm.edu/∼efrom/book1* about installation for other S–PLUS versions.

By downloading the software, the user agrees to consider it as a "black-box" and employ it for educational purposes only.

Setup Information.

1. Prior to downloading the software, you need to create a separate directory, say SE, and make it your working directory. To do this you type (after each line you press Return):

   % mkdir SE

   % cd SE

2. Create a subdirectory of SE called .Data by

   % mkdir .Data

This subdirectory is for use by S–PLUS itself and hence has a UNIX 'dot name' to be hidden from casual inspection. Then type

   % cd .Data

Now you are in the subdirectory where the software will be downloaded.

3. The software is available over the World Wide Web. Point your browser at

http://www.math.unm.edu/∼efrom/book1

The file **book1.tar.gz** contains the compressed version of all the S–PLUS functions. To download this file in Netscape press the right mouse button over the link to the file and choose "save link as" and tell it to save it to your SE/.Data S–PLUS subdirectory. This ends the World Wide Web part of the setup.
4. Return to your SE/.Data subdirectory. To decompress the compressed file, type

% gunzip book1.tar.gz

This should extract a file called **book1.tar**. Then type

% tar -xf book1.tar

This will extract all the individual S–PLUS function files. The software is installed.

Tutorial.
1. To begin S–PLUS session you type

% cd

(after this command you are in the main directory)

% cd SE

(after this command you are in your S–PLUS working directory)

% Splus

Then, after a couple of seconds, you will see the sign > which is the S–PLUS prompt. You are ready to go.
If you would like to use the wavelets package, type

> *module(wavelets)*

To look at graphics you must create a special window by typing

> *motif*()

To interrupt a program hold down the key marked **Control** and hit **c**. This will interrupt the current operation, back out gracefully, and return to the prompt. Another way to do this is again to hold down **Control** and hit **xc**.
2. To repeat a Figure *j.k* (the *k*th figure in the *j*th chapter), whose caption has square brackets (only these figures may be repeated!), type

> *chj(fig=k)*

You will see in the Graphics window a diagram similar to Figure *j.k* only with may be different simulated data. Actually, it is enough to type *chj(f=k)*. For instance, by typing

> *ch4(f=1)*

you repeat Figure 4.1 with the same default arguments shown in the square brackets. If you would like to repeat this particular figure with different arguments from those shown in the square brackets, for instance, you would like to change in Figure 4.1 the sample size n from 50 to 100 and the standard deviation *sigma* from 1 to 2, then type

> ch4(f=1, n=100, sigma=2)

You will see scatter plots with one hundred points overlaid by linear regression lines. Note that an argument may be a numeric value, a string, a vector of numeric values, or a vector of strings. For instance, it may be set equal to 10, *"box"*, *c*(10, 50, 100), or *c*(*"box"*, *"gaussian"*), respectively.

To make a hard copy of a figure shown in the Graphics window you need to drag the mouse in such a way that the arrow is positioned at *Graph* in the Graphics window. Then you push the left button, and while holding down on the button move the cursor down to the option marked *print* and release the button.

Also note that the caption to Figure 2.3 explains how to create a custom-made corner function (or generate a sample according to the density).

To finish the S–PLUS session you type

> q()

References

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statist. Mathematics* **6**, 127–132.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory.* (B.N. Petrov and F. Csaki, eds.). Budapest:Akademio Kiado, 267–281.

Barlow, R.E., Bartholomew, D.J., Bremner, J.H., and Brank, H.D. (1972). *Statistical Inference Under Order Restrictions.* New York: John Wiley & Sons.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-413.

Bary, N.K. (1964). *A Treatise on Trigonometric Series.* Oxford: Pergamon Press.

Baxter, M. and Rennie, A. (1996). *Financial Calculus. An Introduction to Derivative Pricing.* Cambridge: University Press.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). *The New S Language.* Pacific Grove: Wadsworth & Brooks/Cole.

Bentkus, R. (1985). On asymptotic of minimax mean squared risk of statistical estimates of spectral density in L_2. *Lithuanian Mathematical Journal* **25**, 23–42.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Bellman, R.E. (1961). *Adaptive Control Processes.* Princeton: Princeton University Press.

Beran, J. (1994). *Statistics for Long-Memory Processes.* New York: Chapman & Hall.

Bickel, P.J., Klaassen, A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semi-parametric Models.* Baltimore: Johns Hopkins University Press.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. *Festschrift for Lucien Le Cam.* (D. Pollard, ed.). New York: Springer, 55-87.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Monterey: Wadsworth and Brooks/Cole.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods.* Second Edition. New York: Springer-Verlag.

Brown, L.D. and Low, M.L. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384–2398.

Brown, L.D. and Low, M.L. (1996b). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524–2535.

Brown, L.D. and Zhang, C.-H. (1998) Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist.* **26**, 279–287.

Butzer, P.L. and Nessel, R.J. (1971). *Fourier Analysis and Approximations.* New York: Academic Press.

Cai, T.T. (1999) Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, to be published.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* New York: Chapman & Hall.

Carroll, R.J., Ruppert, D., and Stefanskii, L.A. (1995). *Measurement Error in Nonlinear Models.* New York: Chapman&Hall.

Casella, G. and Berger, R. (1990). *Statistical Inference.* New York: Brooks/Cole.

Chentsov, N.N. (1962). Estimation of unknown distribution density from observations. *Soviet Math. Dokl.* **3**, 1559–1562.

Chentsov, N.N. (1980). *Statistical Decision Rules and Optimum Inference.* New York: Springer-Verlag.

Clarke, B. and Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory.* **36**, 453–471.

Cleveland, W.S. (1993). *Visualizing Data.* Summit: Hobart Press.

Collett, D. (1994). *Modeling Survival Data in Medical Research.* London: Chapman & Hall.

D'Agostino, R.B. and Stephens, M.A. (eds.) (1986). *Goodness-Of-Fit Techniques.* New York: Marcel Dekker.

Debnath, L. and Mikusinski, P. (1990). *Introduction to Hilbert Spaces with Applications.* New York: Academic Press.

DeVore, R.A. and Lorentz, G.G. (1993). *Constructive Approximation.* New York: Springer-Verlag.

DeVore, R.A. and Temlyakov, V.N. (1995). Nonlinear approximation by trigonometric sums. *Journal of Fourier Analysis and Applications* **2**, 29–48.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L_1 View.* New York: Wiley.

Devroye, L. (1987). *A Course in Density Estimation.* Boston: Birkhäuser.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Oxford: Oxford Univ. Press.

Donoho, D.L., Liu, R.C., and MacGibbon, B. (1990). Minimax risk over hyperrectangles. *Ann. Statist.* **18**, 1416–1437.

Donoho, D.L. and Liu, R.C. (1991). Geometrizing rate of convergence. III. *Ann. Statist.* **19**, 668–701.

Donoho, D.L. (1994) Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probab. Theory Related Fields* **99**, 145–170.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrica* **81**, 425–455.

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

Donoho, D. (1997). CART and best-ortho-basis: a connection. *Ann. Statist.* **25**, 1870-1911.

Doukhan, P. (1994). *Mixing: Properties and Examples*. New York: Springer-Verlag.

Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. New York: Wiley.

Dym, H. and McKean, H.P. (1972). *Fourier Series and Integrals*. London: Academic Press.

Efromovich, S. (1980a). Information contained in a sequence of observations. *Problems Inform. Transmission* **15**, 178–189.

Efromovich, S. (1980b). On sequential estimation under conditions of local asymptotic normality. *Theory Probab. Applications* **25**, 27–40.

Efromovich, S. and Pinsker M.S. (1981). Estimation of a square integrable spectral density for a time series. *Problems Inform. Transmission* **17**, 50–68.

Efromovich, S. and Pinsker M.S. (1982). Estimation of a square integrable probability density of a random variable. *Problems Inform. Transmission* **18**, 19–38.

Efromovich, S. (1984). Estimation of a spectral density of a Gaussian time series in the presence of additive noise. *Problems Inform. Transmission* **20**, 183–195.

Efromovich, S. and Pinsker M.S. (1984). An adaptive algorithm of nonparametric filtering. *Automation and Remote Control* **11**, 58–65.

Efromovich, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory Probab. Applications* **30**, 557–568.

Efromovich, S. (1986). Adaptive algorithm of nonparametric regression. *Proc. of Second IFAC symposium on Stochastic Control*. Vilnuis: Science, 112–114.

Efromovich, S. and Pinsker M.S. (1986). Adaptive algorithm of minimax nonparametric estimating spectral density. *Problems Inform. Transmission* **22**, 62–76.

Efromovich, S. (1989). On sequential nonparametric estimation of a density. *Theory Probab. Applications* **34**, 228–239.

Efromovich, S. and Pinsker, M.S. (1989). Detecting a signal with an assigned risk. *Automation and Remote Control* **10**, 1383–1390.

Efromovich, S. (1992). On orthogonal series estimators for random design nonparametric regression. *Computing Science and Statistics* **24**, 375–379.

Efromovich, S. (1994a). On adaptive estimation of nonlinear functionals. *Statistics and Probability Letters* **19**, 57–63.

Efromovich, S. (1994b). On nonparametric curve estimation: multivariate case, sharp-optimality, adaptation, efficiency. *CORE Discussion Papers* **9418** 1–35.

Efromovich, S. (1994c) Nonparametric curve estimation from indirect observations. *Computing Science and Statistics* **26**, 196–200.

Efromovich, S. and Low, M. (1994). Adaptive estimates of linear functionals. *Probab. Theory Related Fields* **98**, 261–257.

Efromovich, S. (1995a). Thresholding as an adaptive method (discussion). *Journal of Royal Statistic Society, B* **57**, 343.

Efromovich, S. (1995b). On sequential nonparametric estimation with guaranteed precision. *Ann. Statist.* **23**, 1376–1392.

Efromovich, S. (1996a). On nonparametric regression for iid observations in general setting. *Ann. Statist.* **24**, 1126–1144.

Efromovich, S. (1996b). Adaptive orthogonal series density estimation for small samples. *Computational Statistics and Data Analysis*, **22**, 599–617.

Efromovich, S. (1996c) Can adaptive estimators for Fourier series be of interest to wavelets? Technical Report, University of New Mexico.

Efromovich, S. and Low, M. (1996a). On Bickel and Ritov's conjecture about adaptive estimation of some quadratic functionals. *Ann. Statist.* **24**, 682–686.

Efromovich, S. and Low, M. (1996b). Adaptive estimation of a quadratic functional. *Ann. Statist.* **24**, 1106–1125.

Efromovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica* **6**, 925–945.

Efromovich, S. and Samarov, A. (1996). Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statistics and Probability Letters* **28**, 143–145.

Efromovich, S. and Thomas, E. (1996). Application of nonparametric binary regression to evaluate the sensitivity of explosives. *Technometrics* **38**, 50–58.

Efromovich, S. (1997a). Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association* **92**, 526–535.

Efromovich, S. (1997b). Robust and efficient recovery of a signal passed through a filter and then contaminated by non-Gaussian noise. *IEEE Trans. Inform. Theory* **43**, 1184–1191.

Efromovich, S. (1997c). Quasi-linear wavelet estimation involving time series. *Computing Science and Statistics* **29**, 127–131.

Efromovich, S. (1997d). Density estimation under random censorship and order restrictions: from asymptotic to small sample sizes. Technical report, University of New Mexico.

Efromovich, S. and Koltchinskii, V. (1997). Projection estimation for inverse problems with unknown operator. Technical Report, UNM.

Efromovich, S. (1998a). On global and pointwise adaptive estimation. *Bernoulli* **4**, 273–278.

Efromovich, S. (1998b). Data-driven efficient estimation of the spectral density. *Journal of the American Statistical Association* **93**, 762–770.

Efromovich, S. (1998c). Simultaneous sharp estimation of functions and their derivatives. *Ann. Statist.* **26**, 273–278.

Efromovich, S. (1999a). Quasi-linear wavelet estimation. *Journal of the American Statistical Association* **94**, 189–204.

Efromovich, S. (1999b). How to overcome the curse of long-memory errors. *IEEE Trans. Inform. Theory* **45**, to be published.

Efromovich, S. (1999c). On rate and sharp optimal estimation. *Probab. Theory Related Fields* **113**, 415–419.

Efromovich, S. and Ganzburg, M. (1999). Best Fourier approximation and application in efficient blurred signal reconstruction. *Computational Analysis and Applications* **1**, 43–62.

Efromovich, S. and Samarov, A. (1999). Adaptive estimation of the integral of a squared regression function. *Scandinavian Journal of Statistics* (to be published).

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24**, 2431–2461.

Ermakov, M. (1992). Minimax estimation in a deconvolution problem. *Journal Phys. A: Math. Gen.* **25**, 1273–1282.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel and Dekker.

Eubank, R.L. and Speckman, P. (1990). Curve fitting by polynomial–trigonometric regression. *Biometrica* **77**, 815–827.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications—Theory and Methodologies*. New York: Chapman & Hall.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics (London)*, **10**, 422–429.

Fisher, R.A. (1952). *Contributions to Mathematical Statistics*. New York: Wiley.

Fisher, N.I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Gihman, I.I. and Skorohod, A.V. (1974). *The Theory of Stochastic Processes, I.* New York: Springer-Verlag.

Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regresssion. *Mathematical Methods of Statistics* **6**, 135–170.

Golubev, G.K. (1991). LAN in problems of non-parametric estimation of functions and lower bounds for quadratic risks. *Problems Inform. Transmission* **36**, 152–157.

Golubev, G.K. and Nussbaum, M. (1992). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Applications* **37**, 521–529.

Golubev, G.K. and Levit, B.Y. (1996). On the second order minimax estimation of distribution functions. *Mathematical Methods of Statistics* **5**, 1–31.

Golubev, G.K., Levit, B.Y., and Tsybakov, A.B. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2**, 167–181.

Green P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* London: Chapman & Hall.

Grenander, U. (1981). *Abstract Inference.* New York: Wiley.

Hall, P. and Hart, J.D. (1990). Nonparametric regression with long-range dependence. *Stochastic Process. Appl.* **36**, 339–351.

Hall, P., Kerkyacharian, G., and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922–942.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques With Implementation in S.* New York: Springer-Verlag.

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

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-Of-Fit Tests.* New York: Springer.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hengartner, N.W. (1997). Adaptive demixing in Poisson mixture models. *Ann. Statist.* **25**, 917–928.

Huber, P.J. (1981). *Robust Estimation.* New York: Wiley.

Ibragimov, I.A. and Khasminskii, R.Z. (1973). On the information in a sample about a parameter. *Proc. 2nd International Symp. on Information Theory.* Budapest: Publishing House of the Hungarian Academy of Sciences, 295–309.

Ibragimov, I.A. and Khasminskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory.* New York: Springer.

Ibragimov, I.A. and Khasminskii, R.Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Applications* **29**, 1–32.

Ibragimov, I.A. and Khasminskii, R.Z. (1987). Estimation of linear functionals in Gaussian noise. *Theory Probab. Applications* **32**, 30–39.

Ibragimov, I.A. and Khasminskii, R.Z. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18**, 999–1010.

Ingster, Yu.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives, I, II, III. *Mathematical Methods of Statistics* **2**, 85–114, 171–189, 249–268.

Juditsky, A. (1997). Wavelet estimators: adapting to unknown smoothness. *Mathematical Methods of Statistics* **6**, 1–25.

Kiefer, J. (1982). Optimum rates for non-parametric density and regression estimates under order restrictions. *Statistics and Probability: Essays in Honor of C.R. Rao.* (G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh, eds.). Amsterdam: North-Holland, 419–428.

Kolmogorov, A.N. and Fomin, S.V. (1957). *Elements of the Theory of Functions and Functional Analysis.* Rochester: Graylock Press.

Korostelev, A.P. (1993). Exact asymptotical minimax estimate of a nonparametric regression in the uniform norm. *Theory Probab. Applications* **38**, 875–882.

Korostelev, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image Reconstruction.* New York: Springer-Verlag.

Koshevnik, Yu.A. and Levit, B.A. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Applications* **21**, 738–753.

Krylov, A.N. (1955). *Lectures in Approximate Computation.* Moscow: Science (in Russian).

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation.* New York: Springer.

Lepskii, O.V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Prob. Appl.* **35**, 454–466.

Lepskii, O.V. (1992). On problems of adaptive minimax estimation in Gaussian white noise. *Advances in Soviet Mathematics* **12**, 87–106.

Lepskii, O.V., Mammen, E., and Spokoiny, V. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25**, 929–947.

Levit, B. and Samarov, A. (1978). Estimation of spectral functions. *Problems of Inform. Transmission* **14**, 61–66.

Lorentz, G., Golitschek, M., and Makovoz, Y. (1996). *Constructive Approximation. Advanced Problems.* New York: Springer-Verlag.

Mardia, K.V. (1972). *Statistics of Directional Data.*

Mardia, K.V. (1972). *Statistics of Directional Data.* London: Academic Press.

Marron, J.S. and Tsybakov, A.B. (1995). Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association* **90**, 499–507.

Meyer, Y. (1992). *Wavelets and Operators.* Cambridge: Cambridge Univ. Press.

Morgenthaler,S. and Vardi,Y. (1986). Ranked set samples: a nonparametric approach. *J. Econometrics* **32**, 109–125.

Miller, R.G. (1981) *Survival Analysis.* New York: Wiley.

Mallat, S. (1998). *A Wavelet Tour of Signal Precessing.* Boston: Academic Press.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Berlin: Springer-Verlag.

Nadaraya, E.A. (1989). *Nonparametric Estimation of Probability Density and Regression Curves.* (Translated by S. Kotz.) Boston: Kluwer Academic Publishers.

Nemirovskii, A.S. (1999). *Topics in Non-Parametric Statistics.* New York: Springer.

Nemirovskii, A.S., Polyak, B.T., and Tsybakov, A.B. (1985). Rate of convergence of nonparametric estimators of maximum likelihood type. *Problems Inform. Transm.* **21**, 258–272.

Neyman, J. (1937). "Smooth" test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 149–199.

Nikolskii, S.M. (1975). *Approximation of Functions of Several Variables and Embedding Theorems.* New York: Springer-Verlag.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984–997.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**, 2399–2430.

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

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Pawlak, M. (1994). Discontinuity estimation in nonparametric regression via orthogonal series. *Computing Science and Statistics* **26**, 252–256.

Pinsker, M.S. (1972). Information contained in observations and asymptotically sufficient statistics. *Probl. Inform. Transmission* **8**, 45–61.

Pinsker, M.S. (1980). Optimal filtering a square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16**, 52–68.

Polyak, B.T. and Tsybakov, A.B. (1990). Asymptotic optimality of the C_p test for the orthonormal series estimation of regression. *Theory Probab. Applications* **35**, 293–306.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. New York: Academic Press.

Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals Mathematical Statist.* **27**, 832–837.

Ross, S. (1997). *A First Course in Probability*. Fifth Edition. Upper Saddle River: Prentice Hall.

Rubin, H. and Vitale, R.A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8**, 165–170.

Rubinstein, R.Y. (1997). *Simulation and the Monte Carlo Method*. New York: Wiley.

Rudzkis, R. and Radavicius, M. (1993). Locally minimax efficiency of nonparametric estimates of square-integrable densities. *Lithuanian Mathematical Journal* **33**, 56–75.

Samarov, A. (1977). Lower bound for risk of spectral density estimates. *Problems of Inform. Transmission* **13**, 67–72.

Samarov, A. (1992). On the lower bound for the integral error of density function estimates. *Topics in Nonparametric Estimation. Advances in Soviet Mathematics*. Philadelphia: AMS, 1–6.

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

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Englewood Cliffs: Prentice Hall.

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

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970–983.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators *Ann. Statist.* **8**, 1348–1360.

Tarter, M.E. and Lock, M.D. (1993). *Model-Free Curve Estimation.* London: Chapman and Hall.

Temlyakov, V.N. (1993). *Approximation of Periodic Functions.* New York: Nova Science Publishers.

Tompson, J.R. and Tapia, R.A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation.* Philadelphia: SIAM.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Venables, W.N. and Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS.* Second Edition. New York: Springer.

Vidacovic, B. (1999). *Statistical Modeling by Wavelets.* New York: Wiley.

Vitale, R.A. (1973). An asymptotically efficient estimate in time series analysis. *Quarterly of Applied Mathematics* **17**, 421–440.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

Walter, G.G. (1994). *Wavelets and Other Orthogonal Systems with Applications.* London: CRC Press.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman & Hall.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrica* **82**, 385–397.

Watson, G.S. (1969). Density estimation by orthogonal series. *Annals Mathematical Statist.* **40**, 1496–1498.

Zelen, M. (1974). Problems in cell kinetics and early detection of disease. *Reliability and Biometry.* (F. Proschan and R.J. Serfling, eds.). Philadelphia: SIAM, 57-69.

Zhao, L.H. (1993). Frequentist and Bayesian aspects of some nonparametric estimation problems. Ph.D. thesis, Cornell University.

Author Index

Subject Index