

Multivariate Density Estimation

Multivariate Density Estimation

Theory, Practice, and Visualization

DAVID W. SCOTT

Rice University
Houston, Texas



A Wiley-Interscience Publication

JOHN WILEY & SONS, INC.

New York • Chichester • Brisbane • Toronto • Singapore

A NOTE TO THE READER

This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

In recognition of the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

Copyright © 1992 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Scott, David W., 1950-

Multivariate density estimation: theory, practice, and visualization / David W. Scott

p. cm. — (Wiley series in probability and mathematical statistics)

Includes bibliographical references and indexes.

ISBN 0-471-54770-0 (alk. paper)

I. Estimation theory. 2. Multivariate analysis. I. Title.

II. Series.

QA276.8.S28 1992

91-43950

519.5'35—dc20

CIP

**To Jean, Hilary, Elizabeth, Warren,
and my parents, John and Nancy Scott**

Contents

Preface	xi
1 Representation and Geometry of Multivariate Data	1
1.1 Introduction, 1	
1.2 Historical Perspective, 4	
1.3 Graphical Display of Multivariate Data Points, 5	
1.4 Graphical Display of Multivariate Functionals, 16	
1.5 Geometry of Higher Dimensions, 27	
Problems, 31	
2 Nonparametric Estimation Criteria	33
2.1 Estimation of the Cumulative Distribution Function, 34	
2.2 Direct Nonparametric Estimation of the Density, 35	
2.3 Error Criteria for Density Estimates, 37	
2.4 Nonparametric Families of Distributions, 43	
Problems, 45	
3 Histograms: Theory and Practice	47
3.1 Sturges' Rule for Histogram Bin Width Selection, 47	
3.2 The L_2 Theory of Univariate Histograms, 49	
3.3 Practical Data-Based Bin Width Rules, 72	
3.4 L_2 Theory for Multivariate Histograms, 80	
3.5 Modes and Bumps in a Histogram, 86	
3.6 Other Error Criteria: L_1, L_4, L_6, L_8 , and L_∞ , 90	
Problems, 91	

4 Frequency Polygons	95
4.1 Univariate Frequency Polygons, 95	
4.2 Multivariate Frequency Polygons, 106	
4.3 Bin Edge Problems, 109	
Problems, 111	
5 Averaged Shifted Histograms	113
5.1 Construction, 113	
5.2 Asymptotic Properties, 116	
5.3 The Limiting ASH as a Kernel Estimator, 121	
Problems, 124	
6 Kernel Density Estimators	125
6.1 Motivation for Kernel Estimators, 125	
6.2 Theoretical Properties: Univariate Case, 130	
6.3 Theoretical Properties: Multivariate Case, 149	
6.4 Generality of the Kernel Method, 155	
6.5 Cross-Validation, 160	
6.6 Adaptive Smoothing, 181	
Problems, 190	
7 The Curse of Dimensionality and Dimension Reduction	195
7.1 Introduction, 195	
7.2 Curse of Dimensionality, 198	
7.3 Dimension Reduction, 206	
Problems, 216	
8 Nonparametric Regression and Additive Models	219
8.1 Nonparametric Kernel Regression, 219	
8.2 General Linear Nonparametric Estimation, 226	
8.3 Robustness, 232	
8.4 Regression in Several Dimensions, 236	
8.5 Summary, 244	
Problems, 244	

CONTENTS	ix
9 Other Applications	247
9.1 Classification, Discrimination, and Likelihood Ratios, 247	
9.2 Modes and Bump Hunting, 253	
9.3 Specialized Topics, 257	
Problems, 265	
Appendix A Computer Graphics in \Re^3	267
A.1 Bivariate and Trivariate Contouring Display, 267	
A.2 Drawing 3-D Objects on the Computer, 271	
Appendix B Data Sets	273
B.1 United States Economic Variables Data, 273	
B.2 University Data, 274	
B.3 Blood Fat Concentration Data, 275	
B.4 Penny Thickness Data, 276	
B.5 Gas Meter Accuracy Data, 276	
B.6 Old Faithful Data, 278	
B.7 Silica Data, 279	
B.8 LRL Data, 279	
B.9 Buffalo Snowfall Data, 279	
Appendix C Notation	281
References	285
Author Index	301
Subject Index	305

Preface

With the revolution in computing in recent years, access to data of unprecedented complexity has become commonplace. More variables are being measured, and the sheer volume of data is growing. At the same time, advances in the performance of graphical workstations have given new power to the data analyst. With these changes has come an increasing demand for tools that can detect and summarize the multivariate structure in difficult data. Density estimation is now recognized as a tool useful with univariate and bivariate data; my purpose is to demonstrate that it is also a powerful tool in higher dimensions, with particular emphasis on trivariate and quadrivariate data. I have written this book for the reader interested in the theoretical aspects of nonparametric estimation as well as for the reader interested in the application of these methods to multivariate data. It is my hope that the book can serve as an introductory textbook and also as a general reference.

I have chosen to introduce major ideas in the context of the classical histogram, which remains the most widely applied and most intuitive nonparametric estimator. I have found it instructive to develop the links between the histogram and more statistically efficient methods. This approach greatly simplifies the treatment of advanced estimators, as much of the novelty of the theoretical context has been moved to the familiar histogram setting.

The nonparametric world is more complex than its parametric counterpart. I have selected material that is representative of the broad spectrum of theoretical results available, with an eye on the potential user, based on my assessments of usefulness, prevalence, and tutorial value. Theory particularly relevant to application or understanding is covered, but a loose standard of rigor is adopted in order to emphasize the methodological and application topics. Rather than present a cookbook of techniques, I have adopted a hierarchical approach that emphasizes the similarities among the different estimators. I have tried to present new ideas and practical advice, together with numerous examples and problems, with a graphical emphasis.

Visualization is a key aspect of effective multivariate nonparametric analysis, and I have attempted to provide a wide array of graphic illustrations. All of the

figures in this book were composed using S, S-PLUS, Exponent Graphics from IMSL, and Mathematica. The color plates were derived from S-based software. The color graphics with transparency were composed by displaying the S output using the MinneView program developed at the Minnesota Geometry Project and printed on hardware under development by the 3M Corporation. I have not included a great deal of computer code. A collection of software, primarily Fortran-based with interfaces to the S language, is available by electronic mail at scottdw@rice.edu. Comments and other feedback are welcomed.

I would like to thank many colleagues for their generous support over the past 20 years, particularly Jim Thompson, Richard Tapia, and Tony Gorry. I have especially drawn on my collaboration with George Terrell, and I gratefully acknowledge his major contributions and influence in this book. The initial support for the high-dimensional graphics came from Richard Heydorn of NASA. This work has been generously supported by the Office of Naval Research under grant N00014-90-J-1176 as well as the Army Research Office. Allan Wilks collaborated on the creation of many of the color figures while we were visiting the Geometry Project, directed by Al Marden and assisted by Charlie Gunn, at the Minnesota Supercomputer Center.

I have taught much of this material in graduate courses not only at Rice, but also during a summer course in 1985 at Stanford and during an ASA short course in 1986 in Chicago with Bernard Silverman. Previous Rice students Lynette Factor, Donna Nezames, Rod Jee, and Ferdie Wang all made contributions through their theses. I am especially grateful for the able assistance given during the final phases of preparation by Tim Dunne and Keith Baggerly, as well as Steve Sain, Monnie McGee, and Michael Minnotte. Many colleagues have influenced this work, including Edward Wegman, Dan Carr, Grace Wahba, Wolfgang Härdle, Matthew Wand, Simon Sheather, Steve Marron, Peter Hall, Robert Launer, Yasuo Amemiya, Nils Hjort, Linda Davis, Bernhard Flury, Will Gersch, Charles Taylor, Imke Janssen, Steve Boswell, I.J. Good, Iain Johnstone, Ingram Olkin, Jerry Friedman, David Donoho, Leo Breiman, Naomi Altman, Mark Matthews, Tim Hesterberg, Hal Stern, Michael Trosset, Richard Byrd, John Bennett, Heinz-Peter Schmidt, Manny Parzen, and Michael Tarter. Finally, this book could not have been written without the patience and encouragement of my family.

DAVID W. SCOTT

Houston, Texas
February 1992

Multivariate Density Estimation

Multivariate Density Estimation

DAVID W. SCOTT

Copyright © 1992 by John Wiley & Sons, Inc.

C H A P T E R 1

Representation and Geometry of Multivariate Data

A complete analysis of multidimensional data requires the application of an array of statistical tools—parametric, nonparametric, and graphical. Parametric analysis is the most powerful. Nonparametric analysis is the most flexible. And graphical analysis provides the vehicle for discovering the unexpected.

This chapter introduces some graphical tools for visualizing structure in multidimensional data. One set of tools focuses on depicting the data points themselves, while another set of tools relies upon displaying of functions estimated from those points. Visualization and contouring of functions in more than 2 dimensions is introduced. Some mathematical aspects of the geometry of higher dimensions are reviewed. These results have consequences for nonparametric data analysis.

1.1 INTRODUCTION

Classical linear multivariate statistical models rely primarily upon analysis of the covariance matrix. So powerful are these techniques that analysis is almost routine for data sets with hundreds of variables. While the theoretical basis of parametric models lies with the multivariate Normal density, these models are applied in practice to many kinds of data. Parametric studies provide neat inferential summaries and parsimonious representation of the data.

For many problems second-order information is inadequate. Advanced modeling or simple variable transformations may provide a solution. When no simple parametric model is forthcoming, many researchers have opted for fully “unparametric” methods that may be loosely collected under the heading of exploratory data analysis. Such analyses are highly graphical, but in a complex non-Normal setting, a graph may provide a more concise representation than a parametric model, because a parametric model of adequate complexity may involve hundreds of parameters.

There are some significant differences between parametric and nonparametric modeling. The focus on optimality in parametric modeling does not translate well to the nonparametric world. For example, the histogram might be proved to be an inadmissible estimator, but that theoretical fact should not be taken to suggest histograms should not be used. Quite to the contrary, some methods that are theoretically superior are almost never used in practice. The reason is that the ordering of algorithms is not absolute, but is dependent not only on the unknown density but also on the sample size. Thus the histogram is generally superior for small samples regardless of its asymptotic properties. The exploratory school is at the other extreme, rejecting probabilistic models, whose existence provides the framework for defining optimality.

In this book, an intermediate point of view is adopted regarding statistical efficacy. No nonparametric estimate is considered wrong; only different components of the solution are emphasized. Much effort will be devoted to the data-based calibration problem, but nonparametric estimates can be reasonably calibrated in practice without too much difficulty. The "curse of optimality" might suggest that this is an illogical point of view. However, if the notion that optimality is all important is adopted, then the focus becomes matching the theoretical properties of an estimator to the assumed properties of the density function. Is it a gross inefficiency to use a procedure that requires only 2 continuous derivatives when the curve in fact has 6 continuous derivatives? This attitude may have some formal basis but should be discouraged as too heavy-handed for nonparametric thinking. A more relaxed attitude is required. Furthermore, many "optimal" nonparametric procedures are unstable in a manner that slightly inefficient procedures are not. In practice, when faced with the application of a procedure that requires 6 derivatives, or some other assumption that cannot be proved in practice, it is more important to be able to recognize the signs of estimator failure than to worry too much about assumptions. Detecting failure at the level of a discontinuous fourth derivative is a bit extreme, but certainly the effects of simple discontinuities should be well understood. Thus only for the purposes of illustration are the best assumptions given.

The notions of efficiency and admissibility are related to the choice of a criterion, which can only imperfectly measure the quality of a nonparametric estimate. Unlike optimal parametric estimates that are useful for many purposes, nonparametric estimates must be optimized for each application. The extra work is justified by the extra flexibility. As the choice of criterion is imperfect, so then is the notion of a single optimal estimator. This attitude reflects not sloppy thinking, but rather the imperfect relationship between the practical and theoretical aspects of our methods. Too rigid a point of view leads one to a minimax view of the world where nonparametric methods should be abandoned because there exist difficult problems.

Visualization is an important component of nonparametric data analysis. *Data visualization* is the focus of exploratory methods, ranging from simple

scatterplots to sophisticated dynamic interactive displays. *Function visualization* is a significant component of nonparametric function estimation, and can draw on the relevant literature in the fields of scientific visualization and computer graphics. The focus of multivariate data analysis on points and scatterplots has meant that the full impact of scientific visualization has not yet been realized. With the new emphasis on smooth functions estimated nonparametrically, the fruits of visualization will be attained. Banchoff (1986) has been a pioneer in the visualization of higher-dimensional mathematical surfaces. Curiously, the surfaces of interest to mathematicians contain singularities and discontinuities, all producing striking pictures when projected to the plane. In statistics, visualization of the smooth density surface in 4, 5, and 6 dimensions cannot rely upon projection, as projections of smooth surfaces to the plane show nothing. Instead, the emphasis is on contouring in 3 dimensions and slicing of surfaces beyond. The focus on 3 and 4 dimensions is natural because 1 and 2 are so well understood. Beyond 4 dimensions, the ability to explore surfaces carefully decreases rapidly due to the curse of dimensionality. Fortunately, statistical data seldom display structure in more than 5 dimensions, so guided projection to those dimensions may be adequate. It is these threshold dimensions from 3 to 5 that are and deserve to be the focus of our visualization efforts.

There is a natural flow among the parametric, exploratory, and nonparametric procedures that represents a rational approach to statistical data analysis. Begin with a fully exploratory point of view in order to obtain an overview of the data. If a probabilistic structure is present, estimate that structure nonparametrically and explore it visually. Finally, if a linear model appears adequate, adopt a fully parametric approach. Each step conceptually represents a willingness to more strongly *smooth* the raw data, finally reducing the dimension of the solution to a handful of interesting parameters. With the assumption of Normality, the mind's eye can easily imagine the d -dimensional egg-shaped elliptical data clusters. Some statisticians may prefer to work in the reverse order, progressing to exploratory methodology as a diagnostic tool for evaluating the adequacy of a parametric model fit.

There are many excellent references that complement and expand on this subject. In exploratory data analysis, references include Tukey (1977), Tukey and Tukey (1981), Cleveland and McGill (1988), and Wang (1978). In density estimation, the classic texts of Tapia and Thompson (1978), Wertz (1978), and Thompson and Tapia (1990) first indicated the power of the nonparametric approach for univariate and bivariate data. Silverman (1986) has provided a further look at applications in this setting. Prakasa Rao (1983) has provided a theoretical survey with a lengthy bibliography. Other texts are more specialized, some focusing on regression (Müller, 1988; Härdle, 1990), some on a specific error criterion (Devroye and Györfi, 1985; Devroye, 1987), and some on particular solution classes such as splines (Eubank, 1988; Wahba, 1990). A discussion of additive models may be found in Hastie and Tibshirani (1990).

1.2 HISTORICAL PERSPECTIVE

One of the roots of modern statistical thought can be traced to the empirical discovery of correlation by Galton in 1886 (Stigler, 1986). Galton's ideas quickly reached Karl Pearson. Although best remembered for his methodological contributions such as goodness-of-fit tests, frequency curves, and biometry, Pearson was a strong proponent of the geometrical representation of statistics. In a series of lectures a century ago in November 1891 at Gresham College in London, Pearson spoke on a wide-ranging set of topics (E. S. Pearson, 1938). He discussed the foundations of the science of pure statistics and its many divisions. He discussed the collection of observations. He described the classification and representation of data using both numerical and geometrical descriptors. Finally, he emphasized statistical methodology and discovery of statistical laws. The syllabus for his lecture of November 11, 1891, includes this cryptic note:

Erroneous opinion that Geometry is only a means of popular representation: *it is a fundamental method of investigating and analysing statistical material.* (his italics)

In that lecture Pearson described 10 methods of geometrical data representation. The most familiar is a representation "by columns," which he called the "histogram." (Pearson is usually given credit for coining the word "histogram" later in a 1894 paper.) Other familiar-sounding names include "diagrams," "chartograms," "topograms," and "stereograms." Unfamiliar names include "stigmograms," "euthygrams," "epipedograms," "radiograms," and "hormograms."

Beginning 21 years later, Fisher advanced the numerically descriptive portion of statistics with the method of maximum likelihood, from which he progressed on to the analysis of variance and other contributions that focused on the optimal use of data in parametric modeling and inference. In *Statistical Methods for Research Workers*, Fisher (1932) devotes a chapter entitled "Diagrams" to graphical tools. He begins the chapter with this statement:

The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.

An emphasis on optimization and the efficiency of statistical procedures has been a hallmark of mathematical statistics ever since. Ironically, Fisher was criticized by mathematical statisticians for relying too heavily upon geometrical arguments in proofs of his results.

Modern statistics has experienced a strong resurgence of geometrical and graphical statistics in the form of exploratory data analysis (Tukey, 1977). Given the parametric emphasis on optimization, the more relaxed philosophy of exploratory data analysis has been refreshing. The revolution has been fueled

by the low cost of graphical workstations and microcomputers. These machines have enabled current work on *statistics in motion* (Scott, 1990), that is, the use of animation and kinematic display for visualization of data structure, statistical analysis, and algorithm performance. No longer are static displays sufficient for comprehensive analysis.

All of these events were anticipated by Pearson and his visionary statistical computing laboratory. In his lecture of April 14, 1891, entitled "The Geometry of Motion," he spoke of the "ultimate elements of sensations we represent as motions in space and time." In 1918, after his many efforts during World War I, he reminisced about the excitement created by wartime work of his statistical laboratory:

The work has been so urgent and of such value that the Ministry of Munitions has placed eight to ten computers and draughtsmen at my disposal . . . (E. S. Pearson, 1938, p. 165).

These workers produced hundreds of statistical graphs, ranging from detailed maps of worker availability across England (chartograms) to figures for sighting antiaircraft guns (diagrams). The use of stereograms allowed for representation of data with 3 variables. His "computers," of course, were not electronic but human. Later, Fisher would be frustrated because Pearson would not agree to allocate his "computers" to the task of tabulating percentiles of the *t*-distribution. But Pearson's capabilities for producing high-quality graphics were far superior to those of most modern statisticians prior to 1980. Given Pearson's joint interests in graphics and kinematics, it is tantalizing to speculate on how he would have utilized modern computers.

1.3 GRAPHICAL DISPLAY OF MULTIVARIATE DATA POINTS

The modern challenge in data analysis is to be able to cope with whatever complexities may be intrinsic to the data. The data may, for example, be strongly non-Normal, fall onto a nonlinear subspace, exhibit multiple modes, or be asymmetric. Dealing with these features becomes exponentially more difficult as the dimensionality of the data increases, a phenomenon known as the *curse of dimensionality*. In fact, data sets with hundreds of variables and millions of observations are routinely compiled that exhibit all of these features. Examples abound in such diverse fields as remote sensing, the U.S. Census, geological exploration, speech recognition, and medical research. The expense of collecting and managing these large data sets is often so great that no funds are left for serious data analysis. The role of statistics is clear, but too often no statisticians are involved in large projects and no creative statistical thinking is applied. The goal of statistical data analysis is to extract the maximum information from the data, and to present a product that is as accurate and as useful as possible.

1.3.1 Multivariate Scatter Diagrams

The presentation of multivariate data is often accomplished in tabular form, particularly for small data sets with named or labeled objects. For example, Table 2 in Appendix B contains information on a selected sample of American universities, and Table 1 in Appendix B contains economic data spanning the depression years of the 1930s. It is easy enough to scan an individual column in these tables, to make comparisons of library size, for example, and to draw conclusions *one variable at a time*; see Tufte (1983) and Wang (1978). However, variable-by-variable examination of multivariate data can be overwhelming and tiring, and cannot reveal any relationships among the variables. Looking at all pairwise scatterplots provides an improvement (Chambers et al., 1983). Data on 4 variables of 3 species of iris are displayed in Figure 1.1. [A listing of the Fisher-Anderson iris data, one of the few familiar 4-dimensional data sets, may be found in several references and is provided with the S package (Becker, Chambers, and Wilks, 1988).] What multivariate structure is apparent from this figure? The Setosa variety does not overlap the other 2 varieties. The Versicolor and Virginica varieties are not as well separated, although a close examination reveals that they are almost nonoverlapping. If the 150 observations were unlabeled and plotted with the same symbol, it is likely that only 2 clusters would be observed. Even if it were known a priori that there were 3 clusters, it would still be unlikely that all 3 clusters would be properly identified. These alternative presentations reflect the 2 related problems of discrimination and clustering, respectively.

If the observations from different categories overlap substantially or have different sample sizes, scatter diagrams become much more difficult to interpret properly. The data in Figure 1.2 come from a study of 371 males suffering from chest pain (Scott et al., 1978): 320 had demonstrated coronary artery disease

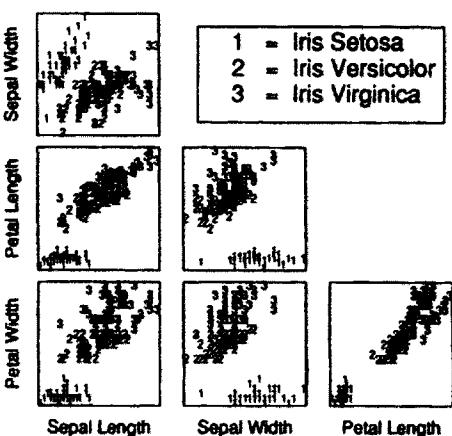


Figure 1.1 All pairwise scatter diagrams of the iris data with the 3 species indicated.

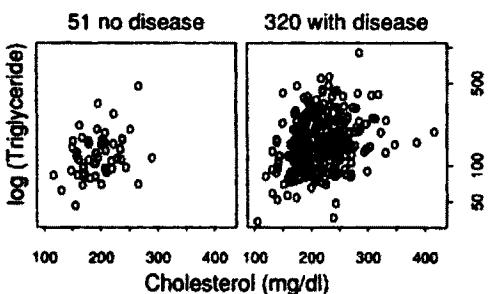


Figure 1.2 Scatter diagram of blood lipid concentrations for 320 diseased and 51 nondiseased males.

(occlusion or narrowing of the heart's own arteries) while 51 had none; see Table 3 in Appendix B. The blood fat concentrations of plasma cholesterol and triglyceride are predictive of heart disease, although the correlation is low. It is difficult to estimate the predictive power of these variables in this setting solely from the scatter diagram. A nonparametric analysis will reveal some interesting nonlinear interactions; see Chapters 5 and 9.

An easily overlooked practical aspect of scatter diagrams is illustrated by these data, which are integer valued. To avoid problems of overplotting, the data have been *jittered* or *blurred* (Chambers et al., 1983); that is, uniform $U(-0.5, 0.5)$ noise is added to each element of the original data. This trick should be regularly employed for data recorded with 3 or fewer significant digits (with an appropriate range on the added uniform noise). Jittering reduces visual miscues that result from the vertical and horizontal synchronization of regularly spaced data.

The visual perception system can easily be overwhelmed if the number of points is more than several thousand. Figure 1.3 displays 3 pairwise scatter-plots derived from measurements taken in 1977 by the LANDSAT remote sensing system over a 5 mile by 6 mile agricultural region in North Dakota with $n = 22,932 = 117 \times 196$ pixels or picture elements, each corresponding to an area approximately 1.1 acres in size (Scott and Thompson, 1983; Scott and Jee, 1984). The LANDSAT instrument measures the intensity of light in 4 spectral bands reflected from the surface of the earth. A principal components transformation gives 2 variables that are commonly referred to as the "brightness" and "greenness" of each pixel. Every pixel is measured at regular intervals of approximately 3 weeks. During the summer of 1977, 6 useful replications were obtained, giving 24 measurements on each pixel. Using an agronomistic growth model for crops, Badhwar, Carnes, and Austin (1982) nonlinearly transformed this 24-dimensional data to 3 dimensions. Badhwar described these synthetic variables, (x_1, x_2, x_3) , as (1) the calendar time at which peak greenness is observed, (2) the length of crop ripening, and (3) the peak greenness value, respectively. The scatter diagrams in Figure 1.3 have also been enhanced by jittering, as the raw data are integers between

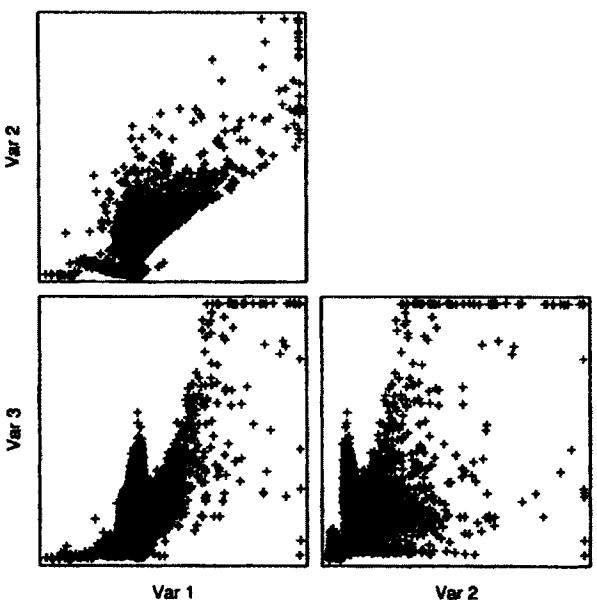


Figure 1.3 Pairwise scatter diagram of transformed LANDSAT data from 22,932 pixels over a 5 by 6 nautical mile region. The range on all the axes is (0, 255).

(0, 255). The use of integers allows compression to 8 bits of computer memory. Only structure in the boundary and tails is readily seen. The overplotting problem is apparent and the blackened areas include over 95% of the data. Other techniques to enhance scatter diagrams are needed to see structure in the bulk of the data cloud, such as plotting random subsets; see Tukey and Tukey (1981).

Pairwise scatter diagrams lack one important property necessary for identifying more than 2-dimensional features—strong interplot linkage among the plots. In principle, it should be possible to locate the same point in each figure, assuming the data are free of ties. But it is not practical to do so for samples of any size. For quadrivariate data, Diaconis and Friedman (1983) proposed drawing lines between corresponding points in the scatterplots of (x_1, x_2) and (x_3, x_4) ; see Problem 2. But a more powerful dynamic technique that takes full advantage of computer graphics has been developed by several research groups (McDonald, 1982; Becker and Cleveland, 1987; see the many references in Cleveland and McGill, 1988). The method is called *brushing* or *painting* a scatterplot matrix. Using a pointing device such as a mouse, a subset of the points in one scatter diagram is selected and the corresponding points are simultaneously highlighted in the other scatter diagrams. Conceptually, a subset of points in \mathbb{R}^d is tagged, for example, by painting the points red or making the points blink synchronously, and that characteristic is inherited by the linked

points in all the “linked” graphs, including not only scatterplots but histograms and regression plots as well. The iris example in Figure 1.1 illustrates the flavor of brushing with 3 tags. Usually the color of points is changed rather than the symbol type. Brushing is an excellent tool for identifying outliers and following well-defined clusters. It is well-suited for conditioning on some variable, for example, $1 < x_3 < 3$.

These ideas are illustrated in Figure 1.4 for the PRIM4 data set (Friedman and Tukey, 1974; the data summarize 500 high-energy particle physics scattering experiments) provided in the S language. Using the brushing tool in S-PLUS (1990), the left cluster in the 1–2 scatterplot was brushed, and then the left cluster in the 2–4 scatterplot was brushed with a different symbol. Try to imagine linking the clusters throughout the scatterplot matrix without any highlighting.

There are limitations to the brushing technique. The number of pairwise scatterplots is $\binom{d}{2}$ so viewing more than 5 or 10 variables at once is impractical. Furthermore, the physical size of each scatter diagram is reduced as more variables are added, so that fewer distinct data points can be plotted. If there are more than a few variables, the eye cannot follow many of the dynamic changes in the pattern of points during brushing, except with the simplest of structure. It is, however, an open question as to the number of dimensions of structure that can be perceived by this method of linkage. Brushing remains an important and well-used tool that has proven successful in real data analysis.

If a 2-D array of bivariate scatter diagrams is useful, then why not construct a 3-D array of *trivariate* scatter diagrams? Navigating the collection of $\binom{d}{3}$ trivariate scatterplots is difficult even with modest values of d . But a single 3-D scatterplot can easily be rotated in real time with significant perceptual gain compared to 3 bivariate diagrams in the scatterplot matrix. Many statistical packages now provide this capability. The program MACSPIN (Donoho, Donoho, and Gasko, 1988) was the first widely used software of this type. The top middle panel in Figure 1.4 displays a particular orientation of a rotating 3-D scatterplot. The kinds of structure available in 3-D data are more complex (and hence more interesting) than in 2-D data. Furthermore, the overplotting problem is reduced as more data points can be resolved in a rotating 3-D scatterplot than in a static 2-D view (although this is resolution dependent—a 2-D view printed by a laser device can display significantly more points than is possible on a computer monitor). Density information is still relatively difficult to perceive, however, and the sample size definitely influences perception.

Beyond 3 dimensions, many novel ideas are being pursued; see Tukey and Tukey (1981). Six-dimensional data could be viewed with 2 rotating 3-D scatter diagrams linked by brushing. Carr and Nicholson (1988) have actively pursued the use of stereography as an alternative and adjunct to rotation. Several computer workstations now provide true stereo viewing as well as rotation with special polarizing glasses. Some workers report that stereo viewing of static data can be more precise than viewing dynamic rotation alone. Unfortunately, many

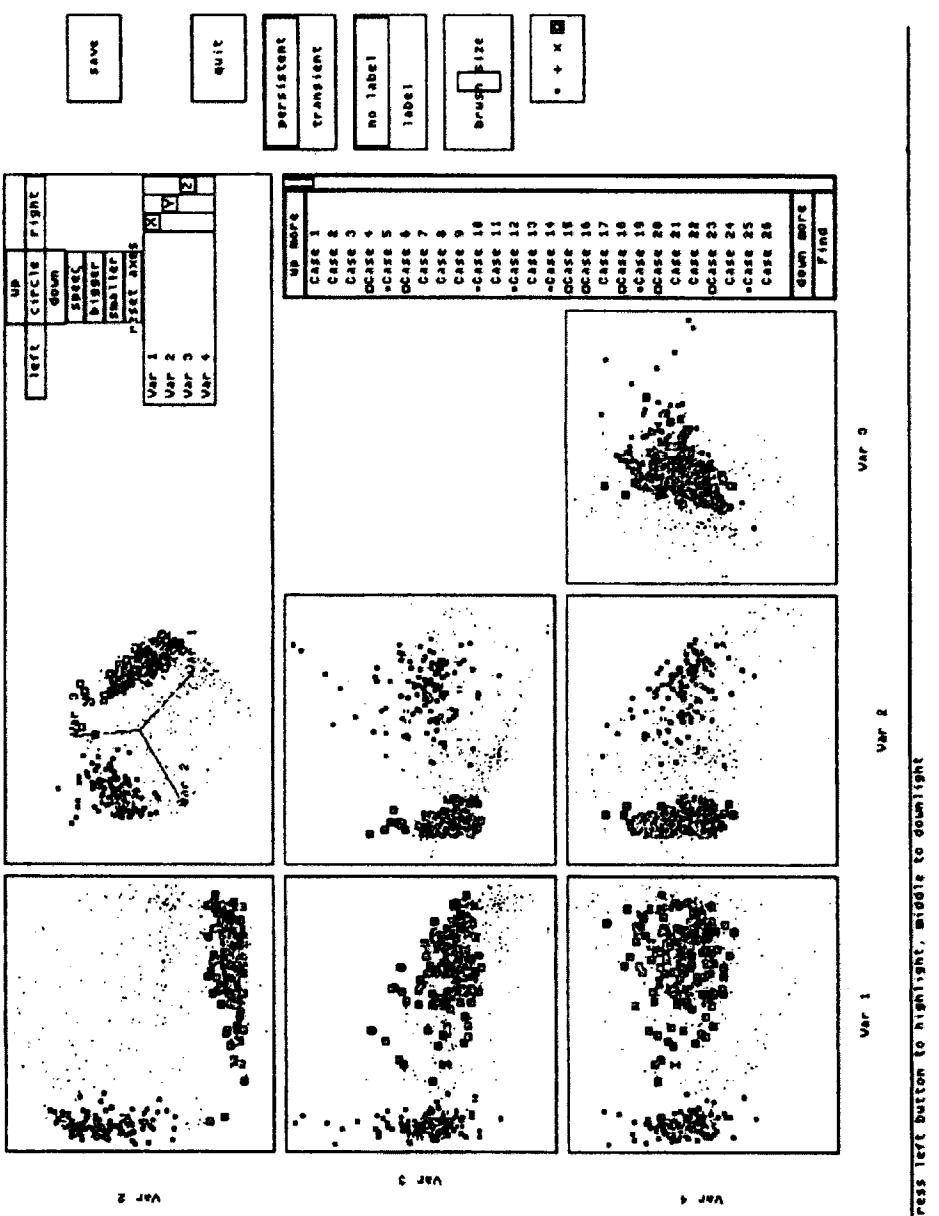


Figure 1.4 Pairwise scatterplots of the transformed PRIM4 data using the “brush” function in S-PLUS (1990). Two clumps of points were highlighted by brushing.

individuals suffer from color blindness and various depth perception limitations, rendering some techniques useless. Nevertheless, it is clear that there is no limit to the possible combinations of ideas one might consider implementing. Such efforts can easily take many months to program without any fancy interface. This state of affairs would be discouraging but for the fact that a LISP-based system for easily prototyping such ideas is now available using object-oriented concepts; see Tierney (1990). A recent collection of articles is devoted to the general topic of animation (Cleveland and McGill, 1988).

The idea of displaying 2- or 3-D arrays of 2- or 3-D scatter diagrams is perhaps too closely tied to the Euclidean coordinate system. It might be better to examine many 2- or 3-D projections of the data. An orderly way to do approximately just that is the "grand tour" discussed by Asimov (1985). Let P be a $d \times 2$ projection matrix, which takes the d -dimensional data down to a plane. The author proposed examining a sequence of scatterplots obtained by a smoothly changing sequence of projection matrices. The resulting kinematic display shows the n data points moving in a continuous (and sometimes seemingly random) fashion. It may be hoped that most interesting projections will be displayed at some point during the first several minutes of the grand tour, although for even 10 variables several hours may be required (Huber, 1985).

Special attention should be drawn to representing multivariate data in the bivariate scatter diagram with points replaced by *glyphs*, which are special symbols whose shapes are determined by the remaining data variables (x_3, \dots, x_d). Figure 1.5 displays the iris data in such a form following Carr et al. (1986). The length and angle of the glyph are determined by the sepal length and width, respectively. Careful examination of the glyphs shows that there is no gap in 4-D between the Versicolor and Virginica species, as the angles and lengths of the glyphs are similar near the boundary. A second glyph representation shown in Color Plate 1 is a 3-D scatterplot omitting one of the 4 variables. This figure clearly depicts the structure in these data. Plotting glyphs in 3-D scatter diagrams with stereography is a more powerful visual tool (Carr and Nicholson,

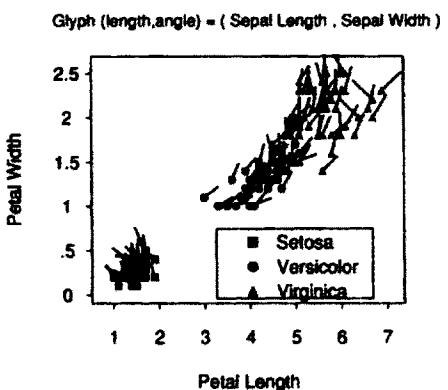


Figure 1.5 Glyph scatter diagram of the iris data.

1988). The glyph technique does not treat variables "symmetrically" and all variable-glyph combinations could be considered. This complaint affects most multivariate procedures (with a few exceptions).

All of these techniques are an outgrowth of a powerful system devised to analyze data in up to 9 dimensions called PRIM-9 (Fisher Keller, Friedman, and Tukey, 1974; reprinted in Cleveland and McGill, 1988). The PRIM-9 system contained many of the capabilities of current systems. The letters are an acronym for "Picturing, Rotation, Isolation, and Masking." The latter two serve to identify and select subsets of the multivariate data. The "picturing" feature was implemented by pressing 2 buttons that cycled through all of the $\binom{9}{2}$ pairwise scatter diagrams in current coordinates. An IBM 360 mainframe was specially modified to drive the custom display system.

1.3.2 Chernoff Faces

Chernoff (1973) proposed a special glyph that associates variables to facial features, such as the size and shape of the eyes, nose, mouth, hair, ears, chin, and facial outline. Certainly, humans are able to discriminate among nearly identical faces very well. Chernoff has suggested that most other multivariate point methods "seem to be less valuable in producing an emotional response" (Wang, 1978, p. 6). Whether an emotional response is desired is debatable. Chernoff faces for the time series data set in Table 1 in Appendix B are displayed in Figure 1.6. (The variable-feature associations are listed in the table.) By carefully studying an individual facial feature such as the smile over the sequence of all the faces, simple trends can be recognized. But it is the overall multivariate impression that makes Chernoff faces so powerful. Variables should be carefully assigned to features. For example, Chernoff faces of the colleges' data in Table 2 might logically assign variables relating to the library to the eyes rather than to the mouth (see Problem 3). Such subjective judgments should not prejudice our use of this procedure.

One early application not in a statistics journal was constructed by Hiebert-Dodd (1982), who had examined the performance of several optimization algorithms on a suite of test problems. She reported that several referees felt this method of presentation was too frivolous. Comparing the endless tables in the paper as it appeared to the Chernoff faces displayed in the original technical report, one might easily conclude the referees were too cautious. On the other hand, when Rice University administrators were shown Chernoff faces of the colleges' data, they were quite open to its suggestions and enjoyed the exercise. The practical fact is that repetitious viewing of large tables of data is tedious and haphazard, and broad-brush displays such as faces can significantly improve data digestion. Several researchers have noted that Chernoff faces contain redundant information because of symmetry. Flury and Riedwyl (1981) have proposed the use of asymmetrical faces, as did Turner and Tidmore (1980), although Chernoff has stated he believes the additional gain does not justify such nonrealistic figures.

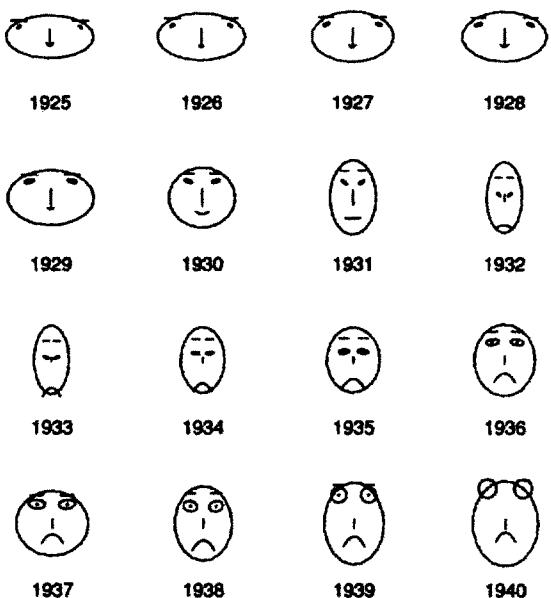


Figure 1.6 Chernoff faces of economic data spanning 1925–1940.

1.3.3 Andrews' Curves and Parallel Coordinate Curves

Three intriguing proposals display not the data points themselves but rather a unique curve determined by the data vector \mathbf{x} . Andrews (1972) proposed representing high-dimensional data by replacing each point in \mathbb{R}^d with a curve $s(t)$ for $|t| < \pi$, where

$$\begin{aligned}s(t | x_1, \dots, x_d) = & \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t \\ & + x_5 \cos 2t + \dots,\end{aligned}$$

the so-called *Fourier series representation*. This mapping provides the first “complete” continuous view of high-dimensional points on the plane, because, in principle, the original multivariate data point can be recovered from this curve. Clearly, an Andrews curve is dominated by the variables placed on the low-frequency terms, so care should be taken to put the most interesting variables early in the expansion (see Problem 4).

A simple graphical device that treats the d variables symmetrically is the star diagram, which is discussed by Fienberg (1979). The d axes are drawn as spokes on a wheel. The coordinate data values are plotted on those axes and connected as shown in Figure 1.7.

Another novel multivariate approach that treats variables in a symmetric fashion is the *parallel coordinates plot*, introduced by Inselberg (1985) in a mathematical setting and extended by Wegman (1990) to the analysis of

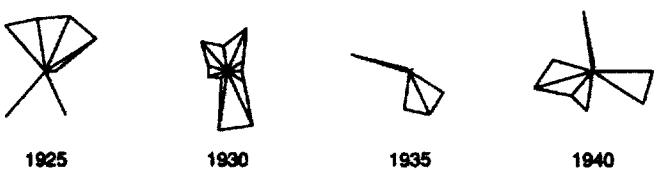


Figure 1.7 Star diagram for 4 years of the economic data shown in Figure 1.6.

stochastic data. Cartesian coordinates are abandoned in favor of d axes drawn parallel and equally spaced. Each multivariate point $\mathbf{x} \in \Re^d$ is plotted as a piecewise linear curve connecting the d points on the parallel axes. For reasons shown by Inselberg and Wegman, there are advantages to simply drawing piecewise linear line segments, rather than a smoother line such as a spline. The disadvantage of this choice is that points that have identical values in any coordinate dimension cannot be distinguished in parallel coordinates. However, with this choice a duality may be deduced between points and lines in Euclidean and parallel coordinates. In the left frame of Figure 1.8, 6 points that fall on a straight line with negative slope are plotted. The right frame shows those same points in parallel coordinates. Thus a scatter diagram of highly correlated Normal points displays a nearly common point of intersection in parallel coordinates. However, if the correlation is positive, that point is not “between” the parallel axes; see Problem 6. The location of the point where the lines all intersect can be used to recover the equation of the line back in Euclidean coordinates; see Problem 8.

A variety of other properties with potential applications are explored by Inselberg and Wegman. One result is a graphical means of deciding if a point $\mathbf{x} \in \Re^d$ is on the inside or the outside of a convex closed hypersurface. If

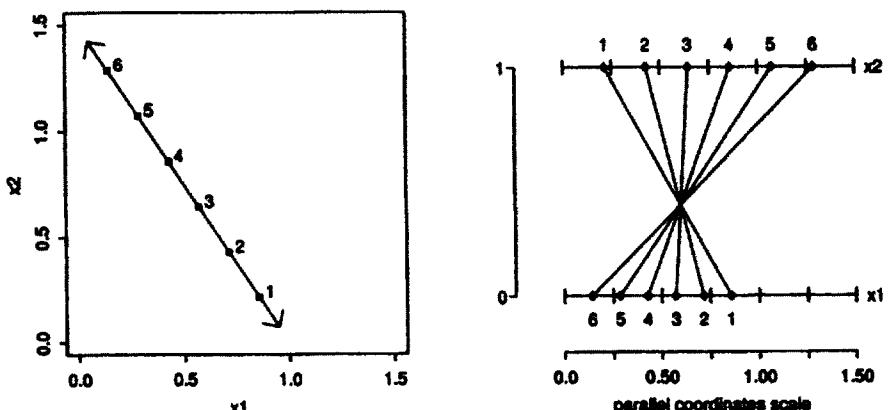


Figure 1.8 Example of duality of points and lines between Euclidean and parallel coordinates. The points are labeled 1 to 6 in both coordinate systems.

all the points on the hypersurface are plotted in parallel coordinates, then a well-defined geometrical outline will appear on the plane. If a portion of the line segments defining the point x in parallel coordinates fall outside the outline, then x is not inside the hypersurface, and vice versa. One of the more fascinating extensions developed by Wegman is a grand tour of all variables displayed in parallel coordinates. The advantage of parallel coordinates is that all d of the rotating variables are visible simultaneously, whereas in the usual presentation, only 2 of the grand tour variables are visible in a bivariate scatterplot.

Figure 1.9 displays parallel coordinate plots of the iris and earthquake data. The earthquake data set represents the epicenters of 473 tremors beneath the Mount St. Helens volcano in the several months preceding its March 1982 eruption (Weaver, Zollweg, and Malone, 1983). Clearly, the tremors are mostly small in magnitude, increasing in frequency over time, and clustered near the surface, although depth is clearly a bimodal variable. The longitude and latitude variables are least effective on this plot, because their natural spatial structure is lost.

1.3.4 Limitations

Tools such as Chernoff faces and scatter diagram glyphs tend to be most valuable with small data sets where individual points are “identifiable” or

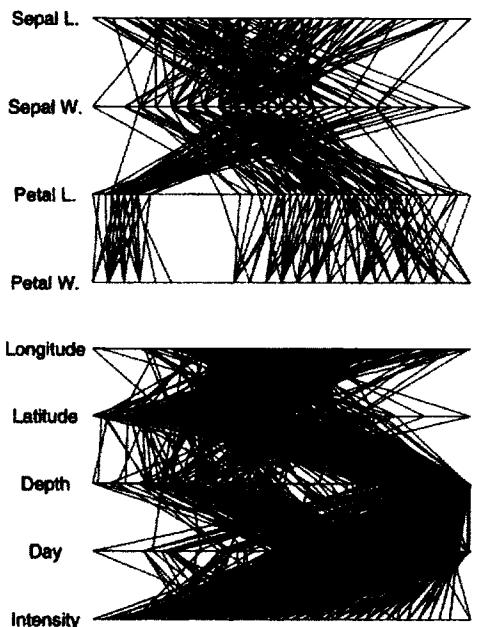


Figure 1.9 Parallel coordinate plots of the iris and earthquake data sets.

interesting. Such individualistic exploratory tools can easily generate “too much ink” (Tufte, 1983) and produce figures with black splotches, which convey little information. Parallel coordinates and Andrews’ curves generate much ink. One obvious remedy is to plot only a subset of the data in a process known as “thinning.” However, plotting random subsets no longer makes optimal use of all the data and does not result in precisely reproducible interpretations. Point-oriented methods typically have a range of sample sizes that is most appropriate: $n < 200$ for faces; $n < 2,000$ for scatter diagrams.

Since none of these displays is truly d -dimensional, each has limitations. All pairwise scatterplots can detect distinct clusters and some 2-dimensional structure (if perhaps in a rotated coordinate system). In the latter case, an interactive supplement such as brushing may be necessary to confirm the nature of the links among the scatterplots (not really providing any higher-dimensional information). On the positive side, variables are treated symmetrically in the scatterplot matrix. But many different and highly dissimilar d -dimensional data sets can give rise to visually similar scatterplot matrix diagrams; hence the need for brushing. However, with increasing number of variables, individual scatterplots physically decrease in size and fill up with ink ever faster. Scatter diagrams provide a highly subjective view of data, with poor density perception and greatest emphasis on the tails of the data.

1.4 GRAPHICAL DISPLAY OF MULTIVARIATE FUNCTIONALS

1.4.1 Scatterplot Smoothing by Density Function

As graphical exploratory tools, each of the point-based procedures has significant value. However, each suffers from the problem of too much ink, as the number of objects (and hence the amount of ink) is linear in the sample size n . To mix metaphors, point-based graphs cannot provide a consistent picture of the data as $n \rightarrow \infty$. As Scott and Thompson (1983) wrote,

the scatter diagram points to the bivariate density function.

In other words, the raw data points need to be smoothed if a consistent view is to be obtained.

A histogram is the simplest example of a *scatterplot smoother*. The amount of smoothness is controlled by the bin width. For univariate data, the histogram with bin width narrower than $\min |x_i - x_j|$ is precisely a univariate scatter diagram plotted with glyphs that are tall, thin rectangles. For bivariate data, the glyph is a beam with a square base. Increasing the bin width, the histogram represents a count per unit area, which is precisely the unit of a probability density. In Chapter 3, the histogram will be shown to provide a consistent estimate of the density function in any dimension.

Histograms can provide a wealth of information for large data sets, even well-known ones. For example, consider the 1979–1981 decennial life table

published by the U.S. Bureau of the Census (1987). Certain relevant summary statistics are well-known: life expectancy, infant mortality, and certain conditional life expectancies. But what additional information can be gleaned by examining the mortality histogram itself? In Figure 1.10, the histogram of age of death for individuals is depicted. Not surprisingly, the histogram is skewed with a short tail for older ages. Not as well-known perhaps is the observation that the most common age of death is 85! The absolute and relative magnitude of mortality in the first year of life is made strikingly clear.

Careful examination reveals two other general features of interest. The first feature is the small but prominent bump in the curve between the ages of 13 and 27. This “excess mortality” is due to an increase in a variety of risky activities, the most notable being obtaining a driver’s license. In the right frame of Figure 1.10, comparison of the 1959-61 (Gross and Clark, 1975) and 1979-81 histograms shows an impressive reduction of death in all preadolescent years. Particularly striking is the 60% decline in mortality in the first year and the 3-year difference in the locations of the modes.

These facts are remarkable when placed in the context of the *mortality histogram* constructed by John Graunt from the Bills of Mortality during the plague years. Graunt (1662) estimated that 36% of individuals died before attaining their sixth birthday! Graunt was a contemporary of the better-known William Petty, to whom some credit for these ideas is variously ascribed, probably without cause. The circumstantial evidence that Graunt actually invented the histogram while looking at these mortality data seems quite strong, although there is reason to infer that Galileo had used histogram-like diagrams earlier. Hald (1990) recounts a portion of Galileo’s *Dialogo*, published in 1632, in which Galileo summarized his observations on the star that appeared in 1572. According to Hald, Galileo noted the symmetry of the “observation errors” and the more frequent occurrence of small errors than large errors. Both points suggest Galileo had constructed a frequency diagram to draw those conclusions.

Many large data sets are in fact collected in binned or histogram form. For example, elementary particles in high-energy physics scattering experiments are

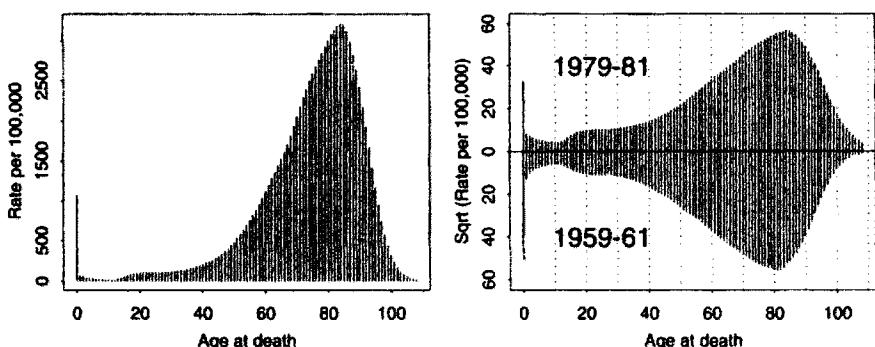


Figure 1.10 Histogram of the U.S. mortality data, 1979–1981. Back-to-back rootograms (histograms plotted on a square-root scale) of the mortality data for 1979-81 and 1959-61.

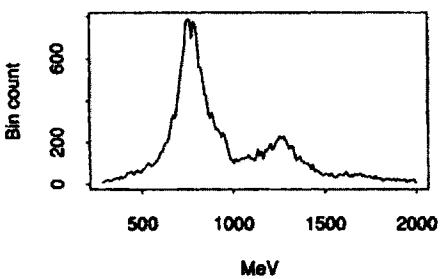


Figure 1.11 Histogram of LRL data.

manifested by small bumps in the frequency curve. Good and Gaskins (1980) considered such a large data set ($n = 25,752$) from the Lawrence Radiation Laboratory (LRL); see Figure 1.11. The authors devised an ingenious algorithm for estimating the odds that a bump observed in the frequency curve was real. This topic is covered in Chapter 9.

Multivariate scatterplot smoothing of time series data is also easily accomplished with histograms. Consider a univariate time series and smooth both the raw data $\{x_t\}$ as well as the lagged data $\{x_t, x_{t+1}\}$. Any strong elliptical structure present in the smoothed lagged-data diagram provides a graphical version of the first-order autocorrelation coefficient. Consider the Old Faithful geyser data from Table 6 in Appendix B. These data are the durations in minutes of 107 eruptions of the Old Faithful geyser (Weisberg, 1985). As there was a gap in the recording of data between midnight and 6 a.m., there are only 99 pairs $\{x_t, x_{t+1}\}$ available. The univariate histogram in Figure 1.12 reveals a simple bimodal structure—short and long eruption durations. The most notable feature in the bivariate (smoothed) histogram is the missing fourth bump corresponding to the short-short duration sequence. Clearly, graphs of $f(x_{t+1} | x_t)$ would be useful for improved prediction compared to a regression estimate.

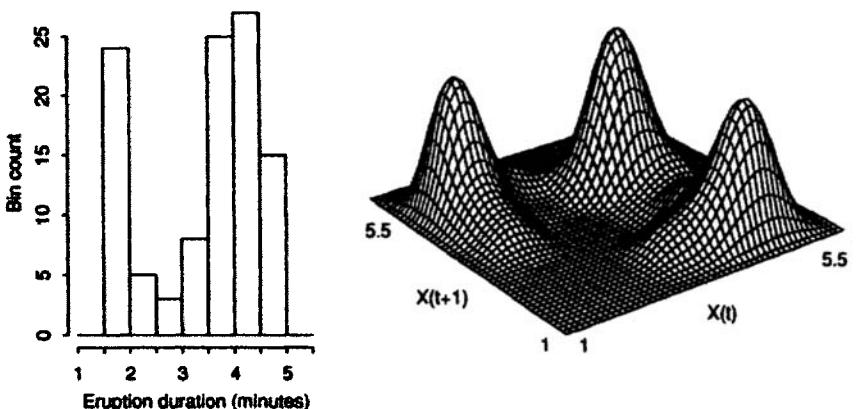


Figure 1.12 Histogram of $\{x_t\}$ for the Old Faithful geyser data, and a smoothed bivariate histogram of the lagged data $\{x_t, x_{t+1}\}$.

For more than 2 dimensions, only slices are available for viewing with histogram surfaces. Consider the LANDSAT data again. Divide the (jittered) data into 4 pieces using quartiles of x_1 , which is the time of peak greenness. Examining a series of bivariate pictures of (x_2, x_3) for each quartile slice provides a crude approximation of the 4-dimensional surface $\hat{f}(x_1, x_2, x_3)$; see Figure 1.13. The histograms are all constructed on the subinterval $[-5, 100] \times [-5, 100]$. Compare this representation of the LANDSAT data to that in Figure 1.3. From Figure 1.3 it is clear that most of the outliers are in the last quartile of x_1 . How well can the relative density levels be determined from the scatter diagrams? Visualization of a smoothed histogram of these data will be considered in Section 1.4.3.

1.4.2 Scatterplot Smoothing by Regression Function

The term *scatterplot smoother* is most often applied to regression data. For bivariate data, either a nonparametric regression line can be superimposed

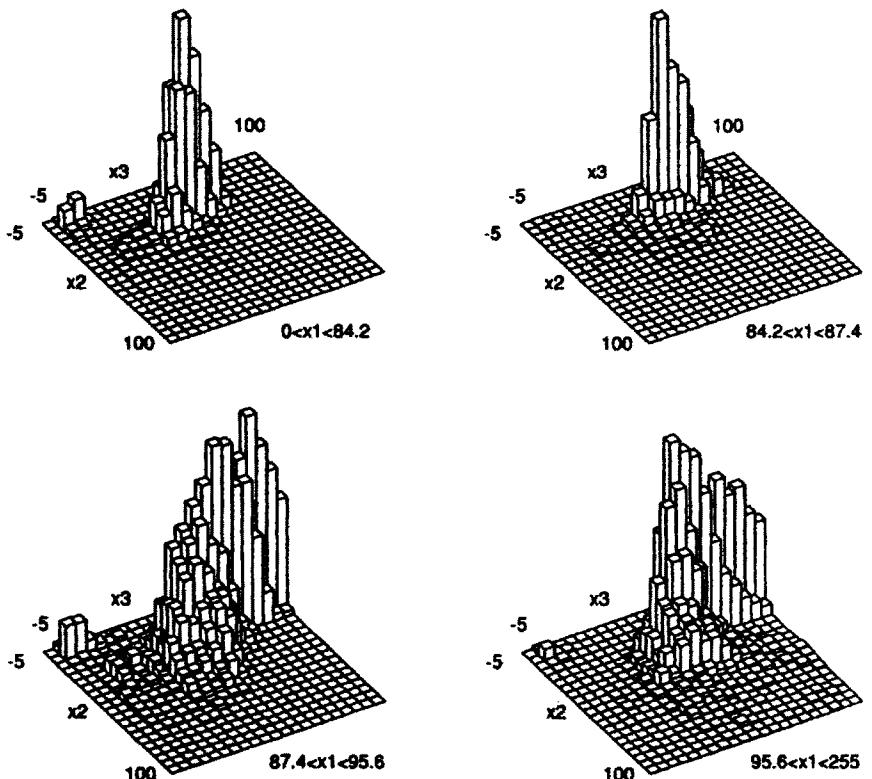


Figure 1.13 Bivariate histogram slices of the trivariate LANDSAT data. Slicing was performed at the quartiles of variable x_1 .

upon the data, or the points themselves can be moved towards the regression line. Tukey (1977) presents the “3R” smoother as an example of the latter. Suppose that the n data points, $\{x_i\}$, are measured on a fixed time scale. The 3R smoothing algorithm replaces each point $\{x_i\}$ with the median of the 3 points $\{x_{i-1}, x_i, x_{i+1}\}$ recursively until no changes occur. This algorithm is a powerful filter that removes isolated outliers effectively. The 3R smoother may be applied to unequally spaced data or repeated data. Tukey also proposes applying a Hanning filter, by which $\bar{x}_i \leftarrow 0.25 \times (x_{i-1} + 2x_i + x_{i+1})$. This filter may be applied several times as necessary. In Figure 1.14, the Tukey smoother (S function *smooth*) is applied to the gas flow data given in the Table 5 in Appendix B. Observe how the single potential outlier at $x = 187$ is totally ignored. The least-squares fit is shown for reference.

The simplest nonparametric regression estimator is the *regressogram*. The x -axis is binned and the sample averages of the responses are computed and plotted over the intervals. The regressogram for the gas flow data is also shown in Figure 1.14. The Hanning filter and regressogram are special cases of nonparametric kernel regression, which is discussed in Chapter 8.

The gas flow data is part of a larger collection taken at 7 different pressures. A stick-pin plot of the complete data set is shown in Figure 1.15 (the 74.6 psia data are second from the right). Clearly, the accuracy is affected by the flow rate, while the effect of psia seems small. These data will be revisited in Chapter 8.

1.4.3 Visualization of Multivariate Functions

Visualization of functions of more than 2 variables has not been common in statistics. The LANDSAT example in Figure 1.13 hints at the potential that

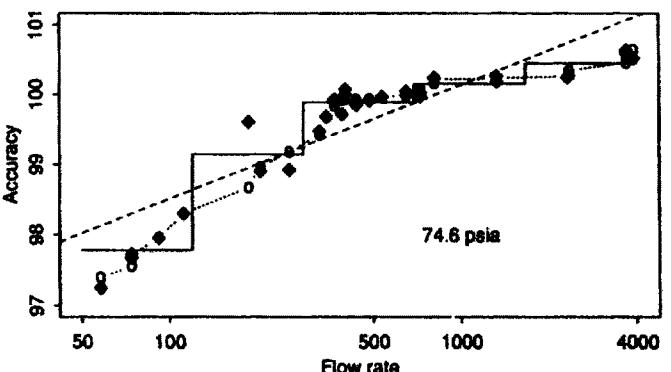


Figure 1.14 Accuracy of a natural gas meter as a function of the flow rate through the value at 74.6 psia. The raw data ($n = 33$) are shown by the filled diamonds. The least-squares fit is the dashed line; the regressogram is the solid line; and the Tukey smoother is shown as circles connected by a dotted line.

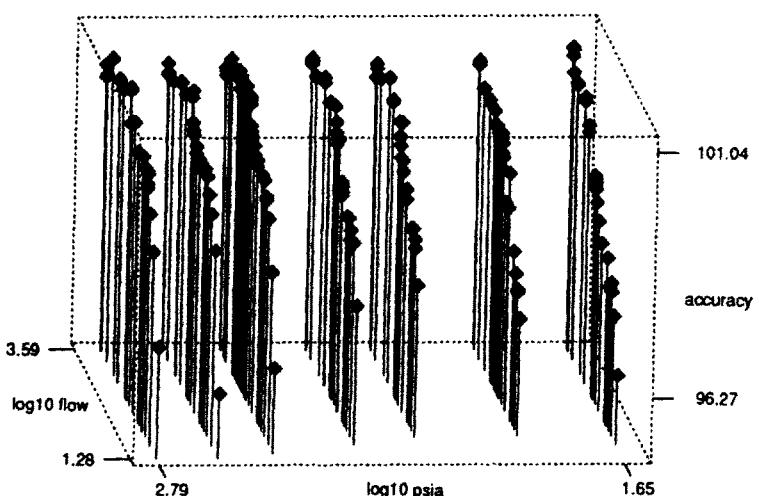


Figure 1.15 Complete 3-D view of the gas flow data.

visualization of 4-D surfaces would bring to the data analyst. In this section, effective visualization of surfaces in more than 3 dimensions is introduced.

Displaying a 3-dimensional perspective plot of the surface $f(x, y)$ of a bivariate function requires 1 more dimension than the corresponding bivariate contour representation; see Figure 1.16. There are trade-offs. The contour representation lacks the exact detail and visual impact available in a perspective plot; however, perspective plots usually have portions obscured by peaks and present less precise height information. One way of expressing the difference is to say that a contour plot displays, loosely speaking, about 2.6–2.9 dimensions of the entire 3-D surface (more, as more contour lines are drawn). Some authors claim that one or the other representation is superior, but it seems clear that both can be useful for complicated surfaces.

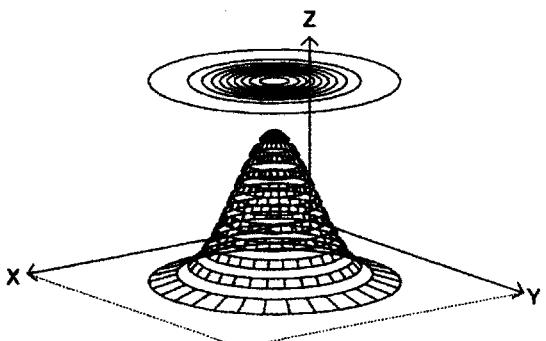


Figure 1.16 Perspective plot of bivariate Normal density with a “floating” representation of the corresponding contours.

The visualization advantage afforded by a contour representation is that it lives in the *same dimension* as the data, whereas a perspective plot requires an additional dimension. Hence with trivariate data, the third dimension can be used to present a 3-D contour. In the case of a density function, the corresponding 3-D contour plot is composed of one or more *α -level contour surfaces*, which are defined for $\mathbf{x} \in \mathbb{R}^d$ by

$$\alpha\text{-Contour: } S_\alpha = \{\mathbf{x}: f(\mathbf{x}) = \alpha f_{\max}\}, \quad 0 \leq \alpha \leq 1,$$

where f_{\max} is the maximum or modal value of the density function.

For Normal data, the general contour surfaces are hyper-ellipses defined by the easily verified equation (see Problem 14)

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = -2 \log \alpha. \quad (1.1)$$

A trivariate contour plot of $f(x_1, x_2, x_3)$ would generally contain several “nested” surfaces, $\{S_{0.1}, S_{0.3}, S_{0.5}, S_{0.7}, S_{0.9}\}$ for example. For the independent standard Normal density, the contours would be nested hyperspheres centered on the mode. In Figure 1.17, 3 contours of the trivariate standard Normal density are shown in stereo. Many if not most readers will have difficulty crossing their eyes to obtain the stereo effect. But even without the stereo effect, the 3 spherical contours are well-represented.

How effective is this in practice? Consider a smoothed histogram $\hat{f}(x, y, z)$ of 1,000 trivariate Normal points with $\boldsymbol{\Sigma} = I_3$. Figure 1.18 shows surfaces of 9 equally spaced bivariate slices of the trivariate estimate. Each slice is approximately bivariate Normal but without rescaling. Of course, the surfaces are not precisely bivariate Normal, due to the finite size of the sample.

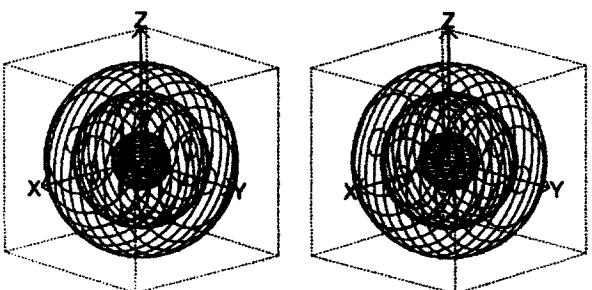


Figure 1.17 Stereo representation of 3 α -contours of a trivariate Normal density. Gently crossing your eyes should allow the 2 frames to fuse in the middle.

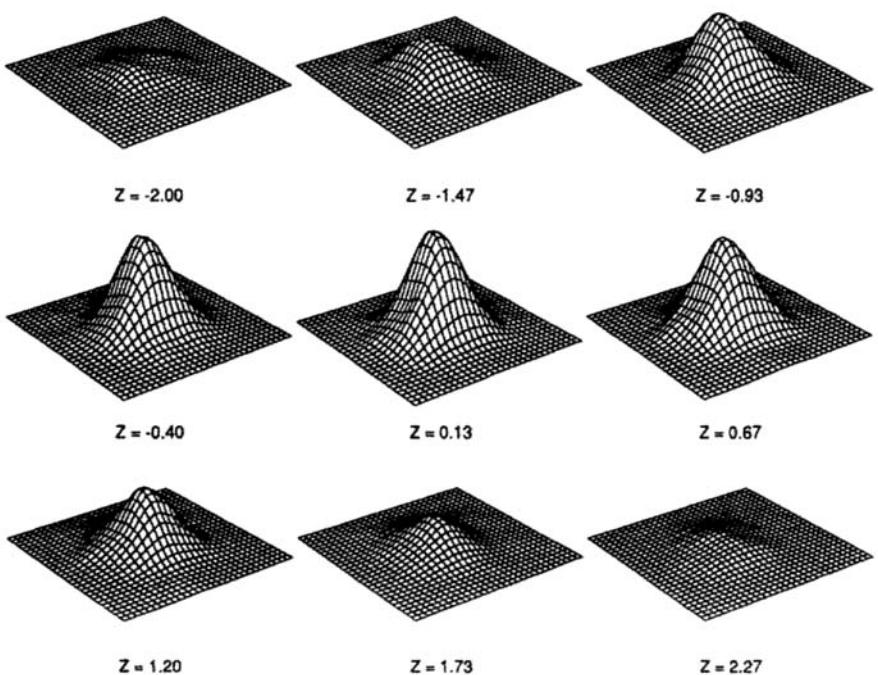


Figure 1.18 Sequence of bivariate slices of a trivariate smoothed histogram.

A natural question to pose is: Why not plot the corresponding sequence of *conditional densities*, $\hat{f}(x, y | z = z_0)$, rather than the *slices*, $\hat{f}(x, y, z_0)$? If this were done, all the surfaces in Figure 1.18 would be nearly identical. If the goal is to understand the 4-D density surface, then the sequence of conditional densities overemphasizes the (visual) importance of the tails and obscures information about the location of the “center” of the data. Furthermore, as nonparametric estimates in the tail will be relatively noisy, the estimates will be especially rough upon normalization; see Figure 1.19. For these reasons it seems best to look at slices and to reserve normalization for looking at conditional densities that are particularly interesting.

Several trivariate contour surfaces of the same estimated density are displayed in Color Plates 3–7. Clearly, the trivariate contours give an improved “big picture”—just as a rotating trivariate scatter diagram improved on 3 static bivariate scatter diagrams. The density estimate is a 4-D surface, and the trivariate contour view in Color Plates 6–7 may present only 3.5 dimensions, while the series of bivariate slices may yield a bit more, perhaps 3.75 dimensions, but without the visual impact. Examine the 3-D contour view for the LANDSAT data in Color Plates 9–10. The structure is quite complex. The presentation of clusters is stunning and shows multiple modes and multiple clusters. This detailed structure is not apparent in the scatterplot in Figure 1.3.

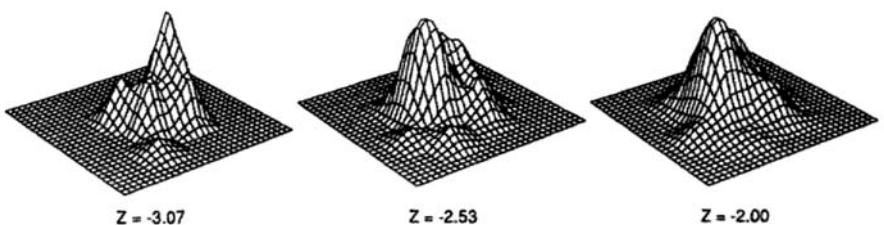


Figure 1.19 Normalized slices in the tail of the smoothed histogram in Figure 1.18.

Depending upon the nature of the variables, slicing can be attempted with 4-, 5-, or 6-dimensional data. Of special importance is the 5-D surface generated by 4-D data, for example, space-time variables such as the Mount St. Helens data in Figure 1.9. These higher-dimensional estimates can be animated in a fashion similar to Figure 1.18; see Scott and Wilks (1990). In the 4-D case, the α -level contours of interest are based on the slices:

$$S_{\alpha,t} = \{(x, y, z): f(x, y, z, t) = \alpha f_{\max}\},$$

where f_{\max} is the global maximum over the 5-D surface. For a fixed choice of α , as the slice value t changes continuously, the contour shells will expand or contract smoothly, finally vanishing for extreme values of t . For example, a single theoretical contour of the $N(0, I_4)$ density would vanish outside a symmetric interval around the origin, but within that interval, the contour shell would be a sphere centered on the origin with greatest diameter when $t = 0$. With several α -shells displayed simultaneously, the contours would be nested spheres of different radii, appearing at different values of t , but of greatest diameter when $t = 0$.

One particularly interesting slice of the smoothed 5-D histogram estimate of the entire iris data set is shown in Color Plate 2. The $\alpha = 5\%$ contour surface reveals two well-separated clusters. However, the $\alpha = 20\%$ contour surface is trimodal, revealing the true structure in this data set even with only 150 points. The Virginica and Versicolor data may not be separated in the point cloud but apparently can be separated in the density cloud.

With more than 4 variables, the most appropriate sequence of slicing is not clear. With 5 variables, bivariate contours of (x_4, x_5) may be drawn; then a sequence of trivariate slices may be examined tracing along one of these bivariate contours. With more than 5 or 6 variables, deciding where to slice at all is a difficult problem because the number of possibilities grows exponentially. That is why projection-based methods are so important; see Chapter 7.

Visualizing Multivariate Regression Functions

The same graphical representation can be applied to regression surfaces. However, the interpretation can be more difficult. For example, if the regression surface is monotone, the α -level contours of the surface will not be “closed”

and will appear to “float” in space. If the regression surface is a simple linear function such as $ax + by + cz$, then a set of trivariate α -contours will simply be a set of parallel planes. Practical questions arise that do not appear for density surfaces. In particular, what is the natural extent of the regression surface; that is, for what region in the design space should the surface be plotted? Perhaps one answer is to limit the plot to regions where there is sufficient data, i.e., where the density of design points is above a certain threshold.

1.4.4 Overview of Contouring and Surface Display

Suppose that a general bivariate function $f(x, y)$ (taking on positive and negative values) is sampled on a regular grid, and the $\alpha = 0$ contour S_0 is desired; that is, $S_0 = \{(x, y): f(x, y) = 0\}$. Label the values of the grid as $+$, 0 , $-$ depending on whether $f > 0$, $f = 0$, and $f < 0$, respectively. Then the desired contour is shown in Figure 1.20. The piecewise linear approximation and the true contour do not match along the bin boundaries since the interpolation is not exact.

However, bivariate contouring is not as simple a task as one might imagine. Usually, the function is sampled on a rectangular mesh, with no gradient information or possibility for further refinement of the mesh. If too coarse a mesh is chosen, then small local bumps or dips may be missed, or two distinct contours at the same level may be inadvertently joined. For speed and simplicity, one wants to avoid having to do any global analysis before drawing contours. A local contouring algorithm avoids multiple passes over the data. In any case, global analysis is based on certain smoothness assumptions and may fail. The difficulties and details of contouring are described more fully in Section A.1.

There are several varieties of 3-D contouring algorithms. It is assumed that the function has been sampled on a lattice, which can be taken to be cubical without loss of generality. One simple trick is to display a set of 2-D contour slices that result from intersecting the 3-D contour shell with a set of parallel planes along the lattice of the data, as was done in Figure 1.17 and Color Plate 2. In this representation, a single spherical shell becomes a set of circular

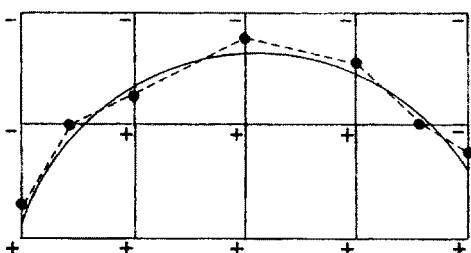


Figure 1.20 A portion of the bivariate contour at the $\alpha = 0$ level of a smooth function measured on a square grid.

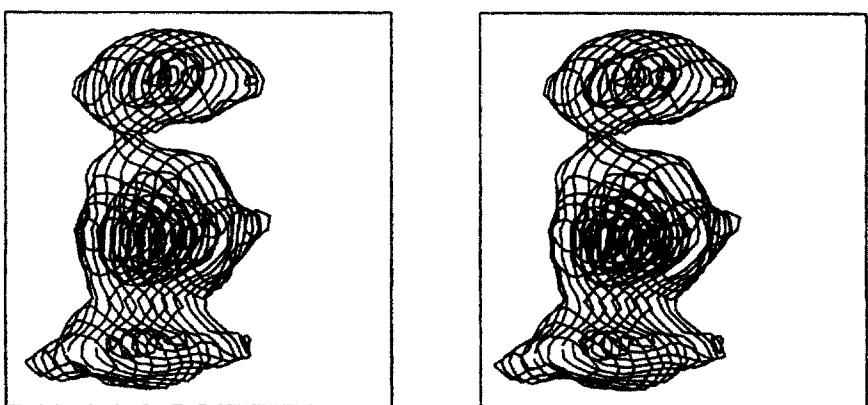


Figure 1.21 Simple stereo representation of four 3-D nested shells of the earthquake data.

contours (Figure 1.21). This approach has the advantage of providing a shell representation that is “transparent” so that multiple α -level contour levels may be visualized. Different colors can be used for different contour levels; see Scott (1983, 1984, 1991a), Scott and Thompson (1983), Härdle and Scott (1988), and Scott and Hall (1989).

More visually pleasing surfaces can be drawn using the *marching cubes* algorithm (Lorensen and Cline, 1987). The overall contour surface is represented by a large number of connected triangular planar sections, which are computed for each cubical bin and then displayed. Depending upon the pattern of signs on the 8 vertices of each cube in the data lattice, up to 6 triangular patches are drawn within each cube; see Figure 1.22. In general, there are 2^8 cases (each corner of the cube being either above or below the contour level). Taking into consideration certain symmetries reduces this number. By scanning through all the cubes in the data lattice, a collection of triangles is found that defines the contour shell. Each triangle has an inner and outer surface, depending upon the gradient of the density function. The inner and outer surfaces may be distinguished by color shading. A convenient choice is various shades of red for surfaces pointing towards regions of higher (hotter) density, and shades of blue toward regions of lower (cooler) density; see the cover jacket of this book for an example. Each contour is a patchwork of several thousand triangles, as in apparent in the color plates. Smoother surfaces may be obtained by using higher-order splines, but the underlying bin structure information would be lost.

In summary, visualizing trivariate functions directly is a powerful adjunct to data analysis. The gain of an additional dimension of visible structure without resort to slices greatly improves the ability of a data analyst to perceive structure. The same visualization applies to slices of density function with more than 3 variables. A demonstration tape that displays 4-D animation of $S_{\alpha,t}$, contours as α and t vary is available (Scott and Wilks, 1990).

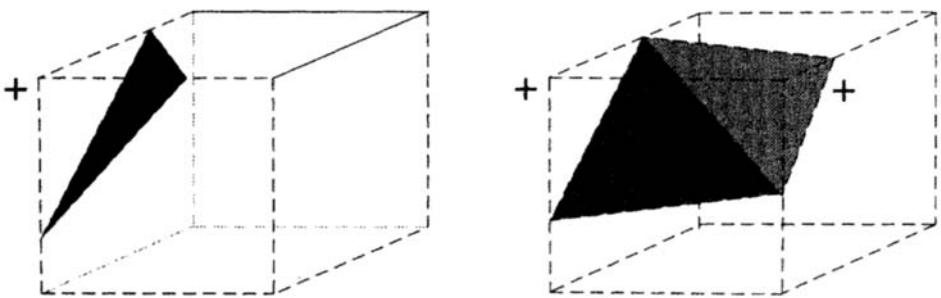


Figure 1.22 Examples of marching cube contouring algorithm. The corners with values above the contour level are labeled with a + symbol.

1.5 GEOMETRY OF HIGHER DIMENSIONS

The geometry of higher dimensions provides a few surprises. In this section, a few standard figures are considered. This material is available in scattered references; see Kendall (1961), for example.

1.5.1 Polar Coordinates in d Dimensions

In d dimensions, a point \mathbf{x} can be expressed in spherical polar coordinates by a radius r , a base angle θ_{d-1} ranging over $(0, 2\pi)$, and $d - 2$ angles $\theta_1, \dots, \theta_{d-2}$ each ranging over $(-\pi/2, \pi/2)$; see Figure 1.23. Let $s_k = \sin \theta_k$ and $c_k = \cos \theta_k$. Then the transformation back to Euclidean coordinates is given by

$$x_1 = r c_1 c_2 \cdots c_{d-3} c_{d-2} c_{d-1}$$

$$x_2 = r c_1 c_2 \cdots c_{d-3} c_{d-2} s_{d-1}$$

$$x_3 = r c_1 c_2 \cdots c_{d-3} s_{d-2}$$

$$\vdots$$

$$x_j = r c_1 \cdots c_{d-j} s_{d-j+1}$$

$$\vdots$$

$$x_d = r s_1 .$$

After some work (see Problem 11), the Jacobian of this transformation may be shown to be

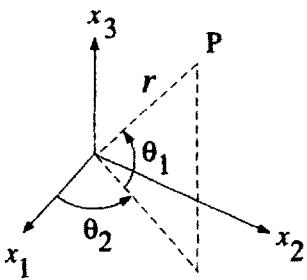


Figure 1.23 Polar coordinates (r, θ_1, θ_2) of a point P in \mathbb{R}^3 .

$$J = r^{d-1} c_1^{d-2} c_2^{d-3} \cdots c_{d-2}. \quad (1.2)$$

1.5.2 Content of Hypersphere

The volume of the d -dimensional hypersphere $\{\mathbf{x} : \sum_{i=1}^d x_i^2 \leq a^2\}$ is given by

$$\begin{aligned} V_d(a) &= \int_{\sum_{i=1}^d x_i^2 \leq a^2} 1 \, d\mathbf{x} \\ &= \int_0^a dr \int_{-\pi/2}^{\pi/2} d\theta_1 \int_{-\pi/2}^{\pi/2} d\theta_2 \cdots \int_0^{2\pi} d\theta_{d-1} r^{d-1} c_1^{d-2} c_2^{d-3} \cdots c_{d-2}. \end{aligned}$$

This can be simplified using the identity

$$\int_{-\pi/2}^{\pi/2} \cos^k \theta \, d\theta = 2 \int_0^{\pi/2} \cos^k \theta \, d\theta = 2 \int_0^{\pi/2} \cos^k \theta \frac{d(\cos^2 \theta)}{-2 \cos \theta \sin \theta},$$

which, using the change of variables $u = \cos^2 \theta$,

$$= \int_0^1 u^{k/2} \frac{du}{u^{1/2}(1-u)^{1/2}} = B\left(\frac{1}{2}, \frac{k+1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k+2}{2}\right)}$$

As $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$,

$$\begin{aligned} V_d(a) &= 2\pi \frac{a^d}{d} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-2}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \cdots \frac{\Gamma\left(\frac{1}{2}\right)\Gamma(1)}{\Gamma\left(\frac{3}{2}\right)} \\ &= \frac{a^d \pi^{d/2}}{\frac{d}{2} \Gamma\left(\frac{d}{2}\right)} = \frac{a^d \pi^{d/2}}{\Gamma\left(\frac{d}{2} + 1\right)}. \end{aligned} \quad (1.3)$$

1.5.3 Some Interesting Consequences

1.5.3.1 Sphere Inscribed in Hypercube

Consider the hypercube $[-a, a]^d$ and an inscribed hypersphere with radius $r = a$. Then the fraction of the volume of the cube contained in the hypersphere is given by

$$f_d = \frac{\text{volume sphere}}{\text{volume cube}} = \frac{\frac{2a^d \pi^{d/2}}{d \Gamma(d/2)}}{(2a)^d} = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)}.$$

For lower dimensions, the fraction f_d is as shown in Table 1.1. It is clear that the center of the cube becomes less important. As the dimension increases, the volume of the hypercube concentrates in its corners. This distortion of space (at least to our 3-dimensional way of thinking) has many potential consequences for data analysis.

1.5.3.2 Hypervolume of a Thin Shell

Wegman (1990) demonstrates the distortion of space in another setting. Consider 2 spheres centered on the origin, one with radius r and the other with slightly smaller radius $r - \epsilon$. Consider the fraction of the volume of the larger sphere in between the spheres. By Equation (1.3),

$$\frac{V_d(r) - V_d(r - \epsilon)}{V_d(r)} = \frac{r^d - (r - \epsilon)^d}{r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d \xrightarrow{d \rightarrow \infty} 1.$$

Hence, virtually all of the content of a hypersphere is concentrated close to its surface, which is only a $(d - 1)$ -dimensional manifold. Thus for data distributed uniformly over both the hypersphere and the hypercube, most of the data fall near the boundary and edges of the volume. Most statistical techniques exhibit peculiar behavior if the data fall in a lower-dimensional subspace. This example illustrates one important aspect of the *curse of dimensionality*, which is discussed in Chapter 7.

1.5.3.3 Tail Probabilities of Multivariate Normal

The preceding examples make it clear that if we are trying to view uniform data over the hypercube in \mathbb{R}^{10} , most (spherical) neighborhoods will be empty!

Table 1.1 Fraction of the Volume of a Hypercube Lying in the Inscribed Hypersphere

Dimension d	1	2	3	4	5	6	7
Fraction Volume f_d	1	0.785	0.524	0.308	0.164	0.081	0.037

Let us examine what happens if the data follow the standard d -dimensional Normal distribution

$$f_d(\mathbf{x}) = (2\pi)^{-d/2} e^{-\mathbf{x}^T \mathbf{x}/2}.$$

Clearly, the origin (mode) is the most likely point and the equiprobable contours are spheres. Consider the spherical contour, $S_{0.01}(\mathbf{x})$, where the density value is only 1% of the value at the mode. Now

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} = e^{-\mathbf{x}^T \mathbf{x}/2} \quad \text{and} \quad -2 \log \frac{f(\mathbf{x})}{f(\mathbf{0})} = \sum_{i=1}^d x_i^2 \sim \chi^2(d);$$

therefore, the probability that a point is *within* the 1% spherical contour may be computed by

$$\Pr\left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq \frac{1}{100}\right) = \Pr\left(\chi^2(d) \leq -2 \log \frac{1}{100}\right). \quad (1.4)$$

Equation (1.4) gives the probability a random point will not fall in the “tails” or, in other words, will fall in the medium- to high-density region. In Table 1.2, these probabilities are tabulated for several dimensions. Around 5 or 6 dimensions, the probability mass of a multivariate Normal begins a rapid migration into the extreme tails. In fact, more than half of the probability mass is in a very low density region for 10-dimensional data. Silverman (1986) has dramatized this in 10 dimensions by noting that $\text{Prob}(\|\mathbf{x}\| \geq 1.6) = 0.99$. In very high dimensions, virtually the entire sample will be in the tails in a sense consistent with low-dimensional intuition. Table 1.2 is also applicable to Normal data with a general full-rank covariance matrix, except that the contour is a hyper-ellipsoid.

1.5.3.4 Diagonals in Hyperspace

Pairwise scatter diagrams essentially project the multivariate data onto all the 2-dimensional faces. Consider the hypercube $[-1, 1]^d$ and let any of the diagonal vectors from the center to a corner be denoted by \mathbf{v} . Then \mathbf{v} is any of the 2^d vectors of the form $(\pm 1, \pm 1, \dots, \pm 1)^T$. The angle between a diagonal vector \mathbf{v} and a Euclidean coordinate axis $\mathbf{e}_j = (0, \dots, 0, 1, 0, \dots, 0)^T$ is given by

$$\cos \theta_d = \frac{\langle \mathbf{v}, \mathbf{e}_j \rangle}{\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle \langle \mathbf{e}_j, \mathbf{e}_j \rangle}} = \frac{\pm 1}{\sqrt{d}} \quad \xrightarrow{d \rightarrow \infty} 0,$$

Table 1.2 Probability Mass Not in the “Tail” of a Multivariate Normal Density

d	1	2	3	4	5	6	7	8	9	10	15	20
p^* 1,000	998	990	973	944	899	834	762	675	582	488	134	20

where $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$, so that $\theta_d \rightarrow \pi/2$ as $d \rightarrow \infty$. Thus the diagonals are nearly orthogonal to all coordinate axes for large d . Hence, any data cluster lying near a diagonal in hyperspace will be mapped into the origin in every paired scatterplot, while a cluster along a coordinate axis should be visible in some plot.

Thus the choice of coordinate system in high dimensions is critical in data analysis and intuition is highly dependent on a good choice. Real data structures may be missed due to overstriking. The general conclusion is that 1- to 2-dimensional intuition is valuable but not infallible when continuing on to higher dimensions.

PROBLEMS

1. (a) Devise a simple way of creating radially symmetric trivariate data with a “hole” in it; i.e., a region where the probability data points lie goes smoothly to zero at the center. *Hint:* Invent a rejection rule based on the distance a trivariate normal point is from the origin.
 (b) Study a pairwise scatter diagram of 5,000 trivariate data with either a “large” or a “small” hole in the middle. When does the hole become difficult to discern? Use “o” and “.” as plotting symbols. Plot a histogram of $(x_1^2 + x_2^2 + x_3^2)^{1/2}$ and see if the hole is apparent.
2. Try the Diaconis-Friedman idea of linked bivariate scatter diagrams using the iris data. Draw the scatterplots side-by-side and try connecting all points or random subsets. Evaluate your findings.
3. Use Chernoff faces on the college data in Table 2 in Appendix B. Try to assign variables to facial features in a memorable way. Compare your subjective choices of variables with those of others. You will notice that if one variable is near the extreme value of the data, it may distort that facial feature to such a degree that it is impossible to recognize the levels of other variables controlling different aspects of that feature. How should that influence your choice of variables for the mouth and eyes?
4. Display Andrews curves for the economic data in Table 1 in Appendix B for several permutations of the variables. How do these curves reflect the onset of the Depression after 1929?
5. Research problem: Generalize Andrews’ representation so that the representation of a multidimensional point is a trajectory in the 3-dimensional rectangle $[-\pi, \pi]^2 \times [0, 1]$.
6. Plot in parallel coordinates random samples of bivariate Normal data with correlations ranging from -1 to 1. When the correlation $\rho = +1$, where

does the point of intersection fall? Can you guess how trivariate correlated Normal data will appear in parallel coordinates? Try it.

7. Investigate the appearance in parallel coordinates of data with clusters. For example, generate bivariate data with clusters centered at $(0, 0)$ and $(3, 3)$. Try centers at $(0, 0)$ and $(3, 0)$. Try centers of 3 clusters at $(0, 0)$, $(1, 0)$, and $(2, 0)$, where the data in each cluster have $\rho = -0.9$. The last example shows the duality between clusters and holes.
8. Prove that points falling on a straight line in Euclidean coordinates intersect in a point in parallel coordinates. What is the one exception? Superimposing Euclidean coordinates upon the parallel axes as shown in the right frame of Figure 1.8, find the (Euclidean) coordinates of the intersection point.
9. Investigate the literature for other ideas of data representation, including the star diagram, linear profile, weathervane, polygon star, and Kleiner-Hartigan faces.
10. What are the possible types of intersection of 2 planes (2-D) in 4-space?
Hint: Consider the 2 planes determined by pairs of coordinate axes; see Wegman (1990).
11. Show that the Jacobian equals what is claimed in Equation (1.2).
12. Verify Equation (1.3) for the well-known cases of a circle and a sphere.
13. Think of another way to represent high-dimensional data. Try using some other set of orthogonal functions for Andrews' curves (step functions, Legendre polynomials, or others). How sensitive is your method to permutations of the coordinate axes?
14. Show that the α -level contours of a multi-Normal density are given by Equation (1.1). Use some of the techniques in Appendix A to display some contours when $d = 3$ with correlated and uncorrelated random variables.
15. (Problem 10 continued) What are the possible types of intersections of a k_1 -dimensional hyperplane and a k_2 -dimensional hyperplane in d dimensions? Think about the intersection of other types of hypersurfaces.
16. What fraction of a d -dimensional hypersphere lies in the inscribed d -dimensional hypercube? Find numerical values for dimensions up to 10.
17. Examine parallel coordinate plots of commonly observed bivariate and trivariate structure, including correlation and clustering. Summarize your findings.

CHAPTER 2

Nonparametric Estimation Criteria

The focus of nonparametric estimation is different from that of parametric estimation. In the latter case, given a parametric density family $f(\cdot|\theta)$, such as the two-parameter Normal family $N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$, the emphasis is on obtaining the best estimator $\hat{\theta}$ of θ . In the nonparametric case, the emphasis is directly on obtaining a good estimate $\hat{f}(\cdot)$ of the *entire* density function $f(\cdot)$. In this chapter, an introduction to nonparametric estimation criteria is given, using only tools familiar from parametric analysis.

R. A. Fisher and Karl Pearson engaged in a lively debate on aspects of parametric estimation, with Pearson arguing in favor of nonparametric curves. Nonparametric curves are driven by structure in the data and are broadly applicable. Parametric curves rely on model building and prior knowledge of the equations underlying the data. Fisher (1922, 1932) called the 2 phases of parametric estimation the problems of *specification* and *estimation*. Fisher focused on achieving optimality in the estimation phase. His only remark on the problem of specification was that it was “entirely a matter for the practical statistician” to choose a form that “we know how to handle” based on experience. Fisher noted that misspecification could be detected by an *a posteriori* test, but offered no further instructions. If a parametric model is overparameterized in an effort to provide greater generality, then many choices of the vector θ will give nearly identical pointwise estimates $\hat{f}(x)$. An incorrectly specified parametric model has a bias that cannot be removed by large samples alone. Determining if the bias is too large to retain the parametric model can be tricky, since goodness-of-fit tests almost always reject quite reasonable models with large samples. The “curse of optimality” is that incorrect application of “optimal” methods is preferred to more general but less efficient methods. What Pearson failed to argue persuasively is that optimal estimators can become inefficient with only small perturbations in the assumptions underlying the parametric model. The modern emphasis on robust estimation correctly sacrifices a small percentage of parametric optimality in order to achieve greater insensitivity to model misspecification. However, in many multivariate situations, only vague prior information on an appropriate

model is available. Nonparametric methods eliminate the need for model specification. The loss of efficiency need not be too large and is balanced by reducing the risk of misinterpreting data due to incorrect model specification.

2.1 ESTIMATION OF THE CUMULATIVE DISTRIBUTION FUNCTION

The simplest function to estimate nonparametrically is the cumulative distribution function (cdf) of a random variable X , defined by

$$F(x) = \Pr(X \leq x).$$

The obvious estimator from elementary probability theory is the *empirical cumulative distribution function* (ecdf), defined as

$$F_n(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{\#x_i \in (-\infty, x]}{n} = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i), \quad (2.1)$$

where the $\{x_1, x_2, \dots, x_n\}$ is a random sample from F and

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

This function has a staircase shape, as shown in Figure 2.1 of the geyser data given in Table 6 in Appendix B. It is easy to see that $F_n(x)$ has excellent mathematical properties for estimating the level of the function $F(x)$ for each fixed ordinate value x :

$$EF_n(x) = EI_{(-\infty, x]}(X) = 1 \times \Pr(X \in (-\infty, x]) = F(x).$$

In fact, $nF_n(x)$ is a binomial random variable, $B(n, p)$, with $p = F(x)$, so that $\text{Var}\{F_n(x)\} = p(1-p)/n$. There are no other unbiased estimators with smaller variance. This result follows since the order statistics form a complete sufficient statistic, and $F_n(x)$ is both unbiased and a function of the sufficient statistic. But notice that while the distribution function is often known to be continuous, the optimal estimator $F_n(x)$ is not.

A reasonable question to ask is whether the distribution function or its associated *probability density function* (pdf), $f(x) = F'(x)$, should be the focus for data analysis. Certainly, the distribution function is quite useful. On the one hand, quite ordinary features such as skewness and multimodality are much more easily perceived in a graph of the density function than in a graph of the distribution function—compare Figures 1.12 and 2.1. On the other hand, it is difficult to ignore the fact that many social scientists prefer to estimate the cdf rather than the pdf.

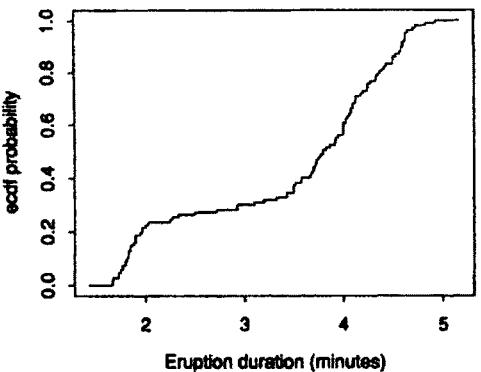


Figure 2.1 Empirical cumulative distribution function of the Old Faithful geyser data.

The distinction between the utility of the cdf and pdf becomes clearer in more than 1 dimension. The definition of a multivariate empirical cdf is simply

$$F_n(\mathbf{x}) = \frac{\#\{\mathbf{x}_i \leq \mathbf{x}\}}{n}, \quad \mathbf{x} \in \mathfrak{R}^d,$$

where the inequality $\{\mathbf{x}_i \leq \mathbf{x}\}$ is applied componentwise for each data point. The same optimality property holds as in the univariate case. However, few statisticians have even *seen* a bivariate empirical cdf. Consider the bivariate ecdf in Figure 2.2 of the geyser data corresponding to the bivariate pdf shown in Figure 1.12. Only a random subset of 30 pairs of the data $\{\mathbf{x}_i, \mathbf{x}_{i+1}\}$ is shown in this figure. The trimodal feature can be recognized with a little reflection, but not easily. The surface is perhaps unexpectedly complex given the small sample size. In particular, the number of jumps in the function is considered in Problem 1. Thus, the multivariate distribution function is of little interest for either graphical or data analytical purposes. Furthermore, ubiquitous multivariate statistical applications such as regression and classification rely on direct manipulation of the density function and not the distribution function.

2.2 DIRECT NONPARAMETRIC ESTIMATION OF THE DENSITY

Following the theoretical relationship $f(x) = F'(x)$, the *empirical probability density function* (epdf) is defined to be the derivative of the empirical cumulative distribution function:

$$f_n(x) = \frac{d}{dx} F_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - x_i), \quad (2.2)$$

where $\delta(t)$ is the Dirac delta function. It is always a discrete Uniform density over the data, that is, the probability mass is n^{-1} at each data point. The epdf

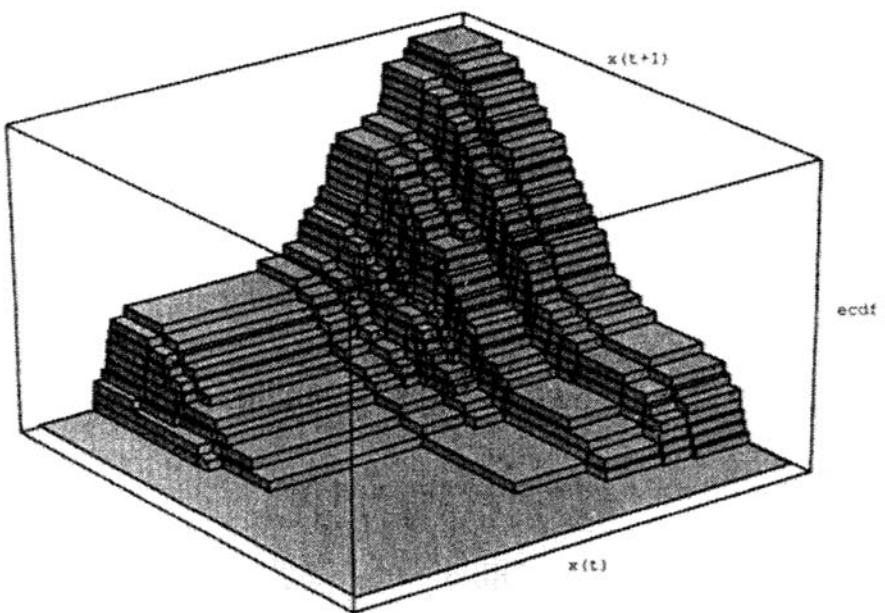


Figure 2.2 Empirical bivariate cdf of a random subset of 30 points of the lagged Old Faithful data $\{x_t, x_{t+1}\}$.

is like a 1-dimensional scatter diagram (dot plot) and is a useless estimate if the density is continuous; see Figure 2.3. The epdf is clearly inferior to the histogram from a graphical point of view. The primary modern use of the epdf has been as the sampling density for the bootstrap (Efron, 1982).

Does a uniformly minimum variance unbiased estimator of $f(x)$ exist? In the first theoretical treatment of the subject, Rosenblatt (1956) proved that no such estimator existed. Let X_n be a random sample of size n from f . Suppose that an estimator $T_n(x; X_n)$ existed such that $E[T_n(x; X_n)] = f(x)$ for all continuous f and for all x and n . The data must appear symmetrically in $T_n(x; X_n)$, as the nonparametric estimate cannot vary with the order in which the sample was collected. Now for all intervals (a, b) ,

$$E \left[\int_a^b T_n(x) dx \right] = \int_a^b f(x) dx = F(b) - F(a) = E[F_n(b) - F_n(a)],$$

using Fubini's theorem to justify switching the order of the expectation and integral operators. Both $T_n(x; X_n)$ and $F_n(b) - F_n(a)$ are functions of the complete sufficient statistics, and since $F_n(b) - F_n(a)$ is the only symmetric unbiased estimator of $F(b) - F(a)$, it follows that

$$F_n(b) - F_n(a) = \int_a^b T_n(x) dx$$

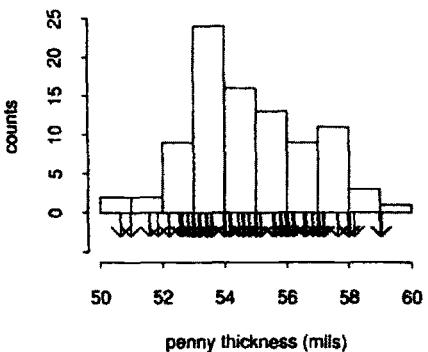


Figure 2.3 A histogram and empirical pdf (pointing down) of the U.S. penny thickness data.

for almost all samples $\{\mathbf{X}_n\}$. This result is a contradiction since the right-hand side is absolutely continuous while the left-hand side is not. At the time, this result was surprising and disappointing in a profession that had become accustomed to the pursuit of optimal unbiased estimators. Today, biased estimators are sometimes preferred in situations where unbiased estimators are available, for example, with shrinkage estimators such as ridge regression (Hoerl and Kennard, 1970) and Stein (1956) estimators.

2.3 ERROR CRITERIA FOR DENSITY ESTIMATES

The desire to compare different estimators and to attempt to identify the best one assumes the specification of a criterion that can be optimized. Optimality is not an absolute concept but is intimately linked to the choice of a criterion. Criterion preference is largely subjective, although certain theoretical or intuitive arguments can be introduced. However, total elimination of the subjective element of nonparametric estimation seems undesirable; for example, it will be shown that the amount of noise in “optimal” histograms can evoke a negative response with very large data sets. In the parametric world, an optimal estimator is likely to be optimal for any related purpose. In the nonparametric world, an estimator may be optimal for one purpose and awful for another. This extra work is a price to be paid for working with a more general class of estimators.

When approximating parameters with biased estimators, the variance criterion is often replaced with the mean squared error (MSE), which is the sum of variance and squared bias. For pointwise estimation of a density function by the estimator $\hat{f}(x)$,

$$\text{MSE}\{\hat{f}(x)\} = E[\hat{f}(x) - f(x)]^2 = \text{Var}\{\hat{f}(x)\} + \text{Bias}^2\{\hat{f}(x)\},$$

where $\text{Bias}\{\hat{f}(x)\} = E[\hat{f}(x)] - f(x)$. This equation treats the nonparametric density estimation problem as a standard point estimation problem with

unknown parameter $\theta = f(x)$. While such pointwise analyses will prove interesting, the clear emphasis in this book will be on estimating and displaying the entire density surface. For some the most intuitively appealing global criterion is the L_∞ norm:

$$\sup_x |\hat{f}(x) - f(x)|.$$

At the other end of the spectrum is the L_1 norm:

$$\int |\hat{f}(x) - f(x)| dx .$$

Neither of these criteria is as easily manipulated as the L_2 norm, which in this context is referred to as the *integrated squared error* (ISE):

$$\text{ISE} = \int [\hat{f}(x) - f(x)]^2 dx . \quad (2.3)$$

Even so, the integrated squared error is a complicated random variable that depends on the true unknown density function, the particular estimator, and the sample size; furthermore, even with these 3 quantities fixed, the ISE is a function of the particular realization of n points. For most purposes, it will be sufficient to examine the average of the ISE over these realizations; that is, the mean of the random variable ISE or *mean integrated squared error* (MISE):

$$\begin{aligned} \text{MISE} &= E[\text{ISE}] = E\left\{\int [\hat{f}(x) - f(x)]^2 dx\right\} \\ &= \int E[\hat{f}(x) - f(x)]^2 dx = \int \text{MSE}\{\hat{f}(x)\} dx = \text{IMSE} , \end{aligned}$$

where the interchange of the integral and expectation operators is justified by an application of Fubini's theorem. The last quantity is the IMSE, which is an abbreviation for the *integrated mean squared error*. Thus the MISE error criterion has two different though equivalent interpretations: it is a measure of both the average global error and the accumulated pointwise error. This criterion could be modified by including a weight function that would emphasize the tails, for example, or perhaps a local interval. Many students may be tempted to compute the definition of the MISE by expanding

$$E\left\{\int [\hat{f}(x) - f(x)]^2 dx\right\} \stackrel{?}{=} \int E[\hat{f}(x) - f(x)]^2 f(x) dx ,$$

which is a weighted average MSE criterion. But this is *wrong!*

Other possible candidates for measuring error include information numbers such as the Kullback-Leibler criterion, which is defined as $\int \hat{f} \log(\hat{f}/f)$,

Hellinger distance, which is defined as $[\int (\hat{f}^{1/p} - f^{1/p})^p]^{\frac{1}{p}}$, Akaike's information criterion, or even other L_p distances. These alternatives are not pursued further here.

2.3.1 MISE for Parametric Estimators

Few statisticians have much intuition about the MISE criterion, since it is almost never discussed in the parametric framework. However, it is entirely possible to evaluate the quality of *parametric* density estimators by the MISE of the entire parametric density estimate rather than the MSE of the parameter alone. In most cases, both the MSE of the parameter and the MISE of the density estimate decrease at the rate $O(n^{-1})$ as the sample size increases.

2.3.1.1 Uniform Density Example

Consider the Uniform density $f = U(0, \theta)$, where θ is estimated by the maximum likelihood estimator $\hat{\theta} = x_{(n)}$, the n th-order statistic. Thus $\hat{f} = U(0, x_{(n)})$. Without loss of generality, choose $\theta = 1$. Following Equation (2.3),

$$\text{ISE} = \left(\frac{1}{x_{(n)}} - 1 \right)^2 \cdot x_{(n)} + (0 - 1)^2 \cdot (1 - x_{(n)}) = \frac{1}{x_{(n)}} - 1.$$

As $f(x_{(n)}) = nx_{(n)}^{n-1}$, $0 < x_{(n)} < 1$, it follows that

$$\text{MISE} = \int_0^1 \left(\frac{1}{x_{(n)}} - 1 \right) \cdot nx_{(n)}^{n-1} dx_{(n)} = \frac{1}{n-1} = O(n^{-1}). \quad (2.4)$$

Consider the one-dimensional family of estimators $\hat{f} = U(0, cx_{(n)})$, where c is a positive constant. There are two cases depending on whether $cx_{(n)} < 1$ or not, but the ISE is easily computed in each instance; see Problem 2. Taking expectations yields

$$\text{MISE}(c) = \begin{cases} \frac{n}{[(n-1)c]} - 1 & c < 1 \\ \frac{[2 - nc^{n-1} + (n-1)c^n]/[(n-1)c^n]}{c} & c > 1. \end{cases} \quad (2.5)$$

As both $\text{MISE}(c)$ and its derivative are continuous at $c = 1$ with $\text{MISE}'(c = 1) = -n/(n-1)$, the minimum MISE is realized for some $c > 1$, namely,

$$c^* = 2^{1/(n-1)} \approx 1 + n^{-1} \log 2.$$

This result is similar to the well-known fact that the MSE of the parametric estimator is minimized when $c = [1 + (n+1)^{-1}]$. However, in this case, the parametric MSE converges at the unusually fast rate $O(n^{-2})$; see Romano and Siegel (1986, p. 212). The $O(n^{-1})$ MISE rate reflects more accurately the average error for the entire density curve and not just its endpoint.

2.3.1.2 General Parametric MISE Method with Gaussian Application

Consider an unbiased estimator $\hat{\theta}$ of a parameter θ . Writing the parametric ISE as $I(\hat{\theta})$ and taking a Taylor's series gives us

$$I(\hat{\theta}) = \int [f(t|\hat{\theta}) - f(t|\theta)]^2 dt = \sum_k \frac{1}{k!} (\hat{\theta} - \theta)^k I^{(k)}(\theta).$$

Now $I(\theta) = 0$ and $I'(\theta) = 2 \int [f(t|\theta) - f(t|\theta)] f'(t|\theta) dt = 0$ as well; hence

$$\text{MISE}(\theta) = E[I(\hat{\theta})] = \frac{1}{2} \text{Var}(\hat{\theta}) I''(\theta) + \dots.$$

Omitting higher-order terms in the MISE expansion leaves the *asymptotic mean integrated squared error* (AMISE), which is given by

$$\text{AMISE}(\theta) = \frac{1}{2} I''(\theta) \text{Var}(\hat{\theta}).$$

This result can easily be extended to a vector of parameters by a multivariate Taylor's expansion. Consider estimation of the 2-parameter Normal density

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right],$$

and an estimator of the form

$$\hat{\phi}(x) = \phi(x; u, v);$$

for example, $\hat{\phi}(x) = \phi(x; \bar{x}, \hat{s}^2)$ used unbiased estimates of μ and σ^2 . If the true density is a standard Normal, then the bivariate ISE may be expanded as

$$\begin{aligned} I(u, v) &\sim I(0, 1) + u I_u + (v - 1) I_v + \frac{u^2}{2} I_{uu} + u(v - 1) I_{uv} \\ &\quad + \frac{(v - 1)^2}{2} I_{vv}, \end{aligned} \tag{2.6}$$

where the partial derivatives ($I_u, I_v, I_{uu}, I_{uv}, I_{vv}$) are evaluated at $(u, v) = (0, 1)$. Then (see Problem 4)

$$\begin{aligned} I_u(u, v) &= 2 \int (\hat{\phi} - \phi) \left(\frac{x - u}{v} \hat{\phi} \right) dx \\ I_{uu}(u, v) &= 2 \int (\hat{\phi} - \phi) \left(\frac{x - u}{v} \hat{\phi} \right)' dx + 2 \int \left(\frac{x - u}{v} \hat{\phi} \right)^2 dx. \end{aligned} \tag{2.7}$$

Now the factor $(\hat{\phi} - \phi) = 0$ when evaluated at $(u, v) = (0, 1)$. The final integral becomes $2 \int [x \phi(x; 0, 1)]^2 dx = 1/(2\sqrt{\pi})$. Similarly, $I_{vv}(0, 1) = 3/(16\sqrt{\pi})$ and $I_{uv} = 0$. Hence, taking the expectation of Equation (2.6) yields

$$\text{AMISE} = \frac{1}{2} I_{uu} \text{Var}(u) + \frac{1}{2} I_{vv} \text{Var}(v) \approx \frac{1}{4n\sqrt{\pi}} + \frac{3}{16n\sqrt{\pi}} = \frac{7}{16n\sqrt{\pi}}, \quad (2.8)$$

since the variances of \bar{x} and s^2 are $1/n$ and $2/(n-1) \approx 2/n$, respectively. It is of interest to note that the AMISE in Equation (2.8) is the sum of

$$\text{AMISE}\{\phi(x; \bar{x}, 1)\} + \text{AMISE}\{\phi(x; 0, s^2)\} = \frac{1}{4n\sqrt{\pi}} + \frac{3}{16n\sqrt{\pi}}. \quad (2.9)$$

Therefore, the parametric MISE is larger if the mean is unknown than if the variance is unknown. This simple calculation has been verified by a Monte Carlo simulation estimate of $1/(4n)$, which is only 1.012 times as large.

2.3.2 The L_1 Criterion

2.3.2.1 L_1 versus L_2

Intuitively, one appeal of the L_1 criterion $\int |\hat{f} - f| dx$ is that it pays more attention to the tails of a density than the L_2 criterion, which de-emphasizes the relatively small density values there by squaring. This de-emphasis can be checked in specific cases and by simulation. Theoretically, the L_1 criterion enjoys several other important advantages. First, consider a dimensional analysis. Since a density function has *inverse length* as its unit, L_1 is a dimensionless quantity after integration. The L_2 criterion, on the other hand, retains the units of inverse length after integration of the squared density error. It is possible to try to make L_2 dimensionless in a manner similar to the construction of such quantities for the skewness and kurtosis, but no totally satisfactory method is available; see Section 7.2.1. Furthermore, L_1 is invariant to monotone continuous changes of scale, as shown below. It is also easy to see that $0 \leq L_1 \leq 2$ while $0 \leq L_2 \leq \infty$; see Problem 7. On the other hand, in practical situations the estimators that optimize these criteria are similar. The point to remember is that there is not a canonical L_1 -method or L_2 -method estimator. Devroye and Györfi (1985) have compiled an impressive theoretical treatment based on L_1 error, but the analytical simplicity of squared error and its adequacy in practical applications makes it the criterion of choice here. Some recent asymptotic results for L_1 estimates by Hall and Wand (1988a) and Scott and Wand (1991) support the notion that the practical differences between L_1 and L_2 criteria are reasonably small except in extreme situations.

2.3.2.2 Three Useful Properties of the L_1 Criterion

Several theoretically appealing aspects of absolute error are discussed in detail below. Each holds in the multivariate setting as well (Devroye, 1987).

The first and most appealing property is its interpretability, since L_1 is dimensionless and invariant under any smooth monotone transformation. Hence it is possible to compare the relative difficulty of estimating different densities. To see this, suppose $X \sim f$ and $Y \sim g$, and define $X^* = h(X)$ and $Y^* = h(Y)$; then $f^* = f[h^{-1}(x^*)] |J|$ with a similar expression for g^* , where J is the Jacobian. A simple change of variables gives

$$\begin{aligned} \int_u |f^*(u) - g^*(u)| du &= \int_u |f[h^{-1}(u)] - g[h^{-1}(u)]| |J| du \\ &= \int_v |f(v) - g(v)| dv. \end{aligned}$$

An incorrect interpretation of this result is to conclude that all measurement scales are equally difficult to estimate or that variable transformation does not affect the quality of a density estimate.

The second result is called Scheffé's lemma:

$$2 \sup_A \left| \int_A f - \int_A g \right| = \int |f - g| = 2 \int_{f > g} (f - g). \quad (2.10)$$

To prove this, consider the set $B = \{x: f(x) > g(x)\}$; then for any set A ,

$$\begin{aligned} 2 \int_A (f - g) &= 2 \int_{AB} (f - g) - 2 \int_{AB^c} (g - f) \leq 2 \int_B (f - g) - 0 \\ &= \int_B (f - g) + \left(1 - \int_{B^c} f\right) - \left(1 - \int_{B^c} g\right) \\ &= \int_B (f - g) + \int_{B^c} (g - f) = \int |f - g|. \end{aligned}$$

which establishes the second equality in the lemma; taking the supremum over A establishes the " \leq " for the first equality. The " \geq " follows directly from

$$2 \sup_A \left| \int_A (f - g) \right| \geq 2 \int_{A=B} (f - g).$$

The equality (2.10) provides a connection with statistical classification. Suppose that data come randomly from 2 densities, f and g , and that a new point x is to be classified. Using a Bayesian rule of the form: assign x to f if $x \in A$, for some set A , and to g otherwise, the probability of misclassification is

$$\begin{aligned} \Pr(\text{error}) &= \frac{1}{2} \int_{A^c} f + \frac{1}{2} \int_A g = \frac{1}{2} \left(1 - \int_A f\right) + \frac{1}{2} \int_A g \\ &= \frac{1}{2} - \frac{1}{2} \int_A (f - g). \end{aligned}$$

Choosing A to minimize this error leads to $A = B$ using the lemma above and gives us the third result

$$\Pr(\text{error}) = \frac{1}{2} - \frac{1}{4} \int |f - g|. \quad (2.11)$$

Thus minimizing the L_1 distance between $g = \hat{f}$ and f is equivalent to maximizing the probability in (2.11); that is, optimizing L_1 amounts to maximizing the *confusion* between \hat{f} and f . This optimization is precisely what is desired. The probability interpretation holds in any dimension. In Problem 8, this expression is verified in the 2 extreme cases.

2.4 NONPARAMETRIC FAMILIES OF DISTRIBUTIONS

2.4.1 Pearson Family of Distributions

Karl Pearson provided impetus for nonparametric estimators in two ways. He coined the word *histogram*, and he studied the density functions that are solutions to the differential equation

$$\frac{d \log f(x)}{dx} = \frac{x - a}{b + cx + dx^2}. \quad (2.12)$$

Pearson (1902a, b) identified 7 types of solutions to this equation, depending upon the roots of the denominator and which parameters were 0. Interestingly, this class contains most of the important classical distributions: Normal, Student's t , Beta, and Snedecor's F ; see Problem 9. Pearson proposed using the first 4 sample moments to estimate the unknown constants (a, b, c, d) in Equation (2.12). Today, maximum likelihood might be used.

Pearson motivated the differential equation (2.12) by appealing to the discrete hypergeometric distribution, which describes a large urn with N balls, pN of which are black. A sample of n balls is drawn, and the number of black balls is recorded as X ; then

$$f(x) = \Pr(X = x) = \binom{Np}{x} \binom{N(1-p)}{n-x} / \binom{N}{n}, \quad x = 0, \dots, n.$$

Pearson's differential equation (2.12) emerges after a bit of computation (see Problem 11), keeping p fixed as the urn grows:

$$\frac{d \log f(x)}{dx} \approx \frac{\Delta f(x)}{f(x)} = \frac{f(x) - f(x-1)}{f(x)} = \frac{x - a}{b + cx + dx^2}, \quad (2.13)$$

where

$$a = b = \frac{-(n + 1)(Np + 1)}{N + 2}; \quad c = \frac{(Np + n + 2)}{N + 2}; \quad d = \frac{-1}{N + 2}. \quad (2.14)$$

Pearson's contribution and insight was to use the data not to fit a particular parametric form of a density but rather to compute the coefficients of the differential equation from the data via the sample moments. Mathematically, the Pearson family could be considered parametric, but philosophically, it is nonparametric. Pearson devoted considerable resources to computing percentiles for his distributions. Given the lack of computing power at that time, it is remarkable that the Pearson family gained wide acceptance. It still appears regularly in practical applications today.

Many other families of distributions, some multivariate, have been proposed. The Johnson (N.L. Johnson, 1949) family is notable together with the multivariate generalizations of Marshall and Olkin (1985).

2.4.2 When Is an Estimator Nonparametric?

A parametric estimator is defined by the model $\hat{f}(x|\theta)$, where $\theta \in \Theta$. It has proven surprisingly difficult to formulate a working definition for what constitutes a nonparametric density estimator. A heuristic definition may be proposed based on the necessary condition that the estimator "work" for a "large" class of true densities. One useful notion is that a nonparametric estimator should have many parameters, in fact, perhaps an infinite number, or a number that diverges as a function of the sample size. The Pearson family does not fully qualify as nonparametric under any of these definitions, although the dimension of the Pearson family is larger than for most parametric families. Tapia and Thompson (1978) tend to favor the notion that the nonparametric estimator should be infinite dimensional. Silverman (1986) simply indicates that a nonparametric approach makes "less rigid assumptions . . . about the distribution of the observed data." But how many parameters does a histogram have? What about the naive orthogonal series estimator with an infinite number of terms that is equal (in distribution) to the empirical density function for any sample size? (See Section 6.1.3.)

A surprisingly elegant definition is implicit in the work of Terrell (Terrell and Scott, 1992), who shows that all estimators, parametric or nonparametric, are *generalized kernel estimators*, at least asymptotically; see Section 6.4. Terrell introduces the idea of the influence of a data point x_i on the point density estimate at x . If $\hat{f}(x)$ is a nonparametric estimator, the influence of a point should vanish asymptotically if $|x - x_i| > \epsilon$ for any $\epsilon > 0$, while the influence of distant points does not vanish for a parametric estimator. Roughly speaking, nonparametric estimators are asymptotically local, while parametric estimators are not. However, nonparametric estimators must not be *too* local in order to be consistent.

PROBLEMS

1. Devise an algorithm for plotting the empirical bivariate cdf. Give a range for the possible number of jumps there can be in this function. Give simple examples for the 2 extreme cases. What is the order of your algorithm (i.e., the number of arithmetic operations as a function of the sample size)?
2. Verify Equation (2.5). Plot it for several sample sizes and compare the actual minimizer to the asymptotic formula.
3. Show that the expected Kullback-Leibler distance for the parametric estimator $\hat{f} = U(0, x_{(n)})$ of $f = U(0, 1)$ is $1/n$.
4. Verify the equations in (2.7) and that $I_{vv}(0, 1) = 3/(16\sqrt{\pi})$.
5. Complete the calculations for the Normal parametric AMISE. Find the optimal AMISE estimators of the form $c\bar{x}$ and $c\hat{s}^2$ for standard Normal data.
6. Assuming standard Normal data, compute the *exact* MISE of the estimator $N(\bar{x}, 1)$, and check that the series approximations match. Research problem: Can you find any similar closed-form MISE expressions for the estimators $N(0, \hat{s}^2)$ and $N(\bar{x}, \hat{s}^2)$?
7. Verify that the ranges of L_1 and L_2 errors are $(0, 2)$ and $(0, \infty)$, respectively. Give examples where the extreme values are realized.
8. Verify that the probability of error given in (2.11) is correct in the 2 extreme cases.
9. Verify directly that the parametric densities mentioned in the text actually satisfy the differential equation for Pearson's system.
10. Use a symbolic program to verify that

$$f(x) \propto (1 + x^2)^{-k} \exp(-\alpha \arctan x)$$

is a possible solution to Pearson's differential equation. Plot $f(x)$.

11. Verify the calculations in Equations (2.13) and (2.14).

CHAPTER 3

Histograms: Theory and Practice

The framework of the classic histogram is useful for conveying the general flavor of nonparametric theory and practice. From squared-error theory to state-of-the-art cross-validation algorithms, these topics should be studied carefully. Discussing these issues in the context of the simplest and best known nonparametric density estimator is of practical interest, and will be of value when developing the theory of more complicated estimators.

The histogram is most often displayed on a nondensity scale: either as bin counts or as a stem-and-leaf plot (Tukey, 1977). Some authors have drawn a distinction between the purposes of a histogram as a density estimator and as a data presentation device (Emerson and Hoaglin, 1983, p. 22). Such a distinction seems artificial, as there is no reason to disregard a histogram constructed in a less than optimal fashion. The examination of both undersmoothed and oversmoothed histograms should be routine. A histogram conveys visual information of both the frequency and relative frequencies of observations; that is the essence of a density function.

3.1 STURGES' RULE FOR HISTOGRAM BIN WIDTH SELECTION

The classical frequency histogram is formed by constructing a complete set of nonoverlapping intervals, called *bins*, and counting the number of points in each bin. In order for the bin counts to be comparable, the bins should all have the same width. If so, then the histogram is completely determined by two parameters, the *bin width*, h , and the *bin origin*, t_0 , which is any conveniently chosen bin interval endpoint. Often the bin origin is chosen to be $t_0 = 0$.

Although the idea of grouping data in the form of a histogram is at least as old as Graunt's work in 1662, no systematic guidelines for designing histograms were given until Herbert Sturges' short note in 1926. His work made use of a device that has been advocated more generally by Tukey (1977); namely, taking the Normal density as a point of reference when thinking about data. Sturges simply observed that the binomial distribution, $B(n, p = 0.5)$, could

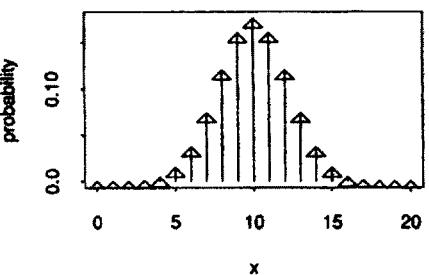


Figure 3.1 Binomial pdf with $p = 0.5$ used by Sturges to determine the number of histogram bins.

be used as a model of an optimally constructed histogram with appropriately scaled Normal data; see Figure 3.1.

Construct a frequency histogram with k bins, each of width 1 and centered on the points $i = 0, 1, \dots, k - 1$. Choose the bin count of the i th bin to be the Binomial coefficient $\binom{k-1}{i}$. As k increases, this ideal frequency histogram assumes the shape of a Normal density with mean $(k - 1)/2$ and variance $(k - 1)/4$. The total sample size is

$$n = \sum_{i=0}^{k-1} \binom{k-1}{i} = (1 + 1)^{k-1} = 2^{k-1}$$

by the Binomial expansion. Sturges' rule follows immediately:

Sturges' number-of-bins rule: $k = 1 + \log_2 n$. (3.1)

In practice Sturges' rule is applied by dividing the sample range of the data into the prescribed number of equal-width bins. Technically, Sturges' rule is a number-of-bins rule rather than a bin-width rule. Much simpler is to adopt the convention that all histograms have an infinite number of bins, only a finite number of which are nonempty. Furthermore, adaptive histograms, which are considered in Section 3.2.8, do not use equal-width bins. Thus the focus on bin width rather than number of bins seems appropriate.

Sturges' rule is widely recommended in introductory statistics texts and is often used in statistical packages as a default. If the data are not Normal, but are skewed or kurtotic, additional bins may be required. For example, Doane (1976) proposed increasing the number of bins in (3.1) by $\log_2(1 + \hat{\gamma}\sqrt{n}/6)$, where $\hat{\gamma}$ is an estimate of the standardized skewness coefficient; see Problem 1.

3.2 THE L_2 THEORY OF UNIVARIATE HISTOGRAMS

3.2.1 Pointwise Mean Squared Error and Consistency

In this section are presented the mean squared error properties of a density histogram. The difference between a frequency histogram and a density histogram is that the latter is normalized to integrate to 1. As noted above, the histogram is completely determined by the sample $\{x_1, \dots, x_n\}$ from $f(x)$ and a choice of mesh $\{t_k, -\infty < k < \infty\}$. Let $B_k = [t_k, t_{k+1})$ denote the k th bin. Suppose that $t_{k+1} - t_k = h$ for all k ; then the histogram is said to have fixed bin width h . A frequency histogram is built using blocks of height 1 and width h stacked in the appropriate bins. The integral of such a figure is clearly equal to nh . Thus a density histogram uses building blocks of height $1/(nh)$, so that each block has area equal to $1/n$. Let v_k denote the bin count of the k th bin, that is, the number of sample points falling in bin B_k ; see Figure 3.2. Then the histogram is defined as

$$\hat{f}(x) = \frac{v_k}{nh} = \frac{1}{nh} \sum_{i=1}^n I_{[t_k, t_{k+1})}(x_i) \quad \text{for } x \in B_k. \quad (3.2)$$

The analysis of the histogram random variable, $\hat{f}(x)$, is quite simple, once it is recognized that the bin counts are Binomial random variables:

$$v_k \sim B(n, p_k), \quad \text{where } p_k = \int_{B_k} f(t) dt.$$

Consider the MSE of $\hat{f}(x)$ for $x \in B_k$. Now $E[v_k] = np_k$ and $\text{Var}[v_k] = np_k(1 - p_k)$. Hence,

$$\text{Var } \hat{f}(x) = \frac{\text{Var } v_k}{(nh)^2} = \frac{p_k(1 - p_k)}{nh^2} \quad (3.3)$$

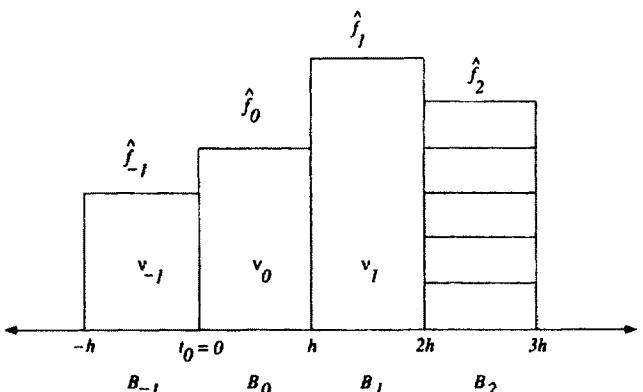


Figure 3.2 Notation for construction of an equally-spaced histogram.

and

$$\text{Bias } \hat{f}(x) = E\hat{f}(x) - f(x) = \frac{1}{nh} E\nu_k - f(x) = \frac{p_k}{h} - f(x). \quad (3.4)$$

To proceed with the fewest assumptions, suppose that $f(x)$ is Lipschitz continuous over the bin B_k .

Definition: A function is said to be Lipschitz continuous over an interval B_k if there exists a positive constant γ_k such that $|f(x) - f(y)| < \gamma_k|x - y|$ for all $x, y \in B_k$.

Then by the mean value theorem (MVT),

$$p_k = \int_{B_k} f(t) dt = hf(\xi_k) \quad \text{for some } \xi_k \in B_k. \quad (3.5)$$

It follows that

$$\text{Var } \hat{f}(x) \leq \frac{p_k}{nh^2} = \frac{f(\xi_k)}{nh} \quad (3.6)$$

and

$$|\text{Bias } \hat{f}(x)| = |f(\xi_k) - f(x)| \leq \gamma_k |\xi_k - x| \leq \gamma_k h;$$

hence, $\text{Bias}^2 \hat{f}(x) \leq \gamma_k^2 h^2$ and

$$\text{MSE } \hat{f}(x) \leq \frac{f(\xi_k)}{nh} + \gamma_k^2 h^2. \quad (3.7)$$

In the literature on smoothing, the bin width h is referred to as a *smoothing parameter*, since it controls the amount of smoothness in the estimator for a given sample of size n . Equation (3.7) summarizes the recurring trade-off between bias and variance as determined by the choice of smoothing parameter. The variance may be controlled by making h large so that the bins are wide and of relatively stable height; however, the bias is large. On the other hand, the bias may be reduced by making h small so that the bins are narrow; however, the variance is large. Note that the bias can be eliminated by choosing $h = 0$, but this very rough histogram is exactly the empirical probability density function $f_n(x)$, which has infinite (vertical) variance. The bias and variance may be controlled simultaneously by choosing an intermediate value of the bin width, and allowing the bin width to slowly decrease as the sample size increases.

Definition: A density estimator is said to be consistent in the mean square if the $MSE\{\hat{f}(x)\} \rightarrow 0$ as $n \rightarrow \infty$.

An optimal smoothing parameter h^* is defined to be that choice that minimizes the (asymptotic) mean squared error. The following results are consequences of Equation (3.7).

Theorem 3.1: Assume that $x \in B_k$ is a fixed point and that f is Lipschitz continuous in this bin with constant γ_k . Then the histogram estimate $\hat{f}(x)$ is mean square consistent if, as $n \rightarrow \infty$, then $h \rightarrow 0$ and $nh \rightarrow \infty$.

The first condition ensures that the bias vanishes asymptotically, while the second condition ensures that the variance goes to zero. Duda and Hart (1973) suggest choosing $h = n^{-1/2}$, for example.

Corollary 3.2: The $MSE(x)$ bound (3.7) is minimized when

$$h^*(x) = \left[\frac{f(\xi_k)}{2\gamma_k^2 n} \right]^{1/3}; \quad (3.8)$$

the resulting $MSE^*(x)$ is $O(n^{-2/3})$.

These results deserve careful examination. The optimal bin width decreases at a rate proportional to $n^{-1/3}$. This rate is much faster than Sturges' rule, which suggests the rate $\log_2^{-1}(n)$; see Table 3.1. The optimal rate of decrease of the MSE does not attain the Cramer-Rao lower bound rate of $O(n^{-1})$ for parametric estimators.

The noise inherent in the histogram varies directly with its height, since $\text{Var}\{\hat{f}(x)\} \approx f(x)/(nh)$ from Equation (3.6). This heteroscedasticity (unequal variance) across the histogram estimate may be eliminated by the use of a variance-stabilizing transformation. Each bin count is approximately a Poisson random variable. It is well-known that the square root is the variance-stabilizing transformation for Poisson data. Suppose that Y_n has moments (μ_n, σ_n^2) with $\mu_n > 0$ and $\sigma_n \rightarrow 0$. Then $\text{Var}\{g(Y_n)\} \approx g'(\mu_n)^2 \sigma_n^2$. Choosing $Y_n = \hat{f}(x)$ so that $\mu_n \approx f(x)$ and $\sigma_n^2 \approx f(x)/(nh)$, then $g(y) = \sqrt{y}$ and $g'(y) = 1/(2\sqrt{y})$ or $1/(2\sqrt{f(x)})$ at $y = \mu_n$. Therefore,

$$\sqrt{\text{Var}\sqrt{\hat{f}(x)}} \approx \frac{1}{2\sqrt{f(x)}} \sqrt{\frac{f(x)}{nh}} = \frac{1}{2\sqrt{nh}}, \quad (3.9)$$

which is independent of the unknown $f(x)$.

Thus plotting the histogram on a square root scale allows for easy comparison of noise in the histogram in regions of high and low density. Tukey (1977) called the resulting estimate the *rootgram*. Of course, the rootgram no longer accurately portrays the relative frequencies of the observations. In more than 1 dimension, the rootgram is still variance-stabilizing. However, since the contours of the bivariate histogram and bivariate rootgram are identical, the practical applications are limited.

One consequence of Corollary 3.2 is that the use of a fixed bandwidth over the entire range of the data is not generally optimal. For example, the bin width should be relatively wider in regions of higher density to reduce the variance in (3.7). Now if the width of bin B_k is sufficiently narrow, then the Lipschitz constant γ_k is essentially the magnitude of the slope of $f(x)$ in that bin. Therefore, from (3.8), the bin width should be narrower in regions where the density is changing rapidly, and vice versa. These notions are confirmed in Section 3.2.8. However, in practice, there are no reliable algorithms for constructing adaptive histogram meshes. Therefore, the study of fixed-width histograms remains important.

3.2.2 Global L_2 Histogram Error

Consistency results based on upper bounds are not useful in practice, since the upper bounds may be quite far from truth. More useful approximations can be made by assuming the existence of derivatives of f . These results can be useful even in the practical situation where f is unknown, by employing a variety of techniques called cross-validation algorithms; see Section 3.3.

Computing the MISE is accomplished by aggregating the MSE over each bin and summing over all bins. Consider first the integrated variance (IV):

$$\text{IV} = \int_{-\infty}^{\infty} \text{Var } \hat{f}(x) dx = \sum_{k=-\infty}^{\infty} \int_{B_k} \text{Var } \hat{f}(x) dx. \quad (3.10)$$

From Equation (3.3), the last integral over B_k is simply $p_k(1 - p_k)/(nh)$. Now $\sum p_k = \int f(x) dx = 1$. Recall that $\sum \phi(\xi_k) \cdot h = \int \phi(x) dx + o(1)$ by standard Riemannian integral approximation. Therefore, using the approximation (3.5) for p_k , $\sum p_k^2 = \sum f(\xi_k)^2 h^2 = h[\int f(x)^2 dx + o(1)]$. Combining, we have

$$\text{IV} = \frac{1}{nh} - \frac{R(f)}{n} + o(n^{-1}), \quad (3.11)$$

where the following notation is adopted for the squared L_2 -norm of ϕ :

$$R(\phi) \equiv \int \phi(x)^2 dx.$$

The squared L_2 -norm is only one possible measure of the *roughness* (R) of the function ϕ . Alternatives include $R(\phi')$ and $R(\phi'')$. $R(\phi)$ in this context refers more to the *statistical roughness* than to the *mathematical roughness* of the function ϕ . The latter usually refers to the number of continuous derivatives in the function. The former refers loosely to the number of wiggles in the function. Statistical roughness does, however, take account of the first few derivatives of the function. In summary, neither definition would describe a Normal density as rough, but the lognormal density is a very rough function from the statistical point of view, even though it is infinitely differentiable. A low-order polynomial density, such as one from the Beta family, is statistically smooth even though it possesses only a few derivatives.

In order to compute the bias, consider a typical bin $B_0 = [0, h]$. The bin probability, p_0 , may be approximated by

$$\begin{aligned} p_0 &= \int_0^h f(t) dt = \int_0^h \left[f(x) + (t - x)f'(x) + \frac{1}{2}(t - x)^2 f''(x) + \dots \right] dt \\ &= hf(x) + h\left(\frac{h}{2} - x\right)f'(x) + O(h^3) \end{aligned}$$

so that

$$\text{Bias } \hat{f}(x) = \frac{p_0}{h} - f(x) = \left(\frac{h}{2} - x\right)f'(x) + O(h^2); \quad (3.12)$$

see Figure 3.3. For future reference, note that (3.12) implies that the bias is of higher-order $O(h^2)$ at the center of a bin, when $x = h/2$.

Using the generalized mean value theorem (GMVT), the leading term of the integrated squared bias for this bin is

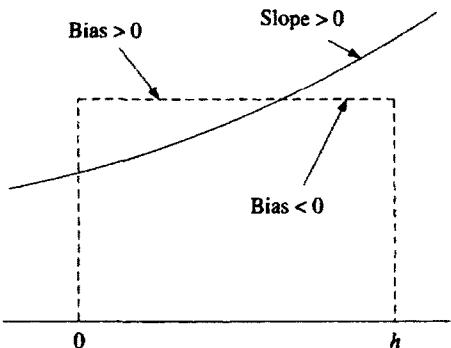


Figure 3.3 Bias of histogram estimator in a typical bin.

$$\int_{B_0} \left(\frac{h}{2} - x \right)^2 f'(x)^2 dx = f'(\eta_0)^2 \int_0^h \left(\frac{h}{2} - x \right)^2 dx = \frac{h^3}{12} f'(\eta_0)^2, \quad (3.13)$$

for some $\eta_0 \in B_0$. This result generalizes to other bins for some collection of points $\eta_k \in B_k$. Hence the total integrated squared bias (ISB) is

$$\text{ISB} = \frac{h^2}{12} \sum_{k=-\infty}^{\infty} f'(\eta_k)^2 \times h = \frac{h^2}{12} \int_{-\infty}^{\infty} f'(x)^2 dx + o(h^2), \quad (3.14)$$

which follows from standard Riemannian convergence of sums to integrals. A note on assumptions: if the total variation of $f'(\cdot)^2$ is finite, then the remainder term in (3.14) becomes $O(h^3)$; see Problem 3. Assuming the existence of an absolutely continuous second derivative, the error term is $O(h^4)$. There is little practical difference among these observations.

Rather than persist in keeping track of the order of the remainder term explicitly, the following notation will be adopted. The main terms in the ISB will be referred to as the *asymptotic integrated squared bias* (AISB). Thus, $\text{ISB} = \text{AISB} + o(h^2)$ and

$$\text{AISB} = \frac{1}{12} h^2 \int_{-\infty}^{\infty} f'(x)^2 dx = \frac{1}{12} h^2 R(f').$$

Similarly, the *asymptotic integrated variance* (AIV) and *asymptotic mean integrated squared error* (AMISE) refer to the main terms in the approximations to the IV and MISE, respectively.

The following theorem, which summarizes Equations (3.11) and (3.14), is due to Scott (1979) and Freedman and Diaconis (1981).

Theorem 3.3: Suppose that f has an absolutely continuous derivative and a square-integrable first derivative. Then the asymptotic MISE is

$$\text{AMISE}(h) = \frac{1}{nh} + \frac{1}{12} h^2 R(f');$$

hence

$$\begin{aligned} h^* &= [6/R(f')]^{1/3} n^{-1/3} \\ \text{AMISE}^* &= (3/4)^{2/3} R(f')^{1/3} n^{-2/3}. \end{aligned} \quad (3.15)$$

Thus the asymptotically optimal bin width depends on the unknown density only through the roughness of its first derivative. This result holds irrespective of the choice of bin origin, which must have a secondary role in MISE compared

to the bin width. The optimal bin width decreases at a relatively slow rate. The corresponding optimal error,

$$\text{AMISE}^* = \text{AMISE}(h^*),$$

decreases at the same rate as the bound in Corollary 3.2, far from the desirable $O(n^{-1})$ rate.

3.2.3 Normal Density Reference Rule

It will be convenient to fix some of these results for the special case of Normal data. If $f = N(\mu, \sigma^2)$, then $R(f') = 1/(4\sqrt{\pi}\sigma^3)$; see Problem 2. Hence, from Theorem 3.3,

$$h^* = (24\sqrt{\pi}\sigma^3/n)^{1/3} \approx 3.5\sigma n^{-1/3}. \quad (3.16)$$

Scott (1979) proposed using the Normal density as a reference density for constructing histograms from sampled data by using the sample standard deviation $\hat{\sigma}$ in (3.16) to obtain the

Normal bin width reference rule:	$\hat{h} = 3.5\hat{\sigma}n^{-1/3}.$	(3.17)
----------------------------------	--------------------------------------	--------

Since $\hat{\sigma} \rightarrow \sigma$ faster than $O(n^{-1/3})$, this rule is very stable. Freedman and Diaconis (1981) proposed a more robust rule, replacing the unknown scale parameter σ by a multiple of the interquartile range (IQ):

$$\hat{h} = 2(\text{IQ})n^{-1/3}.$$

If the data are in fact Normal, the Freedman-Diaconis rule is about 77% of the Scott's rule as the IQ = 1.348 σ in this case.

How does the optimal rule compare to Sturges' rule in the case where the situation is most favorable to the latter? In Table 3.1, the number of bins suggested by the two rules is given, assuming that the data are Normal and that the sample range is $(-3, 3)$. Perhaps the most remarkable observation obtainable from this table is how closely the rules agree for samples between 50 and 500 points. For larger samples, Sturges' rule gives too few bins, corresponding to a much oversmoothed histogram estimate that wastes much of the information in the data. As shown in Section 3.3.1, almost any other non-Normal sampling density will require even more bins. This result is a consequence of the fact that the Normal case is very close to a theoretical lower bound on the number of

Table 3.1 Comparison of Number of Bins from Three Normal Reference Rules

<i>n</i>	Sturges' Rule	Scott's Rule	F-D Rule
50	5.6	6.3	8.5
100	7.6	8.0	10.8
500	10.0	13.6	18.3
1,000	11.0	17.2	23.2
5,000	13.3	29.4	39.6
10,000	14.3	37.0	49.9
100,000	17.6	79.8	107.6

bins, or equivalently, an upper bound on the bin width. The Freedman-Diaconis rule has 35% more bins than Scott's rule and the histogram will be rougher.

Scott (1979) also proposed using the lognormal and *t* distributions as reference densities to modify the Normal rule when the data are skewed or heavy-tailed. Suppose Y is a non-Normal random variable with density $g(y)$ and moments σ_y^2 , β_1 , and β_2 (the standardized skewness and kurtosis, respectively). Consider using the Normal rule with $\sigma = \sigma_y$ compared to the optimal bin width in Theorem 3.3. The ratio of these two bin widths is given by

$$\frac{h_y^*}{h_N} = \left[\frac{R(\phi'(y; 0, \sigma_y^2))}{R(g'(y))} \right]^{1/3}, \quad (3.18)$$

if this ratio is much different from 1, then the Normal reference rule should be modified by this ratio.

For example, let $g(y)$ be the lognormal density of the random variable $Y = \exp(X)$, where $X \sim N(0, \sigma^2)$. Then treating σ^2 as a parameter, we have

$$\begin{aligned} \sigma_y^2 &= e^{\sigma^2}(e^{\sigma^2} - 1); & \beta_1 &= \frac{e^{3\sigma^2} - 3e^{\sigma^2} + 2}{(e^{\sigma^2} - 1)^{3/2}}; \\ R(g') &= \frac{(\sigma^2 + 2)e^{9\sigma^2/4}}{8\sqrt{\pi}\sigma^3}. \end{aligned}$$

Since $R(\phi'(y)) = 1/(4\sqrt{\pi}\sigma^3)$, the Normal rule should be multiplied by the factor in (3.18) given by

$$\text{skewness factor}\{\beta_1(\sigma)\} = \frac{2^{1/3}\sigma}{e^{5\sigma^2/4}(\sigma^2 + 2)^{1/3}(e^{\sigma^2} - 1)^{1/2}}. \quad (3.19)$$

This factor is plotted in Figure 3.4. Clearly, any skewness requires smaller bin widths than given by the Normal reference rule.

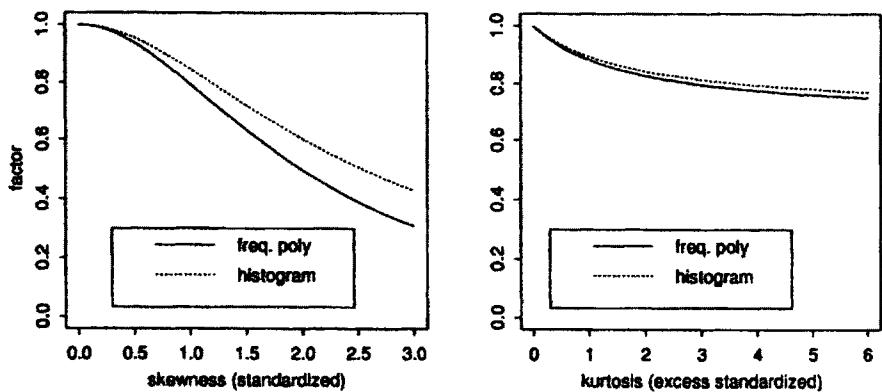


Figure 3.4 Factors that modify the Normal reference bin width rules for the histogram and frequency polygon as a function of the standardized skewness and excess kurtosis.

A similar calculation can be performed for kurtosis assuming that $Y \sim t_\nu$, which is the t distribution with ν degrees of freedom. Treating ν as a parameter yields

$$\sigma_y^2 = \frac{\nu}{\nu - 2}; \quad \beta_2 = 3 + \frac{6}{\nu - 4}; \quad R(g') = \frac{(\nu + 1)^2 B\left(\frac{3}{2}, \frac{2\nu+3}{2}\right)}{\nu^{3/2} B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)^2},$$

where $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$. Substituting into (3.18) gives us

$$\text{kurtosis factor } \left\{ \tilde{\beta}_2 = \frac{6}{\nu - 4} \right\} = \frac{(\nu - 2)^{1/2} B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)^{2/3}}{2^{2/3} \pi^{1/6} (\nu + 1)^{2/3} B\left(\frac{3}{2}, \frac{2\nu+3}{2}\right)^{1/3}}, \quad (3.20)$$

where $\tilde{\beta}_2 = \beta_2 - 3$ is the excess kurtosis; see Figure 3.4. The modification to the bin width is not as great as for large skewness.

3.2.4 Equivalent Sample Sizes

In comparison with parametric estimators, histograms enjoy the advantage of quite general consistency without what Fisher called the “specification problem.” How great a penalty is incurred if an optimal histogram is used rather than the corresponding optimal parametric estimator?

Consider again the example of Section 3.2.3. From Theorem 3.3, it follows that if $f = N(0, 1)$, then the optimal AMISE of the histogram is

$$\text{AMISE}^* = [9/(64\sqrt{\pi})]^{1/3} n^{-2/3} \approx 0.4297 n^{-2/3}.$$

Table 3.2 Equivalent Sample Sizes for Several Normal Density Estimators

AMISE	Estimator				Histogram
	$N(\bar{x}, s^2)$	$N(\bar{x}, 1)$	$N(0, s^2)$		
0.002468	100	57	43		2,297
0.000247	1,000	571	429		72,634

Suppose a sample of size 100 is available for the parametric estimator $N(\bar{x}, \sigma^2)$, for which AMISE = 0.002468 from Equation (2.9). Table 3.2 gives the equivalent sample sizes for the histogram and several parametric estimators. The parametric estimators are the clear winners, more so for smaller errors.

On the other hand, if the true density is only approximately Normal, then the parametric MISE will never be less than the true integrated squared bias level

$$\text{ISB} = \int [\phi(x; \mu_f, \sigma_f^2) - f(x)]^2 dx,$$

where (μ_f, σ_f^2) are the actual moments of the unknown density $f(x)$. The maximum likelihood estimator assuming Normality will asymptotically match those moments. Thus, while the variance of the parameters will asymptotically vanish, the ISB will remain. Testing the goodness-of-fit of a parametric model is necessary and can be especially difficult in the multivariable case; distinguishing between certain parametric families may require large samples.

3.2.5 Sensitivity of MISE to Bin Width

How important is it that the smoothing parameter be optimally chosen? If it happens that virtually any bin width “works,” then less effort would need to be applied to the calibration problem.

3.2.5.1 Asymptotic Case

In practice, given real data from an unknown density, the smoothing parameter chosen will not be h^* , but instead be of the form $h = ch^*$. On average, if $c \ll 1$, then the histogram will have high variance and will be “too rough”; if $c \gg 1$, then the estimate will have high bias or systematic error and will be “too smooth.” How sensitive is the MISE to local deviations of c from 1?

The asymptotic MISE of the histogram in Theorem 3.3 is of the form

$$\text{AMISE}(h) = \frac{a}{nh} + \frac{b}{2} h^2, \quad (3.21)$$

where a, b are positive constants. Rather than considering only this special case, a more general form of the AMISE will be considered:

$$\text{AMISE}(h) = \frac{a}{(d+r)nh^{d+r}} + \frac{b}{2p} h^{2p}, \quad (3.22)$$

where (d, p, r) are positive integers, and a, b are positive constants that depend upon the density estimator and unknown density function. In general, the triple (d, p, r) refers to, respectively: (1) the dimension of the data; (2) the "order" of the estimator's bias; and (3) the order of the derivative being estimated. Comparing (3.21) and (3.22), $(d, p, r) = (1, 1, 0)$ for the histogram. It follows from minimizing (3.22) that

$$h^* = (a/nb)^{1/(d+r+2p)}$$

$$\text{AMISE}^* = \text{AMISE}(h^*) = \left(\frac{d+r+2p}{(d+r)2p} \right) \left(\frac{a^{2p} b^{d+r}}{n^{2p}} \right)^{1/(d+r+2p)}. \quad (3.23)$$

In Problem 5 it is shown that the variance portion of the histogram's AMISE* is twice that of the squared bias. Finally, it is easy to check (see Problem 6) that

$$\frac{\text{AMISE}(ch^*)}{\text{AMISE}(h^*)} = \frac{\frac{2p}{d+r} + c^{2p+(d+r)}}{\left(\frac{2p}{d+r} + 1\right)c^{d+r}}. \quad (3.24)$$

A consequence of this expression is that departures of h from h^* should be measured in a *multiplicative* rather than an *additive* fashion. This may also be clear from a dimensional analysis of h^* in Theorem 3.3. In the author's experience, a 10–15% change in h produces a small but noticeable change in a histogram. Nonetheless, in Table 3.3, it is clear that the histogram, which corresponds to the case $p = 1$, is fairly insensitive to a choice of bin width within 33% of optimal. Notice that the L_2 criterion is less affected by high variance than high bias errors (for example, $c = 1/2$ vs. $c = 2$).

Looking forward, larger values of p correspond to "higher-order" methods that have faster MISE rates. However, the sensitivity to choice of smoothing parameters is much greater. Increased sensitivity is also apparent with increasing dimension d and derivative order r ; see Problem 7.

3.2.5.2 Large-Sample and Small-Sample Simulations

The previous analysis focused only on average behavior. The actual ISE for an individual sample is considered in this section.

Table 3.3 Asymptotic Sensitivity of AMISE to Error in Bin Width Choice $h = ch^*$

$(d = 1, r = 0)$	$p = 1$	$p = 2$	$p = 4$
c	$(c^3 + 2)/(3c)$	$(c^5 + 4)/(5c)$	$(c^9 + 8)/(9c)$
1/2	1.42	1.61	1.78
3/4	1.08	1.13	1.20
1	1	1	1
4/3	1.09	1.23	1.78
2	1.67	3.60	28.89

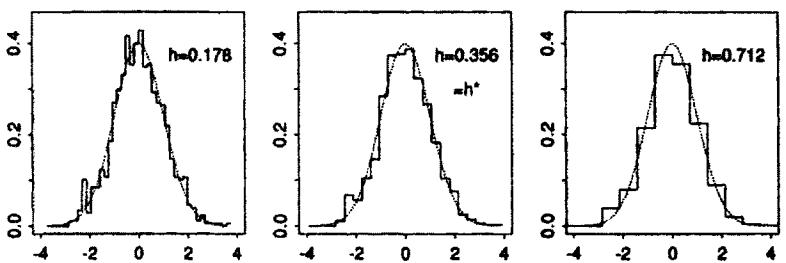


Figure 3.5 Three histograms of 1,000 Normal observations with bin widths $h = (\frac{1}{2}h^*, h^*, 2h^*)$.

Figure 3.5 displays 3 histograms of 1,000 Normal observations. When $h = 2h^*$, the estimate incurs substantial bias because of the broad bins. In informal questioning, most statisticians prefer the histogram with $h = \frac{1}{2}h^*$, even though it contains several spurious small bumps. However, it is easy to visually smooth the local noise on that estimate. The converse is not true, since it is not possible to visualize the lost detail in the histogram with bin width $h = 2h^*$.

Generalizing too much from small-sample experience is a temptation that should be avoided. Figure 3.6 displays 4 histograms of a sample of a million Normal points. The exact ISE is given in each frame. Examine the changes in ISE. Locally, the histogram remains very sensitive to changes in h even when n is large. This may seem counterintuitive initially, especially given the very general conditions required for consistency. But large samples do not automatically guarantee good estimates. The histogram with $h = \frac{1}{2}h^*$ contains many local modes. The behavior of sample modes in a histogram will be reexamined in Section 3.5.2. Even the “optimal” histogram with $h = h^*$ appears noisy (locally) compared to the $h = 2h^*$ histogram. The observation that optimally smoothed estimates appear on the verge of a form of instability seems a common occurrence.

3.2.6 Exact MISE vs. Asymptotic MISE

For a general unequally spaced mesh, where the width of bin B_k is h_k , it is straightforward to show from Equations (3.3) and (3.4) that

$$\text{IV} = \frac{1}{n} \sum_k \frac{p_k(1-p_k)}{h_k} \quad \text{and} \quad \text{ISB} = R(f) - \sum_k \frac{p_k^2}{h_k}, \quad (3.25)$$

exactly (see Problem 8). Now, $\sum_k p_k = 1$; hence, for the special case of an equally spaced mesh, $h_k = h$,

$$\text{MISE}(h, t_0, n) = \frac{1}{nh} - \frac{n+1}{nh} \sum_k p_k^2 + R(f).$$

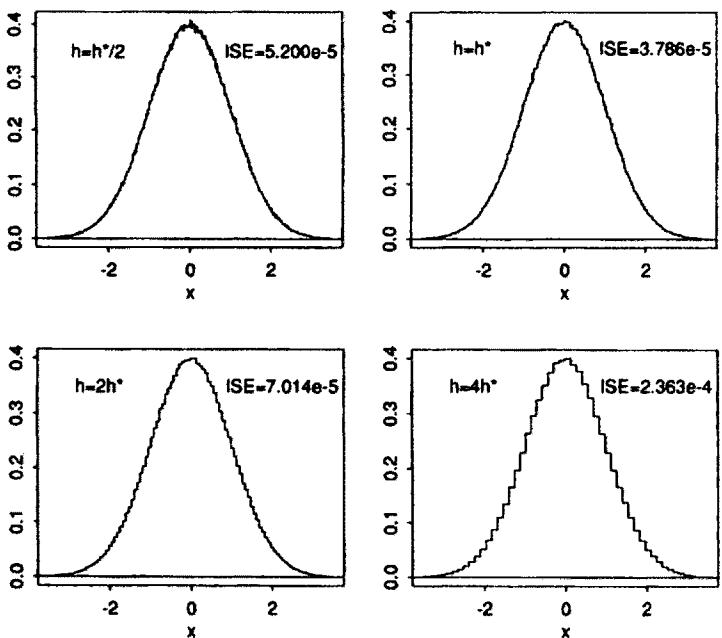


Figure 3.6 Four histograms of a million Normal points with $h = (\frac{1}{2}h^*, h^*, 2h^*, 4h^*)$.

This expression can be used to test the accuracy of the AMISE for small samples, to verify the sensitivity of MISE to misspecification of bin width, and to evaluate the effect of choosing the mesh origin t_0 .

3.2.6.1 Normal Density

In Figure 3.7, the exact and asymptotic MISE of a fixed bin histogram with $t_0 = 0$ are plotted on a log-log scale for standard Normal data with several sample sizes. Clearly, the AMISE adequately approximates the true MISE even with small samples. The approximation error (gap) rapidly narrows as n increases. Of particular interest is the fact that the bin widths minimizing MISE and AMISE are virtually indistinguishable for all n . That both h and MISE decrease as n increases is readily apparent. Furthermore, the gap is explained almost entirely by the term $-R(f)/n$ from Equation (3.11).

In Figure 3.8, the decomposition of MISE into IV and ISB is shown. The IV and ISB lines on the log-log scale are nearly straight lines with slopes as predicted by asymptotic theory in Theorem 3.3. Perhaps most surprising is the single line for the ISB. This feature is easily explained by the fact that the ISB as given in Equation (3.25) is only a function of the partition and not of the sample size. Notice that at the optimal choice of bin width, the contribution of IV exceeds ISB, by a ratio approximately of 2:1, as shown in Problem 5. The L_2 criterion is more sensitive when $h > h^*$.

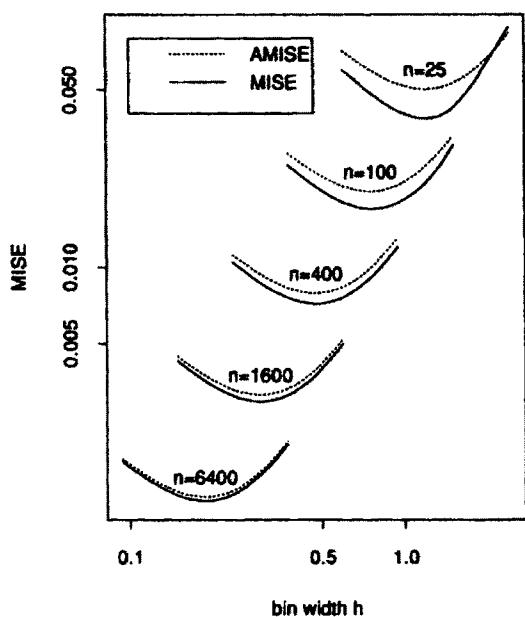


Figure 3.7 AMISE and exact MISE for the $N(0, 1)$ density.

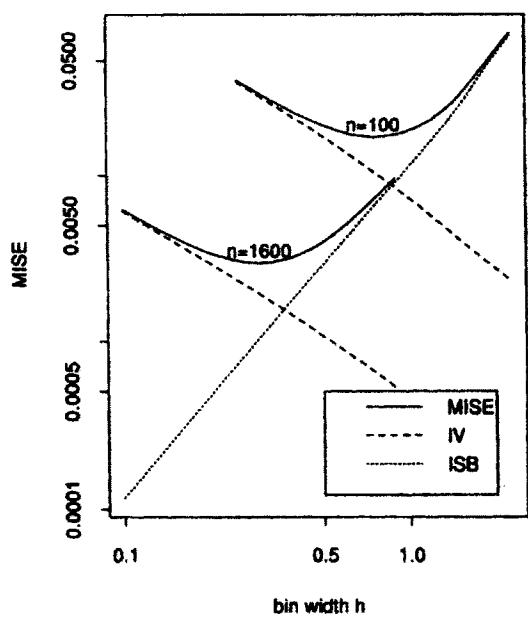


Figure 3.8 Integrated squared-bias/variance decomposition of MISE for the $N(0, 1)$ density.

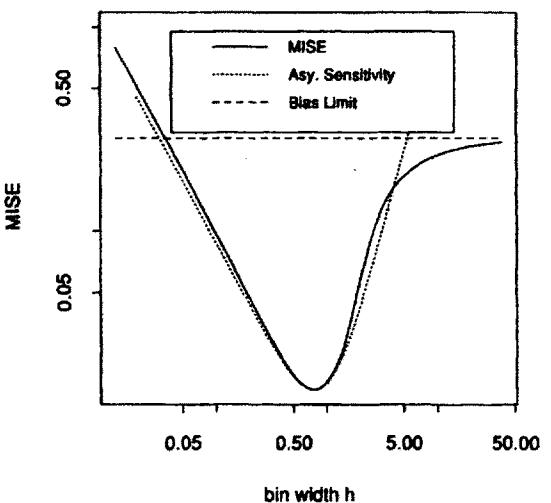


Figure 3.9 Complete MISE curve for the $N(0,1)$ density when $n = 100$. The asymptotic sensitivity relationship holds over a wide range of bin widths.

Figure 3.9 displays a picture of the MISE curve over a much wider range of bin widths when $n = 100$. The $\text{MISE}(h)$ is unbounded as $h \rightarrow 0$, but

$$\lim_{h \rightarrow \infty} \text{MISE}(h) = \lim_{h \rightarrow \infty} \text{ISB}(h) = R(f),$$

as is clearly evident. Asymptotic theory predicts the manner in which MISE varies with h ; see Equation (3.24). The dashed line in Figure 3.9 is a plot of

$$\left(h, \frac{c^3 + 2}{3c} \text{MISE}^* \right),$$

where $c = h/h^*$ and $\text{MISE}^* = \text{MISE}(h^*)$. This approximation is accurate far beyond the immediate neighborhood of h^* .

3.2.6.2 Lognormal Density

Figure 3.10 repeats the computation indicated in Figure 3.7 but with a lognormal density and $t_0 = 0$. While the AMISE retains its parabolic shape, the true MISE curves are quite complicated for $n < 1,000$. When $n = 25$, asymptotic theory predicts $h^* = 0.49$ while $h = 1.28$ is in fact optimal. Also, $\text{AMISE}^* = 0.1218$, which is far greater than the true $\text{MISE}^* = 0.0534$. Furthermore, when $n = 200$, the MISE curve has 2 (!) local minima at $h = 0.245$ and $h = 0.945$. For $n > 1,000$ the asymptotic theory appears to be adequate.

What can explain these observations? First, the lognormal density, while infinitely differentiable, is very statistically rough, particularly near the origin. In fact, 90% of the roughness, $R(f') = 3e^{9/4}/(8\sqrt{\pi})$, comes from the small

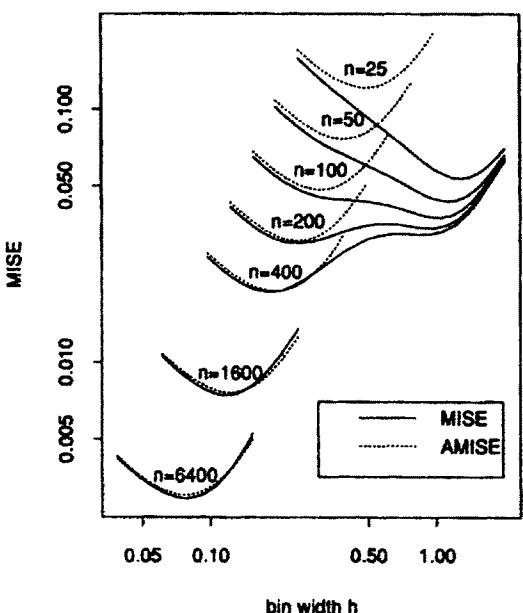


Figure 3.10 AMISE and exact MISE for the lognormal density.

interval $[0, 0.27]$, even though the mode is located at $x = 0.368$ and the 99th percentile is $x = 10.2$. Hence, asymptotically, the optimal bin width must be relatively small in order to track the rapidly changing density near the origin. For very small samples, there is insufficient data to accurately track the rise of the lognormal density to the left of the mode; hence the optimal MISE bandwidth $h > 1$, which is much larger than that predicted by the AMISE. For $n > 1,000$, the optimal bin width satisfies $h < 0.15$ and the rise can be tracked. For $n \approx 200$, the two locally optimal bin widths indicate a situation where the rise can be equally well approximated by either a narrow or wide bin, or almost any bin width in between since the MISE curve is relatively flat.

Figure 3.10 illustrates an important practical and generalizable observation about nonparametric methods. Sample sizes may be grouped into 3 categories: inadequate, transitional, and sufficient. For lognormal data, the inadequate sample sizes are for $100 < n < 400$ with equally spaced histograms, and transitional for $400 < n < 1,000$. For other densities, the transitory region may extend to even larger samples, although the lognormal density is reasonably extreme. Unfortunately, in practice with a particular data set, it is not possible to be certain into which category n falls. However, as can readily be seen, features narrower than h cannot be detected reliably below certain sample sizes. Donoho (1988) neatly summarizes the situation in nonparametric estimation as one-sided inference—only larger sample sizes can definitively answer whether smaller structure exists in the unknown density.

3.2.7 Influence of Bin Edge Location on MISE

3.2.7.1 General Case

The asymptotic MISE analysis in Theorem 3.3 indicates that the choice of the bin origin t_0 is a lower order effect, as long as the conditions of the theorem are satisfied. For $N(0, 1)$ data with optimal smoothing, the exact MISE is minimized if the point $x = 0$ is in the *middle* of a bin rather than at the boundary. The difference in MISE when using these two bin edge locations is minuscule, being 1.09% when $n = 25$ and less than 10^{-5} when $n > 100$.

With lognormal data, the edge effect is also less than 10^{-5} when $n > 400$. But when $n = 25$ and $h = 1.28$, the MISE ranges over $(0.042, 0.149)$ —best for the mesh $(-1.20, 0.08, 1.36, 2.64, \dots)$ and worst for the mesh $(-0.65, 0.63, 1.91, 3.19, \dots)$. Clearly, the choice of bin edge is negligible for sufficiently large sample sizes but can be significant for inadequate sample sizes. Graphically, the choice of t_0 for fixed h provides a range of possible histograms with quite different subjective appearances; see Chapter 5 and Figure 5.1.

3.2.7.2 Boundary Discontinuities in the Density

The approximation properties of a histogram are not affected by a simple jump in the density, if the jump occurs at the boundary of a histogram bin. This follows from a careful examination of the derivation of the MISE in Section 3.2.2. Discontinuities can adversely affect all density estimators. The adverse effect can be demonstrated for the histogram by example.

Consider $f(x) = e^{-x}$, $x \geq 0$ and mesh $t_k = kh$ for integer $k \geq 0$. Then Theorem 3.3 holds on the interval $(0, \infty)$, on which $R(f') = 1/2$. Therefore,

$$h^* = (12/n)^{1/3} \quad \text{and} \quad \text{AMISE}^* = 0.6552n^{-2/3}.$$

Suppose the discontinuity at zero was not known *a priori* and the mesh $t_k = (k - \frac{1}{2})h$ was chosen. Then attention focuses on the bias in bin $B_0 = [-h/2, h/2]$ where the density is discontinuous; see Figure 3.11.

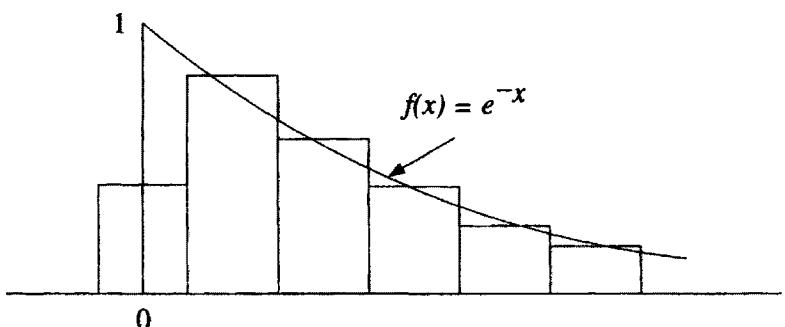


Figure 3.11 Illustration of discontinuity boundary bin problem.

Note that the probability mass in bin $B_0 = (-h/2, h/2]$ is

$$p_0 = \int_0^{h/2} e^{-x} dx = 1 - e^{-h/2}.$$

Now $E\hat{f}(x) = p_0/h$; therefore,

$$\int_{-h/2}^{h/2} \text{Bias}(x)^2 dx = \int_{-h/2}^0 \left(\frac{p_0}{h} - 0\right)^2 dx + \int_0^{h/2} \left(\frac{p_0}{h} - e^{-x}\right)^2 dx, \quad (3.26)$$

which equals $(1 - e^{-h})/2 + (2e^{-h/2} - e^{-h} - 1)/h \approx h/4 - h^2/8 + \dots$. Thus over the interval $(-h/2, \infty)$,

$$\text{ISB} = \frac{h}{4} - \frac{h^2}{8} + O(h^3) + \frac{1}{12}h^2 \int_{h/2}^{\infty} f'(x)^2 dx + o(h^2). \quad (3.27)$$

The worst outcome has been realized—the ISB is entirely dominated by the contribution from the bin containing the discontinuity at $x = 0$. The asymptotic integrated variance is unchanged so that

$$\text{AMISE}(h) = \frac{1}{nh} + \frac{h}{4} \implies h^* = \frac{2}{\sqrt{n}} \quad \text{and} \quad \text{AMISE}^* = \frac{1}{\sqrt{n}}, \quad (3.28)$$

which is significantly worse than $O(n^{-2/3})$. In fact, the rate is as slow as for bivariate data, as will be shown in Section 3.4. Table 3.4 illustrates the costs. The histogram tries to accommodate the discontinuity by choosing a narrower bin width. Compare this situation to that of the very rough behavior of the lognormal density near the origin with small samples.

3.2.8 Optimally Adaptive Histogram Meshes

An optimally calibrated fixed-bin-width histogram with real data often *appears* rough in the tails due to the paucity of data. This phenomenon is one of several reasons for considering histograms with adaptive meshes. Evaluating

Table 3.4 Potential Impact on AMISE of Lack of Knowledge of Boundary Discontinuities

n	$X > 0$ Known		$X > 0$ Unknown		Error Ratio
	h^*	AMISE*	h^*	AMISE*	
10	1.063	0.14116	0.3162	0.31623	2.24
100	0.493	0.03041	0.1	0.1	3.29
1,000	0.229	0.00655	0.0316	0.03162	4.83
10,000	0.106	0.00141	0.01	0.01	7.09
100,000	0.049	0.00030	0.0032	0.00316	10.54

the reduction in MISE with adaptive meshes is the task of this section. Creating adaptive meshes is a familiar exercise for anyone who has ever performed a χ^2 goodness-of-fit test: to satisfy the recommendation that the expected count in every cell exceeds 5, cells in the tails are usually combined or, alternatively, cells are formed so that each contains exactly the same number of points; see Section 3.2.8.4.

3.2.8.1 Bounds on MISE Improvement for Adaptive Histograms

A lower bound, which is asymptotically correct, for the reduction in MISE of an optimally adaptive histogram follows from Equations (3.6) and (3.14). Construct an approximation to the asymptotically adaptive pointwise histogram MSE (AAMSE):

$$\text{AAMSE}(x) \approx \frac{f(x)}{nh} + \frac{1}{12}h^2f'(x)^2. \quad (3.29)$$

Minimize AAMSE(x) for each point x . Therefore,

$$h^*(x) = \left[\frac{6f(x)}{nf'(x)^2} \right]^{1/3} \implies \text{AAMSE}^*(x) = \left[\frac{3f(x)f'(x)}{4n} \right]^{2/3}.$$

Integrating AAMSE $^*(x)$ over x gives the following result.

Theorem 3.4: *Asymptotically, for an optimally adaptive histogram,*

$$\text{AAMISE}^* = (3/4)^{2/3} \left(\int_{-\infty}^{\infty} [f'(x)f(x)]^{2/3} dx \right) n^{-2/3}. \quad (3.30)$$

This result has been discussed by Terrell and Scott (1983, 1991) and Kogure (1987). Comparing Equations (3.15) and (3.30), the improvement of the adaptive mesh is guaranteed if

$$\int [f'(x)f(x)]^{2/3} \leq \left[\int f'(x)^2 \right]^{1/3}$$

or equivalently, if

$$E \left[\frac{f'(X)^2}{f(X)} \right]^{1/3} \leq \left[E \frac{f'(X)^2}{f(X)} \right]^{1/3}; \quad (3.31)$$

but this last inequality follows from Jensen's inequality (the concave function version); see Problem 11. Table 3.5 shows numerical computations of this ratio for some simple densities. At first glance the gains with adaptive meshes are surprisingly small.

Table 3.5 Reduced AMISE Using an Optimally Adaptive Histogram Mesh

Density	$\int (f^2 f'^2)^{1/3} \div [\int f'^2]^{1/3}$
$N(0, 1)$	0.4648/0.5205 = 89.3%
$\frac{3}{4}(1 - x^2)_+$	0.8292/1.1447 = 72.4%
$\frac{15}{16}(1 - x^2)_+^2$	2.1105/2.8231 = 74.8%
$\frac{315}{256}(1 - x^2)_+^4$	1.4197/1.6393 = 86.6%
Cauchy	0.3612/0.4303 = 84.0%
Lognormal	0.6948/1.2615 = 55.1%

3.2.8.2 Some Optimal Meshes

Since the expression for the exact MISE for arbitrary meshes is available in Equation (3.25), the optimal adaptive mesh may be found numerically. Certain features in these optimal meshes may be anticipated [review the discussion following Equation (3.8)]. Compared to the optimal fixed bandwidth, bins will be relatively wider not only in the tails but also near modes where $f'(x)$ is small. In regions where the density is rapidly changing, the bins will be narrower.

This “accordion” pattern is apparent in the following example with a Beta(5, 5) density rescaled to the interval $(-1, 1)$; $f(x) = (315/256)(1 - x^2)_+^4$ (see Figure 3.12). Since the cdf of this density is a polynomial, it is more amenable than the Normal cdf to numerical optimization over the exact adaptive MISE given in Equation (3.25). Clearly, the optimal mesh is symmetric about 0 in this example. But the optimal mesh does not include 0 as a *knot* (mesh node). Forcing 0 to be a knot increases the MISE, for example, by 4.8% when $n = 10,000$; the best MISE for an equally spaced mesh ($h = 0.0515$) is 8.4% greater. These results are very similar to those for the Normal density.

3.2.8.3 Null Space of Adaptive Densities

There exists a class of density functions for which there is no improvement asymptotically with an adaptive histogram procedure. This class may be called

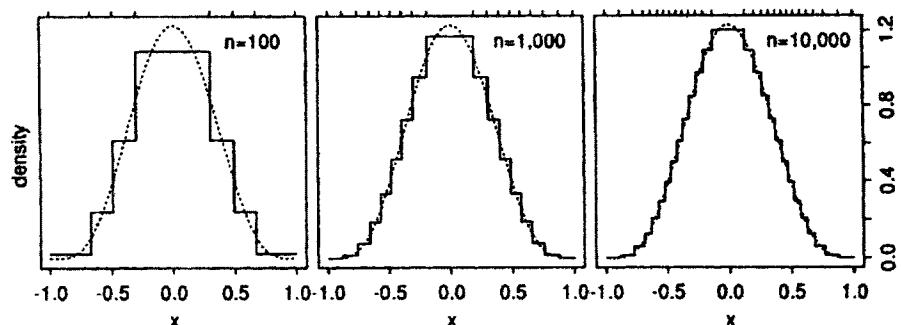


Figure 3.12 Representation of optimal adaptive meshes for the scaled Beta(5,5) pdf, which equals $(315/256)(1 - x^2)_+^4$. The optimal adaptive mesh is also indicated by tick marks above each graph.

the *null space of adaptive densities*. Examining Equation (3.31), any density that results in equality in Jensen's inequality is in the null space. This occurs when the argument in brackets is constant; that is,

$$\frac{f'(x)^2}{f(x)} = c \quad \Rightarrow \quad f(x) = \frac{1}{4}c(x - a)^2, \quad (3.32)$$

where "a" is an arbitrary constant. The densities $f_1(x) = 3(1 - x)^2 I_{[0,1]}(x)$ and $f_2(x) = \frac{3}{2}(1 - |x|)^2 I_{[-1,1]}(x)$ also satisfy the differential equation piecewise. In fact, any density that is obtained by piecing together densities of the form (3.32) in a continuous manner is in the null space; see Figure 3.13. As before, the best adaptive mesh may be found by numerical optimization. However, with $n = 100$ and the density $f_1(x)$ defined above, the computed optimal mesh nodes differ by less than 2% from the equally spaced mesh with $h = 1/6$. And the MISE of the computed optimal mesh is only 0.06% better than the equally spaced mesh.

3.2.8.4 Percentile Meshes or Adaptive Histograms with Equal Bin Counts
 In practice, finding the adaptive mesh described in the preceding section is difficult. Thus some authors have proposed (nonoptimal) adaptive meshes that are easier to implement. One of the more intuitively appealing meshes has an equal number (or equal fraction) of points in each bin. This idea can be modeled by a *percentile mesh*, which is nonstochastic, with m bins, each containing an identical fraction of probability mass equal to $1/m$. The mesh will have $m + 1$ knots at

$$t_k = F_x^{-1}\left(\frac{k}{m}\right), \quad k = 0, \dots, m. \quad (3.33)$$

Before computing the exact MISE for such meshes, it is instructive to give a heuristic derivation. Asymptotically, for any $x \in B_i$, $h_i f(x) \approx p_i$, the probability mass in the i th bin of width h_i . Equation (3.29) suggests

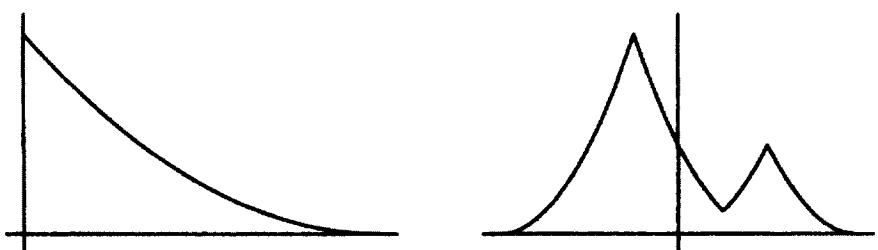


Figure 3.13 Examples of densities for which the asymptotically optimal adaptive mesh is in fact equally spaced.

that $\text{AAMSE}(x) \approx f(x)/(nh_i) + h_i^2 f'(x)^2/12$. If each bin contains exactly k points, then $p_i \approx k/n$ and, therefore, it is reasonable to assign

$$h_i = \frac{k}{nf(x)} \quad \Rightarrow \quad \text{AAMSE}(x) = \frac{f(x)^2}{k} + \frac{1}{12} \frac{k^2}{n^2} \left[\frac{f'(x)}{f(x)} \right]^2. \quad (3.34)$$

Now k is to be chosen not to minimize the pointwise $\text{AAMSE}(x)$, but the global

$$\text{AAMISE}(k) = \frac{R(f)}{k} + \frac{1}{12} \frac{k^2}{n^2} \int \frac{f'(x)^2}{f(x)^2} dx. \quad (3.35)$$

Now $k^* = O(n^{2/3})$ with $\text{AAMISE}^* = O(n^{-2/3})$. But computing the integral in the bias term when $f = \phi$, the standard Normal density, gives

$$\int \frac{f'(x)^2}{f(x)^2} dx = \int \frac{[-x\phi(x)]^2}{\phi(x)^2} dx = \int_{-\infty}^{\infty} x^2 dx = \infty \quad (?) . \quad (3.36)$$

This puzzling result may only be the result of the use of the approximation (3.34) combined with the infinite support of ϕ , or it may indicate that these meshes suffer unexpectedly large biases. The results depicted in Figure 3.14 clearly demonstrate the poor performance of the percentile meshes when compared with fixed-width bins, with the gap increasing with sample size. The consequences, if any, of this negative result for chi-squared tests are unknown; see Problem 13.

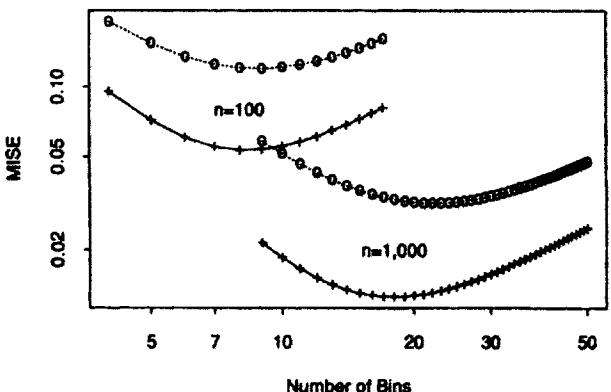


Figure 3.14 Exact MISE for fixed (+++++) and percentile (ooooo) meshes with k bins over $[0, 1]$ for the rescaled Beta(5, 5) density with $n = 100$ and 1,000.

3.2.8.5 Using Adaptive Meshes vs. Transformation

In the univariate setting, adaptive meshes are most useful when the data are skewed or when the data are clustered with different scaling within each cluster. For heavy-tailed data alone, adaptive meshes may provide some visual improvement, but the MISE criterion is largely insensitive to those features. The best recommendation is to use Tukey's (1977) transformation ladder or Box-Cox (1964) methods to reduce skewness before applying a fixed mesh. The Tukey ladder is a subset of the power family of transformations, x^λ , where the transformation is defined to be $\log(x)$ when $\lambda = 0$:

$$\dots, x^{-2}, x^{-1}, x^{-1/2}, x^{-1/4}, \log(x), x^{1/4}, x^{1/2}, x, x^2, x^3, \dots \quad (3.37)$$

The Box-Cox family is similar, but continuous in λ for $x > 0$:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0. \end{cases} \quad (3.38)$$

Simple transformations such \sqrt{x} or $\log(x + c)$ are often sufficient. In the multivariate setting, working with strongly skewed marginal variables results in data clustered near the faces of a surrounding hypercube. Even simple marginal variable transformations can move the data cloud towards the middle of the hypercube, greatly reducing one of the effects of the *curse of dimensionality*; see Chapter 7. In the univariate setting, Wand, Marron, and Ruppert (1991) have proposed finding an optimal transformation (within the Box-Cox transformation family, for example) with respect to AMISE in the transformed coordinates, and then transforming the density estimate back to the original scale. In many applications, the choice is made to work directly in the transformed coordinates. Economists, for example, usually work with $\log(\text{income})$ rather than income to cope with the strong skewness present in income data.

3.2.8.6 Remarks

Adaptive histograms that are optimally constructed *must* be better than non-adaptive fixed-width histograms. But adaptive histograms constructed in an *ad hoc* fashion need not be better and in fact can be much worse. In particular, an adaptive histogram algorithm may be inferior if it does not include fixed bin widths as a special case. Even so, in practice, an adaptive procedure may be worthwhile. Of particular practical concern is the introduction of additional smoothing parameters defining the adaptivity that must be estimated *from the data* by some method such as *cross-validation* (which is introduced below). Given the performance of these algorithms in the simplest case of only one smoothing parameter in the fixed-bandwidth setting, caution seems in order. Other adaptive histogram algorithms have been proposed by Van Ryzin (1973) and Wegman (1970).

3.3 PRACTICAL DATA-BASED BIN WIDTH RULES

While simple ideas such as the Normal reference rule in Equation (3.17) are useful, the quest for data-based procedures that approximately minimize the MISE and/or the ISE is the subject of much research. The following topics indicate the variety of that work.

3.3.1 Oversmoothed Bin Widths

In principle, there is no lower bound on h because the unknown density can be arbitrarily rough, although for finite samples the “optimal” bin width may be surprisingly large, as was the case for lognormal data. On the other hand, it turns out that useful *upper* bounds exist for the bin width, depending upon some data-based knowledge of the scale of the unknown density (Terrell and Scott, 1985).

3.3.1.1 Lower Bounds on the Number of Bins

Reexamining the expression for the minimizer of AMISE in Equation (3.15),

$$h^* = \left(\frac{6}{R(f')} \right)^{1/3} n^{-1/3}, \quad (3.39)$$

it follows that any *lower bound* on $R(f')$ leads to an *upper bound* on the bin width. The simplest prior knowledge of scale is that the density is zero outside an interval (a, b) . This suggests the following optimization problem:

$$\min_f \int_{-\infty}^{\infty} f'(x)^2 dx \quad \text{s/t} \quad \text{support of } f = [-0.5, 0.5]. \quad (3.40)$$

Hodges and Lehmann (1956) solved this optimization problem in another setting and showed that the solution is

$$f_1(x) = \frac{3}{2} (1 - 4x^2) I_{[-0.5, 0.5]}(x) = \frac{3}{2} (1 - 4x^2)_+.$$

It is instructive to show that f_1 indeed minimizes $R(f')$. Define $g(x) = f_1(x) + e(x)$, where $e(x)$ is a perturbation function satisfying the following conditions: (i) $e(x) = 0$ outside the interval $[-0.5, 0.5]$; (ii) $e(\pm 0.5) = 0$ and is continuous, as otherwise $g(x)$ would be discontinuous and $R(g') = \infty$; (iii) $\int e(x) dx = 0$ so that g integrates to 1.

The solution must have support exactly on $(-0.5, 0.5)$ and not on a strict subinterval. Rescaling the density to the interval $(-0.5, 0.5)$ reduces the roughness, which would lead to a contradiction. The solution

must be symmetric around 0. If g is asymmetric, consider the symmetric density $[g(x) + g(-x)]/2$, whose derivative is $[g'(x) - g'(-x)]/2$. Using Minkowski's inequality, which states that $[\int (f + g)^2]^{1/2} \leq [\int f^2]^{1/2} + [\int g^2]^{1/2}$, the roughness of the first derivative of this symmetric density equals

$$\frac{1}{2} \sqrt{\int [g'(x) - g'(-x)]^2 dx} \leq \frac{1}{2} \left[\sqrt{R(g')} + \sqrt{R(g'(-x))} \right] = \sqrt{R(g')} ,$$

which contradicts the optimality of g .

To verify the optimality of f_1 , directly compute

$$\int g'(x)^2 dx = \int f'_1(x)^2 dx + \int e'(x)^2 dx + 2 \int f'_1(x)e'(x) dx . \quad (3.41)$$

The last integral vanishes as

$$\int_{-1/2}^{1/2} f'_1(x)e'(x) dx = f'_1(x)e(x) \Big|_{x=-1/2}^{x=1/2} - \int_{-1/2}^{1/2} f''_1(x)e(x) dx = 0 ,$$

because $f''_1(x) = -12$, a constant. Therefore, from Equation (3.41), $R(g') = R(f'_1) + R(e')$, so that $R(g') \geq R(f'_1)$, which proves the result. Notice that f_1 is the unique quadratic kernel that satisfies all the side conditions.

To apply this result, note that $R(f') = 12$. If f_1 is linearly mapped to the interval (a, b) , then the lower bound is $R(f') \geq 12/(b - a)^3$. Therefore,

$$h^* = \left(\frac{6}{nR(f')} \right)^{1/3} \leq \left(\frac{6(b-a)^3}{n \cdot 12} \right)^{1/3} = \frac{b-a}{\sqrt[3]{2n}} \equiv h_{OS} ,$$

which is the *oversmoothed bandwidth*. Rearranging gives

$$\text{number of bins} = \frac{b-a}{h^*} \geq \frac{b-a}{h_{OS}} = \sqrt[3]{2n} . \quad (3.42)$$

Terrell and Scott (1985) showed that the sample range may be used if the interval (a, b) is unknown or even if $b - a = \infty$ and the tail is not too heavy.

Examples: For the Buffalo snowfall data ($n = 63$; see Table 9 in Appendix B) and the LRL data ($n = 25,752$; see Table 8 in Appendix B), rule (3.42) suggests 5 and 37 bins, respectively. For the snowfall data, the sample range is $126.4 - 25.0 \approx 100$; therefore, the oversmoothed bandwidth $h_{OS} = 20$ inches of snow is suggested; see Figure 3.15. There is no hint of the trimodal behavior that appears to be in the histogram with more bins. The LRL data were

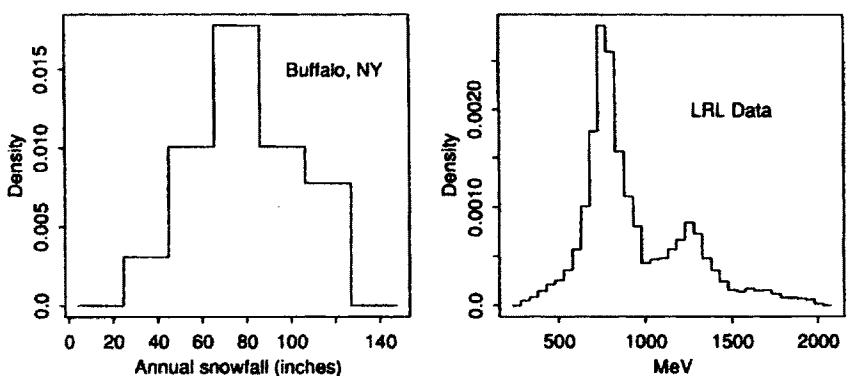


Figure 3.15 Oversmoothed histograms of the Buffalo snowfall and LRL data.

recorded in 172 consecutive bins, so that $h_{os} = (2,000 - 280)/37 = 46.5$. Since the original data were rounded to bins of width of 10 MeV, the only choices available for h_{os} are 40 or 50 MeV. Strictly speaking, combining only 4 adjacent bins ($h = 40$) is not oversmoothing, so the slightly conservative choice $h = 50$ is adopted; see Figure 3.15. At least 3 groups appear in this histogram, the third around $x = 1,700$.

3.3.1.2 Upper Bounds on Bin Widths

The previous result is most notable for its simplicity and mnemonic value. However, the focus is on choosing the smoothing parameter rather than the number of bins. Oversmoothing provides a solution in that setting as well.

Consider the optimization problem (3.40) except with the fixed-range constraint replaced by the constraint that the variance of f equal σ^2 . Terrell (1990) showed that the solution to this problem is

$$f_2(x) = \frac{15}{16\sqrt{7}\sigma} \left(1 - \frac{x^2}{7\sigma^2}\right)^2 I_{[-\sqrt{7}\sigma, \sqrt{7}\sigma]}(x).$$

It can be checked that $R(f'_2) = 15\sqrt{7}/(343\sigma^3)$; hence,

$$h^* \leq \left(\frac{6}{nR(f'_2)}\right)^{1/3} = \left(\frac{686\sigma^3}{5\sqrt{7}n}\right)^{1/3} \approx 3.729\sigma n^{-1/3} \equiv h_{os}. \quad (3.43)$$

Terrell (1990) considered other less common scale constraints. The version based on the interquartile range (IQ) is particularly robust:

$$h^* \leq 2.603(\text{IQ})n^{-1/3} \equiv h_{\text{OS}}. \quad (3.44)$$

EXAMPLE: For $N(\mu, \sigma^2)$ data,

$$h^* = 3.5\sigma n^{-1/3} < 3.729\sigma n^{-1/3} = h_{\text{OS}} \quad (= \text{upper bound}), \quad (3.45)$$

but h^* is only 93.6% of h_{OS} . Apparently, the Normal density is very smooth and using the simple rule (3.17) is further justified. In the case of lognormal data, for which $\sigma^2 = e(e - 1)$,

$$h^* = 1.44n^{-1/3} < 3.729 \times 2.161 \times n^{-1/3} = 8.059n^{-1/3}, \quad (3.46)$$

which is 18% of h_{OS} . On the other hand, for $n = 25$, the best $h \approx 1.28$ from Figure 3.10, which is 46% of $h_{\text{OS}} = 2.756$. For Cauchy data, the bound based on variance is asymptotically meaningless, whereas the oversmoothed rule based on interquartile range still applies.

3.3.2 Biased and Unbiased Cross-Validation

The cross-validation approach to the automatic data-based calibration of histograms is introduced in this section. Cross-validation (CV) algorithms *reuse the data*. The goal is not simply to produce a consistent sequence of bin widths—even Sturges’ rule is consistent. Rather, the goal is to reliably produce bin widths \hat{h}_{CV} that are close to h^* for *finite samples* or perhaps even more optimistically, bin widths that minimize ISE errors for *individual samples*.

3.3.2.1 Biased Cross-Validation

The only unknown quantity in the AMISE is $R(f')$, which may be estimated using the data at hand. An approximation of $f'(t_k)$ is available based on a finite difference of the histogram at the midpoints of bins B_k and B_{k+1} , namely, $\hat{f}'(t_k) = [\nu_{k+1}/(nh) - \nu_k/(nh)]/h$. A potential estimate of $R(f')$ is

$$\hat{R}_1 = \sum_k [\hat{f}'(t_k)]^2 \cdot h = \frac{1}{n^2 h^3} \sum_k (\nu_{k+1} - \nu_k)^2. \quad (3.47)$$

It can be shown (see Problem 16) that

$$\mathbb{E}[\hat{R}_1] = R(f') + 2/(nh^3) + O(h). \quad (3.48)$$

With optimal smoothing, $2/(nh^3)$ converges to $R(f')/3$ by Theorem 3.3. It follows that \hat{R}_1 is a biased estimator of $R(f')$, too large by a factor of a third, so that $\frac{3}{4}\hat{R}_1$ is an asymptotically unbiased estimator of $R(f')$. Alternatively,

$$\hat{R}_h(f') = \frac{1}{n^2 h^3} \sum_k (\nu_{k+1} - \nu_k)^2 - \frac{2}{nh^3}. \quad (3.49)$$

Substituting (3.49) into the AMISE expression (3.15) gives a *biased cross-validation* (BCV) estimate of the $MISE(h)$:

$$BCV(h) = \frac{5}{6nh} + \frac{1}{12n^2 h} \sum_k (\nu_{k+1} - \nu_k)^2, \quad (3.50)$$

where ν_k is recomputed as h and the mesh vary (for definiteness, the bin origin, t_0 , remains fixed). The bias simply refers to the fact that the $AMISE(h)$ is a biased approximation to the true $MISE(h)$. The BCV bin width, \hat{h}_{BCV} , is defined to be the minimizer of $BCV(h)$, subject to the constraint $h \leq h_{OS}$. The theoretical properties of these random variables have been studied by Scott and Terrell (1987), who showed that \hat{h}_{BCV} converges to h^* with a relative error of $O(n^{-1/6})$. This rate of (relative) convergence is especially slow. However, the algorithm seems useful in practice, especially for large samples. Examples are given in Section 3.3.2.4.

3.3.2.2 Unbiased Cross-Validation

Rudemo (1982) undertook the more ambitious task of attempting to minimize the actual L_2 ISE error of a histogram for an individual sample. Expanding yields

$$ISE(h) = \int [\hat{f}(x) - f(x)]^2 dx = R(\hat{f}) - 2 \int \hat{f}(x)f(x) dx + R(f).$$

Rudemo noted that the minimizer of $ISE(h)$ did not depend on the unknown quantity $R(f)$ and that $R(\hat{f})$ could be computed easily. Furthermore, he observed that the second integral could be written as $E[\hat{f}(X)]$, where the expectation is with respect to the point of evaluation and not over the random sample x_1, \dots, x_n . Cross-validation methodology suggests removing one data point and using the remaining $n - 1$ points to construct an estimator of $E[\hat{f}(X)]$. The i th data point is then evaluated in the estimate for the purpose of determining the quality of fit. This step is repeated n times, once for each data point, and the results averaged.

Rudemo considered the histogram $\hat{f}_{-i}(x)$, which is the histogram based on the $n - 1$ points in the sample excluding x_i . It is easy to check that the observable random variable $\hat{f}_{-i}(X_i)$ has the same mean as the unobservable

random variable $E\hat{f}(X)$, although based on a sample of $n - 1$ rather than n points. By leaving out each of the n data points one at a time, Rudemo obtained a stable estimate of $E\hat{f}(X)$ by averaging over the n cases. With this estimate, he proposed minimizing the least-squares CV or *unbiased cross-validation* (UCV) function

$$\text{UCV}(h) = R(\hat{f}) - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(x_i). \quad (3.51)$$

For the histogram, the terms in the UCV are easily evaluated. For example, $\hat{f}_{-i}(x_i) = (\nu_k - 1)/[(n - 1)h]$ if $x_i \in B_k$. The final computational formula for the UCV equals

$$\text{UCV}(h) = \frac{2}{(n - 1)h} - \frac{n + 1}{n^2(n - 1)h} \sum_k \nu_k^2; \quad (3.52)$$

see Problem 17. The similarity of the UCV and BCV formulas is striking and partially justifies the BCV label, since BCV is based on AMISE rather than MISE. Methods relying on AMISE formula are often referred to as *plug-in* (PI) methods. Scott and Terrell (1987) also studied the asymptotic distributional properties of the UCV and \hat{h}_{UCV} random variables. They showed that the UCV and BCV bandwidths were consistent, but that the convergence was slow. Specifically, $\sigma_{h_{\text{UCV}}}/h_{\text{UCV}} = O(n^{-1/6})$. However, Hall and Marron (1987a, b) showed that this is the same rate of convergence of the smoothing parameter that actually minimizes ISE. Hall and Marron (1987c) also studied the estimation of functionals like $R(f')$.

3.3.2.3 End Problems with BCV and UCV

For a fixed sample, $\text{BCV}(h)$ and $\text{UCV}(h)$ exhibit behavior not found in $\text{MISE}(h)$ for large and small h , respectively. For example, as $h \rightarrow \infty$, all bins except one or two will be empty. It is easy to see that

$$\lim_{h \rightarrow \infty} \text{BCV}(h) = \lim_{h \rightarrow \infty} \text{UCV}(h) = 0. \quad (3.53)$$

Clearly, $\text{ISE}(h) \rightarrow R(f)$ as $h \rightarrow \infty$. Hence, for $\text{UCV}(h)$, the limit in (3.53) is correct since the term $R(f)$ is omitted in its definition. However, for $\text{BCV}(h)$, which is asymptotically nonnegative near h^* , the limiting value of 0 means that the global minimizer is actually at $h = \infty$. In practice, \hat{h}_{BCV} is chosen to be a local minimizer constrained to be less than the oversmoothed bandwidth. If there is no local minimizer within that region, the oversmoothed bandwidth itself is the *constrained minimizer*.

On the other hand, as $h \rightarrow 0$, the bin counts should all be zero or one, assuming the data are continuous (no ties). In that case, both $\text{BCV}(h)$ and $\text{UCV}(h)$ approximately equal $1/(nh)$, the correct expression for the integrated variance (IV). Observe that $\text{IV}(h) \rightarrow +\infty$ as $h \rightarrow 0$. However, if the data contain many ties, perhaps the result of few significant digits or rounding, $\text{UCV}(h)$ can diverge. For example, suppose the n data points consist of n/m distinct values, each replicated m times; then $\text{UCV}(h)$ is approximately $(2 - m)/(nh)$, which diverges to $-\infty$ (as opposed to $+\infty$) if $m > 2$ as $h \rightarrow 0$. Thus the global minimizer of $\text{UCV}(h)$ occurs at $h = 0$ in that case. In such situations, the divergence may occur only in a small neighborhood around $h = 0$, and \hat{h}_{UCV} is again defined to be the appropriate local minimum, if it exists, closest to the oversmoothed bandwidth; see, however, Figure 3.18.

3.3.2.4 Applications

Figure 3.16 displays the BCV and UCV curves for a $N(0, 1)$ sample of 1,000 points. The vertical scales are comparable, since UCV is shifted by a fixed constant, $R(f)$. Thus the UCV function turns out to be much noisier than the BCV function; however, for other more difficult data, the minimizer of the BCV function may be quite poor—typically, it is biased towards larger bin widths. The UCV function may be noisy, but its minimizer is correct on average. For these data, the two minimizers are $h_{\text{BCV}} = 0.32$ and $h_{\text{UCV}} = 0.36$, while the oversmoothed bin width is $h_{\text{OS}} = 0.37$, using the variance rule in Equation (3.43). Choosing h_{UCV} from the many strong local minimizers in the UCV function is a practical problem.

Taken together, oversmoothing, BCV, and UCV are a powerful set of tools for choosing a bin width. Specifically, the UCV and BCV curves should be plotted on comparable scales, marking the location of the upper bound given by the oversmoothed bandwidth. A log-log plot is recommended, but UCV is negative, so only $\log(h)$ is plotted. The minimizer is located and the *quality* of that CV is evaluated subjectively by examining how well-articulated the minimizer appears. Watch for obvious failures: no local minimizer in $\text{BCV}(h)$

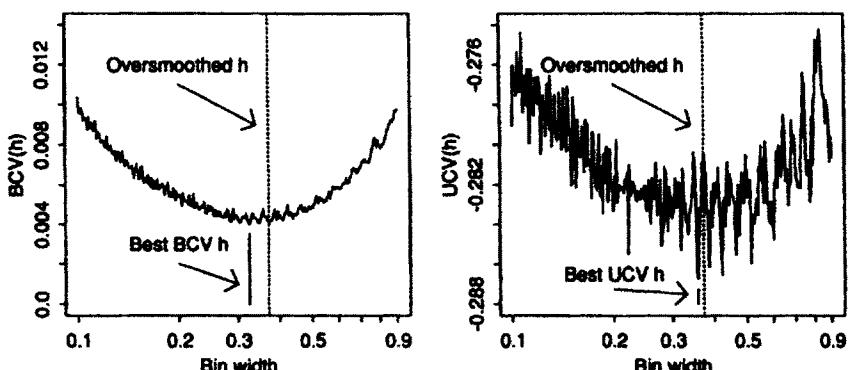


Figure 3.16 BCV and UCV functions for a $N(0, 1)$ sample of 1,000 points.

or the degenerate $h = 0$ UCV solution. If there is no agreement among the proposed solutions, examine plots of the corresponding histograms and choose the one that exhibits a small amount of local noise, particularly near the peaks.

The second example focuses on 1983 income sampled from 5,626 German households (DIW, 1983). The data were transformed to a log scale, as is often done by economists, in order to reduce the MISE penalty due to extreme skewness. Four histograms of these transformed data points are plotted in Figure 3.17. For more details, see Scott and Schmidt (1988, 1989).

Which histogram is best? Sturges' rule suggests only 13 bins, which is clearly oversmoothed. The oversmoothed rule (maximum bin width) gives $h = 0.160$, corresponding to 68 bins over the sample range (not shown). The BCV and UCV functions are plotted in Figure 3.18. The minimizer of the BCV criterion occurs at $h = 0.115$, suggesting 95 bins. The UCV function appears to have its global minimum at $h = 0$, but has no strong local minimum in the region of interest. Indeed, examination of the original data revealed only 3,322 unique incomes, so the duplicated data problem is clearly in effect. It was determined empirically that this effect could be eliminated by adding some small uniform noise to the logarithmic values, as shown in Figure 3.18. However, a wide range of possible minima still exists. Even the small bin width $h = 0.055$ seems feasible, which corresponds to 200 bins. When $U(-0.005, 0.005)$ noise was added, the UCV curve was horizontal, splitting the two curves shown in the Figure. For large amounts of additive noise, the UCV curve was visually

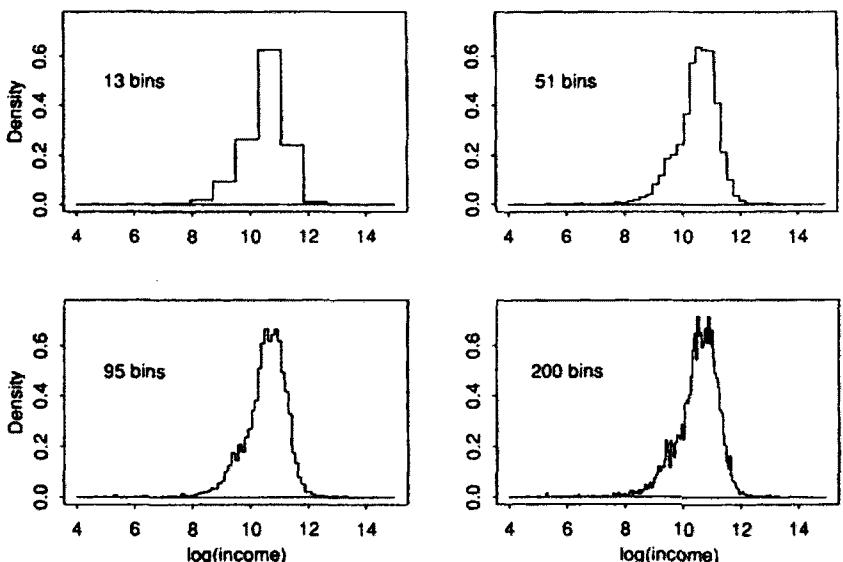


Figure 3.17 Four histograms of the 1983 German income sample of 5,625 households.

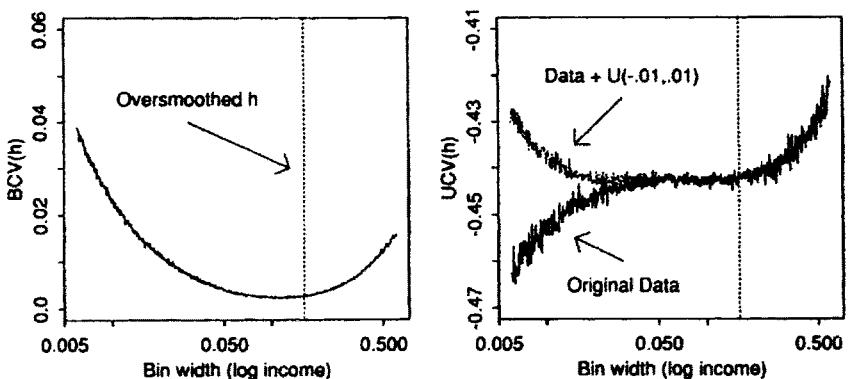


Figure 3.18 BCV and UCV functions of the 1983 German income sample of 5,625 households. A second UCV curve is shown for the blurred logarithmic data.

affected for bin widths near h_{OS} . Thus there is a bit of subjectiveness in the process of removing the duplicate data.

The lesson is that all three algorithms should be examined simultaneously even with very large samples. This topic will be continued in Chapter 6.

3.4 L_2 THEORY FOR MULTIVARIATE HISTOGRAMS

The derivation of the MISE for the multivariate histogram is only slightly more complicated than in the univariate case. Given a sample from $f(\mathbf{x})$, where $\mathbf{x} \in \Re^d$, the histogram is determined by a partition of the space. Consider a regular partition by hyper-rectangles of size $h_1 \times h_2 \times \cdots \times h_d$. Choosing hypercubes as bins would be sufficient if the data were properly scaled, but in general that will not be the case. Further improvements may be obtained by considering nonregular or rotated bins; see Scott (1988a) and Hüsemann and Terrell (1991).

Consider a generic hyper-rectangular bin labeled B_k containing ν_k points. As usual, $\sum_k \nu_k = n$. Then

$$\hat{f}(\mathbf{x}) = \frac{\nu_k}{nh_1h_2\cdots h_d} \quad \text{for } \mathbf{x} \in B_k.$$

The variance of the histogram is constant over each bin and is given by

$$\text{Var } \hat{f}(\mathbf{x}) = \frac{n p_k (1 - p_k)}{(nh_1h_2\cdots h_d)^2} \quad \text{for } \mathbf{x} \in B_k. \quad (3.54)$$

Integrating the variance over the bin simply multiplies (3.54) by the volume of the hyper-rectangle, $h_1 h_2 \cdots h_d$. Summing the variance contributions from all bins gives the integrated variance as

$$\text{IV} = \frac{1}{nh_1 h_2 \cdots h_d} - \frac{R(f)}{n} + o\left(\frac{1}{n}\right), \quad (3.55)$$

where $R(f) = \int_{\mathbb{R}^d} f(\mathbf{x})^2 d\mathbf{x}$. The first term in (3.55) is exact, since $\sum_k p_k = 1$; the remainder is obtained by noting that $p_k^2 \approx [h_1 h_2 \cdots h_d f(\xi_k)]^2$, where $\xi_k \in B_k$, and using standard multivariate Riemannian integration approximations.

The following outline for the bias calculation can be made rigorous. Consider the bin B_0 centered on the origin $\mathbf{x} = \mathbf{0}$. Now

$$f(\mathbf{x}) = f(\mathbf{0}) + \sum_{i=1}^d x_i f_i(\mathbf{0}) + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d x_i x_j f_{ij}(\mathbf{0}) + O(h^3),$$

where $f_i(\mathbf{x}) = \partial f(\mathbf{x}) / \partial x_i$ and $f_{ij}(\mathbf{x}) = \partial^2 f(\mathbf{x}) / \partial x_i \partial x_j$. Hence,

$$p_0 = \int_{-h_d/2}^{h_d/2} \cdots \int_{-h_1/2}^{h_1/2} f(\mathbf{x}) d\mathbf{x} = h_1 h_2 \cdots h_d f(\mathbf{0}) + O(h^{d+2}), \quad (3.56)$$

where $h = \min_i(h_i)$. Computing the $\text{Bias}\{\hat{f}(\mathbf{x})\}$ gives

$$\mathbb{E}\hat{f}(\mathbf{x}) - f(\mathbf{x}) = \frac{p_0}{h_1 h_2 \cdots h_d} - f(\mathbf{x}) = - \sum_{i=1}^d x_i f_i(\mathbf{0}) + O(h^2). \quad (3.57)$$

Squaring and integrating over B_0 , the integrated squared bias for the bin is

$$h_1 h_2 \cdots h_d \left[\sum_{i=1}^d \frac{1}{12} h_i^2 f_i(\mathbf{0})^2 + O(h^4) \right]. \quad (3.58)$$

A similar expression holds for all other bins, with the origin $\mathbf{x} = \mathbf{0}$ replaced by the respective multivariate bin centers. Summing over all bins yields

$$\text{AISB}(\mathbf{h}) = \sum_{i=1}^d h_i^2 R(f_i)/12. \quad (3.59)$$

These approximations are collected in a theorem.

Theorem 3.5: *For a sufficiently smooth density function $f(\mathbf{x})$, the multivariate MISE is asymptotically*

$$\text{AMISE}(\mathbf{h}) = \text{AIV} + \text{AISB} = \frac{1}{nh_1 h_2 \cdots h_d} + \frac{1}{12} \sum_{i=1}^d h_i^2 R(f_i). \quad (3.60)$$

The asymptotically optimal bin widths, h_k^ , and resulting AMISE* are*

$$h_k^* = R(f_k)^{-1/2} \left(6 \prod_{i=1}^d R(f_i)^{1/2} \right)^{1/(2+d)} n^{-1/(2+d)}, \quad (3.61)$$

$$\text{AMISE}^* = \frac{1}{4} 6^{2/(2+d)} \left(\prod_{i=1}^d R(f_i) \right)^{1/(2+d)} n^{-2/(2+d)}. \quad (3.62)$$

EXAMPLE: Suppose that $X \sim N(\mu, \Sigma)$, $\Sigma = \text{Diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2)$. Then, letting $c_d = \sigma_1 \sigma_2 \cdots \sigma_d$, we have

$$R(f_i) = (2^{d+1} \pi^{d/2} \sigma_i^2 c_d)^{-1} \quad (3.63)$$

and

$$\begin{aligned} h_k^* &= 2 \cdot 3^{1/(2+d)} \pi^{d/(4+2d)} \sigma_k n^{-1/(2+d)} \\ \text{AMISE}^* &= 2^{-(1+d)} 3^{2/(2+d)} \pi^{-d^2/(4+2d)} c_d^{-1} n^{-2/(2+d)}. \end{aligned} \quad (3.64)$$

Note that the constant in the bandwidth increases slowly from 3.4908 in one dimension to the limiting value of $2\sqrt{\pi} = 3.5449$ as $d \rightarrow \infty$. Hence a very useful formula to memorize is

Normal reference rule:	$h_k^* \approx 3.5 \sigma_k n^{-1/(2+d)}$.	(3.65)
------------------------	---	--------

3.4.1 Curse of Dimensionality

Bellman (1961) first coined the phrase “curse of dimensionality” to describe the exponential growth in combinatorial optimization as the dimension increases. Here, it is the number of bins that grows exponentially as the dimension

Table 3.6 Example of Asymptotically Optimal Bin Widths and Errors for $f = N(\mathbf{0}, I_d)$

Dimension d	h_d^*	AMISE $_d^*$
1	$3.491n^{-1/3}$	$0.430n^{-2/3}$
2	$3.504n^{-1/4}$	$0.122n^{-2/4}$
3	$3.512n^{-1/5}$	$0.035n^{-2/5}$
4	$3.518n^{-1/6}$	$0.010n^{-2/6}$

increases. It is important to try to see how histograms are affected by this phenomenon.

A relatively simple density to estimate is the multivariate Normal with $\Sigma = I_d$. Following upon the results in the example above, we have Table 3.6. Clearly, the rate of decrease of the MISE with respect to the sample size degrades rapidly as the dimension increases compared to the ideal parametric rate $O(n^{-1})$. A strong advantage of parametric modeling is that the rate of decrease of MISE is independent of dimension. However, several data analysts have used histograms when classifying quadrivariate data from remote sensing with satisfactory results (Wharton, 1983). One possibly optimistic observation is that the constants in the AMISE in Table 3.6 are also decreasing as the dimension increases. Unfortunately, MISE is not a dimensionless quantity and, hence, these coefficients are not directly comparable.

Epanechnikov (1969) described a procedure for comparing histogram errors and performance across dimensions. He considered one possible dimensionless rescaling of MISE:

$$\epsilon_d \equiv \frac{\text{MISE}}{R(f)} \quad \left\{ \approx 2^{-1} 3^{\frac{2}{2+d}} \pi^{\frac{d}{2+d}} n^{-\frac{2}{2+d}} \quad \text{when } f = N(\mathbf{0}, I_d) \right\}. \quad (3.66)$$

Again, for the Normal case, a table of equivalent sample sizes may be computed (Table 3.7). This table graphically illustrates the curse of dimensionality and the intuition that density estimation in more than two or three dimensions will not work. This conclusion is, however, much too pessimistic, and other evidence will be presented in Section 7.2.

Table 3.7 Equivalent Sample Sizes Across Dimensions for the Multivariate Normal Density, Based on Epanechnikov's Criterion

d	Columns of Equivalent Sample Sizes		
1	10	100	1,000
2	22	471	10,155
3	48	2,222	103,122
4	105	10,472	1,047,198
5	229	49,360	10,634,200

Another approach towards understanding this problem is to count bins in a region of interest. For this same Normal example, consider the region of interest to be a sphere with radius r_d , containing 99% of the probability mass. The radius r_d is the solution to

$$\text{Prob}\left(\sum_{i=1}^d Z_i^2 \leq r_d^2\right) = 0.99 \quad \Rightarrow \quad r_d = \sqrt{\chi^2_{.99}(d)}, \quad (3.67)$$

that is, the square root of the appropriate chi-squared quantile. The volume of the sphere is given by Equation (1.3). The number of hypercube bins over the data in this sphere is approximately equal to the volume of the sphere divided by the bin volume, h^d , plus the bins covering the 1% of points outside the sphere. Table 3.8 illustrates these calculations for a sample of size 1,000. In five dimensions, the optimally smoothed histogram has approximately 1,250 bins in the region of interest. Since there are only 1,000 points, it is clear that most of these bins will be empty and that the histogram will be rather rough. Scott and Thompson (1983) have called this the “empty space phenomenon.” As the dimension grows, the histogram provides reasonable estimates only near the mode and away from the tails.

3.4.2 A Special Case: $d = 2$ with Nonzero Correlation

The effects, if any, of correlation have been ignored up to this point. Consider a simple bivariate case. If $f(x_1, x_2) = N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then $R(f_1) = [8\pi(1 - \rho^2)^{3/2}\sigma_1^3\sigma_2]^{-1}$ and $R(f_2) = [8\pi(1 - \rho^2)^{3/2}\sigma_1\sigma_2^3]^{-1}$. From Theorem 3.5,

$$\begin{aligned} h_i^* &= 3.504 \sigma_i (1 - \rho^2)^{3/8} n^{-1/4} \\ \text{AMISE}^* &= \frac{0.122}{\sigma_1 \sigma_2} (1 - \rho^2)^{-3/4} n^{-1/2}. \end{aligned} \quad (3.68)$$

The effect of the correlation ρ is to introduce powers of the quantity $(1 - \rho^2)$ into the equations. Thus, if the data are not independent but are clustering along a line, smaller bin widths are required to “track” this feature. If the

Table 3.8 Approximate Number of Bins in the Region of Interest for a Multivariate Histogram of 1,000 Normal Points

d	h_d^*	r_d	Number of Bins
1	0.35	2.57	15
2	0.62	3.03	75
3	0.88	3.37	235
4	1.11	3.64	573
5	1.30	3.88	1,254

density is degenerate (i.e., $\rho = \pm 1$), then the MISE blows up. This result also indicates that if the data fall onto any lower-dimensional manifold, a histogram will never be consistent! This inconsistency is a second and perhaps more important aspect of the “curse of dimensionality” as it applies to statistics. Therefore, a significant portion of the effort in good data analysis should be to check the (local) rank of the data and identify nonlinear structure falling in lower dimensions; see Chapter 7.

3.4.3 Optimal Regular Bivariate Meshes

Consider tiling the plane with regular polygons as bins for a bivariate histogram. Other than squares, there are only two other regular polygon meshes: equilateral triangles and hexagons; see Figure 3.19. Nonregular tiling of the plane is of course feasible, but is of secondary interest. Scott (1988a) compared the bivariate AMISE for these three regular tile shapes. If the individual bins are parameterized so that each has the same area h^2 , then Scott showed that

$$\text{AMISE}(h) = \frac{1}{nh^2} + ch^2[R(f_1) + R(f_2)], \quad (3.69)$$

where $f_1 = \partial f(x, y)/\partial x$ and $f_2 = \partial f(x, y)/\partial y$, and

$$c = \left[\frac{1}{12}, \frac{1}{6\sqrt{3}}, \frac{5}{36\sqrt{3}} \right] = \left[\frac{1}{12}, \frac{1}{10.39}, \frac{1}{12.47} \right], \quad (3.70)$$

for the square, triangle, and hexagonal tiles, respectively; see Problem 25. Therefore, hexagonal bins are in fact the best, but only marginally. The triangular meshes are quite inferior. Scott also considered meshes composed of right triangles. He noted that these were inferior to equilateral triangles and that the ISB sometimes included cross-product terms such as $\int \int f_1(x, y)f_2(x, y) dx dy$.

Carr et al. (1987) suggested using hexagonal bins for a different reason. They were evaluating plotting bivariate glyphs based on counts in bivariate bins. Using square bins resulted in a visually distracting alignment of the glyphs in the vertical and horizontal directions. This distraction was virtually eliminated

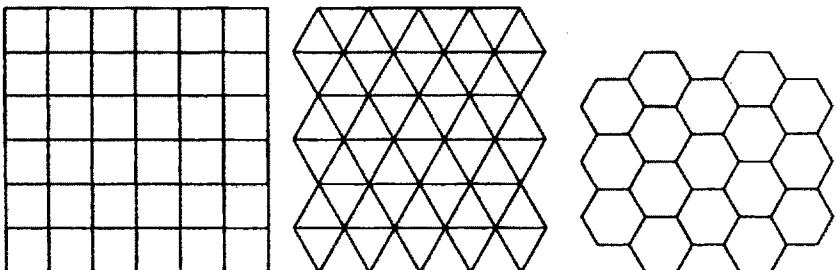


Figure 3.19 The 3 possible regular polygon meshes for bivariate histograms.

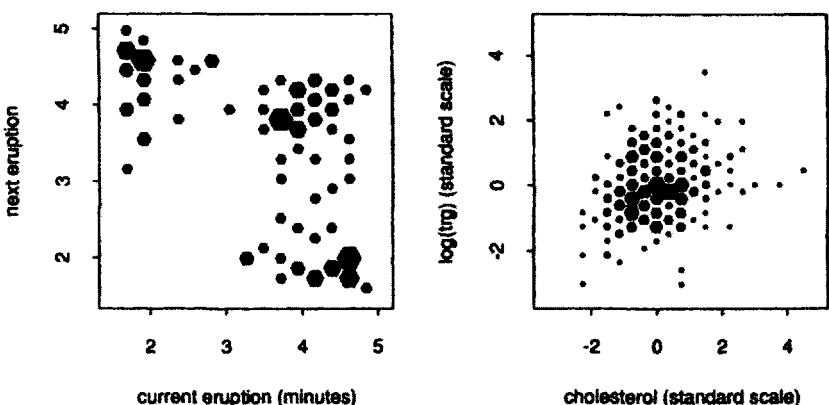


Figure 3.20 Hexagon glyphs for the lagged Old Faithful duration data and for the lipid data for 320 males with heart disease.

when the glyphs were located according to a hexagonal mesh. The elimination is readily apparent in the examples given in Figure 3.20. Note that the size of each hexagon glyph is drawn so that its area is proportional to the number of observations in that bin. The authors note the method is especially useful for very large data sets. Observe that while the MISE criterion does not strongly differentiate between square or hexagonal bins, subjective factors may strongly influence the equation one way or the other.

3.5 MODES AND BUMPS IN A HISTOGRAM

Many of the examples to this point have shown multimodal data. Good and Gaskins (1980) discussed advanced methods for finding modes and bumps in data. In 1-D, a *mode* is a set (a collection of points) where $f'(x) = 0$ and $f''(x) < 0$. A *bump* is a set (a collection of disjoint intervals) where $f''(x) < 0$. Thus bump hunting is more general than estimating the mode. Figure 3.21 shows a histogram which contains one mode and one bump. A bump does not necessarily contain a mode, although a mode is always located in a bump. In practical bump-hunting situations, the density is often thought to be a mixture of several component densities; see Izenman and Sommer (1988), for example. The mixture problem is even more general than bump hunting, since a Normal mixture density such as

$$f(x) = \sum_{i=1}^q w_i \phi(x; \mu_i, \sigma_i^2) \quad \text{where} \quad \sum_{i=1}^q w_i = 1, \quad (3.71)$$

need not exhibit modes or bumps; see Problem 26. However, the estimation of the parameters $\{q, w_i, \mu_i, \sigma_i^2\}$ is often ill-conditioned, especially if q is larger

than the true number of densities; see Day (1969), Everitt and Hand (1981), Hathaway (1982), and Redner and Walker (1984).

With histogram data, it is natural to examine plots of (standardized) first and second differences for evidence of modes and bumps, respectively:

$$\frac{\nu_{k+1} - \nu_k}{\sqrt{\nu_{k+1} + \nu_k}} \quad \text{and} \quad \frac{\nu_{k+1} - 2\nu_k + \nu_{k-1}}{\sqrt{\nu_{k+1} + 4\nu_k + \nu_{k-1}}}, \quad (3.72)$$

since the bin counts can be approximately modeled as independent Poisson random variables. For example, $\text{Var}(\nu_{k+1} - \nu_k) \approx \nu_{k+1} + \nu_k$. Good and Gaskins found 13 bumps in their LRL data; see Figure 3.21. A plot of the standardized second differences with $h = 30$ MeV is shown in Figure 3.22. Recall that these data were collected in bins of width 10 MeV. For narrower bin widths, the figure was too noisy. Both cross-validation criteria give $h = 10$ MeV as the best bin width. Most of the bumps found by Good and Gaskins can be seen in the plot. However, a few bumps are not readily apparent. One large negative portion of the curve near $x = 1,145$ MeV seems to have been missed.

3.5.1 Properties of Histogram “Modes”

Suppose, without loss of generality, that the true density f has a mode at 0. Consider an equally spaced histogram with bin B_0 centered on $x = 0$. The bin count $\nu_0 \sim B(n, p_0) \approx P(\lambda_0)$, which is the Poisson density with $\lambda_0 = np_0$. Asymptotically, the adjacent bin counts, $(\nu_{-k}, \dots, \nu_0, \dots, \nu_k)$, are independent and Normally distributed with $\nu_i \approx N(\lambda_i, \lambda_i)$. Consider the following question: What is the probability that the histogram will have a sample mode in bin B_0 ?

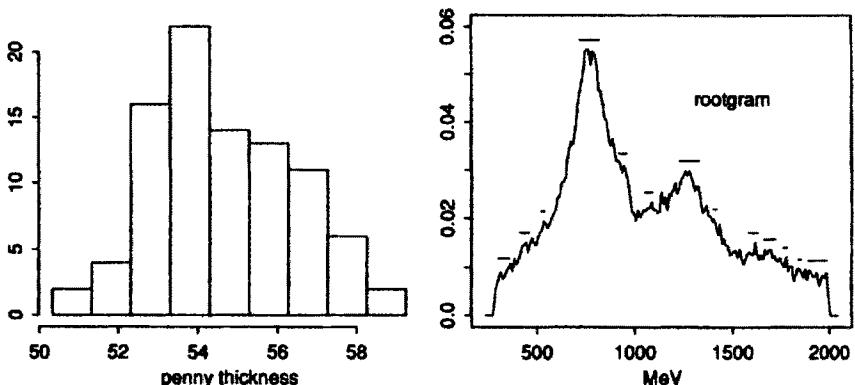


Figure 3.21 Histogram of U.S. penny thickness with one mode and a bump. In the right frame, the histogram of the LRL data with $h = 10$ MeV is plotted on a square root scale; the 13 bumps found by Good and Gaskins are indicated by the line segments above the histogram.

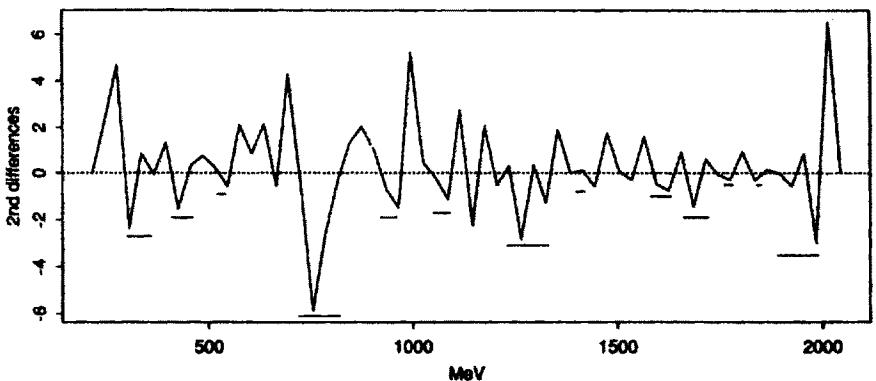


Figure 3.22 Standardized second differences of LRL histogram data with bin width of 30 MeV. The 13 bumps found by Good and Gaskins are indicated as before.

Conditioning on the observed bin count in B_0 , we obtain

$$\begin{aligned} \Pr\left(\nu_0 = \arg \max_{|j| \leq k} \nu_j\right) &= \sum_{x_0} \Pr\left(\nu_0 = \arg \max_{|j| \leq k} \nu_j \mid \nu_0 = x_0\right) f_{\nu_0}(x_0) \\ &= \sum_{x_0} \Pr(\nu_j < x_0; |j| \leq k, j \neq 0) f_{\nu_0}(x_0) \\ &\approx \int_{x_0} \prod_{\substack{j=-k \\ j \neq 0}}^k \Phi\left(\frac{x_0 - \lambda_j}{\sqrt{\lambda_j}}\right) \phi\left(\frac{x_0 - \lambda_0}{\sqrt{\lambda_0}}\right) dx_0, \end{aligned} \quad (3.73)$$

using the Normal approximation for $\Pr(\nu_j < x_0)$ and replacing the sum by an integral. Now $\lambda_j = np_j$. Using the approximation for p_j [since $f'_0 \equiv 0$],

$$\begin{aligned} p_j &= \int_{(j-\frac{1}{2})h}^{(j+\frac{1}{2})h} \left[f_0 + \frac{1}{2}x^2 f''_0 + \dots \right] dx \\ &= hf_0 + \frac{1}{6}h^3 \left(3j^2 + \frac{1}{4} \right) f''_0 + \dots \end{aligned} \quad (3.74)$$

it follows that (3.73) equals

$$\begin{aligned} \Pr\left(\nu_0 = \arg \max_{|j| \leq k} \nu_j\right) &\approx \int_y \prod_{\substack{j=-k \\ j \neq 0}}^k \Phi\left(\frac{\lambda_0 - \lambda_j + y\sqrt{\lambda_0}}{\sqrt{\lambda_j}}\right) \phi(y) dy \\ &\approx \int_y \prod_{\substack{j=-k \\ j \neq 0}}^k \Phi\left(y - \frac{j^2 h^{5/2} \sqrt{n}}{2} \frac{f''(0)}{\sqrt{f(0)}} + \dots\right) \phi(y) dy. \end{aligned} \quad (3.75)$$

In the case of an optimal histogram, $h = cn^{-1/3}$, so that

$$\lim_{n \rightarrow \infty} [h^{5/2} \sqrt{n}] = O(n^{-1/3}) \rightarrow 0;$$

hence,

$$\lim_{n \rightarrow \infty} \Pr \left(\nu_0 = \arg \max_{|j| \leq k} \nu_j \right) = \int_y \Phi(y)^{2k} \phi(y) dy = \frac{1}{2k + 1} \quad (!!) \quad (3.76)$$

from Equation (3.75). The probability $1/(2k + 1)$ is far from a more desirable value close to 1. The correct interpretation of this result is that optimal MISE-smoothing results in bins with widths too narrow to estimate modes. In fact, in a neighborhood of the mode, the density looks flat (and the bins have essentially the same expected height to first order), so that each of the $2k + 1$ bins is equally likely to be the *sample mode*. In a simulation of optimally smoothed Normal data with $n = 256,000$, the average number of sample modes in the histogram was 20! Now admittedly, most of these “modes” were just small aberrations, but the result is most unexpected; see Figure 3.23.

Next, suppose that the origin is not a mode; then $f'(0) \neq 0$. A similar analysis shows the probability that bin B_0 is a mode (which it is not!) converges to a fixed nonzero probability as $n \rightarrow \infty$. The probability is smaller the larger the magnitude of $f'(0)$. If wider bins are used, then the probabilities above can be made to converge to the desired values of 1 and 0, respectively. An interesting special case for the limit in Equation (3.76) follows from the use of a much larger bin width $h = cn^{-1/5}$. This choice will be examined in the next chapter, which deals with the closely related frequency polygon density estimator.

3.5.2 Noise in Optimal Histograms

Consider again the histograms of a million Normal points shown in Figure 3.6. Plotting a rootgram, which is the variance-stabilized histogram plotted on a square root scale, for the cases $h = h^*, h^*/2$ clearly displays the many local false modes; see Figure 3.23. The dominant mode at zero is not, however, lost among all the noisy modes.

The bottom line is that histograms which are “optimal” with respect to MISE may not be “optimal” for other purposes. This specific insight concerning modes should not be surprising given the fixed asymptotic noise in $R(\hat{f}')$ encountered in the biased CV derivation. But this general feature of nonparametric procedures is quite different than in the parametric setting, in which a maximum likelihood density estimate will be optimal for virtually any specific application. In the nonparametric setting, the “optimal” calibration will depend upon the purpose. For small samples, the noise in the histogram is seldom attributed to the correct source and, hence, is not well understood. In any case, optimally smooth histograms do not provide a powerful tool for bump hunting

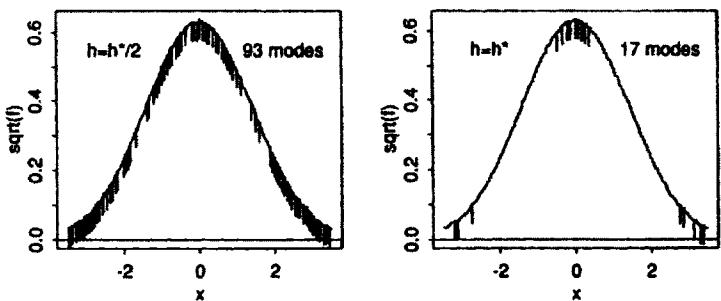


Figure 3.23 Rootgram of 2 histograms of a million Normal points ($h^*/2, h^*$).

(look again at the LRL example). In fact, the second differences of an optimally smoothed histogram diverge from the true second derivative. This result will be easier to see later with estimators that are themselves differentiable.

3.6 OTHER ERROR CRITERIA: L_1, L_4, L_6, L_8 , and L_∞

3.6.1 Optimal L_1 Histograms

Recall the several theoretically appealing aspects of absolute rather than squared error discussed in Section 2.3.2.2. The primary difficulty with absolute error is that it lacks an exact variance–bias decomposition [although Hall and Wand (1988a) have shown how to directly estimate the total error asymptotically]. Hjort (1986) has shown that

$$\begin{aligned} E \int |\hat{f} - f| &\leq E \int |\hat{f} - E\hat{f}| + \int |E\hat{f} - f| \\ &\approx \sqrt{\frac{2}{\pi nh}} \int \sqrt{f} + \frac{1}{4}h \int |f'|. \end{aligned}$$

Therefore, minimizing this asymptotic upper bound for MIAE leads to

$$\begin{aligned} h^* &= 2\pi^{-1/3} \left[\int f^{1/2} \div \int |f'| \right]^{2/3} n^{-1/3} \\ &= 2.717\sigma n^{-1/3} \quad \text{for } N(\mu, \sigma^2) \text{ data.} \end{aligned} \tag{3.77}$$

However, numerical simulations with Normal data show that in fact $h^* \approx 3.37\sigma n^{-1/3}$, which is 3.5% narrower than the optimal L_2 bandwidth. Minimizing upper bounds to MIAE leads to bin widths that can be either larger or smaller than h^* . Therefore, contrary to intuition, absolute error does not always provide wider bins in the tails. What would happen with optimal adaptive L_1

Table 3.9 Optimal Bandwidths for $N(0, 1)$ Data with Different L_p Criteria

Error Criterion	Optimal Bin Width	Expected Error
L_1 (upper bound)	$2.72n^{-1/3}$	$1.6258n^{-1/3}$
L_1 (numerical)	$3.37n^{-1/3}$	$1.1896n^{-1/3}$
L_2	$3.49n^{-1/3}$	$(0.6555n^{-1/3})^2$
L_4	$3.78n^{-1/3}$	$(0.6031n^{-1/3})^4$
L_6	$4.00n^{-1/3}$	$(0.6432n^{-1/3})^6$
L_8	$4.18n^{-1/3}$	$(0.6903n^{-1/3})^8$

meshes is unknown. Recently, Hall and Wand (1988a) have examined the optimal bandwidths for many densities, and found that with some heavier tailed densities, the L_1 bandwidth can be wider than the optimal L_2 bandwidth.

3.6.2 Other L_p Criteria

The error analysis of histograms for any even p is straightforward using higher order Binomial moments and proceeding as before. For example, using L_4 error, the asymptotic mean integrated fourth-power error is (see Problem 27)

$$\text{AMI4E} = \frac{3}{n^2 h^2} \int f(x)^2 + \frac{h}{2n} \int f'(x)^2 f(x) + \frac{h^4}{80} \int f'(x)^4. \quad (3.78)$$

Table 3.9 summarizes some known optimal bandwidths for $N(0, 1)$ data. The L_∞ optimal bandwidth is $O(\log(n)n^{-1/3})$.

Again, any fixed-bandwidth criterion will pay most attention to regions where the density is rough; that region is not necessarily in the tails. The coefficients in the table suggest that L_4 may have some special attraction, but that is an open problem.

PROBLEMS

1. Show how the Normal bin width rule can be modified if f is skewed or kurtotic, as discussed in the introduction to Section 3.3 using other reference densities. Examine the effect of bimodality. Compare your rules to Doane's (1976) extensions of Sturges' rule.
2. Perform a dimensional analysis for the quantities f , f' , f'' , $R(f)$, $R(f')$, and $R(f'')$. Check your results by computing these quantities for the case $f = N(\mu, \sigma^2)$ by tracking the factor σ .

3. An approximation to the error of a Riemannian sum:

$$\begin{aligned} & \left| \sum_{n=-\infty}^{\infty} g(nh) h - \int_{-\infty}^{\infty} g(x) dx \right| = \left| \sum_{n=-\infty}^{\infty} \int_{nh}^{nh+h} [g(nh) - g(x)] dx \right| \\ & \leq \sum_{n=-\infty}^{\infty} \int_{nh}^{nh+h} |g(nh) - g(x)| dx \leq \sum_{n=-\infty}^{\infty} h V_g(nh, nh + h) \leq h V_g(\mathbb{R}^1), \end{aligned}$$

where $V_g(a, b)$ is the total variation of g on $[a, b]$ defined by the $\sup\{\sum_{i=1}^n |g(x_i) - g(x_{i-1})|\}$ over all partitions on $[a, b]$, including $(a, b) = (-\infty, \infty)$. Conclude that if $f'(\cdot)^2$ has finite total variation, then the remainder term in the bias (3.14) is $O(h^3)$.

4. Compute the roughness of several parametric densities: Cauchy, Student's t , Beta, lognormal. For each compute the optimal bin width. Express the optimal bin width in terms of the variance, if it exists. Compare the coefficients in these formulas to the Normal rule in Equation (3.16).
5. Show that when $h = h^*$ for the histogram, the contribution to AMISE of the IV and ISB terms is asymptotically in the ratio 2:1.
6. Verify Equation (3.24). Hint: Substitute $h = ch^*$ into Equation (3.22).
7. Compare the sensitivity of the $\text{AMISE}(ch^*)$ in Equation (3.24) for various combinations of d , p , and r .
8. Prove the exact IV and ISB expressions in Equation (3.25).
9. Show that the ISB in a bin containing the origin of the double exponential density, $f(x) = \exp(-|x|)/2$, is $O(h^3)$; hence, the discontinuity in the derivative of f does not have any asymptotic effect on consistency. Compare when 0 and $h/2$ are bin edges. Formally, if f is the ordinary negative exponential density, $R(f')$ is infinite because of the jump at zero (integral of the square of the Dirac delta function at 0), but $R(f')$ is well-defined for the double exponential.
10. Consider the exact MISE of a histogram when $f = U(0, 1)$.
- (a) If the mesh $t_k = kh$ is chosen where $h = 1/m$, show that $\text{MISE}(m, n) = (m - 1)/n$.
- (b) If $t_k = (k + \frac{1}{2})h$ with $h = 1/m$, show that

$$\text{MISE}(m, n) = \frac{m - 1}{n} + \frac{1}{2mn} + \frac{1}{2m}.$$

- (c) Conclude that $h^* = \sqrt{2/n}$ and $\text{MISE}^* = \sqrt{2/n}$.

11. Verify the inequalities for the adaptive MISE in Equation (3.31).
12. Find the optimal adaptive meshes for a skewed Beta density using a numerical optimization program.
13. Investigate the use of fixed and percentile meshes when applying chi-squared goodness-of-fit hypothesis tests.
14. Apply the oversmoothing procedure to the LRL data. Compare the results of using the range and variance as measures of scale.
15. Find an oversmoothed rule based on the interquartile range as a measure of scale.
16. Use Taylor series and Riemannian integral approximations to verify (3.48).
17. Carefully evaluate and verify the formulas in Equation (3.52). *Hint:* $\hat{f}_{-i}(x_i) = (\nu_k - 1)/[(n - 1)h]$.
18. What are the UCV and BCV bin widths for the LRL and snowfall data?
19. The UCV rule seems to estimate something between ISE and MISE [the middle quantity, $-2 \int \hat{f}(x)f(x) dx$, being the focus]. Using the exact MISE formula with Normal and lognormal data, investigate the behavior of the terms in the UCV approximation.
20. Take a Normal sample of size 1,000. Apply increasing degrees of rounding to the data and compare the resulting UCV curves. At what point does the minimizer of UCV become $h = 0$?
21. In BCV, the bias in the roughness was reduced by subtraction. Alternatively, the roughness could have been multiplied by 3/4. Examine the effect of this idea by example. Where does the factor of 3/4 originate?
22. Develop a BCV formulation based on a central difference estimator of f' given by $(\nu_{k+1}/nh - \nu_{k-1}/nh)/2h$; see Scott and Terrell (1987).
23. Verify the expressions in Equations (3.64) and (3.68) for bin widths and errors when the density is bivariate Normal.
24. For bivariate Normal data, examine the inefficiency of using square bins relative to rectangular bins. Examine combinations of situations where the correlation is or is not 0, and situations where the ratio of the marginal variances is or is not equal to 1.

25. Verify the AMISE expressions given in Equations (3.69) and (3.70) for the 3 regular bivariate meshes. *Hint:* Carefully integrate the bias over the individual bin shape, and then aggregate by using the Riemannian integral approximation.
26. Consider a two-component Normal mixture density, $f(x) = 0.5N(-\mu, 1) + 0.5N(\mu, 1)$. How large must μ be so that the density exhibits two modes and bumps?
27. Compute one of the error criteria based on L_p for some even $p > 2$.

CHAPTER 4

Frequency Polygons

The discontinuities in the histogram limit its usefulness as a graphical tool for multivariate data. The *frequency polygon* (FP) is a continuous density estimator based on the histogram, with some form of linear interpolation. For example, the rootogram of the LRL data in Figure 3.21 is actually a linearly interpolated histogram. With 172 bins in the data, the vertical lines representing the raw histogram overlap, an undesirable feature that can be seen in the 200-bin histogram of the German income data in Figure 3.17. Scott (1985a) examined the theoretical properties of univariate and bivariate frequency polygons and found them to have surprising improvements over histograms. Fisher (1932, p. 37) disapproved of the frequency polygon, ironically for graphical reasons:

The advantage is illusory, for not only is the form of the curve thus indicated somewhat misleading, but the utmost care should always be taken to distinguish the infinitely large hypothetical population from which our sample of observations is drawn, from the actual sample of observations which we possess; the conception of a continuous frequency curve is applicable only to the former, and in illustrating the latter no attempt should be made to slur over this distinction.

Fisher was unaware of any theoretical differences between histograms and frequency polygons and was thinking only of univariate histograms when he wrote this passage. His objection to using a continuous nonparametric density estimator is no longer justified, but his concern about using techniques that totally obscure the statistical noise with mathematical sophistication is worth re-emphasizing. Finally, as a matter of terminology, the distinction between the histogram and frequency polygon in the scientific literature is being blurred, with the histogram label being applied to both.

4.1 UNIVARIATE FREQUENCY POLYGONS

In one dimension, the frequency polygon is the linear interpolant of the midpoints of an equally spaced histogram. As such, the frequency polygon extends

beyond the histogram into an empty bin on each extreme. The frequency polygon is easily verified to be a *bona fide* density function, that is, nonnegative with integral equal to 1; see Problem 1.

4.1.1 MISE

The asymptotic MISE is easily computed on a bin-by-bin basis, by considering a typical pair of histogram bins displayed in Figure 4.1. The frequency polygon connects the two adjacent histogram values, \hat{f}_0 and \hat{f}_1 , between the bin centers, as shown. The FP is described by the equation

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1, \quad -\frac{h}{2} \leq x < \frac{h}{2}. \quad (4.1)$$

The randomness in the frequency polygon comes entirely from the randomness in the histogram levels, $f_i = \nu_i/(nh)$. The “ x ” in $\hat{f}(x)$ is *not random* but is fixed.

As before, using the Taylor's series

$$f(x) = f(0) + xf'(0) + \frac{1}{2}x^2f''(0) + \dots, \quad (4.2)$$

approximations for p_0 and p_1 can be obtained:

$$\begin{aligned} p_0 &= \int_{-h}^0 f(s) \, ds \approx hf(0) - h^2f'(0)/2 + h^3f''(0)/6 \\ p_1 &= \int_0^h f(s) \, ds \approx hf(0) + h^2f'(0)/2 + h^3f''(0)/6. \end{aligned} \quad (4.3)$$

The bias is computed by noting that the pointwise expectation of the FP is a linear combination of the expectations of the two histogram values. As

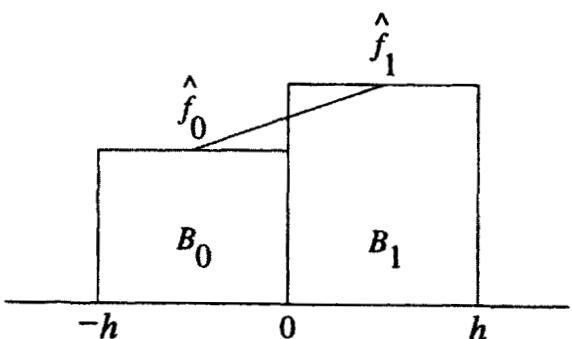


Figure 4.1 The frequency polygon in a typical bin, $(-h/2, h/2)$, which is derived from two adjacent histogram bins.

$E\{\hat{f}_i\} = p_i/h$, then from (4.1) and (4.3), and again noting that “ x ” is not random, we have

$$E\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\frac{p_0}{h} + \left(\frac{1}{2} + \frac{x}{h}\right)\frac{p_1}{h} \approx f(0) + xf'(0) + h^2f''(0)/6.$$

Subtracting (4.2) gives $\text{Bias}\{\hat{f}(x)\} \approx (h^2 - 3x^2)f''(0)/6$. The integral of the squared bias (ISB) over the FP bin $(-h/2, h/2)$ equals $[49h^4f''(0)^2/2, 880] \times h$, with a similar expression for other FP bins. Summing over all of the bins and using standard Riemannian approximation, yields

$$\text{ISB} \approx \sum_k \frac{49}{2,880} h^4 f''(kh) \times h = \frac{49}{2,880} h^4 R(f'') + O(h^6).$$

Evidently the squared bias of the frequency polygon is of significantly lower order than the order $O(h^2)$ of the histogram. For purposes of cross-validation, the bias is driven by the unknown roughness $R(f'')$ rather than by $R(f')$. Recalling Equation (3.12), the FP extends the good property of the histogram at its bin centers, the elimination of $O(h)$ effects, to the entire estimator. The bias is a function of the curvature in the density function rather than the slope as with the histogram.

The variance calculation is similar. From the FP definition in (4.1), the variance of $\hat{f}(x)$ equals

$$\left(\frac{1}{2} - \frac{x}{h}\right)^2 \text{Var } \hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)^2 \text{Var } \hat{f}_1 + 2\left(\frac{1}{4} - \frac{x^2}{h^2}\right) \text{Cov}(\hat{f}_0, \hat{f}_1). \quad (4.4)$$

For the variance and covariance terms, only the most trivial approximation $hf(0)$ is required for p_0 and p_1 . Since the bin counts are Binomial random variables,

$$\text{Var}(\hat{f}_i) = \frac{np_i(1-p_i)}{(nh)^2} \approx \frac{f(0)(1-hf(0))}{nh}$$

and

$$\text{Cov}(\hat{f}_0, \hat{f}_1) = \frac{-np_0p_1}{(nh)^2} \approx -\frac{f(0)^2}{n}.$$

Substituting these approximations into (4.4) gives

$$\text{Var } \hat{f}(x) = \left(\frac{2x^2}{nh^3} + \frac{1}{2nh}\right)f(0) - \frac{f(0)^2}{n} + o(n^{-1}).$$

Integrating over the FP bin $(-h/2, h/2)$ yields $[2f(0)/(3nh) - f(0)^2/n] \times h$. Summing the corresponding expression for all bins and noting that $\int f = 1$

gives

$$\text{IV} \approx \sum_k \left[\frac{2f(kh)}{3nh} - \frac{f(kh)^2}{n} \right] \times h = \frac{2}{3nh} - \frac{1}{n} R(f) + o(n^{-1}).$$

If the optimal histogram bin width were used with a frequency polygon, the asymptotic effect would be to eliminate the bias entirely relative to the variance in the MISE, the orders being $O(n^{-4/3})$ and $O(n^{-2/3})$, respectively. Since the ISB comprises a third of the MISE for a histogram, the reduction would be substantial. But a better FP can be constructed. The improved order in the bias suggests that a larger bin width could be used to reduce the variance, but still with smaller bias than the histogram. In fact, the bin width $h = O(n^{-1/5})$ turns out to be just right. The improvement is substantial, as the following theorem reveals (Scott, 1985a).

Theorem 4.1: Suppose f'' is absolutely continuous and $R(f''') < \infty$. Then

$$\text{AMISE}(h) = \frac{2}{3nh} + \frac{49}{2,880} h^4 R(f''); \quad (4.5)$$

hence,

$$\begin{aligned} h^* &= 2[15/(49R(f''))]^{1/5} n^{-1/5} \\ \text{AMISE}^* &= (5/12)[49R(f'')/15]^{1/5} n^{-4/5}. \end{aligned} \quad (4.6)$$

For example, with 800 Normal data points, the optimal bin width for the FP is 50% wider than the corresponding histogram bin width given in Theorem 3.1. Apparently, in order for the discontinuous histogram to approximate a continuous density, the histogram must be quite rough to track the density function in regions where its level is changing rapidly. The FP is inherently continuous and can approximate the continuous density better with piecewise linear fits over wider bins. The FP does most poorly near peaks where the second derivative and the density are both large in magnitude. The improvement in MISE is reflected not only by a decrease in the constant in front of $n^{-2/3}$, but also by a real decrease in the exponent.

The one situation where FPs are at a disadvantage occurs when the underlying density is discontinuous. A histogram is unaffected by such points if they

are known and placed at bin boundaries. A FP cannot avoid overlapping such points, and the asymptotic theory above does not apply; see Problem 3.

Finally, as was shown in Table 3.3 in the column with $p = 2$, frequency polygons are more sensitive than histograms with respect to errors in choice of bin width, particularly when $h > h^*$. On the other hand, quite a large error in bin width for the FP is required before its MISE is worse than the best histogram MISE; see Figure 4.2.

Example: For the Normal density, $R(\phi'') = 3/(8\sqrt{\pi}\sigma^5)$; hence, from Theorem 4.1,

$$h^* = 2.15 \sigma n^{-1/5} \quad \text{and} \quad \text{AMISE}^* = 0.3870 \sigma^{-1} n^{-4/5}. \quad (4.7)$$

To understand the practical consequences and to see where the FP fits among the parametric estimators and the histogram with respect to sample size, examine Table 4.1, which extends Table 3.2. Clearly, the frequency polygon is not just a theoretical curiosity. The FP is even more data efficient relative to the histogram as the sample size grows. Of course, both nonparametric estimates will be increasingly inferior to the correct parametric fit.

Another way to see the difference between the histogram and the FP for Normal data is shown in Figure 4.2. On a log-log scale, not only are the different rates of convergence easily seen but also the differences in the optimal bin widths. Continuing on to a million Normal points, the optimal bin widths for the histogram and FP are 0.035 and 0.136, respectively. These are in the ratio 4:1, an example of which appears in Figure 3.6 labeled $h = 4h^*$. The stability (small variance) of the histogram with $h = 4h^*$ is evident; however, so too is the bias resulting from the staircase shape. The FP retains the stability of this histogram, while the linear interpolation dramatically reduces the bias. The ISE of the FP in Figure 3.6 is approximately equal to 5.40×10^{-6} , which is 14% of the ISE of the best histogram.

4.1.2 Practical FP Bin Width Rules

To highlight the differences with the corresponding histogram results, some bin width rules for the FP will be presented. The plug-in rule based on (4.7) is

Table 4.1 Sample Sizes Required for $N(0, 1)$ Data So That $\text{AMISE}^* \approx 1/400$ and $1/4,000$

Estimator	Equivalent Sample Sizes
$N(\bar{x}, 1)$	57
$N(\bar{x}, s^2)$	100
Optimal FP	546
Optimal histogram	2,297
	571
	1,000
	9,866
	72,634

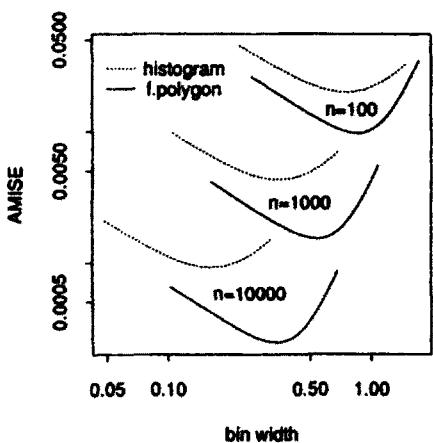


Figure 4.2 AMISE for histogram and frequency polygon for standard Normal density.

$$\text{FP Normal reference rule: } \hat{h} = 2.15 \hat{\sigma} n^{-1/5}, \quad (4.8)$$

where $\hat{\sigma}$ is an estimate, perhaps robust, of the standard deviation. One robust choice appropriate based on the interquartile range is $\hat{\sigma} = \text{IQ}/1.348$, where 1.348 is $\Phi^{-1}(0.75) - \Phi^{-1}(0.25)$. The factors modifying rule (4.8) based on sample skewness and kurtosis were shown in Figure 3.4. The factors are based on the relationship

$$\frac{h_y^*}{h_N} = \left[\frac{R(\phi''; 0, \sigma_y^2)}{R(g''(y))} \right]^{1/5}$$

corresponding to (3.18) and Theorem 4.1. Now $R(\phi'') = 3/(8\sqrt{\pi}\sigma_y^5)$ and the roughness $R(g'')$ of the lognormal and t_ν densities are

$$\frac{(9\sigma^4 + 20\sigma^2 + 12)e^{25\sigma^2/4}}{32\sqrt{\pi}\sigma^5} \quad \text{and} \quad \frac{(\nu + 1)^2(\nu + 3)^2 B\left(\frac{5}{2}, \frac{2\nu+5}{2}\right)}{\nu^{5/2} B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)^2},$$

respectively. Following the notation in Section 3.2.3,

$$\text{skewness factor } \{\beta_1(\sigma)\} = \frac{12^{1/5}\sigma}{e^{7\sigma^2/4}(e^{\sigma^2} - 1)^{1/2}(9\sigma^4 + 20\sigma^2 + 12)^{1/5}}.$$

and

$$\text{kurtosis factor } \{\beta_2\} = \frac{(\nu - 2)^{1/2}3^{1/5}B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)^{2/5}}{2^{3/5}\pi^{1/10}(\nu + 1)^{2/5}(\nu + 3)^{2/5}B\left(\frac{5}{2}, \frac{2\nu+5}{2}\right)^{1/5}},$$

where $\tilde{\beta}_2 = 6/(\nu - 4)$; see Problem 5.

Biased and unbiased cross-validation algorithms are only slightly more complicated to implement for the frequency polygon. For BCV, the following estimate of $R(f'')$ was proposed by Scott and Terrell (1987):

$$\hat{R}(f'') = \frac{1}{n^2 h^5} \sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2 - \frac{6}{nh^5}. \quad (4.9)$$

Plugging this estimate into the AMISE expression (4.6) results in

$$\text{BCV}(h) = \frac{271}{480nh} + \frac{49}{2880n^2h} \sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2.$$

The unbiased cross-validation formula is left as an exercise; see Problem 7.

As an example, consider the German income data displayed in Figures 3.17 and 3.18. The $\text{BCV}(h)$ estimates for the histogram and frequency polygon are shown in Figure 4.3. The BCV estimate of MISE for the FP is 71% lower than that for the histogram. The BCV-optimal FP is constructed from a histogram with 51 bins, which is displayed in Figure 3.17. Examine the shapes of the two BCV curves more closely. For small bin widths, the curves are parallel with slope -1 on the log-log scale, since the integrated variances for the histogram and FP are $1/(nh)$ and $2/(3nh)$, respectively. For large bin widths, the difference in the slopes reflects the different orders in the bias.

Upper bounds for the bin width for a frequency polygon may be obtained by variational methods similar to those used with the histogram; see Scott and Terrell (1987) and Terrell (1990). Examining the expression for the AMISE* in Theorem 4.1, the objective function becomes $R(f'')$, rather than $R(f')$ as

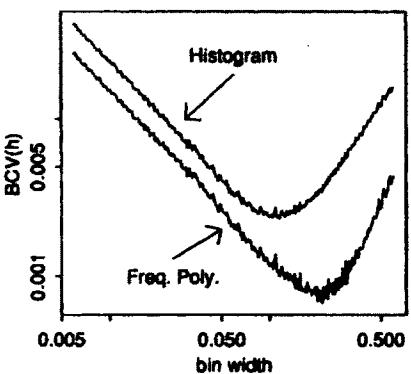


Figure 4.3 BCV for histogram and frequency polygon for German income data.

with the histogram. Subject to the constraint that the range is the interval $[-0.5, 0.5]$, the smoothest density is

$$f_3(x) = \frac{15}{8}(1 - 4x^2)^2 I_{[-0.5, 0.5]}(x) \quad \text{so that} \quad R(f'') \geq \frac{720}{(b - a)^5}$$

when the support interval is the more general (a, b) . Replacing $R(f'')$ in Equation (4.6) for h^* leads to

$$\text{number of bins} = \frac{b - a}{h^*} \geq \left(\frac{147}{2} n \right)^{1/5}. \quad (4.10)$$

For example, with the large LRL data set of 25,752 points, the optimal FP requires at least 18 bins, while the optimal histogram requires at least 37 bins. Given the amount of structure in the LRL data, these are conservative bounds.

A different version of the oversmoothed problem leads to a bin width rule. Among all densities with variance σ^2 , the smoothest density is

$$f_4(x) = \frac{35}{96\sigma} \left(1 - \frac{x^2}{9\sigma^2} \right)^3 I_{[-3\sigma, 3\sigma]}(x) \quad \text{so that} \quad R(f'') \geq \frac{35}{243\sigma^5}.$$

Substituting this inequality into the expression for h^* in Theorem 4.1 leads to the oversmoothed bin width rule:

$$h \leq \left(\frac{23,328}{343} \right)^{1/5} \sigma n^{-1/5} = 2.33 \sigma n^{-1/5} \equiv h_{OS}. \quad (4.11)$$

This bin width is only 108% of the Normal rule, which suggests that using the Normal-based rule in Equation (4.8) will also oversmooth in most practical data situations. In general, a Normal rule may be substituted for an oversmoothed rule whenever the variational problem is too difficult to solve explicitly.

4.1.3 Optimally Adaptive Meshes

Consider the theoretical improvement possible when applying frequency polygons to *adaptive* histogram meshes. Note, however, that connecting histogram midpoints in an adaptive mesh does not lead to an estimate that integrates to 1 except asymptotically. With that caveat, the following results are a consequence of Equation (4.6).

Theorem 4.2: *The asymptotic properties of the optimal adaptive frequency polygon constructed by connecting midpoints of an adaptive histogram are*

$$\text{AMSE}(x) = \frac{2f(x)}{3nh} + \frac{49}{2,880} h^4 f''(x)^2 \quad (4.12)$$

from which it follows that

$$\begin{aligned} h^*(x) &= 2[15f(x)/49f''(x)^2]^{1/5} n^{-1/5} \\ \text{AMSE}^*(x) &= (5/12)[49/15]^{1/5} [f''(x)^2 f(x)^4]^{1/5} n^{-4/5} \\ \text{AAMISE}^* &= (5/12)[49/15]^{1/5} \left\{ \int [f''(x)^2 f(x)^4]^{1/5} dx \right\} n^{-1/5}. \end{aligned} \quad (4.13)$$

Comparing Equations (4.6) and (4.13), we see that

$$\text{AAMISE}^* \leq \text{AMISE}^* \iff \int [f''(x)^2 f(x)^4]^{1/5} dx \leq \left[\int f''(x)^2 dx \right]^{1/5},$$

which is equivalent to the following inequality (which is true by Jensen's inequality):

$$\mathbb{E} \left[\frac{f''(x)^2}{f(x)} \right]^{1/5} \leq \left[\mathbb{E} \frac{f''(x)^2}{f(x)} \right]^{1/5}.$$

Thus, asymptotically, the MISE of an adaptive FP is only 91.5% and 76.7% of the MISE of a fixed-bin-width FP for Normal and Cauchy data, respectively; see Problem 8.

The MISE for the FP of an adaptive histogram can be computed exactly. Since the resulting adaptive FP does not integrate to 1, judgment is reserved as to its practical value; however, there is much of interest to examine in the structure of the optimal mesh. Note that asymptotically, the optimal adaptive FP will integrate to 1 since the underlying adaptive histogram exactly integrates to 1.

The general pattern in an adaptive FP mesh may be inferred from Theorem 4.2. The FP mesh seems out of phase with the optimal adaptive histogram mesh at critical points. The FP bins are widest where the second derivative is small, which is at points of inflection and, to a lesser extent, in the tails. In between, the bins can be quite narrow depending upon the magnitude of $f''(x)$. Consider the optimal adaptive mesh of the Normal-like scaled Beta(5, 5) density in Figure 4.4. In the tails, the optimal bins are not much wider. In fact, the pattern is relatively difficult to see except for the largest sample size. Given not only the complexity of an optimally adaptive mesh but also the relatively modest reduction in MISE, practical adaptive algorithms have been slow to appear. An intermediate strategy would be to perform data transformations to minimize skewness or to handle widely separated clusters individually.

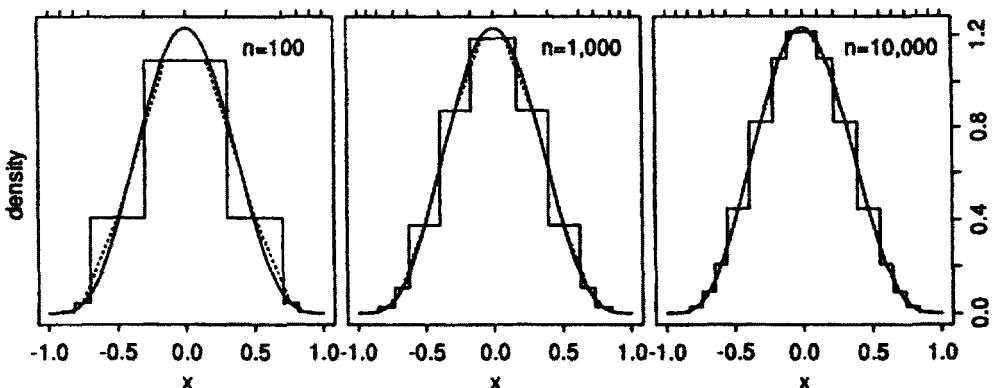


Figure 4.4 Optimal adaptive frequency polygon meshes for the scaled Beta(5, 5) density. The histogram is drawn from which the FP (dotted line) is derived. The tick marks for the adaptive mesh are shown above each figure.

4.1.4 Modes and Bumps in a Frequency Polygon

In Chapter 3, the optimal MISE smoothing for a histogram was observed as unable to provide a reliable estimate of modes or bumps. Since the optimal FP bin widths are wider and of order $O(n^{-1/5})$, the discussion in Section 3.5 shows that the sample modes and bumps in an optimal frequency polygon are more reliable than those found in an optimal histogram.

Continuing the analysis of Section 3.5 and assuming that $x = 0$ is a mode, write

$$h^* = cn^{-1/5} \quad \text{and define} \quad \beta \equiv -\frac{1}{2} c^{5/2} \frac{f''(0)}{\sqrt{f'(0)}}.$$

Then Equation (3.73) becomes

$$\lim_{n \rightarrow \infty} \Pr\left(\nu_0 = \arg \max_{|j| \leq k} \nu_j\right) = \int_y \prod_{\substack{j=-k \\ j \neq 0}}^k \Phi(y + j^2 \beta) \phi(y) dy.$$

This probability is a constant that depends only on β (and a bit on k) but *not on the sample size*. A graph is shown in Figure 4.5 for the choice $k = 4$. When $\beta = 0$ [i.e., $f''(0) = 0$ so that the density function is flat], then the probability that ν_0 is the mode is $1/(2k + 1)$ since the density is locally uniform and all $2k + 1$ bins are equally likely to contain the largest count. As β increases, the probability that ν_0 is the greatest count increases to 1. β , which is dimensionless, measures the “strength” of the mode at $x = 0$. A similar expression may be computed for the probability that each of $\{\nu_\ell, 1 \leq |\ell| \leq k\}$ is the largest bin count (an “error” since the mode is at the center of B_0). The probabilities are symmetric in ℓ . Each is also a function of β alone; see Figure 4.5.

A “confidence interval” for the true mode may be derived from this graph given an estimate of β . For example, if $\beta = 2.7$, then the probability the ν_0 is the greatest count is 95%. Thus the sample bin interval $(-h/2, h/2)$ is a 95% confidence interval for the mode. If $\beta = 0.46$, then the probability is 95%

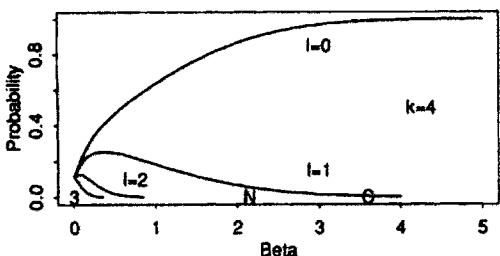


Figure 4.5 Probability distribution of the location of the sample mode as a function of β for the choice $k = 4$.

that the sample mode is in $(-3h/2, 3h/2)$, that is, bins B_{-1}, B_0, B_1 . (However, the derivation assumed that the true mode was at the center of a bin. Hence, to use this result in practice, the mesh should be shifted so that the values of the FP in the bins adjacent to the sample mode are approximately equal in height. This ensures that the sample mode is approximately at the center of a bin.) The sample estimate of β is sure to be underestimated because of the downward bias at modes, so the confidence interval is conservative. Are the estimates consistent if the probabilities are not changing? The answer is yes, because the confidence intervals are stated in terms of multiples of the bin width, which is shrinking towards zero at the rate $n^{-1/5}$. Notice that the calculation does not preclude other smaller sample modes (for example, $\nu_3 > \nu_2$) in this neighborhood, but the interpretation should be acceptable in most situations. A figure similar to Figure 4.5 but with a larger value of k is virtually identical except near $\beta = 0$. For Normal and Cauchy data, $\beta \approx 2.15$ and 3.6, respectively. There is an 89% chance that the sample mode for Normal data is in bin B_0 for Normal data (99% for Cauchy). Locating the modes for these densities is relatively easy.

4.2 MULTIVARIATE FREQUENCY POLYGONS

There are two important ways of defining a linear interpolant of a multivariate histogram with hyper-rectangular bins, where $\mathbf{x} \in \mathbb{R}^d$. The first, considered by Scott (1985a, b), is to interpolate the values at the centers of $d + 1$ adjacent histogram bins in a “triangular” or simplex-like configuration. The resulting collection of triangular pieces of a hyperplane defines a continuous but not differentiable surface in \mathbb{R}^{d+1} . The definition is not unique since several reflections of the basic pattern work; see Figure 4.6.

The second definition for a multivariate FP, which was investigated independently by Terrell (1983) and Hjort (1986), is known as the *linear blend* in the computer graphics literature. A single portion of a linear blend extends over a hyper-rectangle with 2^d vertices, defined by the centers of the 2^d adjacent histogram bins. Any cut of the surface parallel to a coordinate axis gives a linear fit; see Figure 4.7. Certainly, this definition of a multivariate FP is smoother than the first, but the primary advantage of this formulation is the beautifully simple AMISE result.

The linear blend frequency polygon (LBFP) is only slightly more complicated to define than the triangular mesh. Consider a typical LBFP bin,

$$B_{k_1, \dots, k_d} = \prod_{i=1}^d [t_{k_i}, t_{k_i} + h_i).$$

Then for $\mathbf{x} \in B_{k_1, \dots, k_d}$, the LBFP is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{j_1, \dots, j_d \in \{0, 1\}^d} c_{j_1, \dots, j_d} \nu_{k_1 + j_1, \dots, k_d + j_d}, \quad (4.14)$$

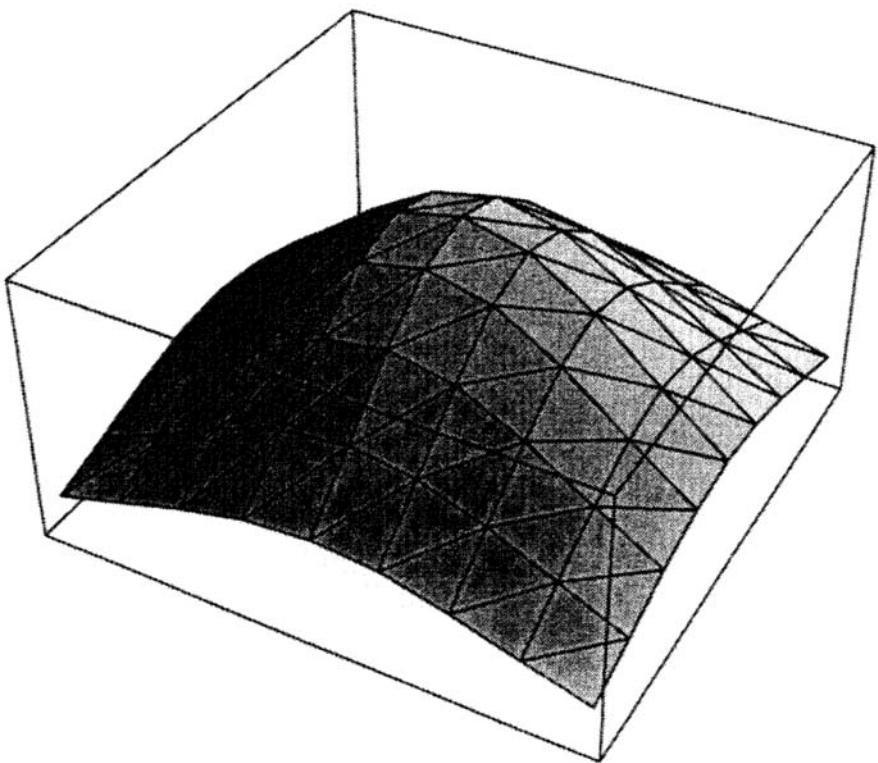


Figure 4.6 An example of the construction of a bivariate frequency polygon using triangular meshes.

where

$$c_{j_1, \dots, j_d} = \prod_{i=1}^d u_i^{j_i} (1 - u_i)^{1-j_i} \quad \text{and} \quad u_i = \frac{x_i - t_{k_i}}{h_i}.$$

Hjort (1986) showed that the LBFP integrates to 1 and that

$$\text{AMISE}(\mathbf{h}) = \frac{2^d}{3^d n h_1 \cdots h_d} + \frac{49}{2,880} \sum_{i=1}^d h_i^4 R(f_{ii}) + \frac{1}{32} \sum_{i < j} h_i^2 h_j^2 R(\sqrt{f_{ii} f_{jj}}),$$

where f_{ij} is the mixed second-order partial derivative. Although this cannot be optimized in closed form except in special cases, it is easy to show that

$$h_i^* = O(n^{-1/(4+d)}) \quad \text{and} \quad \text{AMISE}^* = O(n^{-4/(4+d)}). \quad (4.15)$$

Not only are frequency polygons more efficient than histograms, but the difference in the order of convergence rates across dimensions is significant. If the notion that bivariate histograms “work” is correct, then Table 4.2 suggests that

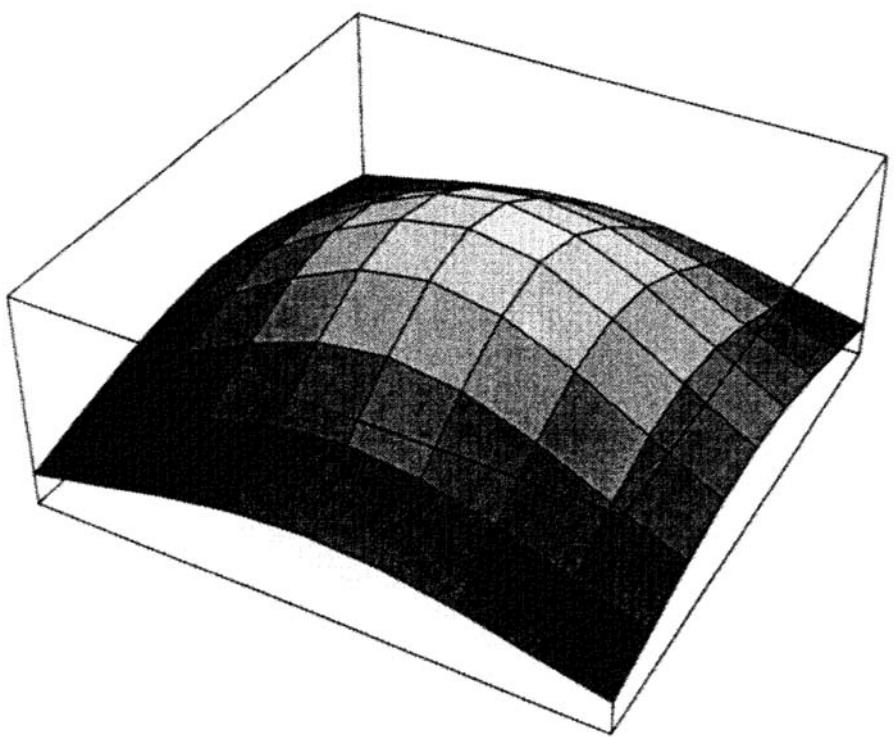


Figure 4.7 An example of a linear blend element.

quadrivariate frequency polygons should work equally well. Some authors have claimed good results for quadrivariate histograms, which would correspond to an eight-dimensional FP in terms of exponent order. On the other hand, the level of detail demanded in higher dimensions must diminish. Working in more than 4 or 5 dimensions is usually done for convenience of interpretation, not for reasons of structure (such as interactions in higher dimensions).

Table 4.2 Order of Decrease of MISE for Multivariate Density Estimators

d	Histogram	Frequency Polygon
1	$n^{-2/3}$	$n^{-4/5}$
2	$n^{-2/4}$	$n^{-4/6}$
3	$n^{-2/5}$	$n^{-4/7}$
4	$n^{-2/6}$	$n^{-4/8}$
5	$n^{-2/7}$	$n^{-4/9}$
6	$n^{-2/8}$	$n^{-4/10}$
7	$n^{-2/9}$	$n^{-4/11}$
8	$n^{-2/10}$	$n^{-4/12}$

For graphical reasons, the first definition of a FP is simpler to work with because the resulting contours are comprised of piecewise polygonal sections, which can be depicted with many CAD-CAM graphics packages. There is little practical difference in the approximation quality of the two estimators and the binning structure is easily apparent in the former and not in the latter. Advanced surface visualization algorithms require the value of the function on a 3-D mesh. The simplifying idea here is that the visualization mesh can be identical to the FP mesh. Usually, there are several interpolation options in visualization programs, including linear blends and piecewise triangular. Thus the choice of interpolation can be thought of as primarily an aesthetic issue of visualization smoothness, and secondarily, as a choice of density quality.

Using the triangular mesh with a bivariate Normal data, Scott (1985a) showed that the optimal bin widths are approximately equal to

$$h_i^* = 2.105 \left(1 - \frac{107}{208} \rho^2 + \dots\right) \sigma_i n^{-1/6}, \quad i = 1, 2.$$

For multivariate Normal data with $\Sigma = I_d$, the optimal smoothing parameters in each dimension are equal with the constant close to 2. Thus Scott also proposed using

Approximate Normal FP reference rule: $h_i = 2 \hat{\sigma}_i n^{-1/(4+d)}$. (4.16)

4.3 BIN EDGE PROBLEMS

As was mentioned in Chapter 3, the mesh is completely determined by the pair (h, t_0) . The asymptotic theory indicates that the choice of bin origin is asymptotically negligible. Consider the Buffalo snowfall data set given in Table 9 in Appendix B. The annual snowfall during 63 winters was recorded from 1910/11–1972/73. Some have argued (Scott, 1980) that the data appear to be trimodal, but Parzen (1979) has suggested the evidence leans towards a unimodal density. One might imagine that the choice of the bin width and not the bin origin would be critical for understanding this issue. Indeed, in Figure 4.8 the histogram with 15 bins of width 10 inches suggests trimodality, while the histogram with 10 bins of width 15 inches suggests unimodality. But in Figure 4.9, the effect of bin origin choice is clearly revealed not to be negligible. One histogram is unimodal. Two histograms are bimodal, but with secondary mode on the left and right, respectively. And remarkably, only one histogram is trimodal. Continuing to the multivariate setting, the effect of the bin origin is more pronounced.

FREQUENCY POLYGONS

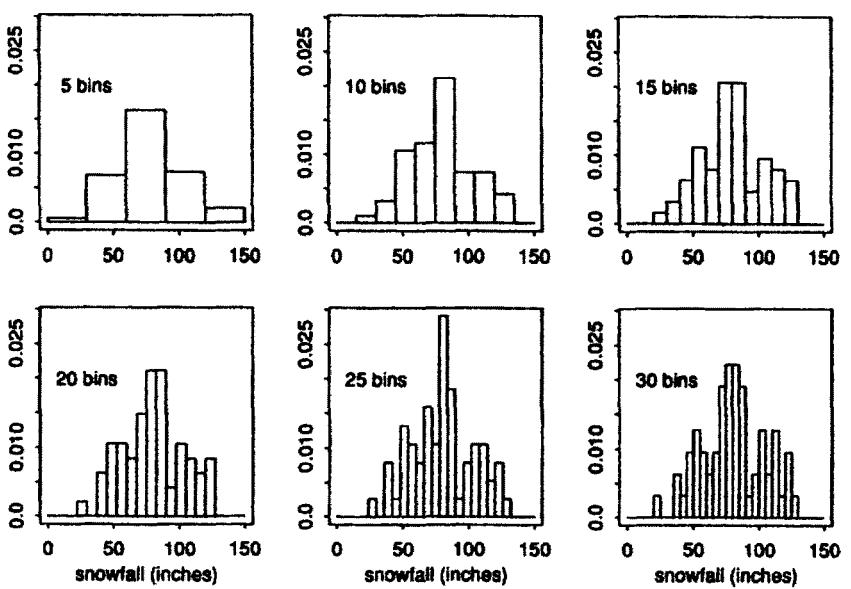


Figure 4.8 Histograms of the Buffalo snowfall data with bin origin $t_0 = 0$, and bin widths of 30, 15, 10, 7.5, 6, and 5 inches.

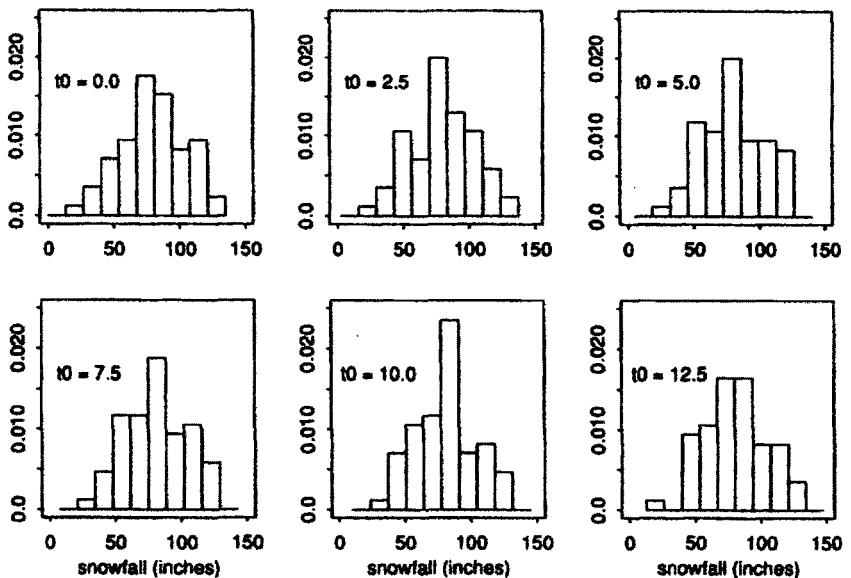


Figure 4.9 Histograms of the Buffalo snowfall data. All have a bin width of 13.5 inches, but different bin origins t_0 .

The statement that the choice of t_0 is asymptotically negligible with respect to MISE is also true for the frequency polygon. But the optimal bin widths for a FP are substantially wider than for the histogram. Thus there are many more possible choices for the bin origin. For a particular choice of bin width, one possible recommendation is to choose the bin origin so that the estimate is “smoothest.” If this could always be done, then the amount of roughness in the estimate would always be determined by the bin width as much as possible, and not by the bin origin, t_0 , which may be thought of as a nuisance parameter. The next chapter introduces an ingenious device that eliminates the effect of this nuisance parameter entirely.

PROBLEMS

1. Demonstrate that the frequency polygon which interpolates the histogram at the midpoints of an equally spaced bins integrates to 1. Investigate alternative definitions of a FP derived from an adaptive or unequally spaced histogram.
2. Verify the derivations of the bias and variance of the FP in Section 4.1.1.
3. Consider the FP when the data come from the negative exponential density function, $f(x) = e^{-x}$, $x \geq 0$. Using the histogram mesh $(-h, 0, h, 2h, \dots)$, compute the contribution to the bias from the bins adjacent to $x = 0$ and show that the total integrated squared bias over $(-h/2, h/2)$ is no longer $O(h^4)$, but rather $17h/24 + O(h^2)$. Compare this result to the corresponding result for the histogram. *Hint:* Use a symbolic computer package to compute the probability exactly and take a Taylor’s series at the end.
4. One suggested fix to the boundary problem is to reflect the data around 0, that is, compute the FP using the data $-x_n, \dots, -x_1, x_1, \dots, x_n$ and then doubling the estimate for $x \geq 0$. Consider again the negative exponential density.
 - (a) Show that using the same mesh as in Problem 3 gives a “flat” histogram-like estimate over $(0, h/2)$, which contributes a term of order h^2 to the integrated squared bias.
 - (b) Show that the histogram mesh $(-3h/2, -h/2, h/2, 3h/2, \dots)$ with reflected data leads to a contribution towards the ISB from the bin $(0, h)$ equal to $h^3/48 + O(h^4)$, which is between the usual histogram and FP exponent orders.
5. Find some simple approximation to the skewness and kurtosis factors in Section 4.1.2. Try them on some simulated data.
6. Consider the FP roughness estimate given in Equation (4.9).
 - (a) Show that it is unbiased to first order.

(b) Alternatively, show that

$$\left[80/(129n^2h^5)\right] \sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2$$

is also unbiased to first order.

(c) Construct the two BCV estimators that follow from these two estimators of $R(f'')$ and compare them empirically on simulated data.

7. Find the UCV expression for a FP and try it on the LRL data.
8. Compute the asymptotic efficiency of an optimal adaptive mesh relative to a fixed mesh for Normal and Cauchy data.
9. How do k th-nearest-neighbor meshes (equal number of points in each bin) perform for the FP? Make a figure when $f(x) = \text{Beta}(5, 5)$.

CHAPTER 5

Averaged Shifted Histograms

A simple device has been proposed for eliminating the bin edge problem of the frequency polygon while retaining many of the computational advantages of a density estimate based on bin counts. Scott (1983, 1985b) considered the problem of choosing among the collection of multivariate frequency polygons, each with the same smoothing parameter but differing bin origins. Rather than choosing the “smoothest” such curve or surface, he proposed averaging several of the shifted frequency polygons. As the average of piecewise linear curves is also piecewise linear, the resulting curve appears to be a frequency polygon as well. If the weights are nonnegative and sum to 1, the resulting “averaged shifted frequency polygon” (ASFP) is nonnegative and integrates to 1.

A nearly equivalent device is to average several shifted histograms, which is just as general but simpler to describe and analyze. The result is the “averaged shifted histogram” (ASH). Since the average of piecewise constant functions such as the histogram is also piecewise constant, the ASH appears to be a histogram as well. In practice, the ASH is made continuous using either of the linear interpolation schemes described for the frequency polygon in Chapter 4 and will be referred to as the frequency polygon of the averaged shifted histogram (FP-ASH). The ASH is the practical choice for computationally and statistically efficient density estimation. Algorithms for its evaluation are described in detail.

5.1 CONSTRUCTION

Consider a collection of m histograms, $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m$, each with bin width h , but with bin origins

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \dots, \frac{(m-1)h}{m}, \quad (5.1)$$

respectively. The (*naive* or unweighted) averaged shifted histogram is defined as

$$\hat{f}(\cdot) = \hat{f}_{\text{ASH}}(\cdot) = \frac{1}{m} \sum_{i=1}^m f_i(\cdot). \quad (5.2)$$

Observe that the ASH is piecewise constant over intervals of width $\delta \equiv h/m$, as the bin origins in (5.1) differ by this amount.

Reexamine the sequence of shifted histograms of the Buffalo snowfall data shown in Figure 4.9. Each shifted histogram has bin width $h = 13.5$. In Figure 5.1, a series of ASHs using this same bin width are shown for an increasing sequence in the parameter m . Although the ordinary histogram (ASH with $m = 1$) displays a second bump to the right of the mode, every ASH with $m > 1$ reveals the presence of larger third bump to the left of the mode. The third bump was masked by the larger bump at the mode. The appearance of these additional bumps is not an artifact of the ASH algorithm, but rather the result of a significantly improved signal-to-noise ratio obtained by averaging out the nuisance parameter t_0 . In a sense, the parameter t_0 has been replaced by a different parameter m that must be specified; however, the improvement over the ordinary histogram justifies any additional work.

Multivariate ASHs are constructed by averaging shifted multivariate histograms, each with bins of dimension $h_1 \times h_2 \times \dots \times h_d$. If every possible multivariate histogram is constructed by coordinate shifts that are multiples of $\delta_i \equiv h_i/m_i$, $i = 1, \dots, d$, then the multivariate ASH is the average of $m_1 \times m_2 \times \dots \times m_d$ shifted histograms. In the bivariate case, the ASH is

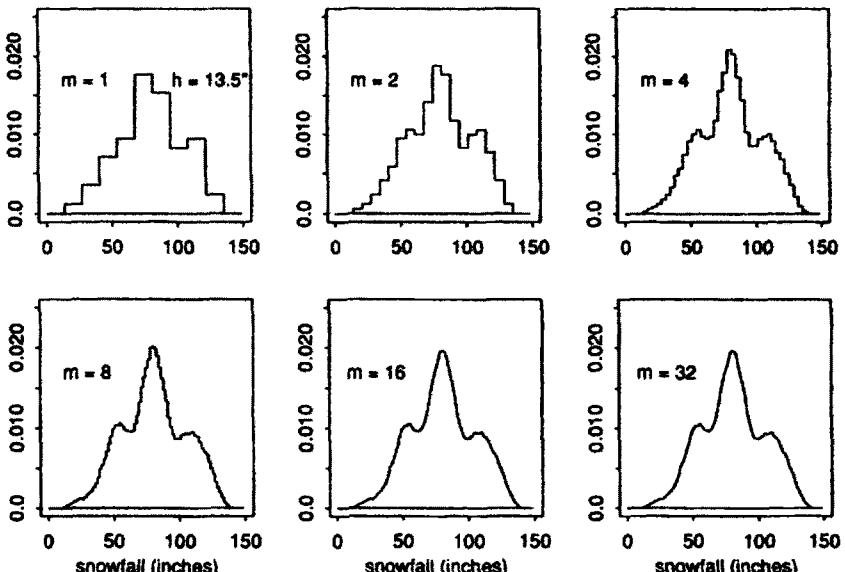


Figure 5.1 Naive averaged shifted histograms of the Buffalo snowfall data with bin width $h = 13.5$ inches.

given by

$$\hat{f}(\cdot, \cdot) = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} f_{ij}(\cdot, \cdot), \quad (5.3)$$

where the bin origin for the bivariate shifted histogram \hat{f}_{ij} is the point $(x, y) = ((i - 1)\delta_1, (j - 1)\delta_2)$. Figure 5.2 displays several bivariate ASHs of the plasma lipid data (see Table 3 in Appendix B) with $m_1 = 1, 2$, and 3 . Only a few shifts along each axis are required to provide a smoother estimate. The underlying histogram bin size is the same for all three estimates, although the apparent number of bins in the ASH grows from 7^2 to 14^2 to 21^2 . A contour plot of the linear interpolant of the ASH with $m_1 = m_2 = 3$ suggests a multimodal structure not apparent in the original histogram. Only the bivariate glyph histogram in Figure 3.20 hints at the structure in this data set.

The recognition of the need to smooth a bivariate histogram is not new. In 1886, Galton performed bivariate bin smoothing on a cross-tabulation of 928 adult children and their parents' average height (Stigler, 1986, p. 285). Galton further adjusted all female heights upward by a factor of 1.08 to account for

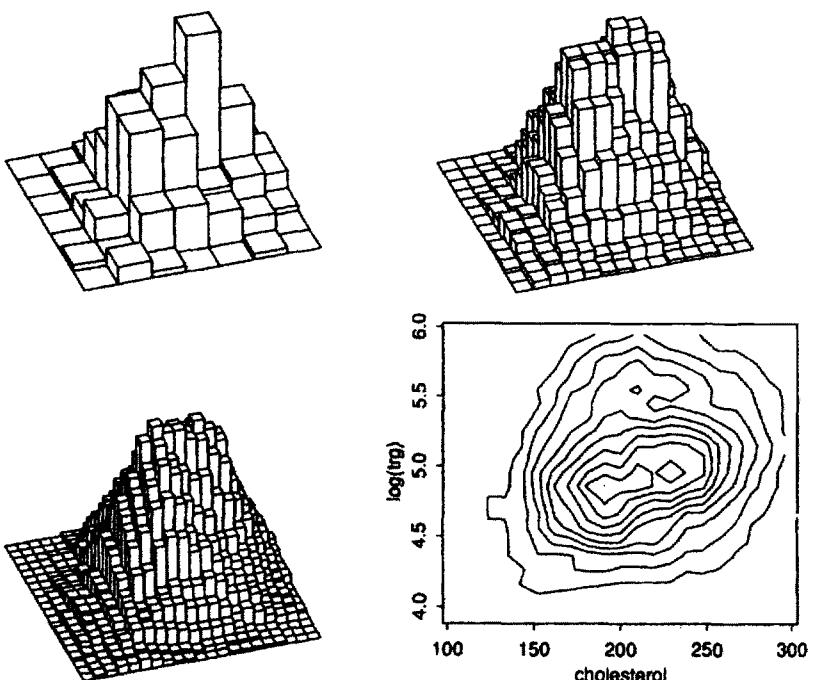


Figure 5.2 Bivariate averaged shifted histograms of the lipid data for 320 diseased males; see the text.

male-female height differences. Stigler quotes Galton's description of how he then smoothed his raw bin counts:

... by writing at each intersection of a horizontal column with a vertical one, the sum of the entries in the four adjacent squares, and using these to work upon.

This smoothing sharpened the elliptical shape of the contours of the data. Galton's smoothing corresponds roughly to the bivariate ASH with $m_1 = m_2 = 2$; see Problem 1.

5.2 ASYMPTOTIC PROPERTIES

As the univariate ASH is piecewise constant over the intervals $[k\delta, (k + 1)\delta)$ where $\delta \equiv h/m$, it is convenient to refer to this *narrower interval* as the bin B_k and let

$$\nu_k = \text{bin count in bin } B_k, \quad \text{where } B_k \equiv [k\delta, (k + 1)\delta).$$

With this new definition of the bin intervals, the bin count for an ordinary histogram may be obtained by adding m of the adjacent bin counts $\{\nu_k\}$ from the finer grid.

Consider the ASH estimate in bin B_0 . The height of the ASH in B_0 is the average of the heights of the m shifted histograms, each of width $h = m\delta$, in bin B_0 :

$$\frac{\nu_{1-m} + \cdots + \nu_0}{nh} \cdot \frac{\nu_{2-m} + \cdots + \nu_0 + \nu_1}{nh}, \dots, \frac{\nu_0 + \cdots + \nu_{m-1}}{nh}.$$

Hence, a general expression for the naive ASH in Equation (5.2) is

$$\begin{aligned}\hat{f}(x; m) &= \frac{1}{m} \sum_{i=1-m}^{m-1} \frac{(m - |i|)\nu_{k+i}}{nh} \\ &= \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) \nu_{k+i} \quad \text{for } x \in B_k.\end{aligned}\quad (5.4)$$

The weights on the bin counts in Equation (5.4) take on the shape of an isosceles triangle with base $(-1, 1)$. Other shapes may be contemplated, such as uniform weights or perhaps smoother (differentiable) shapes. The general ASH uses arbitrary weights, $w_m(i)$, and is defined by

General ASH: $\hat{f}(x; m) = \frac{1}{nh} \sum_{|i| \leq m} w_m(i) \nu_{k+i} \quad \text{for } x \in B_k.$ (5.5)

In order that $\int \hat{f}(x; m) dx = 1$, the weights must sum to m ; see Problem 2. An easy way to define the general weights is

$$w_m(i) = m \times \frac{K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)} \quad i = 1 - m, \dots, m - 1, \quad (5.6)$$

where K is a continuous function defined on $(-1, 1)$. K is often chosen to be a probability density function, such as

$$K(t) = \frac{15}{16} (1 - t^2)_+^2 = \frac{15}{16} (1 - t^2)^2 I_{[-1, 1]}(t), \quad (5.7)$$

which is called the *biweight kernel* or quartic kernel.

The computational algorithm for the generalized ASH is quite simple. Construct an equally spaced mesh of width δ over the interval (a, b) , and compute the corresponding bin counts $\{\nu_k, k = 1, \dots, nbin\}$ for the n data points. Typically, $\delta \ll h$, and $nbin$ refers to the number of bins width δ . This computation is accomplished by the BIN1 algorithm given in the box below.

BIN1($x, n, a, b, nbin$) Algorithm: (* Bin univariate data *)

```

 $\delta = (b - a)/nbin$ 
for  $k = 1, nbin \{ \nu_k = 0 \}$ 
for  $i = 1, n \{
    k = (x_i - a)/\delta + 1 \quad (* \text{integer part} *)
    if (k \in [1, nbin]) \nu_k = \nu_k + 1 \}
return (\{\nu_k\})$ 
```

Next, compute the weight vector, $\{w_m(i)\}$, as in Equation (5.6). Then the univariate ASH estimates, $\{f_k, k = 1, \dots, nbin\}$, over the $nbin$ intervals may be computed in an efficient manner by reordering the operations indicated in formula (5.5). Rather than computing the ASH estimates individually in each bin by sweeping through the $2m - 1$ adjacent bin counts, a single pass is made through the bin counts, with a weighted count applied to the $2m - 1$ adjacent ASH estimates. This modification avoids repeated weighting of empty bins; see the ASH1 algorithm given in the box. The algorithm assumes that there are at least $m - 1$ empty bins on each end. Observe that the amount of work is determined by m and by the number of nonempty bins. The algorithm is quite efficient even when $n > 10^6$, in which case most of the work involves tabulating the several hundred bin counts.

ASH1($m, \nu, nbin, a, b, n, w_m$) Algorithm: (* Univariate ASH *)

```

 $\delta = (b - a)/nbin$ 
 $h = m\delta$ 
for  $k = 1, nbin \{ f_k = 0 \}$ 
for  $k = 1, nbin \{$ 
    if ( $\nu_k = 0$ ) next  $k$ 
    for  $i = \max(1, k - m + 1), \min(nbin, k + m - 1) \{$ 
         $f_i = f_i + \nu_k w_m(i - k) \}$ 
    for  $k = 1, nbin \{ f_k = f_k/(nh); t_k = a + (k - 0.5)\delta \}$ 
return ( $x = \{t_k\}$ ,  $y = \{f_k\}$ ) (* Bin centers and ASH heights *)

```

In Figure 5.3, examples of the ASH using the biweight kernel are shown. For the Buffalo snowfall data, observe how the use of the biweight kernel weights rather than the isosceles triangle weights results in a visually smoother curve, with less local noise in the estimate. (As the variances of the triangle and biweight kernels are $1/6$ and $1/7$, respectively, a bin width of $h = 13.5 \times \sqrt{7/6} = 14.6$ inches was applied with the biweight kernel. This rescaling is justified in Section 6.2.3.3.) For a large data set such as the German household income data, the additional visual smoothness is still apparent, even when the smoothing parameters are small enough to reveal any possible feature (compare to Figure 3.17).

In practice, the narrow bin width δ is usually fixed first by choosing between 50 and 500 bins over the sample range (extended by 5–10% to include some empty bins on both sides). Since $h = m\delta$, only values of the smoothing parameter h that are integer multiples of δ may be considered, although it is easy to remove this restriction (see Problem 4). On the other hand, if h is known, then δ may be computed as $h/5$ or $h/10$. This case is rare. Many large

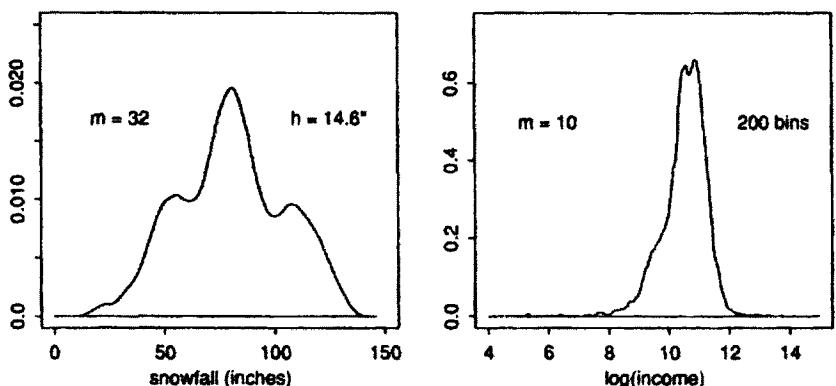


Figure 5.3 Examples of ASH with biweight kernel applied to the Buffalo snowfall and German household income data sets.

data sets are prebinned; that is, the raw data are not recorded, only the bin counts. If the width of those bins is called δ and h^* turns out to be close to δ , then no additional smoothing can be applied since $m = 1$ is the only option. Careful planning can avoid such an unfortunate outcome. For example, using the oversmoothed FP bin width rule in Equation (4.11), choose δ sufficiently small or n sufficiently large so that $\delta < h_{\text{OS}}/25$ or $\delta < h_{\text{OS}}/50$. Only a small pilot study is required to estimate the variance of the data to use in the bin width rule for oversmoothed frequency polygon.

The derivation of the AMISE for the naive (isosceles triangle weight function) ASH is similar to previous calculations and is not given here. The corresponding result for the general weighted ASH is much more complicated. Scott (1985b) proved the following result.

Theorem 5.1: *For the naive ASH with the isosceles triangle kernel,*

$$\begin{aligned} \text{AMISE} = & \frac{2}{3nh} \left(1 + \frac{1}{2m^2} \right) + \frac{h^2}{12m^2} R(f') \\ & + \frac{h^4}{144} \left(1 - \frac{2}{m^2} + \frac{3}{5m^4} \right) R(f''). \end{aligned} \quad (5.8)$$

The first term in the AMISE gives the error due to the integrated variance. The ISB or bias portion of the AMISE combines terms involving $R(f')$ and $R(f'')$, which were found in the ISB of the histogram and frequency polygon, respectively.

It is easy to check that the first 2 terms in this result match the ordinary histogram result in Theorem 3.1 when $m = 1$. On the other hand, as $m \rightarrow \infty$, the second histogram-like bias term disappears and the bias is similar to that for a frequency polygon in Theorem 4.1. Usually, for $m \geq 10$, the middle term is negligible compared to the last term, which may be taken to equal $h^4/144$. Comparing Equations (4.6) and (5.8), the IV terms are identical while the ISB for the ASH is 41% of the ISB for the FP. The optimal bin width for the naive ASH as $m \rightarrow \infty$ is simply

$$h_{m=\infty}^* = [24/(nR(f''))]^{1/5} \quad (= 2.576 \sigma n^{-1/5} \text{ if } f(x) = N(\mu, \sigma^2)).$$

The sample sizes in Table 5.1 summarize the efficiency of the ASH and other estimators with Normal data. The ASH requires 80% of the samples required by the FP to achieve the same MISE. In fact, this figure of 80% holds true for any sampling density, asymptotically (compare Theorems 4.1 and 5.1). In some difficult situations, such as a small sample from a rough density, the

Table 5.1 Equivalent Sample Sizes Required for AMISE $\approx 1/400$ for $N(0, 1)$ Data

Estimator	$N(\bar{x}, s^2)$	ASH	FP-ASH	FP	Histogram
Sample Size	100	436	436	546	2,297

histogram may actually be competitive with the ASH. But asymptotically the efficiency of the histogram will be 0 relative to the ASH or FP, because of the different rates of convergence of MISE. Of course, the improvement of the ASH relative to the FP is not as dramatic as the improvement of the FP relative to the histogram, as diminishing returns begin to take effect.

The expression for the asymptotic L_2 error of the FP-ASH or linear interpolant of the naive ASH is much simpler than for the naive ASH itself.

Theorem 5.2: *For the frequency polygon interpolant of the naive ASH,*

$$\text{AMISE} = \frac{2}{3nh} + \frac{h^4}{144} \left(1 + \frac{1}{m^2} + \frac{9}{20m^4} \right) R(f''). \quad (5.9)$$

Notice that the histogram-like bias term involving $R(f')$ has vanished. Furthermore, the dependence of the remaining terms on the choice of m is greatly reduced. Usually, $m \geq 3$ is sufficient to achieve the 20% improvement in AMISE over the frequency polygon, and not $m \geq 10$ as recommended for the ASH itself.

The multivariate FP-ASH has been studied by Scott (1985b) using a triangular mesh, but the linear blend results of Hjort (1986) are more elegant and are reported here. Let subscripts on f denote partial derivatives.

Theorem 5.3: *The AMISE of the multivariate linear blend of the naive ASH equals*

$$\frac{2^d}{3^d n h_1 \cdots h_d} + \frac{1}{720} \sum_{i=1}^d \delta_i^4 R(f_{ii}) + \frac{1}{144} \int_{\mathbb{R}^d} \left[\sum_{i=1}^d h_i^2 \left(1 + \frac{1}{2m_i^2} \right) f_{ii} \right]^2. \quad (5.10)$$

Except in special circumstances, closed-form expressions for the optimal smoothing parameters are not available. Rather they must be obtained by solving a system of nonlinear equations. If $\delta_i \approx 0$ in (5.10), then $h_i^* = O(n^{-1/(4+d)})$

and $\text{AMISE}^* = O(n^{-4/(4+d)})$, which are comparable to the results for the multivariate frequency polygon in Equation (4.15). While the rates are the same, the multivariate FP is inferior by a fixed amount.

The BIN2 and ASH2 algorithms for $d = 2$ are given below. Note that the parameters in the univariate ASH become vectors in the bivariate algorithm. The BIN2 and ASH2 algorithms are easily extended to the cases $d = 3$ and 4 by increasing the dimensions on the vectors and matrices. For dimensions greater than 4 , it is generally not possible to fit arrays of sufficient dimension directly in computer memory. In those cases, the ASH algorithm may be modified to compute only 2- or 3-dimensional slices of the higher-dimensional ASH.

BIN2($x, n, a, b, nbins$) Algorithm:	(* Bin bivariate data *)
<pre> for j = 1, 2 { δ_j = (b_j - a_j)/nbins_j } for k₁ = 1, nbins₁ { for k₂ = 1, nbins₂ { ν_{k₁,k₂} = 0 } } for i = 1, n { for j = 1, 2 { k_j = 1 + (x_{ij} - a_j)/δ_j } (* integer part *) ν_{k₁,k₂} = ν_{k₁,k₂} + 1 } return ({ν_{k₁,k₂}}) </pre>	

ASH2($m, ν, nbins, a, b, n, w_{m_1}, w_{m_2}$) Algorithm:	(* Bivariate ASH *)
<pre> for i = 1 - m₁, m₁ - 1 { for j = 1 - m₂, m₂ - 1 { w_{ij} = w_{m₁}(i)w_{m₂}(j) } for j = 1, 2 { δ_j = (b_j - a_j)/nbins_j; h_j = m_jδ_j } for k = 1, nbins₁ { for ℓ = 1, nbins₂ { f_{kℓ} = 0 } } for k = 1, nbins₁ { for ℓ = 1, nbins₂ { if (ν_{kℓ} = 0) next ℓ for i = max(1, k - m₁ + 1), min(nbins₁, k + m₁ - 1) { for j = max(1, ℓ - m₂ + 1), min(nbins₂, ℓ + m₂ - 1) { f_{ij} = f_{ij} + ν_{kℓ}w_{ij} } } } for k = 1, nbins₁ { for ℓ = 1, nbins₂ { f_{kℓ} = f_{kℓ}/(nh₁h₂) } } for k = 1, nbins₁ { t_{1k} = a₁ + (k - 0.5)δ₁ } (* Bin centers x-axis *) for k = 1, nbins₂ { t_{2k} = a₂ + (k - 0.5)δ₂ } (* Bin centers y-axis *) return (x = {t_{1k}}, y = {t_{2k}}, z = {f_{kℓ}}) (* z = ASH *) </pre>	

5.3 THE LIMITING ASH AS A KERNEL ESTIMATOR

The parameter m in the ASH is a nuisance parameter, but much less so than the bin origin. The precise choice of m is unimportant as long as it is greater than 2 and h is well-chosen. Then why study the limiting behavior of the ASH

as $m \rightarrow \infty$, where the ASH loses computational efficiency? The limit is in a class of nonparametric estimators that has been extensively studied since the pioneering works of Fix and Hodges (1951), Rosenblatt (1956), and Parzen (1962).

With h and n fixed and m increasing, it is easy to isolate the effect of a single data point x_j on the ASH estimate $\hat{f}(x)$, at a fixed point x . If $x \in B_k$ and $x_j \in B_{k+i}$, where the index labeling of the bins changes as m increases, then from Equation (5.4) the influence of x_j on x is proportional to

$$1 - \frac{|i|}{m} = 1 - \frac{|i| \cdot \delta}{m \cdot \delta} = 1 - \frac{|x - x_j|}{h} + O\left(\frac{\delta}{h}\right), \quad \text{if } |x - x_j| < h. \quad (5.11)$$

If x_j is not in the interval $(x - h, x + h)$, then the influence is 0. Note that the number of bins between x and x_j is approximately i , since these points are in bins B_k and B_{k+i} , respectively; hence, $|x - x_j| \approx |i| \cdot \delta$. Equation (5.4) may be reexpressed as

$$\lim_{m \rightarrow \infty} \hat{f}(x; m) = \frac{1}{nh} \sum_{j=1}^n \left(1 - \frac{|x - x_j|}{h}\right) I_{[-1, 1]} \left(\frac{x - x_j}{h}\right), \quad (5.12)$$

where the sum is over the number of data points rather than the number of bins. Defining a *kernel function* $K(\cdot)$ to be an isosceles triangle density,

$$K(t) = (1 - |t|) I_{[-1, 1]}(t), \quad (5.13)$$

the limiting ASH may be written as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right). \quad (5.14)$$

Formula (5.14) also defines the *general kernel density estimator* with kernel K , corresponding to the generalized ASH in Equation (5.5). Apparently, the kernel estimate is simply a mixture density, which has n identical component densities centered on the data points. The component densities are the kernel functions. Any probability density may be chosen for the kernel, and sometimes kernels that are not densities are used. The ASH kernel always has finite support, but an infinite-support kernel such as the Normal density is often chosen in (5.14). The isosceles triangle kernel density estimator could be described as an *indifferent histogram*, where the reference is to the uniform weighting over all possible choices for the bin origin of a histogram. Kernel estimators are studied in detail in Chapter 6.

Graphically, the kernel estimate places a probability mass of size $1/n$ in the shape of the kernel, which has been scaled by the smoothing parameter h , centered on each data point. These probability masses are then added vertically to give the kernel estimate. In contrast, the histogram uses a rectangular kernel

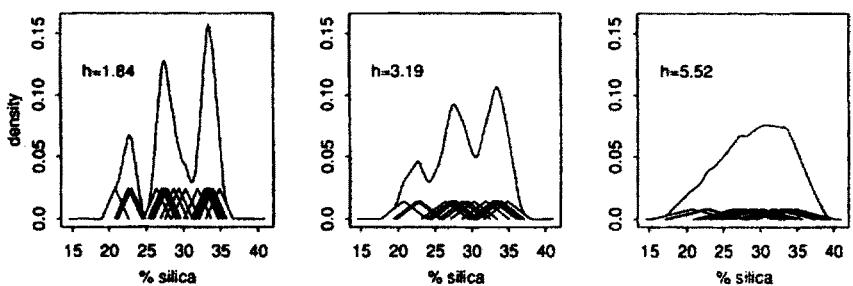


Figure 5.4 Triangle kernel estimates of the silica data showing the individual kernels.

but does not center these kernels on the data points; rather, these kernels are placed in a rigid mesh. In Figure 5.4, this process is illustrated with the silica data (in Table 7 in Appendix B) for several choices of the smoothing parameter and the isosceles triangle kernel. The 22 kernels for the individual data points are shown correctly scaled in each panel.

Of particular interest is the multivariate kernel corresponding to the multivariate naive ASH. Some algebra reveals that as $m_i \rightarrow \infty$,

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 h_2 \cdots h_d} \sum_{i=1}^n \left\{ \prod_{j=1}^d K\left(\frac{x_j - x_{ij}}{h_j} \right) \right\}, \quad (5.15)$$

where K is the univariate isosceles triangle kernel (5.13). This special form of the multivariate kernel function is called the *product kernel* and the estimate (5.15) the *product kernel estimator*. Although the individual multivariate product kernel does factor (implying that the coordinates are independent), the resultant density estimate does not factor, as is apparent from the examples displayed in Figure 5.2.

Thus the ASH provides a direct link to the better known kernel methods. However, kernel estimators are notoriously slow to compute, and many faster numerical approximations have been considered. The ASH is a *bona fide* density estimator and a natural candidate for computation. The ASH uses a discrete convolution to perform smoothing, a device well-known in spectral density estimation. The ASH construction was described independently by Chamayou (1980). The ASH is a special case of a more general framework called WARPing (weighted average of shifted points) developed by Härdle and Scott (1988) where the computational efficiency of the ASH is discussed in more detail. Wegman (1990) has used the ASH to address the problem of too much ink in the parallel coordinates plot discussed in Chapter 1. He proposed plotting the line segments as a series of points on a fine vertical mesh and plotting the contours of a bivariate ASH of those points.

PROBLEMS

1. Consider Galton's bivariate smoothing scheme, which placed equal weights on the counts in only 4 of the 8 bins surrounding the bin of interest, and no weight on the count in the central bin. What are the weights on these 9 bins with the bivariate naive ASH with $m_1 = m_2 = 2$?
2. Prove that if the weights $\{w_m(i)\}$ in Equation (5.5) sum to 1, then the ASH integrates to 1.
3. Prove Theorem 5.2.
4. Generalize the ASH1 algorithm to handle noninteger values of m , which is the situation when the smoothing parameter h is not an integer multiple of δ (Scott, 1991c).
5. What is the kernel corresponding to the averaged shifted frequency polygon?
6. Show that the limiting form of the bivariate naive ASH is in the form (5.15).

CHAPTER 6

Kernel Density Estimators

It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950s, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. In a little-known paper, Fix and Hodges (1951) introduced the basic algorithm of nonparametric density estimation. They addressed the problem of statistical discrimination when the parametric form of the sampling density was not known. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt (1956), Parzen (1962), and Cencov (1962). There followed a second wave of important and primarily theoretical papers by Watson and Leadbetter (1963), Loftsgaarden and Quesenberry (1965), Schwartz (1967), Epanechnikov (1969), Tarter and Kronmal (1970), and Wahba (1971). The natural multivariate generalization was introduced by Cacoullos (1966). Finally, in the 1970s came the first papers focusing on the practical application of these methods: Scott et al. (1978) and Silverman (1978b). These and later multivariate applications awaited the computing revolution.

The basic kernel estimator may be written compactly as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i), \quad (6.1)$$

where $K_h(t) = K(t/h)/h$ [a notation introduced by Rosenblatt (1956)]. The kernel estimator can be motivated not only as the limiting case of the averaged shifted histogram as in (5.14), but also by other techniques demonstrated in Section 6.1. In fact, virtually all nonparametric algorithms are asymptotically kernel methods, a fact demonstrated empirically by Walter and Blum (1979) and proved rigorously by Terrell and Scott (1992). Woodroffe (1970) called the general class “delta sequences.”

6.1 MOTIVATION FOR KERNEL ESTIMATORS

From the vantage point of a statistician or instructor, the averaging of shifted histograms seems the most natural motivation for kernel estimators. However,

following other starting points in numerical analysis, time series, and signal processing provides a deeper understanding of kernel methods. When trying to understand a particular theoretical or practical point concerning a nonparametric estimator, not all approaches are equally powerful. For example, Fourier analysis provides sophisticated tools for theoretical purposes. The bias-variance trade-off can be recast in terms of low-pass and high-pass filters in signal processing. Each is describing the same entity but with different mathematics.

6.1.1 Numerical Analysis and Finite Differences

The kernel estimator originated as a numeric approximation to the derivative of the cumulative distribution function (Rosenblatt, 1956). The empirical probability density function, which was defined in Equation (2.2) as the formal derivative of the empirical cdf $F_n(x)$, is a sum of Dirac delta functions, which is useless as an estimator of a smooth density function. Consider, however, a one-sided finite difference approximation to the derivative of $F_n(\cdot)$:

$$\begin{aligned}\hat{f}(x) &= \frac{F_n(x) - F_n(x - h)}{h} \\ &= \frac{1}{nh} \sum_{i=1}^n I_{[x-h, x)}(x_i) = \frac{1}{nh} \sum_{i=1}^n I_{(0, 1]} \left(\frac{x - x_i}{h} \right),\end{aligned}\quad (6.2)$$

which from Equation (6.1) is clearly a kernel estimator with $K = U(0, 1]$. As $E[F_n(x)] = F(x)$ for all x , then with the Taylor's series

$$F(x - h) = F(x) - hf(x) + \frac{1}{2}h^2f'(x) - \frac{1}{6}h^3f''(x) + \dots,$$

the bias is easily computed as

$$\text{Bias}\{\hat{f}(x)\} = E[\hat{f}(x)] - f(x) = -\frac{1}{2}hf'(x) + O(h^2).$$

Thus the integrated squared bias is $h^2 R(f')/4$, which is comparable to the order of the ISB of the histogram in Theorem 3.1 rather than the $O(h^4)$ ISB of the frequency polygon. Furthermore, the ISB of (6.2) is 3 times larger than the ISB of the histogram. (The integrated variances are identical; see Problem 1.) Thus, the one-sided kernel estimator (6.2) is inferior to a histogram.

Without comment, Rosenblatt proposed a two-sided or central difference estimator of f :

$$\hat{f}(x) = \frac{F_n\left(x + \frac{h}{2}\right) - F_n\left(x - \frac{h}{2}\right)}{h}.\quad (6.3)$$

The bias of (6.3) turns out to be $h^2 f''(x)/24$; see Problem 2. Thus the squared bias is $O(h^4)$, matching that of the FP. The corresponding kernel is $K =$

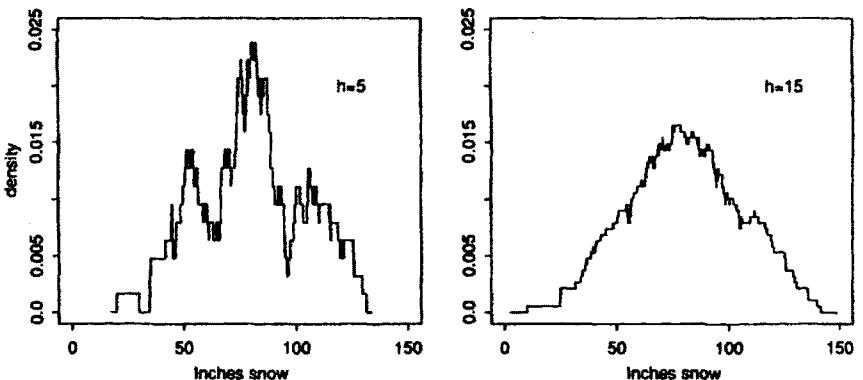


Figure 6.1 Central difference estimates of the Buffalo snowfall data.

$U(-0.5, 0.5)$. Recall that the histogram placed a rectangular block into the bin where each data point fell. The one-sided estimator (6.2) places the left edge of a rectangular block at each data point, whereas the two-sided estimator (6.3) places the center of a rectangular block at each data point; see Tarter and Kronmal (1976).

Figure 6.1 displays 2 central difference estimates of the Buffalo snowfall data. Compared with the FP, these estimates are inferior graphically, as the estimate contains $2n$ jumps that are not even equally spaced. However, most criteria such as MISE are not particularly sensitive to such local noisy behavior.

Quasi-Newton optimization codes routinely make use of numerical central difference estimates of derivatives. Some codes use even “higher-order” approximations to the first derivative; see Section 6.2.3.1.

6.1.2 Smoothing by Convolution

An electrical engineer facing a noisy function will reach into a grab bag of convolution filters to find one which will smooth away the undesired high-frequency components. The convolution operation replaces the value of a function by a local weighted average of the function’s values, according to a weight function $w(\cdot)$ that is usually symmetric and concentrated around 0. Statisticians rely on the operation of averaging to reduce variance. Therefore, the empirical density function, which is too noisy, may be filtered, with the result that

$$\left[\frac{dF_n}{dx} \right] * w = \int_{-\infty}^{\infty} \left[\frac{1}{n} \sum_{i=1}^n \delta(t - x_i) \right] w(x - t) dt = \frac{1}{n} \sum_{i=1}^n w(x - x_i), \quad (6.4)$$

which is the second kernel form given in (6.1) but without the smoothing parameter h . In general, the shape and extent of the convolution filter weight function w will depend on the sample size. The kernel estimator (6.1) uses a single “shape” for all sample sizes, and the width of the kernel is explicitly

controlled through the smoothing parameter h . The literature on filter design often uses different terminology. For example, the width of the filter w is controlled by the half-power point, where the filter reaches half its value at the origin.

6.1.3 Orthogonal Series Approximations

The heuristic introduction to smoothing by convolution may be formalized by an orthogonal series approximation argument. For simplicity, suppose that the density function f is periodic on the interval $[0, 1]$ so that the ordinary Fourier series basis, $\phi_\nu(t) = \exp(2\pi i \nu t)$, is appropriate. Every function, even noisy functions, may be expressed in terms of the basis functions as

$$f(x) = \sum_{\nu=-\infty}^{\infty} f_\nu \phi_\nu(x) \quad \text{where } f_\nu = \langle f, \phi_\nu \rangle = \int_0^1 f(x) \phi_\nu^*(x) dx. \quad (6.5)$$

The basis functions are orthonormal, that is, $\int \phi_\nu^*(x) \phi_\mu(x) dx = \delta_{\mu\nu}$, where $\delta_{\mu\nu}$ is the Kronecker delta function and ϕ^* denotes complex conjugate. As f is a density function, the coefficient f_ν in Equation (6.5) may be expressed in statistical terms as

$$f_\nu = E[\phi_\nu^*(X)]; \quad \text{hence} \quad \hat{f}_\nu = \frac{1}{n} \sum_{i=1}^n \phi_\nu^*(x_i) \quad (6.6)$$

is an unbiased and consistent estimator of the Fourier coefficient f_ν . As an extreme example, consider the Fourier coefficients of the empirical probability density function ((2.2)):

$$f_\nu = \int_0^1 \left[\frac{1}{n} \sum_{i=1}^n \delta(x - x_i) \right] \phi_\nu^*(x) dx = \frac{1}{n} \sum_{i=1}^n \phi_\nu^*(x_i) = \hat{f}_\nu,$$

where \hat{f}_ν is defined in (6.6). Since \hat{f}_ν and f_ν for the empirical pdf are *identical*, the following is formally true for any sample $\{x_i\}$:

$$\sum_{\nu=-\infty}^{\infty} \hat{f}_\nu \phi_\nu(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - x_i), \quad (6.7)$$

which is the empirical probability density function!

Cencov (1962), Kronmal and Tarter (1968), and Watson (1969) suggested smoothing the empirical density function by including only a few selected terms from Equation (6.7). Excluding terms of the form $|\nu| > k$ corresponds to what the engineers call “boxcar” filter weights

$$w_\nu(k) = \begin{cases} 1 & |\nu| \leq k \\ 0 & \text{otherwise.} \end{cases} \quad (6.8)$$

As the Fourier transform of the boxcar function is the *sinc* function, $\sin(x)/(\pi x)$, the estimate will be rough and will experience “leakage”; that is, sample points relatively distant from a point x will influence $\hat{f}(x)$. Wahba (1977) suggests applying a smooth tapering window to this series, which provides more fine tuning of the resulting estimate. She introduces 2 parameters, λ and p , that control the shape and extent of the tapering window:

$$w_\nu(\lambda, p) = \frac{1}{1 + \lambda(2\pi\nu)^{2p}} \quad \text{for } |\nu| \leq n/2. \quad (6.9)$$

Both forms of the weighted Fourier estimate may be written explicitly as

$$\hat{f}(x) = \sum_\nu w_\nu \left[\frac{1}{n} \sum_{i=1}^n \phi_\nu^*(x_i) \right] \phi_\nu(x) = \frac{1}{n} \sum_{i=1}^n \left[\sum_\nu w_\nu \phi_\nu^*(x_i) \phi_\nu(x) \right], \quad (6.10)$$

where the order of summations has been exchanged. With the Fourier basis, the orthogonal series estimator (6.10) equals

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \left[\sum_\nu w_\nu e^{2\pi i \nu(x - x_i)} \right].$$

This estimator is now in the convolution form (6.4) of a fixed kernel estimator, with the filter (or kernel) defined by the quantity in brackets. Some examples of these kernel functions are shown in Figure 6.2. Wahba's equivalent kernels are smoother and experience less leakage. The parameters λ and p can be shown to have interpretations corresponding to the smoothing parameter and to the order of the finite difference approximation, respectively.

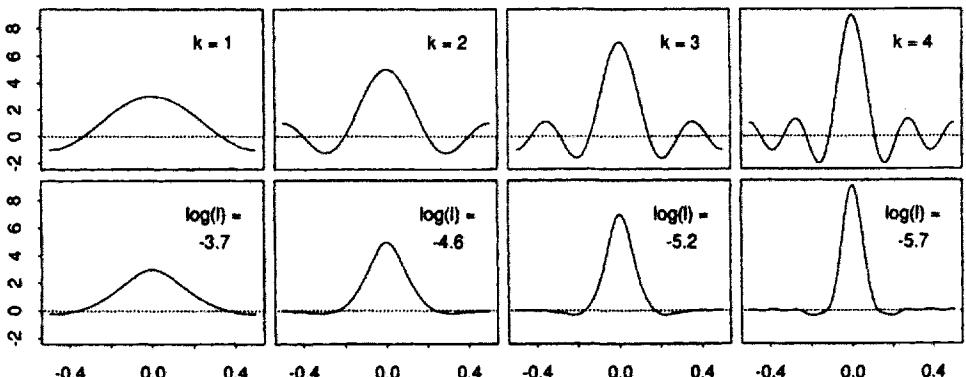


Figure 6.2 Examples of equivalent kernels for orthogonal series estimators. The 4 Wahba kernels (bottom row) have been selected to match the peak height of the corresponding Kronmal-Tarter-Watson kernels (top row). The Kronmal-Tarter-Watson kernels are independent of sample size; the Wahba examples are for $n = 16$.

6.2 THEORETICAL PROPERTIES: UNIVARIATE CASE

6.2.1 MISE Analysis

The statistical analysis of kernel estimators is much simpler than for histograms, as the kernel estimator (6.1) is the *arithmetic mean* of n independent and identically distributed random variables,

$$K_h(x, X_i) = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Therefore,

$$\mathbb{E}\{\hat{f}(x)\} = \mathbb{E} K_h(x, X) \quad \text{and} \quad \text{Var}\{\hat{f}(x)\} = \frac{1}{n} \text{Var} K_h(x, X).$$

The expectation equals

$$\begin{aligned} \mathbb{E} K_h(x, X) &= \int \frac{1}{h} K\left(\frac{x - t}{h}\right) f(t) dt = \int K(w) f(x - hw) dw \\ &= f(x) \int K(w) - hf'(x) \int w K(w) + \frac{1}{2} h^2 f''(x) \int w^2 K(w) + \dots, \end{aligned} \tag{6.11}$$

and the variance is given by

$$\text{Var} K_h(x, X) = \mathbb{E}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right]^2 - \left[\mathbb{E} \frac{1}{h} K\left(\frac{x - X}{h}\right)\right]^2. \tag{6.12}$$

The second term in (6.12) was computed in (6.11) and is approximately equal to $[f(x) \int K(w) + \dots]^2$, while the first term may be approximated by

$$\int \frac{1}{h^2} K\left(\frac{x - t}{h}\right)^2 f(t) dt = \int \frac{1}{h} K(w)^2 f(x - hw) dw \approx \frac{f(x)R(K)}{h}. \tag{6.13}$$

From Equation (6.11), if the kernel K satisfies

$$\int K(w) = 1, \quad \int w K(w) = 0, \quad \text{and} \quad \int w^2 K(w) = \sigma_K^2 > 0,$$

then the expectation of $\hat{f}(x)$ will equal $f(x)$ to order $O(h^2)$. In fact,

$$\text{Bias}(x) = \frac{1}{2} \sigma_K^2 h^2 f''(x) + O(h^4) \implies \text{ISB} = \frac{1}{4} \sigma_K^4 h^4 R(f'') + O(h^6). \tag{6.14}$$

Similarly, from (6.12), (6.13), and (6.11),

$$\text{Var}(x) = \frac{f(x)R(K)}{nh} - \frac{f(x)^2}{n} + O\left(\frac{h}{n}\right) \implies \\ \text{IV} = \frac{R(K)}{nh} - \frac{R(f)}{n} + \dots \quad (6.15)$$

These results are summarized in the following theorem.

Theorem 6.1: *For a nonnegative univariate kernel density estimator,*

$$\begin{aligned} \text{AMISE} &= \frac{R(K)}{nh} + \frac{1}{4} \sigma_K^4 h^4 R(f'') \\ h^* &= \left[\frac{R(K)}{\sigma_K^4 R(f'')} \right]^{1/5} n^{-1/5} \\ \text{AMISE}^* &= \frac{5}{4} [\sigma_K R(K)]^{4/5} R(f'')^{1/5} n^{-4/5}. \end{aligned} \quad (6.16)$$

The conditions under which the theorem holds have been explored by many authors, including Parzen (1962). A simple set of conditions is that the kernel K be a continuous probability density function with finite support, $K \in L_2$, $\mu_K = 0$, $0 < \sigma_K^2 < \infty$, and that f'' be absolutely continuous and $f''' \in L_2$ (Scott, 1985b).

It is easy to check that the ratio of IV to ISB in the AMISE* is 4:1. That is, the ISB comprises only 20% of the AMISE. The similarity to the FP results in Theorem 4.1 is clear. If K is an isosceles triangle, then the results in Theorem 6.1 match those for the naive ASH with $m = \infty$ in Equation (5.8); see Problem 3. Since $R(\phi''(x; 0, \sigma^2)) = 3/(8\sqrt{\pi}\sigma^5)$, the Normal reference rule bandwidth with a Normal kernel is

$$\text{Normal reference rule: } h = (4/3)^{1/5} \sigma n^{-1/5} \approx 1.06 \hat{\sigma} n^{-1/5}. \quad (6.17)$$

6.2.2 Estimation of Derivatives

Occasionally, there arises a need to estimate the derivatives of the density function; for example, when looking for modes and bumps. Derivatives of an ordinary kernel estimate behave consistently if the kernel is sufficiently differentiable and if wider bandwidths are selected. Larger smoothing parameters are required as the derivative of a function is noisier than the function itself. Take as an estimator of the r th derivative of f the r th derivative of the kernel estimate:

$$\hat{f}^{(r)}(x) = \frac{d^r}{dx^r} \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) = \frac{1}{nh^{r+1}} \sum_{i=1}^n K^{(r)}\left(\frac{x - x_i}{h}\right). \quad (6.18)$$

A calculation similar to that leading to Equation (6.13) shows that

$$\text{Var}\{\hat{f}^{(r)}(x)\} \approx \frac{n}{(nh^{r+1})^2} E\left\{K^{(r)}\left(\frac{x - X}{h}\right)^2\right\} \approx \frac{f(x)R(K^{(r)})}{nh^{2r+1}},$$

hence, the asymptotic integrated variance of $\hat{f}^{(r)}$ is $R(K^{(r)})/(nh^{2r+1})$. After an expansion similar to (6.11) to find the bias, the expectation of the first derivative estimator is

$$E\hat{f}'(x) = \frac{1}{h} \left[f_x \int K' - hf'_x \int wK' + \frac{h^2}{2} f''_x \int w^2 K' - \frac{h^3}{6} f'''_x \int w^3 K' + \dots \right],$$

where $f_x^{(r)} \equiv f^{(r)}(x)$. Assuming K is symmetric, $\int w^r K' = 0$ for even r . Integrating by parts, $\int wK' = -1$ and $\int w^3 K' = -3\sigma_K^2$. Hence, the pointwise bias is of order h^2 and involves the third derivative of f . A general theorem is easily given (see Problem 6).

Theorem 6.2: *Based on a nonnegative univariate kernel density estimator \hat{f} ,*

$$\text{AMISE}(\hat{f}^{(r)}) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{1}{4} h^4 \sigma_K^4 R(f^{(r+2)}), \quad (6.19)$$

$$h_r^* = \left[\frac{(2r+1)R(K^{(r)})}{\sigma_K^4 R(f^{(r+2)})} \right]^{1/(2r+5)} n^{-1/(2r+5)}$$

$$\text{AMISE}^*(\hat{f}^{(r)}) = \frac{2r+5}{4} R(K^{(r)})^{\frac{4}{2r+5}} [\sigma_K^4 R(f^{(r+2)})/(2r+1)]^{\frac{2r+1}{2r+5}} n^{\frac{-4}{2r+5}}.$$

While the order of the bias term remains $O(h^4)$, each additional derivative order introduces 2 extra powers of h in the variance. The optimal smoothing parameters h^* for the first and second derivatives are $O(n^{-1/7})$ and $O(n^{-1/9})$, respectively, while the AMISE* is $O(n^{-4/7})$ and $O(n^{-4/9})$. If the optimal density rate $h^* = O(n^{-1/5})$ is used in the estimate of the second derivative, the asymptotic IV in the AMISE does not vanish, since $nh^5 = O(1)$. The estimation of an additional derivative is more difficult than estimating an additional dimension. For example, the optimal AMISE rate for the second derivative is $O(n^{-4/9})$, which is the same (slower) rate as for the optimal AMISE of a 5-D multivariate frequency polygon density estimator.

6.2.3 Choice of Kernel

Much of the first decade of theoretical work focused upon various aspects of estimation properties relating to the characteristics of a kernel. The quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter, and only in a minor way by the choice of kernel. Thus the topic could be de-emphasized. However, there has been a recent spurt of useful research on kernel design in special situations. While many potential hazards face the user of density estimation (for example, underestimating the smoothness of the unknown density), the specification of desired properties for the kernel is entirely at the disposal of the worker, who should have a good understanding of the following results.

6.2.3.1 Higher-Order Kernels

Bartlett (1963) considered the possibility of carefully choosing the kernel to further reduce the contribution of the bias to the MISE. If the requirement that the kernel estimate should itself be a true density is relaxed, then it is possible to achieve significant improvement in the MISE. Suppose a kernel of order p is chosen so that

$$\int K = 1; \quad \int w^i K = 0, \quad i = 1, \dots, p - 1; \quad \text{and} \quad \int w^p K \neq 0, \quad (6.20)$$

then continuing the expansion in equation (6.11), the pointwise kernel bias becomes [letting $\mu_i = \int w^i K(w) dw$]

$$\text{Bias}\{\hat{f}(x)\} = \frac{1}{p!} h^p \mu_p f^{(p)}(x) + \dots$$

Since the formulas for the pointwise and integrated variances are unchanged, the following theorem may be proved.

Theorem 6.3: Assuming that f is sufficiently differentiable and that the kernel K is of order p ,

$$\begin{aligned} \text{AMISE}(h) &= \frac{R(K)}{nh} + \frac{1}{(p!)^2} h^{2p} \mu_p^2 R(f^{(p)}) \\ h^* &= \left[\frac{(p!)^2 R(K)}{2p \mu_p^2 R(f^{(p)})} \right]^{1/(2p+1)} n^{-1/(2p+1)} \\ \text{AMISE}^* &\propto \left[\mu_p^2 R(K)^{2p} R(f^{(p)}) \right]^{1/(2p+1)} n^{-2p/(2p+1)}. \end{aligned} \quad (6.21)$$

Table 6.1 Some Simple Polynomial Higher-Order Kernels

p	K_p on $[-1, 1]$	$N(0, 1)$ AMISE*
2	$\frac{3}{4}(1 - t^2)$	$0.320n^{-4/5}$
4	$\frac{15}{32}(1 - t^2)(3 - 7t^2)$	$0.482n^{-8/9}$
6	$\frac{105}{256}(1 - t^2)(5 - 30t^2 + 33t^4)$	$0.581n^{-12/13}$
8	$\frac{315}{4096}(1 - t^2)(35 - 385t^2 + 1,001t^4 - 715t^6)$	$0.681n^{-16/17}$

Asymptotically, this result indicates it is possible to approach the usual parametric rate of $O(n^{-1})$ for the AMISE. However, the width of the optimal bandwidths increases as the order of the kernel p increases, suggesting that much of the benefit may be quite asymptotic for large p .

Table 6.1 shows some higher-order kernels, together with the optimal AMISE for the case of standard Normal data. These kernels are shown in Figure 6.3. Selecting representative higher-order kernels is difficult, for reasons given in Section 6.2.3.2. Only *even* values of p are considered, because all *odd* "moments" of symmetric kernels vanish. Each increase of 2 in p adds another (even) moment constraint in Equation (6.20). If the kernels are polynomials, then the degree of the polynomial must also increase so that there are sufficient degrees of freedom to satisfy the constraints. The kernels in Table 6.1 begin with the so-called Epanechnikov kernel and are the unique continuous polynomial kernels of degree p that satisfy the constraints *and* have their support on the interval $[-1, 1]$.

The plots of the AMISE for Normal data in Figure 6.3 suggest that the higher-order kernels require several thousand data points before a substantial gain may be realized. For rougher data, the gains are even more asymptotic. The improvement made possible by going to a higher-order kernel is not simply a constant multiplicative factor but rather an exponential change in the order of convergence of n . Of course, for small samples, the difference between

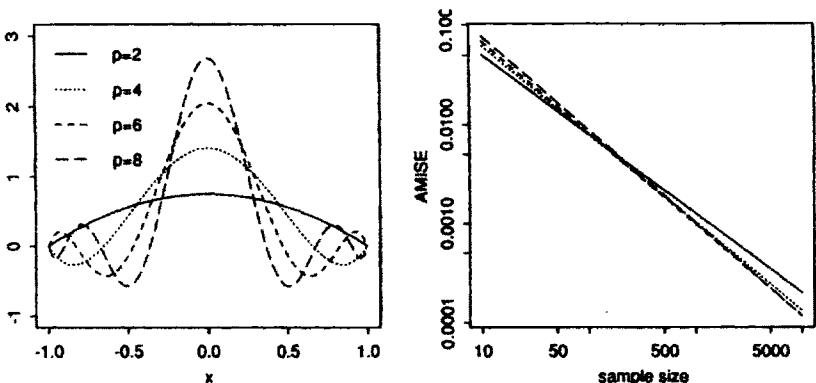


Figure 6.3 Examples of higher-order kernels that are low-order polynomials. The right panel shows the corresponding $N(0, 1)$ AMISE* curves on a log-log scale.

MISE and AMISE may be substantial, particularly for higher-order kernels. The exact MISE may be obtained by numerical integration of the bias and variance equations. For Normal data, the exact MISE was obtained for several sample sizes with the kernels in Table 6.1 plus the histogram. The results are depicted in Figure 6.4. The individual MISE curves are plotted against h/h^* , so

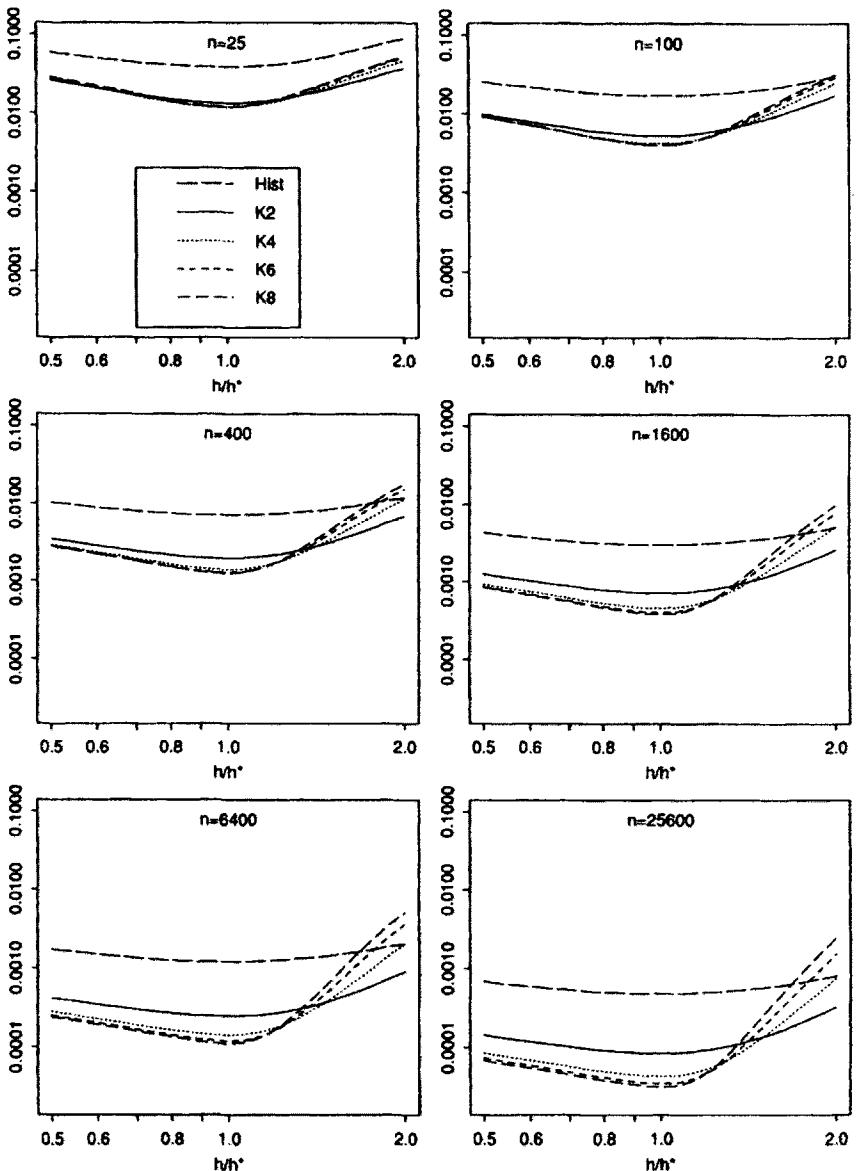


Figure 6.4 Exact MISE using higher-order kernels with Normal data for several sample sizes. The histogram MISE is included for reference.

that the minimum is centered on 1. These figures suggest that in most practical situations, kernels of order 2 and 4 are sufficient. The largest gain in MISE is obtained when going from the histogram to the order-2 kernel, with diminishing returns beyond that. Higher-order kernels also seem sensitive to oversmoothing. The order-8 kernels are inferior to the histogram if $h > 2h^*$.

These higher-order kernels have negative components. They will be referred to as "negative kernels," although the more accurate phrase is "not nonnegative." The introduction of negative kernels does provide improvement in the MISE but at the cost of having to provide special explanations. This negativity is particularly a nuisance in multiple dimensions where the regions of negative estimate can be scattered all over the domain. Statisticians may be comfortable ignoring such features, but care should be taken in their actual use. In practice, negative portions of the estimate could be clipped at 0. Clipping introduces discontinuities in the derivative of the estimate, and the modified density estimate now integrates to slightly more than 1.

In Figure 6.5, ASH estimates using the order 2 and 4 kernels in Table 6.1 are applied to the steel surface data (Bowyer, 1980; Silverman, 1986). The data are measurements from an arbitrary origin of the actual height of a machined flat surface at a grid of 15,000 points. The bandwidths were selected so that the values at the mode matched. This example clearly indicates the reason many statisticians are willing to use negative kernels with large data sets. The negativity problem is quite marginal (minimum value -0.00008) and there is some improvement in the appearance of the peak (simultaneously lower bias and variance).

A second problem introduced by the use of negative kernels involves the subjective choice of bandwidth. With negative kernels, the estimates can appear rough for moderately oversmoothed values of h , and not just for small, undersmoothed values, as with positive kernels. This dual roughness can be a problem for the novice, especially given the promise of higher-order "better"

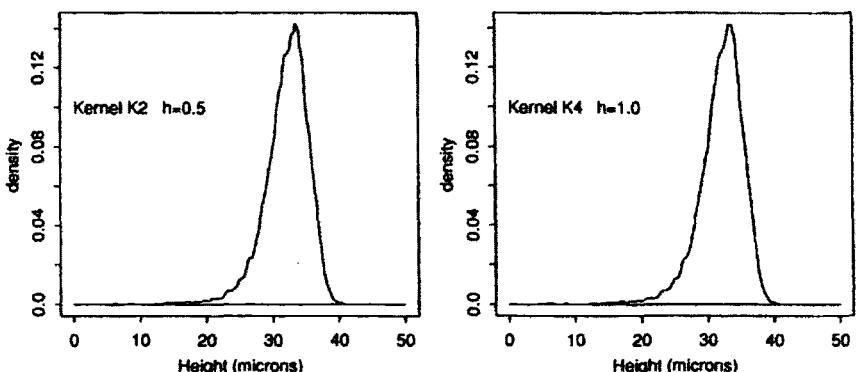


Figure 6.5 Positive and negative kernel estimates of the steel surface data. Kernels used were K_2 and K_4 from Table 6.1.

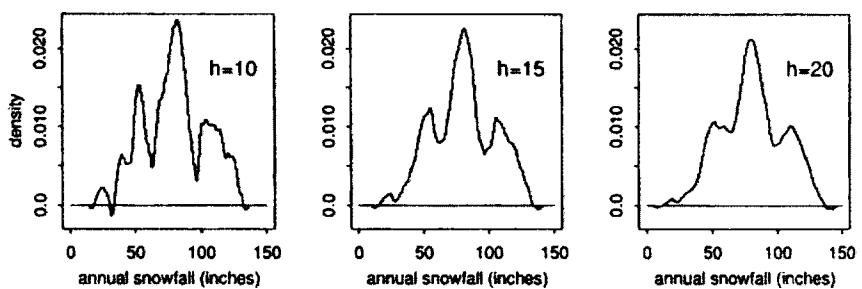


Figure 6.6 Kernel K_4 applied to the Buffalo snowfall data with three smoothing parameters. The ASH estimate is depicted in its histogram form.

estimates. This phenomenon is easy to demonstrate; see Figure 6.6 with the snowfall data.

It is interesting to speculate about the relative merits of using a higher-order kernel versus using a lower-order kernel with an adaptive procedure. (An adaptive kernel behaves much like an adaptive frequency polygon, as in Theorem 4.2.) There is reason to believe that there is a role for both. Asymptotically, higher-order kernels outperform adaptive procedures. The sensitivity to errors in the optimal bandwidth grows with the kernel order, suggesting that adaptivity is more important with higher-order kernels. Consider the exact MSE(h, x) for the simple $N(0, 1)$ density with Normal kernel. A plot of this surface for $x > 0$ is shown in Figure 6.7 (see Fryer, 1976). This surface is a remarkably complex given the simplicity of the density. There are some worrisome anomalies in the surface. For points in the tail, the “optimal” adaptive bandwidths are ridiculously large. The reason is that by averaging over a very large neighborhood, the bias may be eliminated by artificially matching the true density. There is a “local” best adaptive bandwidth (which is the correct target), but on the whole, good adaptive estimation is a difficult task. Schucany (1989) has reported some success in this task.

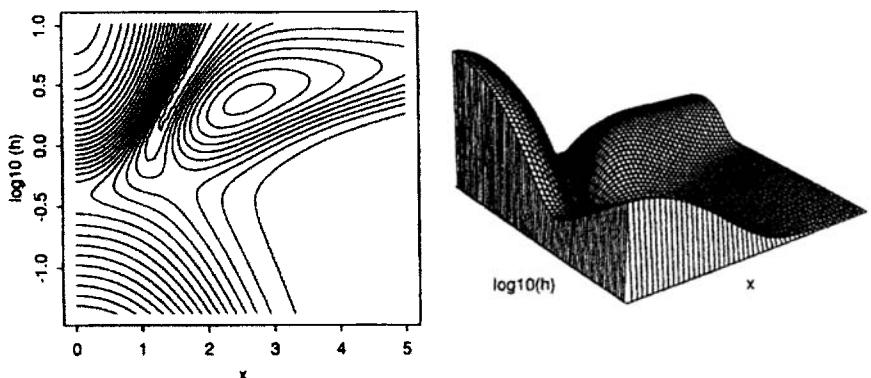


Figure 6.7 Exact MSE as a function of x and h for $f = N(0, 1)$ and $n = 100$.

While taking the limit of kernels as $p \rightarrow \infty$ may not seem wise, Davis (1975) investigated the properties of a particular “ $p = \infty$ ” kernel, the “sinc” function $\sin(x)/(\pi x)$. She showed that the MISE* = $O(n/\log n)$. Marron and Wand (1991) have examined the MISE of a variety of more complex densities and kernels in the Normal mixture family and have computed sample sizes required to justify the use of a higher-order kernel.

Finally, higher-order kernels can be used towards estimating the derivative of a density. However, given the nonmonotone appearance of such kernels, the derivative estimates are likely to exhibit kernel artifacts and should be reserved for data-rich situations.

Terrell and Scott (1980), using a generalized jackknife technique similar to that of Schucany and Sommers (1977), proposed an alternative method of reducing the bias. The jackknife method reduces bias by playing 2 estimators against each other. In density estimation, the procedure involves constructing the ratio of two positive kernel estimators with different bandwidths; for example,

$$\hat{f}(x) = \hat{f}_h(x)^{4/3} \div \hat{f}_{2h}(x)^{1/3}. \quad (6.22)$$

The result follows from jackknifing the $O(h^2)$ bias term in a Taylor's series expansion of the log bias. The expectation E_h of the usual kernel estimate is $E_h = E[\hat{f}_h(x)] = f(x)[1 + c_2 h^2/f(x) + c_4 h^4/f(x) + \dots]$, where $c_2 = h^2 \sigma_K^2 f''(x)/2$, etc. Then by Taylor's expansion,

$$\log E_h = \log f(x) + c_2 h^2/f(x) + (c_4 f(x) - c_2^2/2)h^4/f(x)^2 + \dots$$

$$\log E_{2h} = \log f(x) + 4c_2 h^2/f(x) + 16(c_4 f(x) - c_2^2/2)h^4/f(x)^2 + \dots$$

Then $\frac{4}{3} \log E_h - \frac{1}{3} E_{2h} = \log f(x) + O(h^4)$, which, after taking exponentials, suggests the form given in Equation (6.22); see Problem 13 for details. The authors show that $h^* = 1.42 \sigma n^{-1/9}$ in the case $f = N(0, 1)$ and $K = N(0, \sigma^2)$. The resulting estimate is nonnegative and continuous, but its integral is usually slightly greater than 1. Generally, exceeding the rate $n^{-4/5}$ requires violating 1 of the 2 posits of a density function: nonnegativity and total probability mass of 1.

6.2.3.2 Optimal Kernels

The kernel density estimate inherits all the properties of its kernel. Hence, it is important to note that the naive Rosenblatt kernel is discontinuous, the ASH triangle kernel has a discontinuous derivative, and the Cauchy kernel has no moments. A conservative recommendation is to choose a smooth, clearly unimodal kernel that is symmetric about the origin. However, strange kernel shapes are seldom visible in the final estimate, except perhaps in the tails, because of all the averaging.

The question of finding an optimal kernel for nonnegative estimates was considered by Epanechnikov (1969); the same variational problem was considered by Bartlett (1963) and in another context by Hodges and Lehmann (1956). From Equation (6.16), the kernel's contribution to the optimal AMISE is the following dimensionless factor:

$$\text{AMISE}^* \propto [\sigma_K R(K)]^{4/5}. \quad (6.23)$$

The problem of finding the smoothest density for the oversmoothed bandwidth problem is similar to the problem of minimizing (6.23), which may be written as

$$\min_K R(K) \quad \text{s/t} \quad \sigma_K^2 = \sigma^2.$$

The solution is a scaled version of the so-called Epanechnikov's kernel:

$$K_2^*(t) = \frac{3}{4}(1 - t^2)I_{[-1,1]}(t).$$

It is interesting that the optimal kernel has finite support. The optimal kernel is not differentiable at $t = \pm 1$.

The variance of K_2^* is $1/5$ and $R(K_2^*) = 3/5$ in (6.23). Since the AMISE is also proportional to $n^{-4/5}$, other kernels require (see Problem 7)

$$\frac{\sigma_K R(K)}{\sigma_{K_2^*} R(K_2^*)} = \frac{\sigma_K R(K)}{3/(5\sqrt{5})} \quad (6.24)$$

times as much data to achieve the same AMISE as the optimal Epanechnikov kernel. Table 6.2 lists many commonly used kernels and computes their asymptotic relative efficiency. The optimal kernel shows only modest improvement. Therefore, the kernel can be chosen for other reasons (ease of computation, differentiability, etc.) without undue concern for loss of efficiency. It is somewhat surprising that the popular Normal kernel is so wasteful. Given the computational overhead computing exponentials, it is difficult to recommend the actual use of the Normal kernel except as a point of reference.

In the last part of the table a few absurd kernels are listed to illustrate that a very large inefficiency is still less than 2. From Theorem 4.1, the frequency polygon estimator belongs in the same grouping as the positive kernel estimators. The entries in the table by the FP were obtained by matching the AMISE expressions in Theorems 4.1 and 6.1. The conclusion is that the FP is indeed in the same class, but inefficient. Finally, note that the limiting kernel of the averaged shifted histogram (isosceles triangle) is superior to the limiting kernel of the averaged shifted frequency polygon (indifferent FP), although the FP itself is superior to the histogram.

Table 6.2 Some Common and Some Unusual Kernels and Their Relative Efficiencies^a

Kernel	Equation	$R(K)$	σ_k^2	$\sigma_k R(K)$	Eff.
Uniform	$U(-1, 1)$	1/2	1/3	0.2887	1.0758
Triangle	$(1 - t)_+$	2/3	1/6	0.2722	1.0143
Epanechnikov	$\frac{3}{4}(1 - t^2)_+$	3/5	1/5	0.2683	1
Biweight	$\frac{15}{16}(1 - t^2)_+^2$	5/7	1/7	0.2700	1.0061
Triweight	$\frac{35}{32}(1 - t^2)_+^3$	$\frac{350}{429}$	1/9	0.2720	1.0135
Normal	$N(0, 1)$	$1/2\sqrt{\pi}$	1	0.2821	1.0513
Cosine arch	$\frac{\pi}{4} \cos \frac{\pi}{2} t$	$\pi^2/16$	$1 - 8/\pi^2$	0.2685	1.0005
Indifferent FP	See Problem 16	11/20	1/4	0.2750	1.0249
Dble. exp.	$\frac{1}{2} e^{- t }, t \leq \infty$	1/4	2	0.3536	1.3176
Skewed	$2,860(t + \frac{2}{7})_+^3 (\frac{5}{7} - t)_+^9$	$\frac{7,436}{3,059}$	2/147	0.2835	1.0567
Dble. Epanech.	$3 t (1 - t)_+$	3/5	3/10	0.3286	1.2247
Shifted exp.	$e^{-(t+1)}, t > -1$	1/2	1	0.5743	1.8634
FP	See Theorem 4.1	2/3	$7/12\sqrt{5}$	0.3405	1.2690

^a All kernels are supported on $[-1, 1]$ unless noted otherwise.

The symmetric Beta density functions, when transformed to the interval $(-1, 1)$ so that the mean is 0, are a useful choice for a class of kernels:

$$K_k(t) = \frac{(2k+1)!!}{2^{k+1} k!} (1 - x^2)_+^k, \quad (6.25)$$

where the double factorial notation means $(2k+1)!! = (2k+1)(2k-1)\cdots 5 \cdot 3 \cdot 1$. The Epanechnikov and biweight kernels are in this class. So is the Normal density as $k \rightarrow \infty$; see Problem 17.

The search for optimal high-order kernels is quite different and not so fruitful. Suppose that K_4^*, K_6^*, \dots are the optimal order-4, order-6, \dots kernels, respectively. From Theorem 6.2, the kernel's contribution to the AMISE* for order-4 kernels is the following nonnegative and dimensionless quantity:

$$\text{AMISE}_4^* \propto [R(K)^8 \mu_4^2]^{1/9}. \quad (6.26)$$

Consider the following fourth-order kernel, which is a mixture of K_4^* and K_6^* :

$$K_\epsilon(t) = \epsilon K_4^*(t) + (1 - \epsilon) K_6^*(t), \quad 0 \leq \epsilon \leq 1.$$

K_ϵ has finite roughness but its fourth moment vanishes as $\epsilon \rightarrow 0$. Thus, the fourth moment of K_4^* must be 0 and the criterion in (6.26) equals 0 at the solution. But by definition, then, K_4^* is no longer a fourth-order kernel. As many kernels have zero fourth "moment" but finite roughness, the lower bound of 0 is achieved by many kernels, none of which are in any sense order-4 or

interesting in the sense of Epanechnikov. Of course, the AMISE would not in fact be zero, but would involve higher-order terms.

Choosing among higher-order kernels is quite complex and it is difficult to draw guidelines. In practice, second- and fourth-order methods are probably the most one should consider, as kernels beyond the order-4 provide little further reduction in MISE. For very large samples, where higher-order methods do provide a substantial *fractional* reduction, the *absolute* MISE may already be so small that any practical advantage is lost.

The general advice on choosing a kernel based on these observations is to choose a symmetric kernel that is a low-order polynomial. Gasser, Müller, and Mammitzsch (1985) have developed a smooth hierarchy of higher-order kernels. They show that their kernels are optimal but in a different sense: these kernels have minimum roughness subject to a fixed number of sign changes in the kernel. Such justification does not really warrant the label “optimal” in the sense of Epanechnikov. However, they have provided a valuable formula for low-order polynomial kernels appropriate for various combinations of the kernel order p , and for the r th derivative of the density,

$$K_{(k,p,r)}^* = \sum_{i=0}^{k+2(r-1)} \lambda_i t^i I_{[-1,1]}(t),$$

where $k \geq p + 2$ and (k,p) are both odd or both even, with

$$\lambda_i = \frac{(-1)^{\frac{i+p}{2}} (k+p+2r)! (k-p) (k+2r-i) (k+i)!}{i! (i+p+1) 2^{2(k+r)+1} \left(\frac{k-p}{2}\right)! \left(\frac{k+p+2r}{2}\right)! \left(\frac{k+2r-i}{2}\right)! \left(\frac{k+i}{2}\right)!}$$

if $k+i$ is even, and 0 otherwise; see Müller (1988).

Given the availability of symbolic manipulation programs, it is probably sufficient to solve the set of linear equations governing the particular application simply. A “designer kernel” approach allows the addition of any linear conditions and results in a new kernel of higher polynomial degree. Again, the choice of kernel is not a critical matter.

6.2.3.3 Equivalent Kernels

For a variety of reasons, there is no single kernel that can be recommended for all circumstances. One serious candidate is the Normal kernel; however, it is relatively inefficient and has infinite support. The optimal Epanechnikov kernel is not continuously differentiable and cannot be used to estimate derivatives. In practice, the ability to switch between different kernels without having to reconsider the calibration problem at every turn is convenient. This task is easy to accomplish, *but only for kernels of the same order*. As Scott (1976) noted, if h_1 and h_2 are smoothing parameters to be used with kernels K_1 and K_2 ,

respectively, then Theorem 6.1 implies that asymptotically

$$\frac{h_1^*}{h_2^*} = \left[\frac{R(K_1)/\sigma_{K_1}^4}{R(K_2)/\sigma_{K_2}^4} \right]^{1/5} = \frac{\sigma_{K_2}}{\sigma_{K_1}} \left[\frac{\sigma_{K_1} R(K_1)}{\sigma_{K_2} R(K_2)} \right]^{1/5}. \quad (6.27)$$

Table 6.3 gives a summary of factors for equivalent smoothing bandwidths among popular kernels.

The term in brackets on the right in Equation (6.27) is the ratio of dimensionless quantities for each kernel. Those quantities $\sigma_k R(K)$ are almost equal to each other, as may be seen from Equation (6.24) and in Table 6.2. Thus the term in Equation (6.27) in brackets can be set to 1, so that the task of choosing equivalent smoothing parameters for different kernels can be accomplished by scaling according to the standard deviations:

Equivalent kernel rescaling:	$h_2^* \approx \frac{\sigma_{K_1}}{\sigma_{K_2}} h_1^*. \quad (6.28)$
------------------------------	---

For example, the “exact” factor going from a Normal to triweight bandwidth in Table 6.3 is 2.978 while the approximate rule (6.28) gives the factor 3.

Equivalent bandwidth scaling provides nearly identical estimates not only for optimal smoothing parameters but for nonoptimal values as well. This rescaling is often used when computing and plotting a biweight kernel estimate, but using a smoothing parameter derived from a Normal kernel cross-validation rule.

If all kernels were presented with equal variances, then no changes in smoothing parameters would be required. However, it would be extremely difficult to remember the formulas of those kernels, as the variances would be incorporated into the kernel forms. On balance, it seems easier to write kernels in a parsimonious form. Marron and Nolan (1988) have proposed scaling all kernels to their “canonical” form having equivalent bandwidths to a Normal

Table 6.3 Factors for Equivalent Smoothing Among Popular Kernels^a

From\To	Normal	Uniform	Epan.	Triangle	Biwt.	Triwt.
Normal	1	1.740	2.214	2.432	2.623	2.978
Uniform	0.575	1	1.272	1.398	1.507	1.711
Epanech.	0.452	0.786	1	1.099	1.185	1.345
Triangle	0.411	0.715	0.910	1	1.078	1.225
Biwt.	0.381	0.663	0.844	0.927	1	1.136
Triwt.	0.336	0.584	0.743	0.817	0.881	1

^a To go from h_1 to h_2 , multiply h_1 by the factor in the table in the row labeled K_1 and in the column labeled K_2 .

kernel. This proposal is slightly different than the variance rescaling proposal, since the kernel roughness is also taken into account.

For higher-order kernels K_{p1} and K_{p2} of order p , a similar rescaling follows from Theorem 6.3 based on the appropriate higher-order “moment” (see Problem 19):

$$\frac{h_{p1}^*}{h_{p2}^*} = \left[\frac{\mu_{p2}}{\mu_{p1}} \right]^{1/p} \left\{ \left[\frac{\mu_{p1}}{\mu_{p2}} \right]^{1/p} \frac{R(K_{p1})}{R(K_{p2})} \right\}^{1/(2p+1)} \approx \left[\frac{\mu_{p2}}{\mu_{p1}} \right]^{1/p}, \quad (6.29)$$

since the quantity in brackets is dimensionless and approximately equal to 1 for most symmetric kernels. When $p = 2$, (6.29) and (6.28) agree. Some of the higher-order moments are *negative*, but their ratio is positive in (6.29).

6.2.3.4 Higher-Order Kernels and Kernel Design

Between kernels of different order, there is no similar notion for choosing bandwidths that give “equivalent smoothing.” Furthermore, the lack of a true “optimal kernel” beyond order 2 is troubling. In this section, higher-order kernels are reintroduced from two other points of view. The results have some curious implications for bandwidth selection and kernel design, and suggest ways in which further work may be helpful.

The first alternative approach appeals to the numerical analysis argument first introduced by Rosenblatt (1956). Using forward and central difference approximations for the derivative, it was shown in Section 6.1.1 that the equivalent kernels are the order-1 and order-2 kernels $U(0, 1)$ and $U(-0.5, 0.5)$, respectively. Using the notation

$$\Delta F_n(x, c) \equiv F_n(x + c) - F_n(x - c),$$

the 2-point Rosenblatt estimator with kernel $U(-1, 1)$ is $\hat{f}_2(x) = \Delta F_n(x, h)/(2h)$. Consider the following 4-, 6-, and 8-point derivative estimators (see Problem 20):

$$\begin{aligned} f_4(x) &= \frac{8\Delta F_n(x, h) - \Delta F_n(x, 2h)}{12h} \\ \hat{f}_6(x) &= \frac{45\Delta F_n(x, h) - 9\Delta F_n(x, 2h) + \Delta F_n(x, 3h)}{60h} \\ \hat{f}_8(x) &= \frac{224\Delta F_n(x, h) - 56\Delta F_n(x, 2h) + \frac{32}{3}\Delta F_n(x, 3h) - \Delta F_n(x, 4h)}{280h}. \end{aligned} \quad (6.30)$$

The biases for these estimators are $-h^4 f^{(4)}(x)/30$, $-h^6 f^{(6)}(x)/140$, and $-h^8 f^{(8)}(x)/630$, respectively. The equivalent kernels are easily visualized by plotting $\hat{f}_r(x)$ with one data point at 0 with $h = 1$ as in Figure 6.8.

In Figure 6.9, the $p = 2$ naive kernel estimate is plotted for 320 cholesterol levels of patients with heart disease. As $\hat{\sigma} = 43$, the bandwidth chosen was $h =$

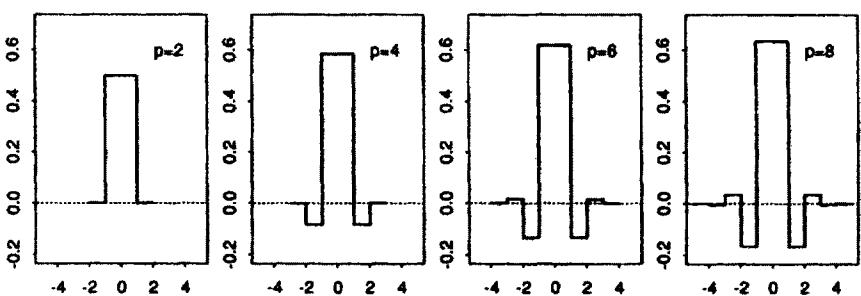


Figure 6.8 Higher-order boxcar or naive kernels based on finite differences.

25, which is the equivalent bandwidth rule for a $U(-1, 1)$ kernel to the Normal reference rule (6.17). Specifically, the equivalent bandwidth is computed as $h \approx 3^{1/2}[1.06 \times 43 \times 320^{-1/5}]$, where $3^{1/2}$ is the standard deviation of the $U(-1, 1)$ kernel. The bandwidths for the $p = 4$ th, 6th, and 8th-order kernels were chosen so that the levels at the modes were equal. The negative side lobes are easily seen. The second- and fourth-order estimates are substantially different. The horizontal line shows the bandwidth h , which is half the width for the central lobe of the kernel. Thus for K_8 , the support is 8 times wider than h ; the influence of each data point extends 216 units (mg/dl) in both directions. For the order-2, 4, and 6 kernels, the extent is 25, 92, and 153, respectively. Higher-order kernels even with finite support are not local.

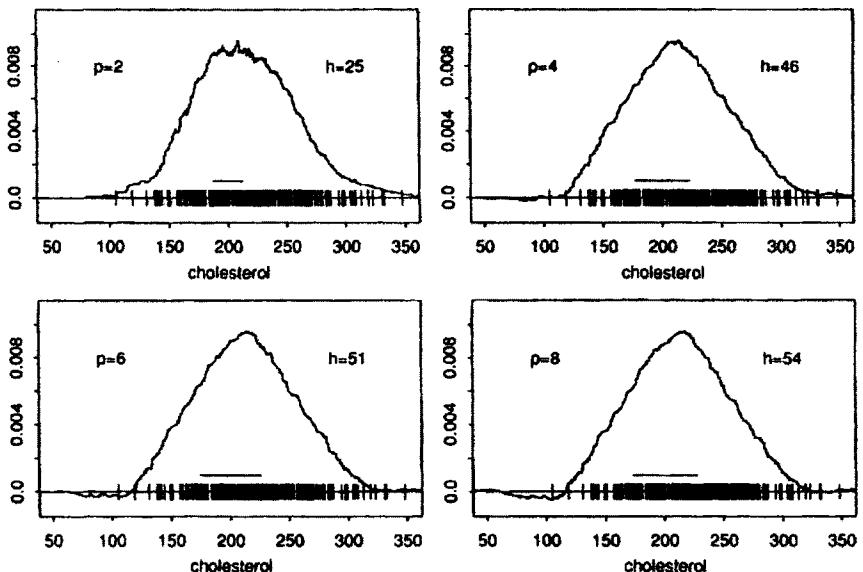


Figure 6.9 Higher-order naive kernels applied to the cholesterol data for diseased males ($n = 320$). Bandwidths are indicated by the horizontal lines.

An empirical observation is that when a higher-order kernel fails to provide further reduction in ISE, then the optimal higher-order density estimates are all very similar (except for small local noise due to the added roughness in the tails of the kernel). This similarity occurs when the central lobe of the kernel [over the interval $(-h, h)$] remains of fixed width even as the order of the kernel grows. The sequence of bandwidths is 25, 46, 51, and 54 for the cholesterol data. Thus $p = 4$ seems a plausible choice for these data, possibly with a narrower bandwidth. As p increases, the negative lobes have an unfortunate tendency to grow and spread out.

Alternatively, the bandwidths for the naive higher-order kernels could have been chosen to increase the estimate at the mode by an estimate of the bias there; for example, for the order-2 kernel

$$\text{Bias}\{\hat{f}(x)\} = \frac{1}{2} h^2 \sigma_K^2 f''(x). \quad (6.31)$$

If the resulting estimate (with the order-4 kernel) is much rougher or if a bandwidth cannot be found that accomplishes the desired increase, then a reasonable conclusion is that the maximum feasible kernel order has been exceeded. The resulting bandwidth for the naive higher-order kernel may be transformed by (6.29) to a smoother higher-order kernel in Table 6.1.

The second alternative introduction to negative kernels suggests that the problem of bandwidth selection is even more complicated than is apparent. Beginning with a positive kernel estimate and the bias estimate given above, the idea is to estimate and remove the bias pointwise by using a kernel estimate of the second derivative. For clarity, a possibly different bandwidth g is used to estimate $f''(x)$:

$$\hat{f}''(x) = \frac{1}{ng^3} \sum_{i=1}^n K''\left(\frac{x - x_i}{g}\right). \quad (6.32)$$

Consider the “bias-corrected” kernel estimate, which is obtained by combining Equations (6.31) and (6.32):

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) - \frac{1}{2} h^2 \sigma_K^2 \times \frac{1}{ng^3} \sum_{i=1}^n K''\left(\frac{x - x_i}{g}\right). \quad (6.33)$$

As $\hat{f}''(x)$ integrates to 0, this modified kernel estimate still integrates to 1. By inspection, (6.33) is itself in the form of a kernel estimator with kernel

$$K_{h,g}(t) = K_h(t) - \frac{h^2 \sigma_K^2}{2g^2} K_g''(t). \quad (6.34)$$

A straightforward calculation verifies that this new kernel has a vanishing second moment, so that this constructive procedure in fact results in a order-4

kernel; see Problem 21. However, this approach suggests that $g \neq h$, since the bandwidth for the second derivative should be wider than for the density estimate itself. If so, then a well-constructed higher-order kernel should slowly “expand” as the sample size increases. Given the relatively slow changes in the bandwidths, there may not be any practical improvement over allowing $g = h$.

6.2.3.5 Boundary Kernels

When the unknown density is discontinuous, kernel estimates suffer the same dramatic loss of MISE efficiency as for the frequency polygon. In the case where the location of the discontinuity is known, an elegant fix is available. Without loss of generality, suppose that the discontinuity occurs at zero and that the density vanishes for $x < 0$. Careful examination of the theoretical argument which led to the elimination of the $O(h)$ bias term for a kernel estimate reveals that the critical requirement is that the kernel satisfy $\int tK(t) dt = 0$. The fundamental requirement is not that the kernel be symmetric; symmetry is only a simple way that the kernel may satisfy the integral equation.

The task, then, is to design finite-support kernels for use with samples in the boundary region $x_i \in (0, h)$. In order that the kernel estimate vanish for $x < 0$, the kernel for a sample point $x_i \in [0, h]$ should cover the interval $[0, x_i + h]$ rather than the interval $(x_i - h, x_i + h)$. As the interval $[0, x_i + h]$ is narrower than $2h$, the roughness of the kernels (and hence the IV) will increase rather dramatically. In a regression setting, Gasser and Müller (1979) and Rice (1984a) have suggested using the wider interval $(0, 2h)$ for every $x_i \in [0, h]$. This suggestion is equivalent to choosing kernels supported on the interval $(c, c + 2)$, for $-1 \leq c \leq 0$, that satisfy $\int_c^{c+2} tK(t) dt = 0$. An attempt to allow the interval width to vary so as to achieve “equivalent smoothing” using the full rule (6.27) seems doomed to failure because a wider interval cannot usually be found with equivalent smoothing. Thus the simple choice of $2h$ is a reasonable compromise.

A designer boundary kernel is described. Assume that the desired boundary kernel is to be a modification of the ordinary biweight kernel, with similar properties at the right-hand endpoint $x = c + 2$. This suggests looking at a designer boundary kernel of the form

$$K_c(t) = [c_1 + c_2(t - c)^2] \cdot [t - (c + 2)]^2, \quad -1 \leq c \leq 0,$$

where the design constants c_1 and c_2 are determined by the two constraints $\int K_c(t) dt = 1$ and $\int tK_c(t) dt = 0$. The form for $K_c(\cdot)$ ensures that the two constraints are linear in the unknowns c_1, c_2 . Solving those equations gives

$$K_c(t) = \frac{3}{4} \left[(c + 1) + \frac{5}{4}(1 + 2c)(t - c)^2 \right] [t - (c + 2)]^2 I_{[c, c+2]}(t). \quad (6.35)$$

Figure 6.10 shows some examples of these kernels in two different ways. First, the kernels are shown centered on the sample (taken to be the origin).

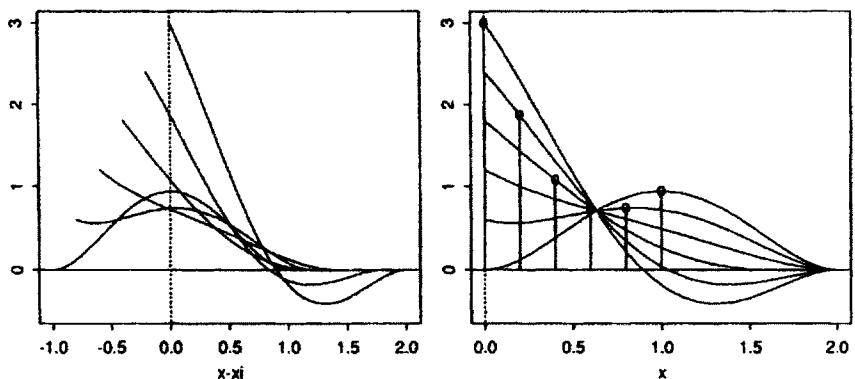


Figure 6.10 Examples of the “floating” boundary kernels $K_c(t)$ for $t = 0, 0.2, 0.4, 0.6, 0.8, 1$. On the right, these kernels are drawn centered on the data point, which is shown by the vertical line.

Second, the kernels are shown as they would be placed on top of the samples, so that they *begin* at zero. Note that if $x_i \in [0, h]$, then the kernel K_c with $c = (0 - x_i)/h$ should be used instead of the ordinary biweight kernel.

Clearly, Figure 6.10 indicates that these are *floating boundary kernels*, meaning that the value of the kernel floats at the left boundary. An example of the use of these kernels with a sample of 100 points from the negative exponential density illustrates the effectiveness of the floating boundary kernels; see Figure 6.11. Notice how the unmodified biweight kernel estimate is quite biased in the interval $(0, h)$ and spills into the $x < 0$ region. However, without any checking or indication of a boundary problem, the unmodified kernel estimate appears quite smooth. (This smoothness should serve as a warning to check for errors resulting from the lack of prior knowledge of the existence of a boundary problem.) For the negative exponential density, the asymptotic theory holds if the roughness is computed on the interval $(0, \infty)$: $\int_0^\infty f''(x)^2 dx = 1/2$, so that $h^* = (70/n)^{1/5}$ for the biweight kernel by Theorem 6.1.

The density estimate of the negative exponential data shown as a big-dashed line in Figure 6.11 was constructed using the *zero boundary kernel*,

$$K_c^0(t) = \frac{15}{16} (t - c)^2 (2 + c - t)^2 [(7c^2 + 14c + 8) - 7t(c + 1)] I_{[c, c+2]}(t). \quad (6.36)$$

This modified biweight kernel was designed with the additional constraint that the boundary kernel and its derivative vanish at the left boundary $x = c$ rather than float as before. This modification was accomplished at the design stage by including the factor $(t - c)^2$, which ensures that the kernel and its derivative vanish at the left-hand endpoint $t = c$. Figure 6.12 displays these kernels in two ways. Clearly, the kernel is inappropriate for negative exponential data, inducing a large, but smooth, oscillation. However, the zero boundary kernel is appropriate for data from the Beta(3, 9) density; see Figure 6.11, which shows

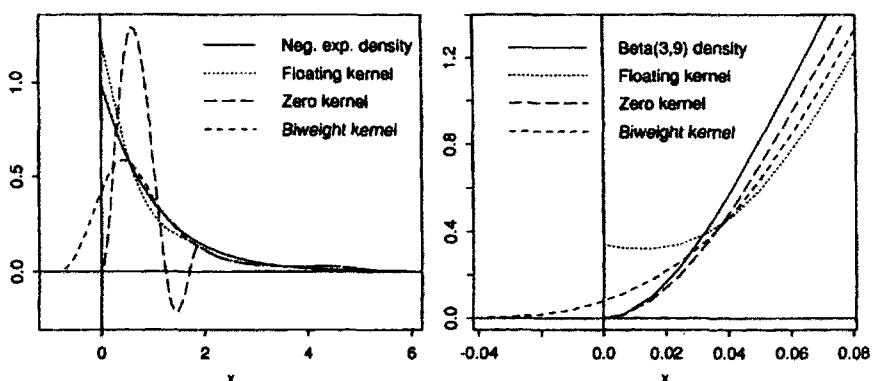


Figure 6.11 *Left frame:* Example with negative exponential data—with and without boundary modification for $n = 100$ and $h = 0.93$. The “floating” and “zero” boundary kernels are defined in (6.35) and (6.36), respectively. *Right frame:* Example with Beta(3,9) density in a neighborhood of 0 for $n = 100$ and $h = 0.11$.

the application of these kernels to a sample of 100 Beta(3,9) points in the vicinity of the origin. For this density, $R(f'') \approx 24,835$ and $h^* \approx (0.269/n)^{1/5}$. The ordinary estimate spills into the negative region and the “floating” kernel estimate lives up to its name.

While boundary kernels can be very useful, there are potentially serious problems with real data. There are an infinite number of boundary kernels reflecting the spectrum of possible design constraints, and these kernels are not interchangeable. Severe artifacts can be introduced by any one of them in inappropriate situations. Very careful examination is required to avoid being victimized by the particular boundary kernel chosen. Artifacts can unfortunately be introduced by the choice of the support interval for the boundary kernel. Little is known about the best way to avoid this situation, but the Rice-Müller

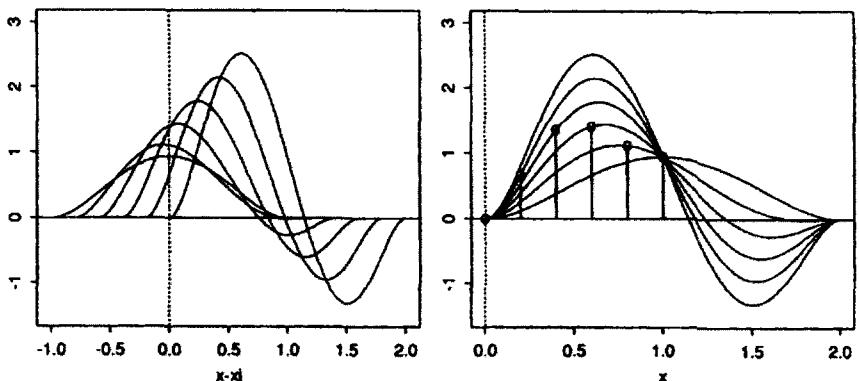


Figure 6.12 Examples of “zero” boundary kernels as in Figure 6.10.

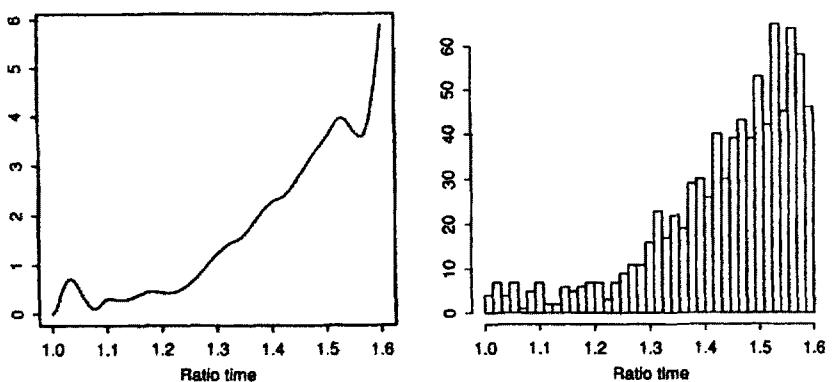


Figure 6.13 Density estimate of 857 fastest times in the 1991 Houston Tenneco Marathon. The data are the ratio to the leader's time for the race. Different boundary kernels were used on each extreme. A histogram is shown for comparison.

solution seems the best of several possible alternatives that have been attempted. Finally, while the boundary kernels seem to cover a wide region of the density estimate, the effect is generally limited to the interval $(0, h/2)$ for appropriately smoothed estimates.

Some data may require a different boundary kernel at each end. For example, the fastest 857 times in the January 20, 1991, Houston Tenneco Marathon were recorded as a ratio to the winning time. Clearly, the features in the density are expected to differ at the two boundaries. An estimate was constructed using K_c^0 on the left and the floating K_c on the right with $h = 0.05$; see Figure 6.13. The clump among the leaders is real; however, the extra bump on the right appears to be more of an artifact.

Reflection Boundary Technique

There is a more conservative technique that can replace the “floating” kernel. If the data are nonnegative and the discontinuity is at $x = 0$, an ordinary kernel estimate is computed but on the augmented data $(-x_n, \dots, -x_1, x_1, \dots, x_r)$. The final estimate is obtained by doubling this estimate for $x \geq 0$. The bandwidth should be based on the sample size n and not $2n$. This technique avoids the pitfalls of negative boundary kernels, but is generally of lower-order consistency; see Problem 4 in Chapter 4 in the context of the frequency polygon.

6.3 THEORETICAL PROPERTIES: MULTIVARIATE CASE

The theoretical analysis of multivariate kernel estimators is the same as for frequency polygons save for a few details. The initial discussion will be limited to product kernel density estimators. The general kernel analysis will be considered afterwards.

6.3.1 Product Kernels

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^n \left\{ \prod_{j=1}^d K\left(\frac{x_i - x_{ij}}{h_j}\right) \right\}. \quad (6.37)$$

The same (univariate) kernel is used in each dimension but with a different smoothing parameter for each dimension. The data x_{ij} come from a $n \times d$ matrix. The estimate is defined pointwise, where $\mathbf{x} = (x_1, \dots, x_n)^T$. Geometrically, the estimate places a probability mass of size $1/n$ centered on each sample point, exactly as in the univariate case. Recall that the limiting form of the naive multivariate ASH is a product triangle kernel estimator. Several bivariate product kernels are displayed in Figure 6.14.

Consider the pointwise bias of the multivariate estimator. Clearly,

$$\begin{aligned} E\hat{f}(\mathbf{x}) &= E \prod_{j=1}^d \frac{1}{h_j} K\left(\frac{x_j - X_j}{h_j}\right) = \int_{\mathbb{R}^d} \prod_{j=1}^d \frac{1}{h_j} K\left(\frac{x_j - t_j}{h_j}\right) f(\mathbf{t}) d\mathbf{t} \\ &= \int_{\mathbb{R}^d} \prod_{j=1}^d K(w_j) f(x_1 - h_1 w_1, \dots, x_n - h_n w_n) d\mathbf{w} \\ &\approx \int_{\mathbb{R}^d} \prod_{j=1}^d K(w_j) \left[f(\mathbf{x}) - \sum_{r=1}^d h_r w_r f_r(\mathbf{x}) + \sum_{r,s=1}^d \frac{h_r h_s}{2} w_r w_s f_{rs}(\mathbf{x}) \right] d\mathbf{w} \\ &= f(\mathbf{x}) + \frac{1}{2} \sigma_K^2 \sum_{j=1}^d h_j^2 f_{jj}(\mathbf{x}) + O(h^4). \end{aligned} \quad (6.38)$$

As before, the $O(h)$ bias terms vanish if the univariate kernels have zero mean. Similarly, the $h_r h_s$ terms vanish; see Problem 23. It follows that the integrated squared bias is as given in the theorem below. The pointwise variance is $f(x)R(K)^d/(nh_1 h_2 \cdots h_n)$, from which the integrated variance follows easily.

Theorem 6.4: *For a multivariate product kernel estimator, the components of the AMISE are*

$$\begin{aligned} \text{AISB} &= \frac{1}{4} \sigma_K^4 \left[\sum_{i=1}^d h_i^4 R(f_{ii}) + \sum_{i \neq j} h_i^2 h_j^2 \int_{\mathbb{R}^d} f_{ii} f_{jj} d\mathbf{x} \right] \\ \text{AIV} &= \frac{R(K)^d}{nh_1 h_2 \cdots h_d} - \frac{R(f)}{n} + O\left(\frac{h}{n}\right). \end{aligned} \quad (6.39)$$

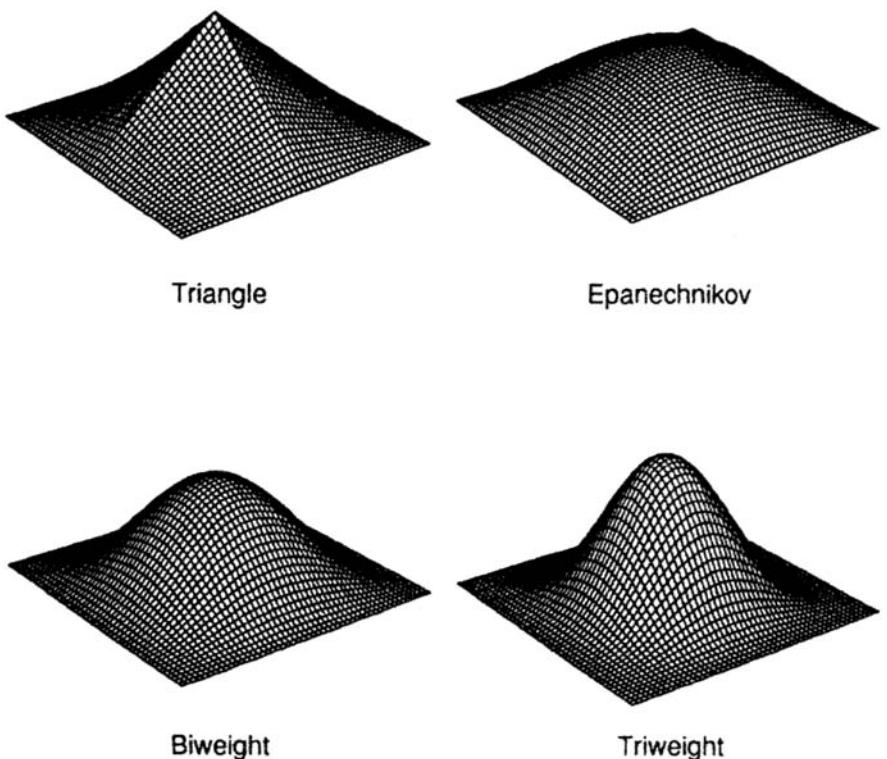


Figure 6.14 Product kernel examples for 4 kernels.

The order of the optimal smoothing parameters is precisely the same as for the multivariate FP: $h_i^* = O(n^{-1/(4+d)})$ and $\text{AMISE}^* = O(n^{-4/(4+d)})$.

It is a minor inconvenience, but no general closed-form expression for the optimal smoothing parameters exists, save as the solution to d nonlinear equations. Solutions may be found in special cases, for example, when $d \leq 2$ or if $h_i = h$ for all i . For example, with general bivariate Normal data and a Normal kernel, straightforward integration shows that

$$\begin{aligned} R(f_{11}) &= 3 \left[16 \pi (1 - \rho^2)^{5/2} \sigma_1^5 \sigma_2 \right]^{-1}, \\ R(f_{22}) &= 3 \left[16 \pi (1 - \rho^2)^{5/2} \sigma_1 \sigma_2^5 \right]^{-1}, \\ \int_{\mathbb{R}^2} f_{11} f_{22} dx_1 dx_2 &= (1 + 2\rho^2) \left[16 \pi (1 - \rho^2)^{5/2} \sigma_1^3 \sigma_2^3 \right]^{-1}. \end{aligned}$$

From Theorem 6.4, the AMISE is minimized when

$$\begin{aligned} h_i^* &= \sigma_i (1 - \rho^2)^{5/12} (1 + \rho^2/2)^{-1/6} n^{-1/6} \\ &\approx \sigma_i (1 - \rho^2/2 - \rho^4/16 - \dots) n^{-1/6} \quad i = 1, 2, \end{aligned} \quad (6.40)$$

for which

$$\text{AMISE}^* = \frac{3}{8\pi} (\sigma_1 \sigma_2)^{-1} (1 - \rho^2)^{-5/6} (1 + \rho^2/2)^{1/3} n^{-2/3}.$$

Observe that the AMISE diverges to infinity when the data are perfectly correlated (the *real* curse of dimensionality). In comparison to other bivariate estimates, if $\rho = 0$ and $\sigma_i = 0$, then the bivariate AMISE is equal to 1/400 when $n = 302$. A bivariate FP and histogram require $n = 557$ and $n = 4,244$, respectively.

A second example of a special case is the multivariate Normal, where all the variables are independent. If a Normal kernel is used, then a short calculation with Theorem 6.4 gives the

Normal reference rule:
$$h_i^* = \left(\frac{4}{d+2} \right)^{1/(d+4)} \sigma_i n^{-1/(d+4)}. \quad (6.41)$$

As the dimension d varies, the constant in Equation (6.41) ranges over the interval (0.924, 1.059), with a limit equal to 1. The constant is exactly 1 in the bivariate case and smallest when $d = 11$. Hence, an easy-to-remember data-based rule is

Scott's rule in \mathbb{R}^d :
$$\hat{h}_i = \hat{\sigma}_i n^{-1/(d+4)}. \quad (6.42)$$

For other kernels, the equivalent kernel smoothing parameter may be obtained by dividing by the standard deviation of that kernel. Just as in one dimension, these formulas can be used in place of more precise oversmoothing values as independent normal data are very smooth. Any special structure will require narrower bandwidths. For example, the modification based on skewness and kurtosis in \mathbb{R}^1 are identical to the factors for the frequency polygon in Section 4.1.2. If the data are not full-rank, kernel methods perform poorly. Dimension reduction techniques will be considered in Chapter 7.

6.3.2 General Multivariate Kernel MISE

In practice, product kernels are recommended. However, for various theoretical studies, general multivariate kernels will be required. This section presents a brief summary of those studies.

The general multivariate kernel estimator will include not only an arbitrary multivariate density as a kernel but also an arbitrary linear transformation of the data. Let H be a $d \times d$ nonsingular matrix and $K : \mathbb{R}^d \rightarrow \mathbb{R}^+$ be a kernel satisfying conditions given below. Then the general multivariate kernel estimator is

$$\hat{f}(\mathbf{x}) = \frac{1}{n|H|} \sum_{i=1}^n K(H^{-1}(\mathbf{x} - \mathbf{x}_i)). \quad (6.43)$$

It should be apparent from Equation (6.43) that the linear transformation H could be incorporated into the kernel definition. For example, it is equivalent to choose K to be $N(\mathbf{0}, \Sigma)$ with $H = I_d$, or to choose K to be $N(\mathbf{0}, I_d)$ with $H = \Sigma^{1/2}$; see Problem 24. Thus it is possible to choose a multivariate kernel with a simple covariance structure without loss of generality. It will not, however, be sufficient to consider only product kernels, as that would limit the discussion to multivariate kernels that are independent (and not just uncorrelated) and to kernels that are supported on a rectangular region.

The multivariate kernel will be assumed hereafter to satisfy three moment conditions (note these are matrix equations):

$$\begin{aligned} \int_{\mathbb{R}^d} K(\mathbf{w}) d\mathbf{w} &= 1 \\ \int_{\mathbb{R}^d} \mathbf{w} K(\mathbf{w}) d\mathbf{w} &= \mathbf{0} \\ \int_{\mathbb{R}^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w}) d\mathbf{w} &= I_d. \end{aligned} \quad (6.44)$$

If K is indeed a multivariate probability density, then the last two equations summarize many assumptions about the *marginal kernels*, $\{K_i(w_i), i = 1, \dots, d\}$. The second equation says that the means of the marginal kernels are all zero. The third equation states that the marginal kernels are all pairwise uncorrelated and that each has unit variance. Thus any simple linear transformation is assumed to be captured entirely in the matrix H and not in the kernel.

In matrix notation, it is straightforward to compute the error of the multivariate kernel estimator. For letting $\mathbf{w} = H^{-1}(\mathbf{x} - \mathbf{y})$,

$$\begin{aligned} E\hat{f}(\mathbf{x}) &= \int_{\mathbb{R}^d} K(H^{-1}(\mathbf{x} - \mathbf{y})) f(\mathbf{y}) d\mathbf{y} / |H| \\ &= \int_{\mathbb{R}^d} K(\mathbf{w}) f(\mathbf{x} - H\mathbf{w}) d\mathbf{w} \\ &= \int_{\mathbb{R}^d} K(\mathbf{w}) \left[f(\mathbf{x}) - \mathbf{w}^T H \nabla f(\mathbf{x}) + \frac{1}{2} \mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H \mathbf{w} \right] d\mathbf{w} \end{aligned} \quad (6.45)$$

to second order. Further simplification is possible using the following property of the trace (tr) of a matrix: $\text{tr}\{AB\} = \text{tr}\{BA\}$, assuming that the matrices A and

B have dimensions $r \times s$ and $s \times r$, respectively. Now the quadratic form in Equation (6.45) is a 1×1 matrix, which trivially equals its trace. Hence, using the trace identity and exchanging the trace and integral operations yields

$$\mathbb{E}\hat{f}(\mathbf{x}) = f(\mathbf{x}) - 0 + \frac{1}{2} \operatorname{tr} \left\{ \int_{\mathbb{R}^d} \mathbf{w} \mathbf{w}^T K(\mathbf{w}) d\mathbf{w} \cdot H^T \nabla^2 f(\mathbf{x}) H \right\}.$$

As the covariance matrix of K is I_d by assumption (6.44), the integral factor in the trace vanishes. Therefore,

$$\operatorname{Bias}\{\hat{f}(\mathbf{x})\} = \frac{1}{2} \operatorname{tr}\{H^T \nabla^2 f(\mathbf{x}) H\} = \frac{1}{2} \operatorname{tr}\{HH^T \nabla^2 f(\mathbf{x})\}.$$

Next, define the scalar $h > 0$ and the $d \times d$ matrix A to satisfy

$$H = hA \quad \text{where } |A| = 1.$$

Choosing A to have determinant equal to 1 means that the elliptical shape of the kernel is entirely controlled by the matrix AA^T and the size of the kernel is entirely controlled by the scalar h . Observe that this parameterization is entirely general and permits different smoothing parameters for each dimension. For example, if

$$H = \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_d \end{pmatrix}; \quad \text{then} \quad H = h \cdot \begin{pmatrix} h_1/h & & 0 \\ & \ddots & \\ 0 & & h_d/h \end{pmatrix},$$

where $h = (h_1 h_2 \cdots h_d)^{1/d}$ is the geometric mean of the d smoothing parameters. Check that $|A| = 1$.

It follows that

$$\operatorname{Bias}\{\hat{f}(\mathbf{x})\} = \frac{1}{2} h^2 \operatorname{tr}\{AA^T \nabla^2 f(\mathbf{x})\}, \quad (6.46)$$

so that

$$\text{AISB} = \frac{1}{4} h^4 \int_{\mathbb{R}^d} [\operatorname{tr}\{AA^T \nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x}.$$

As usual, the variance term is dominated by $\mathbb{E}K_H(\mathbf{x} - \mathbf{x}_i)^2$; therefore,

$$\operatorname{Var}\{\hat{f}(\mathbf{x})\} = \frac{f(\mathbf{x})}{n|H|} \int_{\mathbb{R}^d} K(\mathbf{w})^2 d\mathbf{w} \quad \Rightarrow \quad \text{AIV} = \frac{R(K)}{nh^d}. \quad (6.47)$$

Together, these results may be summarized in a theorem.

Theorem 6.5: *For a general multivariate kernel estimator (6.43) parameterized by $H = hA$,*

$$\text{AMISE} = \frac{R(K)}{nh^d} + \frac{1}{4} h^4 \int_{\mathbb{R}^d} [\text{tr}\{AA^T \nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x}. \quad (6.48)$$

In spite of appearances, this is not using the same bandwidth in each dimension, but rather is applying a general elliptically shaped kernel at an arbitrary rotation.

6.3.3 Boundary Kernels for Irregular Regions

Staniswalis, Messer, and Finston (1990) showed that a boundary kernel for an arbitrarily complicated domain may be constructed by a simple device. Suppose that an estimate at \mathbf{x} is desired. They propose using a kernel with spherical support, with radius h . Only samples \mathbf{x}_i in the sphere of radius h around \mathbf{x} influence the estimate. For each sample \mathbf{x}_i in that region, determine if the diameter on which it falls (the center being the estimation point \mathbf{x}) intersects the boundary. If it does, construct a 1-dimensional boundary kernel along that diameter. Repeating this construction for all samples in the sphere, the authors prove that the resulting estimate retains the correct order of bias.

6.4 GENERALITY OF THE KERNEL METHOD

6.4.1 Delta Methods

Walter and Blum (1979) catalogued the common feature of the already growing list of different density estimators. Namely, each could be reexpressed as a kernel estimator. Such a demonstration for orthogonal series estimators was given in Section 6.1.3. Reexamining Equation (3.2), even the histogram can be thought of as a kernel estimator. Surprisingly, this result was shown to hold even for estimators that were solutions to optimization problems. For example, consider one of the several maximum penalized likelihood (MPL) criteria suggested by Good and Gaskins (1972):

$$\max_f \sum_{i=1}^n \log f(x_i) - \alpha \int_{-\infty}^{\infty} f'(x)^2 dx \quad \text{for some } \alpha > 0. \quad (6.49)$$

Without the *roughness penalty term* in (6.49), the solution would be the empirical pdf. The many MPL estimators were shown to be kernel estimators by de Montricher, Tapia, and Thompson (1975) and Klonias (1982). The form

of the kernel solutions differs in that the weights on the kernels were not all equal to $1/n$. For some other density estimation algorithms, the equivalent kernel has weight $1/n$ but has an adaptive bandwidth. A simple example of this type is the k th nearest-neighbor (NN) estimator. The k -NN estimate at x is equivalent to a histogram estimate with a bin centered on x with bin width sufficiently large so that the bin contains k points (in 2 and 3 dimensions, the histogram bin shape is a circle and a sphere, respectively). Thus the equivalent kernel in \mathbb{R}^d is simply a uniform density over the unit ball in \mathbb{R}^d , but with bin widths that adapt to x .

6.4.2 General Kernel Theorem

There is a theoretical basis for the empirical observations of Walter and Blum that many algorithms may be viewed as generalized kernel estimates. Terrell provided a theorem to this effect that contains a constructive algorithm for obtaining the generalized kernel of any density estimator; see Terrell and Scott (1992). The construction is not limited to nonparametric estimators, a fact that is exploited below.

Theorem 6.6. *Any density estimator that is a continuous and Gâteaux differentiable functional on the empirical distribution function may be written as*

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K(x, x_i, F_n), \quad (6.50)$$

where K is the Gâteaux derivative of \hat{f} under variation of x_i .

The Gâteaux derivative of a functional T at the function ϕ in the direction of the function η is defined to be

$$DT(\phi)[\eta] = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} [T(\phi + \epsilon \eta) - T(\phi)]. \quad (6.51)$$

Theorem 6.6, which is proved below, has an analogous multivariate version (Terrell and Scott, 1992). The kernel K simply measures the influence of x_i on $\hat{f}(x)$. As F_n converges to F , then asymptotically, the form of K is independent of the remaining $n - 1$ observations. Thus, any continuous density estimator may be written (asymptotically) as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K(x, x_i, n) = \frac{1}{n} \sum_{i=1}^n K_n(x, x_i). \quad (6.52)$$

6.4.2.1 Proof of General Kernel Result

The empirical cdf in Equation (2.1) can be written in the unusual form

$$F_n(\cdot) = \frac{1}{n} \sum_{i=1}^n I_{[x_i, \infty)}(\cdot). \quad (6.53)$$

Write the density estimator as an operator $\hat{f}(x) = T_x\{F_n\}$. Define

$$\begin{aligned} K(x, y, F_n) &\equiv \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} [T_x\{(1 - \epsilon)F_n + \epsilon I_{[y, \infty)}\} - (1 - \epsilon)T_x\{F_n\}] \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} [T_x\{F_n + \epsilon(I_{[y, \infty)} - F_n)\} - T_x\{F_n\}] + T_x\{F_n\} \\ &= DT_x(F_n)[I_{[y, \infty)} - F_n] + \hat{f}(x), \end{aligned} \quad (6.54)$$

where $DT(\phi)[\eta]$ is the Gâteaux derivative of T at ϕ in the direction η . Proposition (2.7) of Tapia (1971) shows that the Gâteaux derivative is linear in its second argument, so

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n K(x, x_i, F_n) &= \frac{1}{n} \sum_{i=1}^n DT_x(F_n)[I_{[x_i, \infty)} - F_n] + \hat{f}(x) \\ &= DT_x(F_n) \left[\frac{1}{n} \sum_{i=1}^n I_{[x_i, \infty)} - F_n \right] + \hat{f}(x) \\ &= 0 + \hat{f}(x). \end{aligned}$$

Note that by linearity, the Gâteaux variation in the direction 0, $DT(\phi)[0]$, is identically 0. This concludes the proof.

6.4.2.2 When Is an Estimator Nonparametric?

An estimator is defined to be nonparametric when it is consistent in the mean square for a large class of density functions. With a little effort, this definition translates into specific requirements for the equivalent kernel, $K_n(x, y)$. From the many previous examples, a nonparametric estimator that is consistent must be *local*; that is, the influence of sample points outside an ϵ -neighborhood of x must vanish as $n \rightarrow \infty$. As suggested by the term “delta function sequences” coined by Watson and Leadbetter (1963),

$$\lim_{n \rightarrow \infty} K_n(x, x_i, F_n) = \delta(x - x_i).$$

But K_n must not equal $\delta(x - x_i)$ for finite n , as was the case in Section 6.1.3 with the orthogonal series estimator with all \hat{f}_ν coefficients included.

Beginning with the bias of the asymptotic kernel form in (6.52)

$$\begin{aligned} E \hat{f}(x) &= E K_n(x, X) = \int K_n(x, y) f(y) dy \\ &= \int K_n(x, y) [f(x) + (y - x)f'(x) + (y - x)^2 f''(\xi_y)/2] dy \end{aligned}$$

by the exact version of Taylor's theorem where $\xi_y \in (x, y)$. To be asymptotically unbiased, all 3 terms in the following must vanish:

$$\begin{aligned} \text{Bias}\{\hat{f}(x)\} &= f(x) \left[\int K_n(x, y) dy - 1 \right] + f'(x) \int K_n(x, y) (y - x) dy \\ &\quad + \frac{1}{2} \int K_n(x, y) (y - x)^2 f''(\xi_y) dy. \end{aligned} \quad (6.55)$$

Therefore, the first condition is that

$$\lim_{n \rightarrow \infty} \int K_n(x, y) dy = 1 \quad \forall x \in \mathbb{R}^1.$$

Assume that the estimator has been rescaled so that the integral is exactly 1 for all n . Therefore, define the random variable Y to have pdf $K_n(x, \cdot)$. The second condition is that

$$\begin{aligned} \lim_{n \rightarrow \infty} \int K_n(x, y) y dy &= \int K_n(x, y) x dy = x \quad \forall x \\ \implies \lim_{n \rightarrow \infty} K_n(x, y) &= \delta(y - x) \end{aligned}$$

as Watson and Leadbetter (1963) and Walter and Blum (1979) suggested. The precise behavior of the bias is determined by the rate at which this happens. For example, suppose that $\int K_n(x, y) y dy = x$ for all n and that

$$\sigma_{x,n}^2 = \int K_n(x, y) (y - x)^2 dy \neq 0 \quad (6.56)$$

for finite n so that the first two moments of the random variables $Y \sim (x, \sigma_{x,n}^2)$ and $T \equiv (Y - x)/(\sigma_{x,n}) \sim (0, 1)$. Suppose that the density function of T , which is a simple linear transformation of K_n , converges to a nice density:

$$\tilde{L}_n(x, t) = K_n(x, x + \sigma_{x,n} t) \sigma_{x,n} \rightarrow L(x, t) \quad \text{as } n \rightarrow \infty.$$

Then the last bias term in (6.55) may be approximated by

$$\begin{aligned} \frac{1}{2} f''(x) \int K_n(x, y) (y - x)^2 dy \\ = \frac{1}{2} f''(x) \int K_n(x, x + \sigma_{x,n} t) (\sigma_{x,n} t)^2 \sigma_{x,n} dt \\ = \frac{1}{2} \sigma_{x,n}^2 f''(x) \int t^2 \tilde{L}(x, t) dt \\ \approx \frac{1}{2} \sigma_{x,n}^2 f''(x) \int t^2 L(x, t) dt = \frac{1}{2} \sigma_{x,n}^2 f''(x) \end{aligned}$$

since the $\text{Var}(T) = 1$, so that the bias is $O(\sigma_{x,n}^2)$, the familiar rate for second-order kernels. Thus the third condition is that $\sigma_{x,n} \rightarrow 0$ as $n \rightarrow \infty$.

In order that the variance vanish asymptotically, consider

$$\begin{aligned} \text{Var}\{\hat{f}(x)\} &= \frac{1}{n} \text{Var}\{K_n(x, X)\} \leq \frac{1}{n} E[K_n(x, X)^2] = \frac{1}{n} \int K_n(x, y)^2 f(y) dy \\ &= \frac{1}{n} \int K_n(x, x + \sigma_{x,n} t)^2 f(x + \sigma_{x,n} t) \sigma_{x,n} dt \\ &= \frac{1}{n \sigma_{x,n}} \int \tilde{L}_n(x, t)^2 [f(x) + \dots] dt \approx \frac{f(x) R[L(x, \cdot)]}{n \sigma_{x,n}}. \end{aligned}$$

Thus the fourth condition required is that the variance of the equivalent kernel satisfy $n\sigma_{x,n} \rightarrow \infty$ as $n \rightarrow \infty$.

These findings may be summarized in a theorem, the outline of which is presented in Terrell (1984).

Theorem 6.7: Suppose \hat{f} is a density estimator with asymptotic equivalent kernel $K_n(x, y)$ and that $\sigma_{x,n}^2$ defined in (6.56) is bounded and nonzero. Then \hat{f} is a nonparametric density estimator if, for all $x \in \mathfrak{R}^1$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \int K_n(x, y) dy &= 1 \\ \lim_{n \rightarrow \infty} \int K_n(x, y) y dy &= x \\ \lim_{n \rightarrow \infty} \sigma_{x,n} &= 0 \\ \lim_{n \rightarrow \infty} n \sigma_{x,n} &= \infty. \end{aligned} \tag{6.57}$$

6.4.2.3 Equivalent Kernels of Parametric Estimators

Theorem 6.6 shows how to construct the kernel for any density estimator, parametric or nonparametric. For example, consider the parametric estimation

of $f = N(\mu, 1) = \phi(x; \mu, 1)$ by $\hat{\mu} = \bar{x}$. Thus $T_x(F_n) = \phi(x; \bar{x}, 1)$. Examining the argument in the first line in Equation (6.54) and comparing it to the definition of the ecdf in (6.53), it becomes apparent that the empirical pdf $n^{-1} \sum \delta(x - x_i)$ is being replaced by

$$\frac{1 - \epsilon}{n} \sum_{i=1}^n \delta(x - x_i) + \epsilon \delta(x - y),$$

which is the original empirical pdf with a small portion ϵ of the probability mass proportionally removed and placed at $x = y$. The sample mean of this perturbed epdf is $(1 - \epsilon)\bar{x} + \epsilon y$. Thus the kernel may be computed directly from Equation (6.54) by

$$\begin{aligned} K(x, y, F_n) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} [\phi\{x; (1 - \epsilon)\bar{x} + \epsilon y, 1\} - (1 - \epsilon)\phi\{x; \bar{x}, 1\}] \\ &= \frac{1 + (y - \bar{x})(x - \bar{x})}{\sqrt{2\pi}} e^{-\frac{1}{2}(x - \bar{x})^2}; \end{aligned} \quad (6.58)$$

see Problem 27. The asymptotic equivalent kernel is

$$K(x, y) = \lim_{n \rightarrow \infty} K(x, y, F_n) = \frac{1 + (y - \mu)(x - \mu)}{\sqrt{2\pi}} e^{-\frac{1}{2}(x - \mu)^2}.$$

This kernel is never *local* and so the estimator is *not* nonparametric (if there was any doubt). Note that the *parametric kernel estimator* with kernel (6.58) is quite good, as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1 + (x_i - \bar{x})(x - \bar{x})}{\sqrt{2\pi}} e^{-(x - \bar{x})^2/2} = \frac{1}{\sqrt{2\pi}} e^{-(x - \bar{x})^2/2}.$$

Of course, just as in the parametric setting, this “kernel” estimator will always be $\phi(x; \bar{x}, 1)$ for all data sets, no matter how non-Normal the underlying density is.

6.5 CROSS-VALIDATION

6.5.1 Univariate Data

The goal is to go beyond the theoretical results for optimal bandwidth specification and to achieve practical data-based algorithms. The oversmoothed and Normal reference rules provide reasonable initial choices, together with the simple modifications based on sample skewness and kurtosis given in Section 4.1.2. (The bandwidth modification factors for positive kernels and the frequency polygon are identical). The unbiased and biased cross-validation

algorithms for the histogram are easily extended to both kernel and averaged shifted histogram estimators. However, the development of bandwidth selection algorithms for kernel estimators has progressed beyond those for the histogram.

The importance of choosing the optimal bandwidth is easily overstated. From the sensitivity results in Table 3.3, any choice of h within 15–20% of h^* will often suffice. Terrell (1990) has even suggested that an oversmoothed rule be generally used. With real data, it is relatively easy to examine a sequence of estimates based on the sequence of smoothing parameters

$$h = \hat{h}_{OS}/1.05^k \quad \text{for } k = 0, 1, 2, \dots,$$

starting with the sample oversmoothed bandwidth (\hat{h}_{OS}), and stopping when the estimate displays some instability and very local noise near the peaks. Silverman (1978a) has characterized the expected amount of local noise present in $\hat{f}''(x)$ in the L_∞ norm; see Equation (6.62). He suggested examining plots of $\hat{f}''(x)$ interactively in addition to $\hat{f}(x)$, a procedure referred to as the *test graph method*. The biased cross-validation algorithm, which estimates $R(\hat{f}'')$ from a kernel estimate, attempts to correct the overestimation by $R(\hat{f}''(\cdot))$ that is the direct result of presence of the local noise; see Section 6.5.1.3. However, the power of the interactive approach to bandwidth selection should not be underestimated. The interactive process can be quite painless in systems supporting animation, such as LISP-STAT (Tierney, 1990).

It should be emphasized that from an exploratory point of view, all choices of the bandwidth h lead to useful density estimates. Large bandwidths provide a picture of the global structure in the unknown density, including general features such as skewness, outliers, clusters, and location. Small bandwidths, on the other hand, reveal local structure which may or may not be present in the true density. Furthermore, the optimality of h is dependent not only on the choice of metric L_p , but also on the feature in the density to be emphasized (F, f, f', f'', \dots).

However, the desire for fully automatic and reliable bandwidth selection procedures has led inevitably to a series of novel algorithms. These objective (not subjective) procedures often have the stated goal of finding a bandwidth that minimizes the actual L_2 error rather than using the bandwidth that minimizes the expected L_2 error (MISE). In an early paper, Wahba (1981) expressed the expectation that her generalized cross-validation algorithm would accomplish this goal. The best L_2 bandwidth, h_{ISE} , remained the target in unbiased cross-validation for Hall and Marron (1987a,b). Scott and Factor (1981) had expressed the view that h_{MISE} was an appropriate target. The MISE-optimal bandwidth depends only on (f_n) , whereas the ISE-optimal bandwidth depends on the sample as well, $(f, x, \{x_i\})$. However, it has been shown that the sample correlation between h_{ISE} and σ approached -0.70 for Normal data (Scott and Terrell, 1987; Scott, 1988b). Given such a large negative correlation with the scale of the data, tracking \hat{h}_{ISE} closely would require guessing whether $\hat{\sigma} > \sigma$, or vice versa, a difficult task.

Of some interest is the fact that while $\hat{h}_{\text{ISE}} \approx h_{\text{MISE}}$,

$$\frac{\sigma_{\hat{h}_{\text{ISE}}}}{h_{\text{MISE}}} = O(n^{-1/10}),$$

so that the ISE-optimal bandwidth is only slowly converging to h_{MISE} , as was shown by Hall and Marron (1987a). Scott and Terrell (1987) showed that UCV and BCV bandwidths converged to h_{MISE} at the same slow rate. Some of the more recent extensions have been able to push the relative convergence rate all the way to $O(n^{-1/2})$, which is the best rate possible (Hall and Marron, 1989). These procedures require the introduction of 1 or 2 auxiliary smoothing parameters. Given the slow rate of convergence of the unattainable \hat{h}_{ISE} , it is perhaps unclear whether there is a practical advantage to be had in the faster rates.

In a recent empirical study of the performance of 9 cross-validation algorithms and 4 sampling densities, Jones and Kappenman (1992) report the “broad equivalence of almost all” of these algorithms with respect to the observed ISE. Other simulations (Scott and Factor, 1981; Bowman, 1985, Scott and Terrell, 1987; and Park and Marron, 1990) have reported less equivalence among the estimated smoothing parameter values themselves. Jones and Kappenman reported that the fixed AMISE bandwidth h^* outperformed all 9 CV algorithms with respect to ISE. Their results reinforce the suggestion that h^* is an appropriate target and that any choice within 15–20% of h^* should be adequate. Most efforts now focus on h_{MISE} as the target bandwidth.

6.5.1.1 Early Efforts in Bandwidth Selection

The earliest data-based bandwidth selection ideas came in the context of orthogonal series estimators, which were discussed in Section 6.1.3. Using the Tarter-Kronmal weights (6.8) and the representation in Equation (6.5), the pointwise error is

$$\hat{f}(x) - f(x) = \sum_{\nu=-m}^m \hat{f}_\nu \phi(x) - \sum_{\nu=-\infty}^{\infty} f_\nu \phi(x).$$

As the basis functions are orthonormal,

$$\text{ISE} = \sum_{\nu=-m}^m \|\hat{f}_\nu - f_\nu\|^2 + \sum_{\nu \notin [-m, m]} \|f_\nu\|^2.$$

Recall that $\text{MISE} = E(\text{ISE})$. Tarter and Kronmal’s selection procedure provided unbiased estimates of the increment in MISE in going from a series with $m - 1$ terms to one with m terms (noting the equality of the $\pm \nu$ MISE terms):

$$\text{MISE}(m) - \text{MISE}(m - 1) = 2\{E\|\hat{f}_m - f_m\|^2 - \|f_m\|^2\}. \quad (6.59)$$

Unbiased estimates of the two terms on the right-hand side may be obtained for the Fourier estimator in Equation (6.6), as $E \hat{f}_\nu = f_\nu$ and $\text{Var}(\hat{f}_\nu) E \hat{f}_\nu \hat{f}_\nu^* = (1 - |\hat{f}_\nu|^2)/n$, where \hat{f}_ν^* denotes the complex conjugate of \hat{f}_ν . The data-based choice for m is achieved when the increment becomes *positive*. Notice that accepting the inclusion of the m th coefficient in the series estimator is the result of the judgment that the additional variance of \hat{f}_m is less than the reduction in bias $\|\hat{f}_m\|^2$. Usually, fewer than 6 terms are chosen, so that only relatively coarse adjustments can be made to the smoothness of the density estimator. Sometimes the algorithm misses higher-order terms. But the real significance of this algorithm lies in its claim to be the first unbiased cross-validation algorithm for a density estimator.

Likewise, the credit for the first biased cross-validation algorithm goes to Wahba (1981) with her generalized cross-validation algorithm. She used the same unbiased estimates of the Fourier coefficients as Tarter and Kronmal, but with her smoother choice of weights, she lost the simplicity of examining the incremental changes in MISE. However, those same unbiased estimates of the Fourier coefficients lead to a good estimate of the AMISE. By ignoring all the unknown Fourier coefficients for $|\nu| > n/2$, a small bias is introduced. Both groups recommend plotting the estimated risk function in order to find the best data-based smoothing parameter rather than resorting to (blind) numerical optimization.

The earliest effort at choosing the kernel smoothing parameter in a fully automatic manner was a modified maximum likelihood algorithm due to Habbema, Hermans, and Van Der Broek (1974) and Duin (1976). While it has not withstood the test of time, it is significant for having introduced a leave-one-out modification to the usual maximum likelihood criterion. Choosing the bandwidth h to maximize the usual maximum-likelihood criterion results in the (rough) empirical pdf:

$$0 = \arg \max_h \sum_{i=1}^n \log \hat{f}(x_i; h) \quad \Rightarrow \quad \hat{f}(x; h = 0) = \frac{1}{n} \sum_{i=1}^n \delta(x - x_i),$$

a solution with “infinite” likelihood. The problem arises since, as $h \rightarrow 0$, the contribution to the likelihood at $x = x_i$ from the point x_i itself becomes infinite. The authors sought to eliminate that “self-contribution” by modifying the ML criterion:

$$\max_h \sum_{i=1}^n \log \hat{f}_{-i}(x_i; h),$$

where $\hat{f}_{-i}(x_i; h)$ is a kernel estimator based on the $n - 1$ data points excluding x_i and then evaluated there. In spite of some promising empirical small sample and consistency results (Chow, Geman, and Wu, 1983), the algorithm was shown to be overly influenced by outliers and tight clusters (Scott and Factor, 1981; Schuster and Gregory, 1981). With a finite support kernel, for example,

the bandwidth cannot be less than $x_{(2)} - x_{(1)}$, which is the distance between the first 2 order statistics; for many densities the distance between these order statistics does not converge to 0 and so the bandwidth does not converge to 0 as required.

A simple fixed-point algorithm was suggested by Scott, Tapia, and Thompson (1977). For kernel estimates, the only unknown in h^* in (6.16) is $R(\hat{f}''_h)$. If a Normal kernel is used, then a straightforward calculation finds that

$$R(\hat{f}''_h) = \frac{3}{8\sqrt{\pi} n^2 h^5} \sum_{i=1}^n \sum_{j=1}^n \left(1 - \Delta_{ij}^2 + \frac{1}{12} \Delta_{ij}^4 \right) e^{-\frac{1}{4}\Delta_{ij}^2}, \quad (6.60)$$

where $\Delta_{ij} = (x_i - x_j)/h$. Following Equation (6.16), the search for a fixed-point value for h^* is achieved by iterating

$$h_{k+1} = \left[\frac{R(K)}{n\sigma_K^4 R(\hat{f}''_{h_k})} \right]^{1/5},$$

with h_0 chosen to be the Normal reference bandwidth. As the ratio of the optimal bandwidths for estimating f and f'' diverges as $n \rightarrow \infty$, it is clear that the algorithm is not consistent. That the algorithm worked well for small samples (Scott and Factor, 1981) is not surprising since the optimal bandwidths are reasonably close to each other for small samples. Notice that this algorithm as stated provides no estimate of the MISE. It is a simple matter to use the roughness estimate (6.60) in Equation (6.16), following the lead of Wahba, to obtain

$$\widehat{\text{AMISE}}(h) = \frac{R(K)}{nh} + \frac{1}{4} \sigma_K^4 h^4 R(\hat{f}''_h). \quad (6.61)$$

Finding the minimizer of Equation (6.61) not only provides a data-based bandwidth estimate, but also an estimate of the MISE. This idea was resurrected with biased cross-validation using a consistent estimator for $R(f'')$ rather than Equation (6.60). Alternatively, a much wider bandwidth appropriate to f'' rather than f might be used in the iteration. Sheather (1983) proposed such a scheme while trying to estimate the density at the origin, providing an example of the plug-in estimators discussed by Woodroffe (1970). Sheather's motivation is discussed in Sheather and Jones (1991). In a talk in 1991, Gasser also reported success in this way for choosing a global bandwidth by inflating the bandwidth used in $R(\hat{f}''_h)$ by the factor $n^{-1/10}$.

For a Normal kernel, Silverman (1978a) proved that

$$\frac{\sup |\hat{f}'' - E\hat{f}''|}{\sup |E\hat{f}''|} \approx 0.4. \quad (6.62)$$

He proposed choosing the bandwidth where it appears that the ratio of the noise to the signal is 0.4. This ratio is different for other kernels. The *test graph* procedure can be used in the bivariate case as well.

6.5.1.2 Oversmoothing

The derivation of the oversmoothed rule for kernel estimators will be constructive, unlike the earlier derivations of the histogram and frequency polygon oversmoothed rules. The preferred version uses the variance as a measure of scale. Other scale measures have been considered by Scott and Terrell (1987) and Terrell (1990).

Consider the variational problem

$$\min_f \int_{-\infty}^{\infty} f''(x)^2 dx \quad \text{s/t} \quad \int f = 1 \text{ and } \int x^2 f = 1. \quad (6.63)$$

Clearly, the solution will be symmetric. The associated Lagrangian is

$$L(f) = \int_{-\infty}^{\infty} f''(x)^2 dx + \lambda_1 \left(\int f - 1 \right) + \lambda_2 \left(\int x^2 f - 1 \right).$$

At a solution, the Gâteaux variation, defined in Equation (6.51), of the Lagrangian in any "direction" η must vanish. For example, the Gâteaux variation of $\Psi(f) = \int f''(x)^2$ is

$$\begin{aligned} \Psi'(f)[\eta] &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left[\int [(f + \epsilon \eta'')^2] - \int [f'']^2 \right] \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left[\int 2\epsilon f'' \eta'' + \epsilon^2 \eta''^2 \right] = \int 2f'' \eta''. \end{aligned}$$

Computing the Gâteaux variation of $L(f)$, we have

$$\begin{aligned} 0 &= L'(f)[\eta] = \int 2f''(x)\eta''(x) + \lambda_1 \int \eta(x) + \lambda_2 \int x^2 \eta(x) \\ &= 2f''(x)\eta'(x) \Big|_{-c}^c - 2f'''(x)\eta(x) \Big|_{-c}^c + \int [2f^{iv}(x) + \lambda_1 + \lambda_2 x^2]\eta(x) \end{aligned} \quad (6.64)$$

after integrating by parts twice and where c is the boundary, possibly infinite, for f . Now $\eta(\pm c)$ must vanish so that there is not a discontinuity in the solution; therefore, the second term vanishes. The remaining 2 terms must vanish for all feasible choices for η ; therefore, $f''(\pm c)$ must vanish, leaving only the integral. It follows that the integrand must vanish and that f is a sixth-order polynomial with only even powers (by symmetry). Therefore, the solution must take the form

$$f(x) = a(x - c)^3(x + c)^3$$

so that $f''(\pm c) = 0$. The 2 constraints in (6.63) impose 2 linear conditions on the unknowns (a, c) , with the result that

$$f^*(x) = \frac{35}{69,984} (9 - x^2)_+^3 \quad \text{and} \quad R[(f^*)''] = \frac{35}{243}.$$

A simple change of variables shows that $R(f'') \geq 35/(243\sigma^5)$; therefore,

$$h^* = \left[\frac{R(K)}{n\sigma_K^4 R(f'')} \right]^{1/5} \leq \left[\frac{243\sigma^5 R(K)}{35n\sigma_K^4} \right]^{1/5} \implies \quad (6.65)$$

Oversmoothing rule: $h_{OS} = 3 \left[\frac{R(K)}{35\sigma_K^4} \right]^{1/5} \sigma n^{-1/5}$.

(6.66)

For the Normal kernel, $h_{OS} = 1.144 \sigma n^{-1/5}$. For the biweight kernel, it is exactly $h_{OS} = 3 \sigma n^{-1/5}$. The rule is exactly 1.08 times wider than the Normal reference rule.

6.5.1.3 Unbiased and Biased Cross-Validation

The presentation in the section will rely heavily on the references for certain details. The overall flavor is the same as in the application to the histogram and frequency polygon.

The remarkable fact is that the UCV justification for the histogram is entirely general. Thus the definition in (3.51) applies in the case of a kernel estimator. For the case of a Normal kernel, Rudemo (1982) and Bowman (1984) showed that (replacing $n \pm 1$ with n for simplicity)

$$\text{UCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2h\sqrt{\pi}} \sum_{i < j} (e^{-\Delta_{ij}^2/14} - \sqrt{8}e^{-\Delta_{ij}^2/2}), \quad (6.67)$$

which is a special case of the general formula (Scott and Terrell, 1987)

$$\text{UCV}(h) = \frac{R(K)}{nh} + \frac{2}{n^2h} \sum_{i < j} \gamma(\Delta_{ij}),$$

where

$$\gamma(\Delta) = \int K(w)K(w + \Delta) dw.$$

The BCV algorithm follows from the result (Scott and Terrell, 1987) that

$$ER(\hat{f}_h'') = R(f'') + \frac{R(K'')}{nh^5} + O(h^2),$$

where $R(K'')/(nh^5)$ is asymptotically a constant, representing the fixed but finite noise that exists in the kernel estimate. Therefore, $R(\hat{f}_h'') - R(K'')/(nh^5)$ is an asymptotically unbiased estimator for the unknown roughness $R(f'')$. Substituting into Equation (6.16), the estimate of the MISE becomes

$$\text{BCV}(h) = \frac{R(K)}{nh} + \frac{\sigma_K^4}{2n^2h} \sum_{i < j} \phi(\Delta_{ij}), \quad (6.68)$$

where

$$\phi(\Delta) = \int K''(w)K''(w + \Delta) dw.$$

The similarity of the general UCV and BCV formulas is remarkable, given their quite different origins. In the case of a Normal kernel,

$$\text{BCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2h\sqrt{\pi}} \sum_{i < j} (\Delta_{ij}^4 - 12\Delta_{ij}^2 + 12)e^{-\Delta_{ij}/4}. \quad (6.69)$$

As before, $\lim_{h \rightarrow \infty} \text{BCV}(h) = 0$; therefore, \hat{h}_{BCV} is taken to be the largest local minimizer of $\text{BCV}(h)$ less than or equal to the oversmoothed bandwidth.

The asymptotic theory of the CV criteria is straightforward once it is recognized that the stochastic part is all contained in the so-called U -statistics, which are double sums of the form

$$U_n = \sum_{i < j} H_n(X_i, X_j).$$

Hall (1984) proved that if the function H_n is symmetric and the random variable $E[H_n(X, Y)|X] = 0$, then together with a certain moment condition,

$$U_n = \text{AN}\left(0, \frac{1}{2} n^2 E H_n^2\right).$$

The asymptotic Normality of UCV and BCV is not obvious because the number of terms in the double sum effectively shrinks, as the bandwidth is a decreasing function of n . Hall used an argument from degenerate Martingale theory to prove this theorem.

Scott and Terrell (1987) then proved that for a fixed bandwidth, h , the UCV and BCV functions were (1) both asymptotically Normal; (2) both converged to AMISE(h); and (3) the asymptotic (vertical) variances of UCV and BCV at h are

$$\frac{2R(\gamma)R(f)}{n^2h} \quad \text{and} \quad \frac{\sigma_k^8 R(\phi)R(f)}{8n^2h}, \quad (6.70)$$

respectively. For kernels in the symmetric Beta family of the form (6.25), the (vertical) variance of UCV is at least 80 times greater than for the BCV criterion. This smaller vertical variance suggests that the actual minimizer of the BCV criterion will have smaller variance than \hat{h}_{UCV} . The variances in Equation (6.70) are of order $O(n^{-9/5})$ if $h = O(n^{-1/5})$. However, this rapid rate of decrease can be explained by the fact that the CV functions are themselves going to 0 at the rate $O(n^{-4/5})$. Thus the relevant quantity is the coefficient of variation (C.V.)

$$C.V. = \frac{\sqrt{\text{Var UCV}(h)}}{O[\text{UCV}(h)]} = \frac{O(n^{-9/10})}{O(n^{-4/5})} = O(n^{-1/10}),$$

as was claimed earlier. The astute reader will note that the bias has not been counted in this error in the BCV case; however, the bias turns out to be $O(n^{-1})$, and hence the squared bias is $O(n^{-2})$, which is of lower order than the variance.

Using a delta argument outlined below, Scott and Terrell (1987) showed that \hat{h}_{UCV} and \hat{h}_{BCV} converged to h_{AMISE} and were asymptotically Normal with respective variances given by

$$\frac{2R(f)R[\Delta\gamma'(\Delta)]}{25n^2(h^*)^7\sigma_k^4R(f'')^2} \quad \text{and} \quad \frac{R(f)R[\Delta\phi'(\Delta)]}{200n^2(h^*)^7R(f'')^2}. \quad (6.71)$$

Observe that if $h^* = O(n^{-1/5})$, then these variances are $O(n^{-3/5})$, from which the (horizontal) coefficient of variation is found to still be of $O(n^{-1/10})$:

$$C.V. = \frac{\sqrt{\text{Var } \hat{h}_{UCV}}}{O(\hat{h}_{UCV})} = \frac{O(n^{-3/10})}{O(n^{-1/5})} = O(n^{-1/10}).$$

Again, for densities in the symmetric Beta family, the UCV variance is at least 16 times that of the BCV. These results were confirmed in a simulation study. However, it was noted that BCV performed poorly for several difficult densities without a very large data set. This finding is not surprising, given that the basis for the BCV formula is AMISE, while the exact MISE is the basis of UCV.

It is instructive to outline the derivation of Equation (6.71). Clearly, the BCV bandwidth satisfies

$$\frac{d}{dh} [\text{BCV}(h)] \Big|_{h=\hat{h}_{BCV}} = 0.$$

Noting that $\Delta'_{ij} = -(x_i - x_j)/h^2 = -\Delta_{ij}/h$, the derivative of BCV as defined in (6.68) equals

$$\frac{-R(K)}{nh^2} - \frac{\sigma_K^4}{2n^2h^2} \sum_{i < j} \phi(\Delta_{ij}) + \frac{\sigma_K^4}{2n^2h} \sum_{i < j} \phi'(\Delta_{ij}) \frac{-\Delta_{ij}}{h} = 0$$

or

$$\sum_{i < j} [\phi(\Delta_{ij}) + \Delta_{ij} \phi'(\Delta_{ij})] \Big|_{h=\hat{h}_{BCV}} = -\frac{2nR(K)}{\sigma_K^4}.$$

Define $\psi(\Delta) = \Delta \phi'(\Delta)$; then computing approximations to the moments of ϕ and ψ [a nontrivial calculation given in Section 9 of Scott and Terrell (1987)], it follows that

$$\sum_{i < j} [\phi(\Delta_{ij}) + \psi(\Delta_{ij})] = AN\left(-2n^2h^5R(f''), \frac{1}{2} n^2hR(f)R(\psi)\right).$$

Hence, combining these two results and rearranging,

$$-2n^2R(f'')\hat{h}_{BCV}^5 = AN(-2nR(K)/\sigma_K^4, n^2h^*R(f)R(\psi)/2)$$

or

$$\hat{h}_{BCV}^5 = AN\left(\frac{R(K)}{\sigma_K^4 n R(f'')}, \frac{h^* R(f) R(\psi)}{8n^2 R(f'')^2}\right). \quad (6.72)$$

Clearly, the asymptotic mean of \hat{h}_{BCV}^* is $(h^*)^5$. The random variable \hat{h}_{BCV} is the 1/5 power of that given in (6.72). Applying the delta method, it may be concluded that \hat{h}_{BCV} is AN with mean h^* and variance which may be computed by the formula

$$\text{Var}\{g(h)\} = \left(\frac{dg}{dh}\right)_{h=h^*}^2 \text{Var}\{h\}.$$

Now $g(h) = h^5$ and $g'(h) = 5h^4$, so that

$$\text{Var}\{\hat{h}_{BCV}\} = \text{Var}\{\hat{h}_{BCV}^5\}/[25(h^*)^8]. \quad (6.73)$$

The variance (6.71) follows immediately combining (6.72) and (6.73).

In spite of these apparently favorable findings, BCV does not qualify as a general replacement for UCV. UCV may be noisier but it tends to produce nearly unbiased smoothing parameters. However, there is a need for an auxiliary CV criterion since UCV is susceptible to certain problems. Clearly, BCV has its own set of limitations. But by carefully examining of the trio of smoothing parameters suggested by UCV, BCV, and OS as well as the shapes of the UCV and BCV curves, good bandwidths should be reliably available.

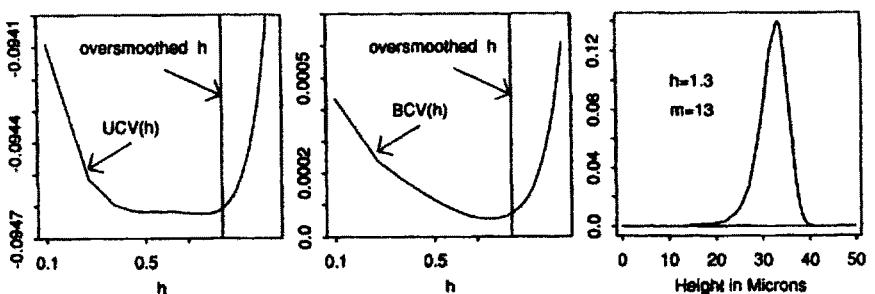


Figure 6.15 UCV and BCV estimates of the steel surface data ($n = 15,000$) using the ASH triweight kernel. The BCV and UCV bandwidths were 1.3 and 1.2, respectively, and gave virtually identical estimates.

As a practical matter, both UCV and BCV involve $O(n^2)$ computation due to the double sums. If the Normal kernel versions are used, the work can easily be prohibitive for $n > 500$. This work can be substantially reduced by using an ASH implementation. The computational details are given in Scott and Terrell (1987) and are not repeated here. Of course, code is available from several sources to perform these and other computations. The ASH implementation is illustrated in Figure 6.15 for the steel surface data. The standard deviation of these data is 3.513, so that the oversmoothed bandwidth for the triweight kernel is $\hat{h}_{os} = 1.749$ from Equation (6.66). Observe that the UCV estimate is flat over a relatively wide interval even with such a large data set. However, the minima of the two criteria are virtually identical, $h_{BCV} = 1.2$ and $h_{UCV} = 1.3$ for the triweight kernel. (The triweight estimate with equivalent smoothing parameter $h = 0.67$ is shown in Figure 6.5.) The original data were binned into 500 intervals, so that the use of the ASH implementation of the CV and estimation functions is natural.

6.5.1.4 Bootstrapping Cross-Validation

Taylor (1989) investigated a data-based algorithm for choosing h based on bootstrap estimates of the $MSE\{\hat{f}(x)\}$ and $MISE\{\hat{f}\}$. The bootstrap resample is not taken from the empirical pdf f_n as in the *ordinary bootstrap*, but rather the bootstrap sample $\{x_1^*, x_2^*, \dots, x_n^*\}$ is a random sample from the candidate kernel density estimate $\hat{f}(x; h)$ itself. Such a resample is called a *smoothed bootstrap sample* and is discussed further in Chapter 9. Letting E_* represent expectation with respect to the smoothed bootstrap random sample, Taylor examined

$$\begin{aligned} BMSE_*(x; h) &= E_*[\hat{f}_*(x; h) - \hat{f}(x; h)]^2 \\ &= E_*\left[\frac{1}{n} \sum_{i=1}^n K_h(x - x_i^*) - \frac{1}{n} \sum_{i=1}^n K_h(x - x_i)\right]^2 \quad (6.74) \\ BMISE_*(h) &= \int_x BMSE_*(x; h) dx . \end{aligned}$$

The interesting result is that if the resample comes from the empirical density function, then the bootstrap quantities in Equation (6.74) estimate only the variance and not the bias. The bias introduced by the smoothed bootstrap is precisely what is needed to mimic the true unknown bias for that choice of the smoothing parameter h . A bit of extra variance is introduced by the smoothed bootstrap, but it is easily removed.

The algebra involved in computing the BMSE_* is perhaps unfamiliar but straightforward. For example, $E_* \hat{f}(x) = E_* K_h(x - x^*)$ and

$$E_* K_h(x - x^*) = \int K_h(x - y) \hat{f}(y) dy = \frac{1}{n} \sum_{i=1}^n \int K_h(x - y) K_h(y - x_i) dy.$$

The computation of $E_* K_h(x - x^*)^2$ is a trivial extension. For the particular case of the Normal kernel, the convolutions indicated in the bootstrap expectation may be computed in closed form. After a few pages of work using the Normal kernel and adjusting the variance, Taylor proposes minimizing

$$\begin{aligned} \text{BMISE}_*(h) &= \frac{1}{2nh\sqrt{\pi}} \left[1 + \frac{\sqrt{2}}{n} \sum_{i < j} \left(\sqrt{2} e^{-\Delta_{ij}^2/4} - \frac{4}{\sqrt{3}} e^{-\Delta_{ij}^2/6} + e^{-\Delta_{ij}^2/8} \right) \right]. \\ &\quad (6.75) \end{aligned}$$

Taylor shows that $\text{BMISE}_*(h)$ has the same order variance as $\text{UCV}(h)$ and $\text{BCV}(h)$, but with a smaller constant asymptotically.

In the Figure 6.16, the 3 CV functions, with comparable vertical scalings, are plotted for the snowfall data along with the corresponding density estimates using the Normal kernel. Many commonly observed empirical results are depicted in these graphs. The BCV and bootstrap curves are similar, although $\text{BCV}(h)$ has a sharper minimum. Both the biased and bootstrap CV functions have minima at bandwidths greater than the oversmoothed bandwidth $h_{\text{OS}} = 11.9$. This serves to emphasize how difficult estimating the bias is with small samples. The unbiased CV function better reflects the difficulty in precisely estimating the bias by presenting a curve that is flat over a wide interval near the minimum. There is some visual evidence of 3 modes in this data. Since these CV functions do not take account of the time series nature of these data and the first-order autocorrelation is -0.6 , a smaller bandwidth is probably justifiable; see Chiu (1989) and Altman (1990).

These computations are repeated for the Old Faithful geyser data set, which is clearly bimodal, and the results depicted in Figure 6.17. It is interesting to note that both the biased and bootstrap CV functions have two local

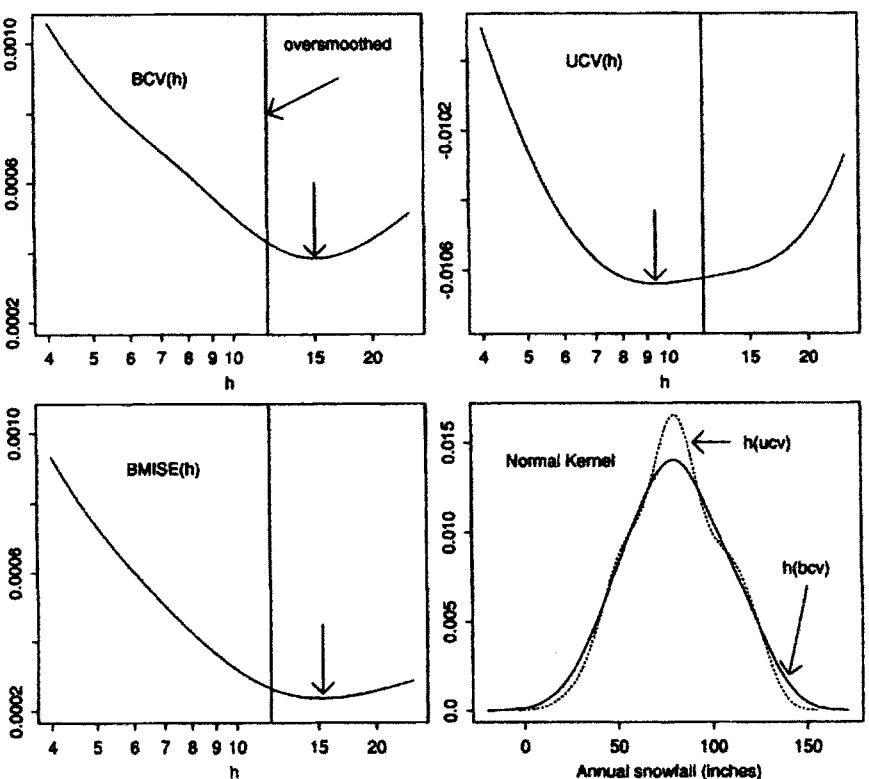


Figure 6.16 Normal kernel cross-validation algorithms and density estimates for the snowfall data ($n = 63$). The CV bandwidths are indicated by arrows and the oversmoothed bandwidth by a solid line. The UCV and BCV density estimates are represented by the dashed and solid lines, respectively.

minima. Fortunately, only one local minima is smaller than the oversmoothed upper bound $h_{OS} = 0.47$, although the bootstrap curve barely exhibits the second local minimum. The UCV function leads to a narrow bandwidth and a density estimate that seems clearly undersmoothed given the sample size. These observations seem to recur with many “real” data sets with modest sample sizes. Concordance among a subset of these rather differently behaving criteria should be taken seriously.

6.5.1.5 Faster Rates and Plug-In Cross-Validation

In a series of papers, several authors have worked to improve the relatively slow $O(n^{-1/10})$ convergence of CV bandwidth algorithms. Along the way to the best $O(n^{-1/2})$ rate, algorithms were proposed with such interesting rates as $O(n^{-4/13})$. All share some features: for example, all have U -statistic formulas similar to UCV and BCV (Jones and Kappenman, 1992). All strive to improve estimates of quantities such as $R(f'')$ and $R(f''')$. The improvements come from

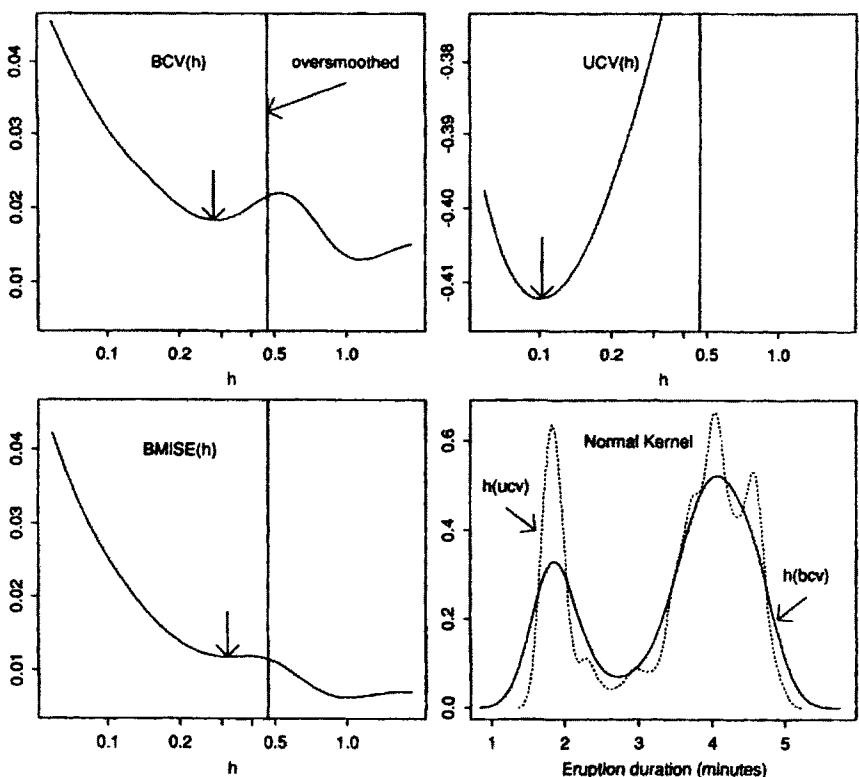


Figure 6.17 Normal kernel cross-validation algorithms and density estimates for the geyser data ($n = 107$). The CV bandwidths are indicated by arrows and the oversmoothed bandwidth by a solid line. The UCV and BCV density estimates are represented by the dashed and solid lines, respectively.

expected sources—the use of higher-order kernels, for example. Some recent examples include Chiu (1991) and Sheather and Jones (1991). For purposes of illustration, the discussion is limited to one particular $O(n^{-1/2})$ plug-in algorithm due to Hall, Sheather, Jones, and Marron (1991). They discovered that simply improving the estimation of $R(f'')$ was not sufficient; rather, a more accurate approximation for the AMISE is required (see Problem 9):

$$\text{AMISE}(h) = \frac{R(K)}{nh} - \frac{R(f)}{n} + \frac{1}{4} h^4 \mu_2^2 R(f'') - \frac{1}{24} h^6 \mu_2 \mu_4 R(f'''),$$

which is accurate to $O(n^{-7/5})$. The second term, which is constant, may be ignored. The two unknown roughness functionals are estimated not as in BCV,

but rather with two auxiliary smoothing parameters, λ_1 and λ_2 , so that

$$\widehat{\text{AMISE}}(h) = \frac{R(K)}{nh} + \frac{1}{4} h^4 \mu_2^2 \hat{R}_{\lambda_1}(f'') - \frac{1}{24} h^6 \mu_2 \mu_4 \hat{R}_{\lambda_2}(f'''). \quad (6.76)$$

Since there is no simple formula linking the 3 smoothing parameters in this formula, the authors propose selecting λ_1 and λ_2 based on a robust version of the Normal reference rule. Several practical observations may be made. First, since the selection of the two auxiliary bandwidths is a onetime choice, the total computational effort is much less than for the BCV or UCV approaches. Second, this new approximation to the AMISE diverges to $-\infty$ as $h \rightarrow \infty$; hence, the plug-in (PI) rule is also looking for a local minimizer less than the oversmoothed bandwidth. However, given the simple form of Equation (6.76), it is relatively easy to provide a closed-form, asymptotic approximation to its (local) minimizer (see Problem 31):

$$\hat{h}_{\text{PI}} = \left[\frac{\hat{j}_1}{n} \right]^{\frac{1}{5}} + \left[\frac{\hat{j}_1}{n} \right]^{\frac{3}{5}} \hat{j}_2; \quad \hat{j}_1 = \frac{R(K)}{\mu_2^2 \hat{R}_{\lambda_1}(f'')}, \quad \hat{j}_2 = \frac{\mu_4 \hat{R}_{\lambda_2}(f''')}{\mu_2 \hat{R}_{\lambda_1}(f'')}. \quad (6.77)$$

A portion of the details of a particular implementation given in the authors' paper is outlined in the next paragraph. The plug-in bandwidths computed in this fashion indeed have rapidly vanishing noise.

The estimates of $R(f'')$ and $R(f''')$ follow from the identities

$$\begin{aligned} \int f''(x)^2 &= + \int f^{iv}(x)f(x) \\ &= + E f^{iv}(X) \leftarrow \frac{2}{n(n-1)\lambda_1^5} \sum_{i < j} K^{iv}\left(\frac{x_i - x_j}{\lambda_1}\right) \\ \int f'''(x)^2 &= - \int f^{vi}(x)f(x) \\ &= - E f^{vi}(X) \leftarrow \frac{2}{n(n-1)\lambda_2^7} \sum_{i < j} \phi^{vi}\left(\frac{x_i - x_j}{\lambda_2}\right). \end{aligned}$$

These 2 estimates are suggested by the UCV estimator. The authors chose a particular order-4 polynomial kernel for $K(x)$, which is supported on the

interval $(-1, 1)$, whose fourth derivative is equal to

$$K^{iv}(x) = \frac{135,135}{4,096} (1 - x^2)(46,189x^8 - 80,036x^6 + 42,814x^4 - 7,236x^2 + 189).$$

Of course, for the Normal kernel, $\phi^{vi}(x) = (x^6 - 15x^4 + 45x^2 - 15)\phi(x)$. Using Normal reference rules and the interquartile range (IQR) as the measure of scale, the formulas for the auxiliary smoothing parameters are calculated to be

$$\hat{\lambda}_1 = 4.29 \text{ IQR } n^{-1/11} \quad \text{and} \quad \hat{\lambda}_2 = 0.91 \text{ IQR } n^{-1/9}.$$

While the rate of convergence of the algorithm is impressive, the bias occasionally seems to be troublesome and is due in part to the sensitivity of the plug-in bandwidth to the choice of the 2 auxiliary bandwidths. This effect may be illustrated by multiplying the true IQR for a data set by a factor. If the estimated roughness is insensitive to the perturbed value of the IQR, then the exact choice for λ_i should not matter. However, the empirical evidence does not support this hypothesis. For example, Figure 6.18 illustrates the sensitivity for a random sample of 200 Normal data points, for which $R(f'') = 3/(8\sqrt{\pi})$ and $R(f''') = 15/(16\sqrt{\pi})$. Clearly, the roughness estimates follow changes in the IQR in an exponential manner. As the IQR varies by a factor of 9 [that is, is multiplied by factors between $(1/3, 3)$], the estimate of $I_2 = \hat{R}(f'')$ varies by a factor of 195.7, and the estimate of $I_3 = \hat{R}(f''')$ varies by a factor of 28,650.0. Thus the auxiliary smoothing problem merits further research.

For small samples, there may be no minimizer of the plug-in AMISE estimate given in Equation (6.76); see Figure 6.19. The lack of a local minimum is a feature observed for small samples with the BCV criterion. It would appear that for small samples, the plug-in formula (6.77) is essentially returning a version of the Normal reference rule rather than a true minimizer of the risk

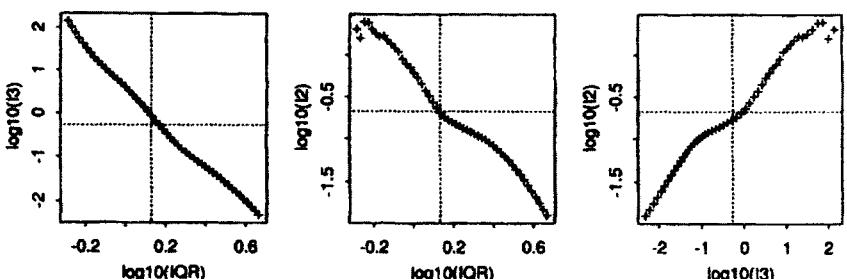


Figure 6.18 Changes in the roughness estimates as a function of changes in the auxiliary smoothing parameters around their default settings. The data are a random sample of 200 Normal points. The true values are indicated by dotted lines.

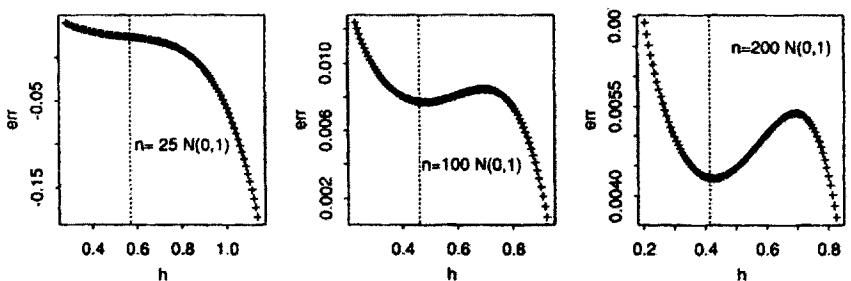


Figure 6.19 Examples of AMISE approximation of plug-in rule with $N(0, 1)$ data and a Normal kernel. The plug-in bandwidth is shown by a vertical dotted line.

function. As the sample size increases, the “strength” of the plug-in estimate grows as the bowl shape of the AMISE estimate widens.

The plug-in AMISE function estimates for the Buffalo snowfall data ($n = 63$) and the Old Faithful eruption duration data ($n = 107$) are shown in Figure 6.20. Neither estimate has a local minimum, although the plug-in formula gives reasonable results. This illustrates the danger inherent in not plotting the risk function. It also suggests that the excellent small sample behavior of plug-in estimators involves subtle factors. In any case, the greater the number of reasonable bandwidth algorithms available to attack a data set, the better.

For a very large data set, such as the steel surface data, the plug-in, BCV, and UCV bandwidths are often identical. The plug-in risk curve is shown in Figure 6.20. The value of $\hat{h}_{PI} = 0.424$ for the Gaussian kernel. The conversion factor to the triweight kernel ($\sigma_K = 1/3$) is 3 or $\hat{h} = 1.27$, which is identical to the UCV and BCV predictions in Figure 6.15. The plug-in AMISE curve for the bimodal geyser data has no local minima. The plug-in bandwidth is $\hat{h}_{PI} = 0.472$ (or $3 \times 0.472 = 1.42$ on the triweight scale). Thus the plug-in bandwidth is well beyond the oversmoothed bandwidth and misses the local bandwidth found by the BCV and BMISE curves in Figure 6.17. Improved

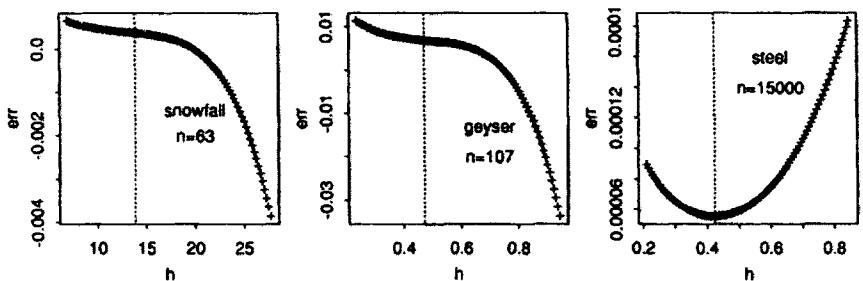


Figure 6.20 Plug-in cross-validation curves for the snowfall data ($n = 63$), the geyser data ($n = 107$) and the steel surface data ($n = 15,000$) for the Normal kernel. The plug-in bandwidth obtained by formula (6.77) is indicated by the dotted line.

plug-in algorithms are under development, using generalized scale measures appropriate for multimodal data, and rapid changes and new successes can be expected. In any case, the experienced worker can expect to be able to judge the success or failure of any cross-validation bandwidth readily with modern interactive computing.

6.5.1.6 Constrained Oversmoothing

Oversmoothing has been presented strictly as a means of bounding the unknown smoothing parameters by choosing a measure of scale, such as the sample range or standard deviation, for the data. It is easy to find cases where the oversmoothed bandwidths are much too wide. If the sampling density is in fact the oversmoothed density, then the bandwidth \hat{h}_{OS} is not strictly an upper bound, as \hat{h}_{OS} varies about h^* . In most instances, the 2 bandwidths will be within a few percentage points of each other.

The variational problems considered in oversmoothing can be generalized to provide much more relevant bandwidths in almost every situation. The basic idea is to add constraints in addition to the one measure of overall scale. The new proposal for constrained oversmoothing (CO) is to require that several of the percentiles in the oversmoothed density match the sample percentiles in the data. Specifically,

$$f_{CO}^* = \arg \min_f \int f''(x)^2 dx \quad s/t \quad \int_{-\infty}^{\alpha_i} f(x) dx = F_n(\alpha_i), \quad i \in [1, k],$$

where $k \geq 2$. In other words, the cdf of the oversmoothed distribution should match the empirical cdf in several intermediate locations. This problem has already been solved in 2 instances when $k = 2$, once with the range constraint and again with the interquartile range constraint. The new suggestion is to choose to match the 10th, 30th, 50th, 70th, and 90th sample percentiles, for example. The resulting constrained oversmoothed density may or may not be close to the true density, but computing the roughness of the CO density provides a significantly improved *practical* estimate for use in the usual asymptotic bandwidth formula. With f_{CO}^* in hand, the constrained oversmoothed smoothing parameter is found as

$$\hat{h}_{CO} = \left[\frac{R(K)}{n\sigma_K^4 R[(f_{CO}^*)'']^2} \right]^{-1/5}.$$

In Figure 6.21 for a sample of 1,000 $N(0, 1)$ points, several possible solutions to the variational problem are displayed along with the computed roughness. The location of the constraints is indicated by the dashed lines. The solution is a quartic spline. The relevant quantity is the fifth root of the ratio of that roughness to the true roughness, $R(\phi'') = 3/(8\sqrt{\pi}) = 0.2115$. The first CO solution is based on matching the 1%, 50%, and 99% sample percentiles. Now $(0.2115/0.085)^{1/5} = 1.20$, so that the constrained oversmoothed bandwidth is

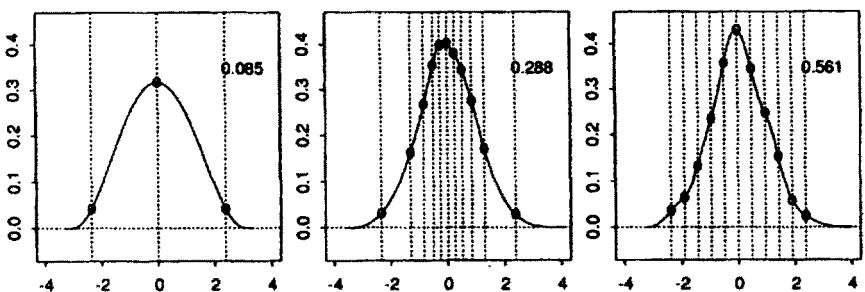


Figure 6.21 Constrained oversmoothed density solutions for a $N(0, 1)$ sample of 1,000 points. The true roughness is 0.212.

1.2 times wider than h^* . Thus $h^* < h_{CO}$ as usual. However, by including more constraints, this inequality will not be strictly observed. For example, two solutions were found with 11 constraints. One used the 1%, 10%, 20%, ..., 90%, and 99% sample percentiles. Since the roughness of this solution is slightly *greater* than the true roughness, the CO bandwidth will be *smaller* than h^* . In fact, $h_{CO} = 0.94 h^*$. A second problem with 11 constraints used a fixed width mesh as shown. The sample percentiles matched in this mesh were 1.0%, 3.4%, 7.8%, 16.6%, 30.6%, 49.9%, 68.9%, 82.7%, 92.5%, 97.3%, and 99.0%. For this solution, $h_{CO} = 0.82 h^*$. Matching 5 to 10 percentiles would seem adequate in practice. For example, the percentile mesh with 7 constraints (not shown) has a roughness of 0.336, which corresponds to $h_{CO} = 0.91 h^*$.

This procedure was repeated for a sample of 1,000 points from a mixture density, $\frac{3}{4}\phi(x; 0, 1) + \frac{1}{4}\phi(x; 2, 1/9)$; see Figure 6.22. The true roughness is 3.225. Solutions for a fixed bin width and 2 percentile meshes are shown. For each solution $h^* < h_{CO}$ as the roughness of each solution is less than the true roughness. The difference is greatest with the solution in the last frame, for which $h_{CO} = 1.25 h^*$. The ordinary oversmoothed rule, which is based on the variance 55/36, leads to a lower bound of 0.0499 for the roughness and $hos = 2.30 h^*$ from Equation (6.65). Thus the constrained oversmoothed procedure gives a conservative bandwidth, but not so conservative. The solution in the middle frame indicates how the constrained oversmoothed density solution may

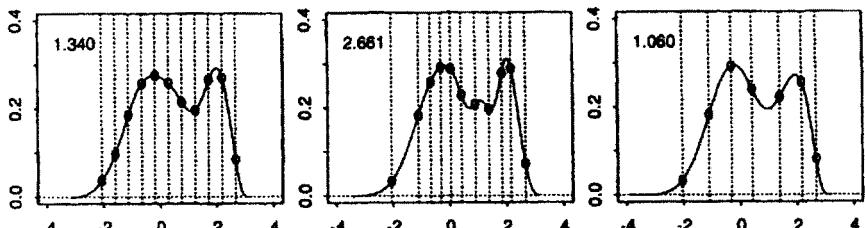


Figure 6.22 Constrained oversmoothed density solutions for a sample of 1,000 points from the mixture density $0.75N(0, 1) + 0.25N(0, 1/9)$. The true roughness is 3.225.

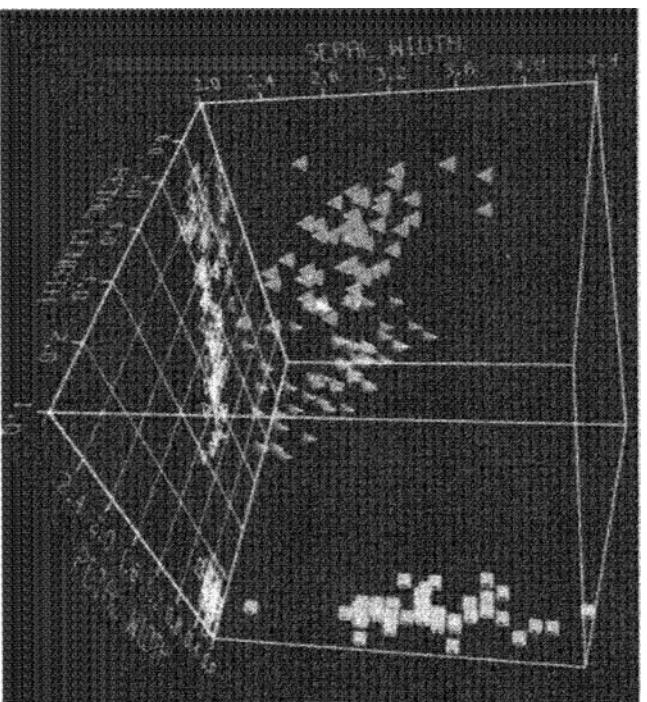


Plate 1. Three-dimensional scatter diagram of the Fisher-Anderson iris data, omitting the sepal length variable. From left to right, the 50 points for each of the three varieties of Virginica, Versicolor, and Setosa are distinguished by symbol type and color. The coloring is required to indicate the presence of three clusters rather than only two. The same basic picture results from any choice of three variables from the full set of four variables. Composed by Dr. Stanley Grotz of Lawrence Livermore Labs and used with permission. See Section 1.3.1.

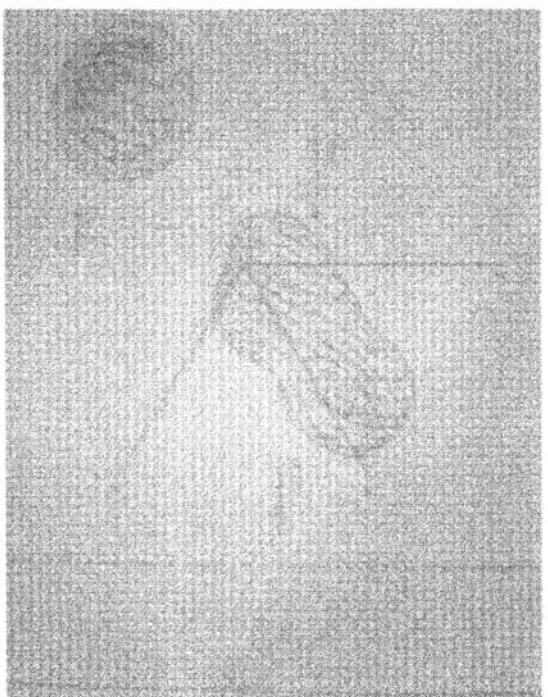


Plate 2. Two α -level contour surfaces from a slice of a four-dimensional averaged shifted histogram estimate, based on all 150 iris data points. The displayed variables x , y , and z are sepal length, petal width and length, respectively, with the sepal width variable sliced at $t = 3.4$ cm. The blue $\alpha = 5\%$ contour reveals only two clusters, while the red $\alpha = 20\%$ contour reveals the three clusters. Adapted from Scott (1996), p. 241 by courtesy of Marcel Dekker, Inc. See Sections 1.4.3 and 1.4.4.

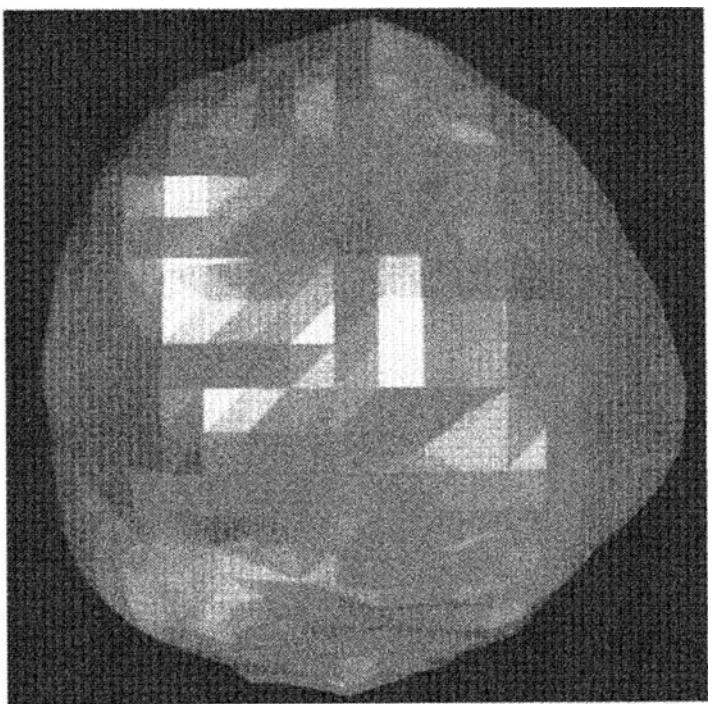


Plate 3. A closeup view of a single $c = 10\%$ level contour surface from an averaged shifted histogram estimate based upon 1,000 pseudo-random points from a standard trivariate Normal density. Observe the many triangular patches from which the surface is composed. Also, the underlying trivariate bin structure of the ASH is apparent. See Section 1.4.3.

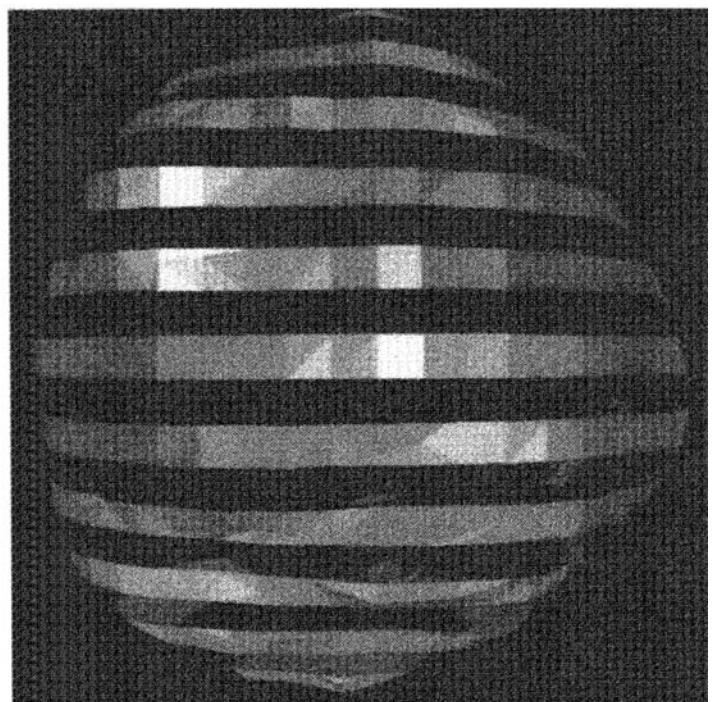


Plate 4. The contour surface in Plate 3 with the triangular patches omitted in alternating horizontal bins. The resulting contour "ribbons" give a stronger three-dimensional impression than the representation in Plate 3. This contour surface representation choice is related to that used in Plate 2. See Section 1.4.3.

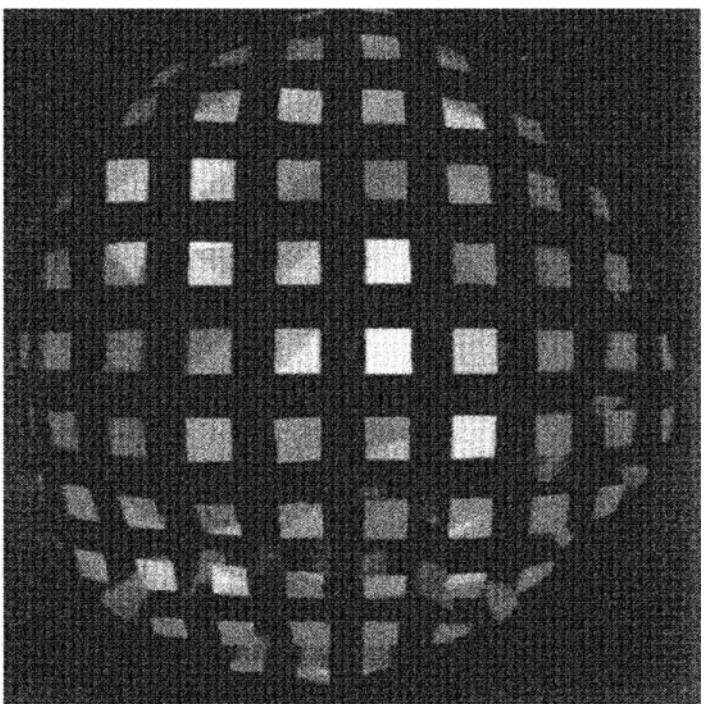


Plate 5. The contour surface in Plate 3 with the triangular patches omitted in alternating bins along all three coordinate axes. See Section 1.4.3.

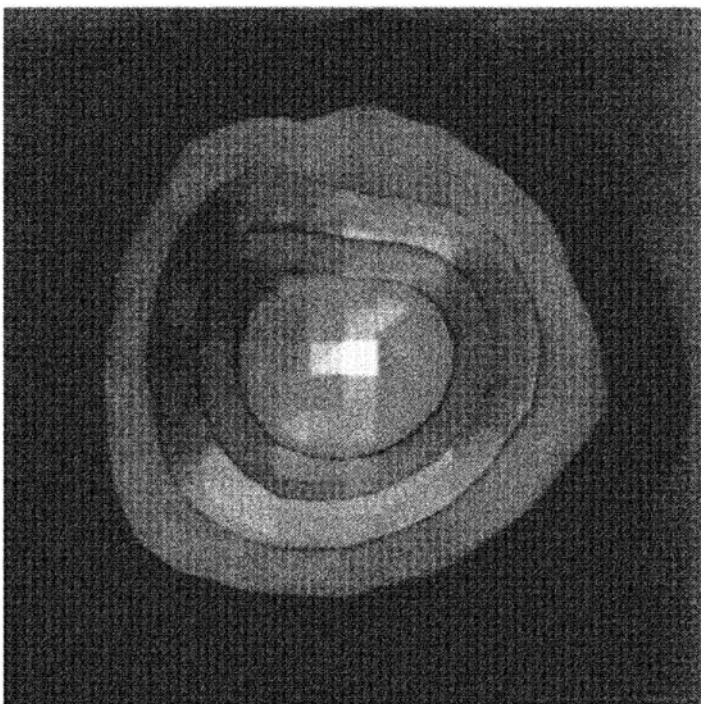


Plate 6. The contour surface in Plate 3 as displayed on a Silicon Graphics Computer with the transparency option, together with four additional α -contour surfaces at α -levels of 30%, 50%, 70%, and 90%. The depth cue is relatively weak in this figure, but much stronger when rotated in real time on the computer. See Section 1.4.3.

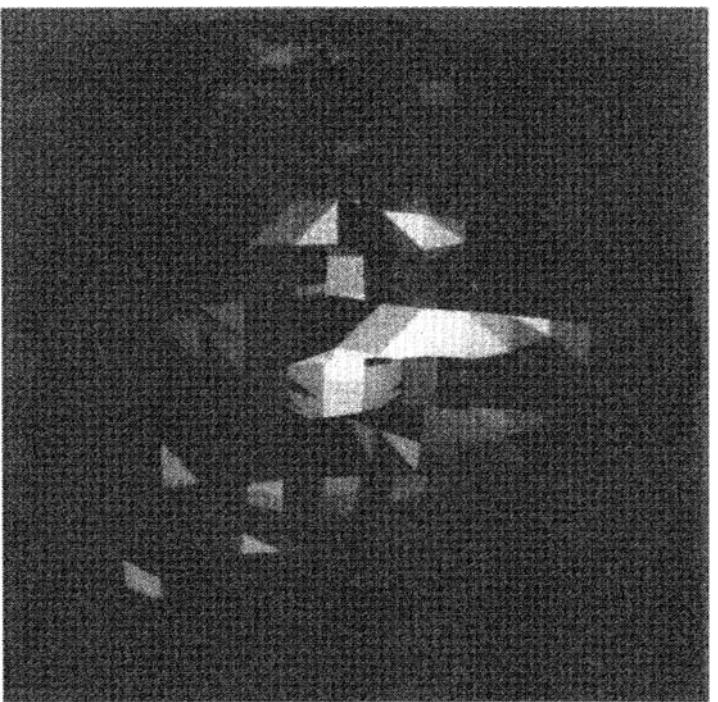


Plate 7. Another view of the 10%, 50%, and 90% α -level contours shown in Plate 6 using the “ribbon” representation as in Plate 4. See Section 1.4.3.

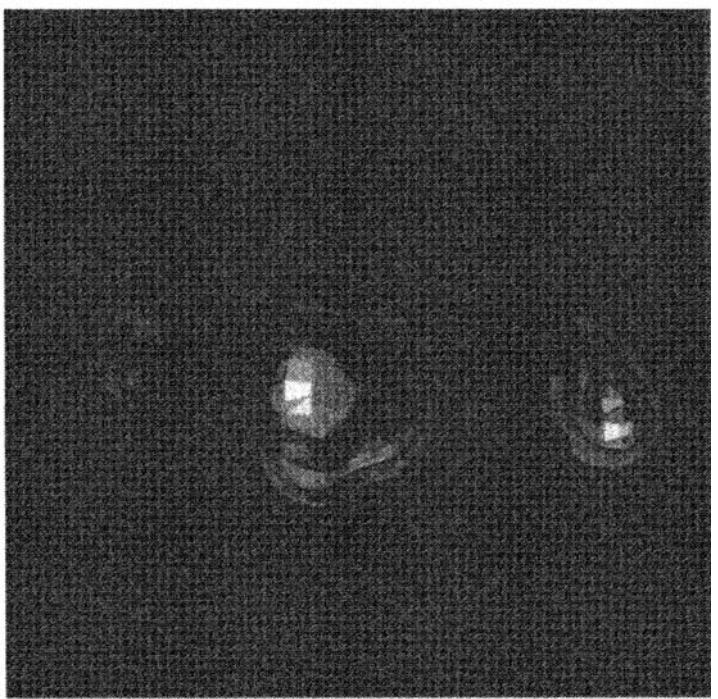


Plate 8. Five transparent α -level contours of the averaged shifted histogram of the Mount St. Helens earthquake epicenter data. The vertical axis is log-depth and the view is towards the southwest. See Section 7.2.3.



Plate 9. The $\alpha = 1\%$ level contour of the trivariate averaged shifted histogram of the LANSAT data set of 22,932 points. The small disjoint second shell in the bottom right corner represents some of the 114 outliers in the data set. The outliers resulted from singularities in the model-based data transformation algorithm from the original 24-dimensional LANSAT data and were recorded at the minimum or maximum values. See Sections 1.4.3 and 7.3.4.

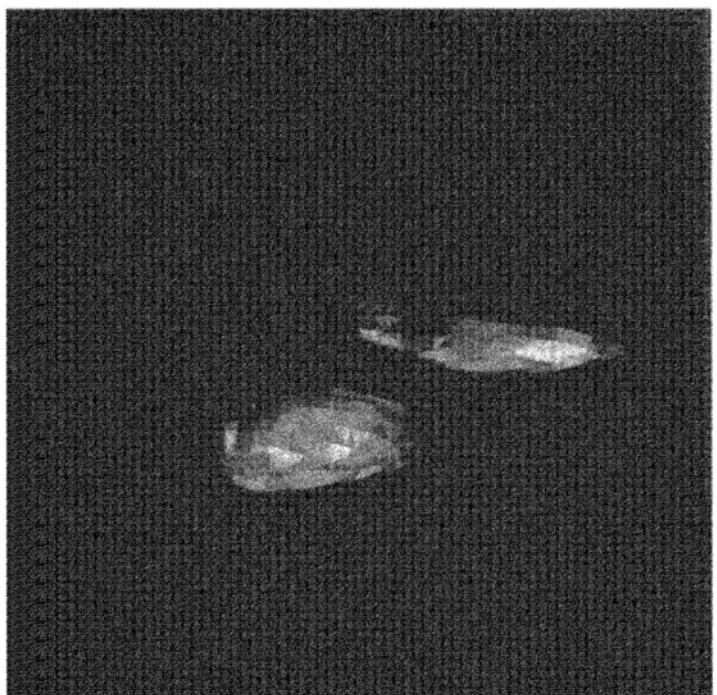


Plate 10. An enhanced view of the density estimate shown in Plate 9. An additional three clusters are clearly depicted. An examination of the crops being grown in each acre of the region reveals that the small cluster on the far left represents sugar beets; the tall cluster in the middle represents sunflowers; and the largest cluster on the right represents small grains including wheat. See Sections 1.4.3 and 7.3.4.



Plate 11. Three contour shells ($\alpha = 10\%$, 30% , and 60%) of a slice of the averaged shifted histogram of the four-dimensional PRIM4 data set with 500 points. These variables are heavily skewed, and the resulting density estimation problem more difficult. See Section 7.2.3.

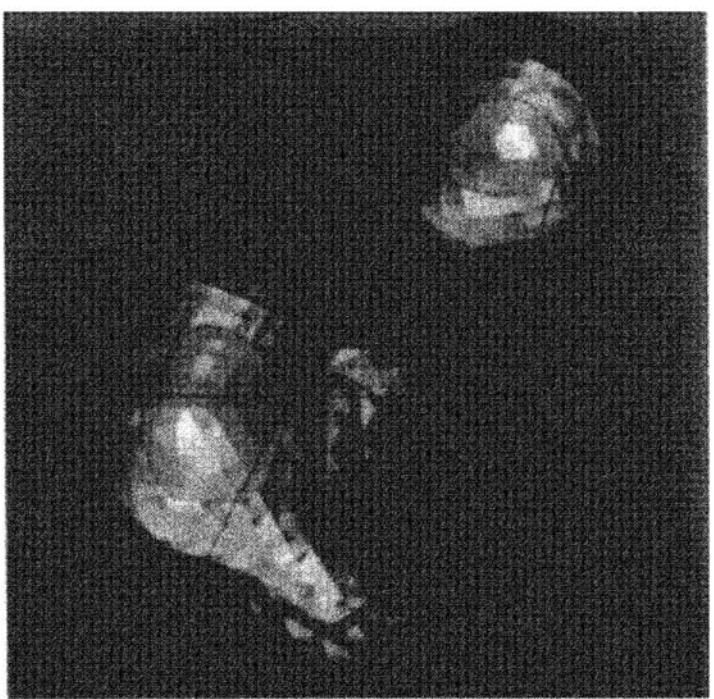


Plate 12. Three contour shell levels as in Plate 11 based on an ASH of the transformed PRIM4 data. The transformation was chosen to reduce skewness in each marginal variable. Such marginal transformations can greatly improve the quality of density estimation in multiple dimensions. See Section 7.2.3.

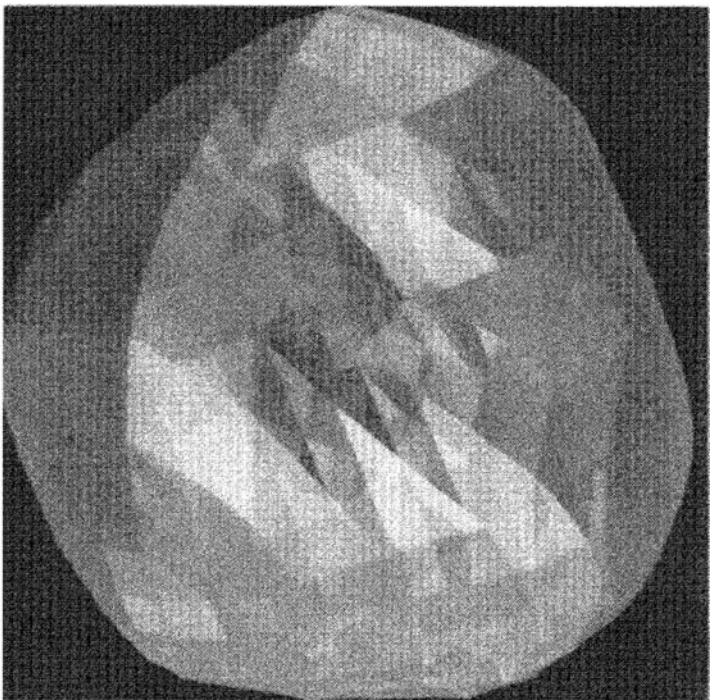


Plate 13. Contour shell derived from averaged shifted histogram estimate derived from a pseudo-random sample of 5,000 points from a trivariate density with a “hole” in the middle. The single α -level contour displayed is a pair of nested and nearly spherical shells. At values of α lower than shown, the inner shell shrinks and then vanishes. See Section 9.3.4.

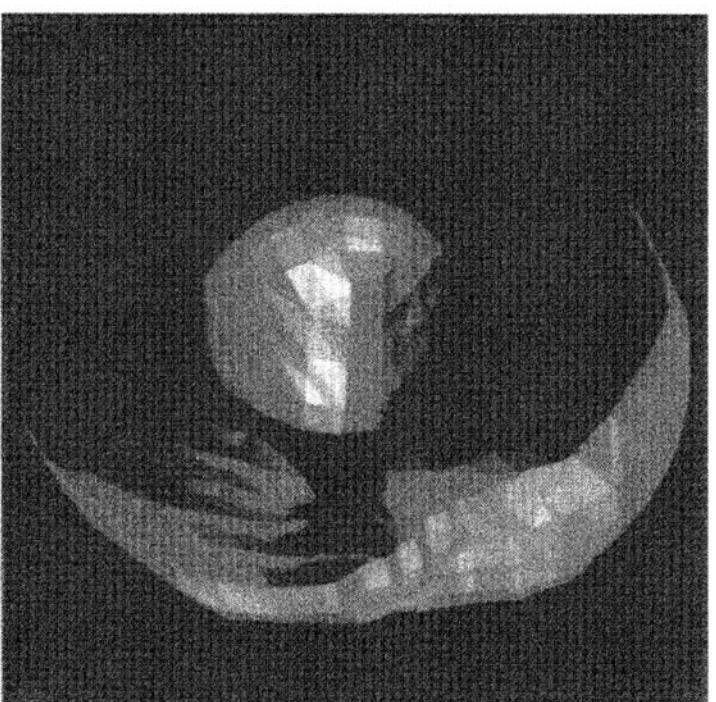


Plate 14. Contour shell at a slightly higher value of α than chosen in Plate 13. The two contours have merged. The front portion of the shell has been cut off to reveal how the inner and outer shells have joined together. See Section 9.3.4.

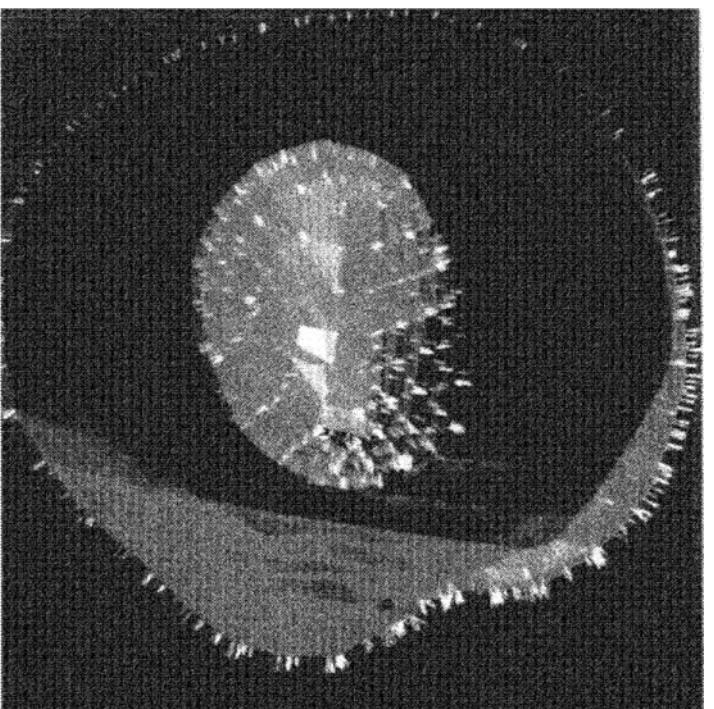


Plate 15. A slightly rotated view of the portion of the contour shell shown in Plate 14. Every contour shell has “inside” and “outside” surfaces, which point towards regions of higher and lower probability density, respectively. The white line segments depict “outer normal vectors” and point towards regions of relatively lower density. In this case the region of relatively high density is *between* or actually *within* this contour shell. Observe the blue hole where the two contours have joined. See Section 9.3.4.

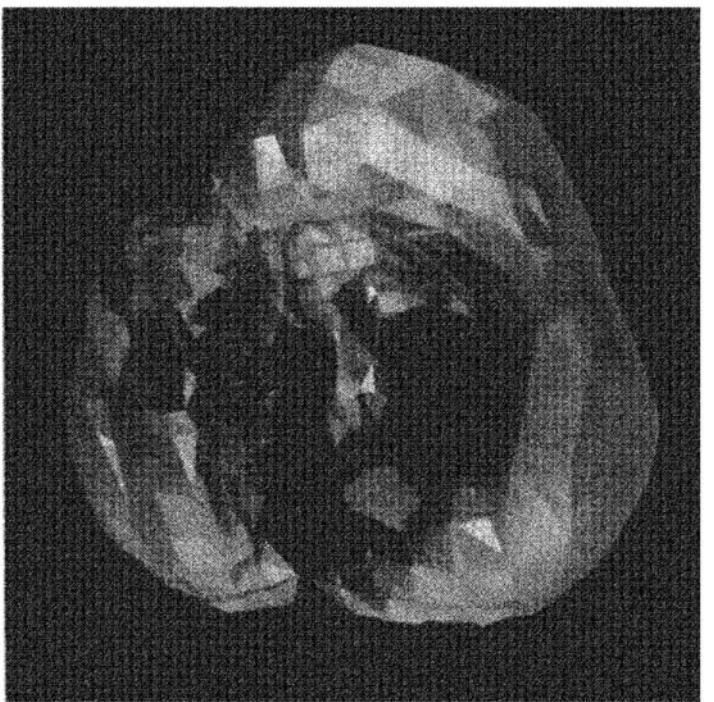


Plate 16. Contour shell at a higher value of α than shown in Plate 14. This contour shell assumes a complicated and noisy shape. By symmetry, the “mode” of the theoretical density is the surface of the unit sphere. However, these 5,000 data points happen to have concentrated in three regions near the unit sphere. This contour surface breaks up into three separate parts as α is increased further. See Section 9.3.4.

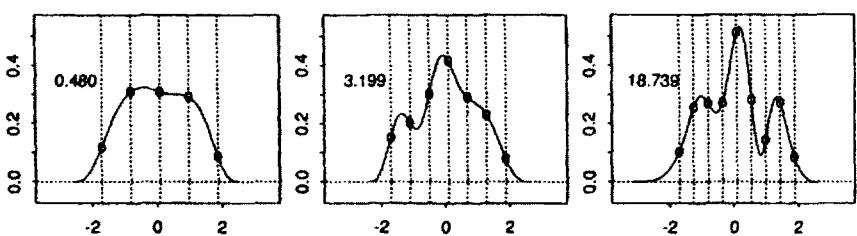


Figure 6.23 Constrained oversmoothed density solutions for Buffalo snowfall data.

contain features not in the true density (the small bump at $x = 1$), but still be conservative ($h_{CO} = 1.04 h^*$).

The application to real data is promising. (In all the examples, the original data were centered and standardized.) Several constrained oversmoothed solutions are shown in Figure 6.23 for the Buffalo snowfall data. In this case, all equally spaced meshes were selected. For small samples, this selection seems to work better than percentile meshes.

The final application is to a large data set. Several constrained oversmoothed solutions are shown in Figure 6.24 for the steel surface data. Again, all equally spaced meshes were selected. Clearly, a roughness of about 0.8 is indicated—this leads to a bandwidth that is 60% of $(0.063/0.8)^{1/5}$, the usual oversmoothing shown in the first frame.

Clearly, the number and location of the constraints serve as surrogate smoothing parameters. However, if not too many are chosen, the solution should still serve as a more useful point of reference (upper bound). It is possible to imagine extending the problem to adaptive meshes.

6.5.2 Multivariate Data

The efficacy of univariate cross-validation algorithms is still under investigation. Thus any multivariate generalizations will be somewhat speculative. In principle, it is straightforward to extend each of the algorithms in the preceding section to the multivariate setting. For example, the bootstrap

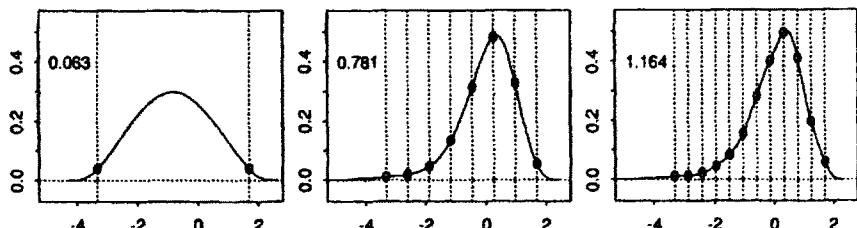


Figure 6.24 Constrained oversmoothed density solutions for steel surface data. The bin count data were jittered.

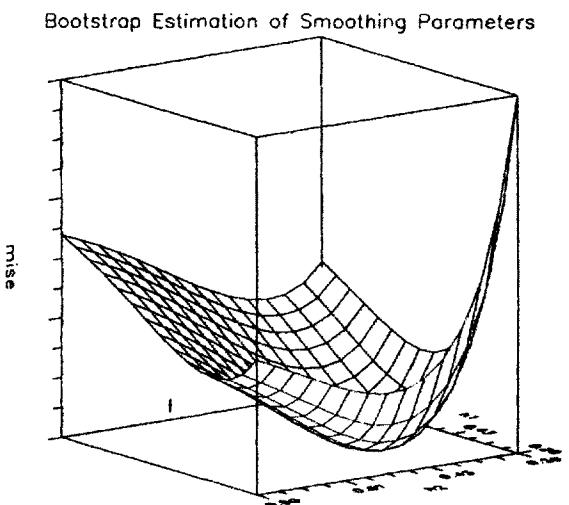


Figure 6.25 Estimated bivariate MISE using the bootstrap algorithm on 500 uncorrelated bivariate Normal points.

algorithm is easily extended (Sain, Baggerly, and Scott, 1992). An example of $\text{BMISE}_*(h_1, h_2)$ based on 500 $N(0, I_2)$ points is shown in Figure 6.25. However, the $\text{BMISE}_*(h_1, h_2)$ estimate for the lipid values ($n = 320$) showed no local minimum. A comparable BCV surface did show a local minimum, but it was oversmoothed. In general, it may be expected that the performance of all multivariate CV algorithms is more asymptotic and hence slower to converge in practice.

Retaining a smoothing parameter for each coordinate direction is important. Even if the data have been standardized, it is unlikely that $h_i = h_j$ will be satisfactory; see Nezames (1980) and Wand and Jones (1991). Density estimation is quite difficult if the data cloud is highly skewed or falls (nearly) onto a lower-dimensional manifold. This topic is taken up in detail in Chapter 7. However, marginal transformations using the Tukey ladder, for example, are always recommended. But as in the univariate setting, an absolute choice of bandwidth is not critical for exploratory purposes.

The extension of the oversmoothing bandwidth to \Re^d has been solved by Terrell (1990). The easiest multivariate density is spherically symmetric. Thus the general kernel formulation is required with the constraint that all the marginal bandwidths are equal. By symmetry, finding the form of the multivariate oversmoothed density along the x -axis, for example, will be sufficient. The variational problem turns out to be identical to the 1-dimensional problem in Equation (6.64), but in polar coordinates: $2f^{iv} + \lambda_1 + \lambda_2 r^2 = 0$, which implies that $f(r) = a(c^2 - r^2)^3$. The general form of the solution has

been given in Theorem 3 of Terrell (1990). The result is that

$$h_{OS} = \left[\frac{R(K)d}{nC_f} \right]^{1/(d+4)}, \quad \text{where } C_f = \frac{16\Gamma\left(\frac{d+8}{2}\right)d(d+2)}{(d+8)^{(d+6)/2}\pi^{d/2}},$$

for kernels that have an identity covariance matrix. The constants for $d \geq 1$ are 1.14, 1.08, 1.05, 1.04, 1.03, etc. They decrease to 1.02 when $d = 9$ and slowly increase thereafter. For $1 \leq d \leq 10$, the oversmoothed rule is between 8–10% wider than the Normal reference rule given in Equation (6.41). Using that easy-to-remember formula may be sufficient. Rescaling to other product kernels is easily accomplished by applying the rescaling rule dimension by dimension.

6.6 ADAPTIVE SMOOTHING

6.6.1 Variable Kernel Introduction

Consider the multivariate fixed kernel estimator

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i).$$

The most general adaptive estimator within this simple framework allows the bandwidth h to vary not only with the point of estimation but also with the particular realization from the unknown density f :

$$h \leftarrow h(\mathbf{x}, \mathbf{x}_i, \{\mathbf{x}_j\}) \approx h(\mathbf{x}, \mathbf{x}_i, f).$$

The second form indicates that asymptotically, the portion of the adaptive bandwidth formula dependent upon the whole sample can be represented as a function of the true density. Furthermore, it may be assumed that the optimal adaptive bandwidth function, $h(\mathbf{x}, \mathbf{x}_i)$, is smooth and a slowly varying function. Thus, for finite samples, it will be sufficient to consider 2 distinct approaches towards adaptive estimation of the density function:

$$\hat{f}_1(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{h_x}(\mathbf{x} - \mathbf{x}_i) \quad \text{where } h_x \equiv h(\mathbf{x}, \mathbf{x}, f) \quad (6.78)$$

or

$$\hat{f}_2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{h_i}(\mathbf{x} - \mathbf{x}_i) \quad \text{where } h_i \equiv h(\mathbf{x}_i, \mathbf{x}_i, f). \quad (6.79)$$

In the first case, a fixed bandwidth is used for all n data points, but that fixed bandwidth changes for each estimation point \mathbf{x} . In the second case, a different bandwidth is chosen for each \mathbf{x}_i , and then applied to estimate

the density globally. Each is justified asymptotically by the local smoothness assumption on $h(\mathbf{x}, \mathbf{x}_i, f)$, since asymptotically only those data points in a small neighborhood of \mathbf{x} contribute to the density value there. Presumably, all the optimal bandwidths in that neighborhood are sufficiently close so that using just one value is adequate. The choice of \hat{f}_1 or \hat{f}_2 depends upon the practical difficulties in specifying the adaptive bandwidth function. For small samples, one may expect some difference in performance between the two estimators. Jones (1990) has given a useful graphical demonstration of the differences between the 2 adaptive estimators.

Examples of \hat{f}_1 include the k -NN estimator of Loftsgaarden and Quesenberry (1965) (see Section 6.4.1) with h_x equal to the distance to the k th nearest sample point:

$$h_x = d_k(\mathbf{x}, \{\mathbf{x}_i\}) \approx \left(\frac{k}{nV_d f(\mathbf{x})} \right)^{1/d}, \quad (6.80)$$

where the stochastic distance is replaced by a simple histogram-like formula. The second form was introduced by Breiman, Meisel, and Purcell (1977), who suggested choosing

$$h_i = h \cdot d_k(\mathbf{x}_i, \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) \approx h \cdot \left(\frac{k}{nV_d f(\mathbf{x})} \right)^{1/d}. \quad (6.81)$$

The similarity of these two particular proposals is evident. However, the focus on the use of the k -NN distance is simply a matter of convenience.

When estimated optimally point by point, the first form provides the asymptotically best possible estimate, at least from the MISE point of view. However, the estimator is not by construction a density function. For example, the k -NN estimator is easily seen to integrate to ∞ (see Problem 33). The second estimator, on the other hand, is by construction a *bona fide* density estimator for nonnegative kernels.

In practice, adaptive estimators in the very general forms given in Equations (6.78) and (6.79) can be very difficult to specify. The "adaptive smoothing function" h_x for \hat{f}_1 is ∞ -dimensional. The specification for \hat{f}_2 is somewhat easier, since the "adaptive smoothing vector" $\{h_i\}$ is only n -dimensional. As usual, the "correct" choices of these quantities rely on further knowledge of unknown derivatives of the density function.

In practice, adaptive estimators are made feasible by significantly reducing the dimension of the adaptive smoothing function. One simple example is to incorporate the distance function $d_k(\cdot, \cdot)$ as in Equations (6.80) and (6.81). Abramson (1982a) proposed a variation on the Breiman et al. formula (6.81):

$$h_i = h/\sqrt{f(\mathbf{x}_i)} \quad \text{for } \mathbf{x}_i \in \mathbb{R}^d.$$

In a companion paper, Abramson (1982b) proves that using a nonadaptive pilot estimate for f is adequate. Observe that these two proposals agree when $d = 2$, where Breiman, Meisel, and Purcell (1977) discovered empirically that their formula worked well.

In the univariate setting, the simple idea of applying a fixed kernel estimator to transformed data and then backtransforming falls into the second category as well. If u is a smooth monotone function selected from a transformation family such as Box-Cox (1964), then when the fixed kernel estimate of $w = u(x)$ is retransformed back to the original scale, the effect is to implicitly specify the value of h_i , at least asymptotically. The transformation approach has a demonstrated ability to handle skewed data well, and symmetric kurtotic data to a lesser extent; see Wand, Marron, and Ruppert (1991). The transformation technique does not work as well with multimodal data.

In each case, the potential instability of the adaptive estimator has been significantly reduced by “stiffening” the smoothing function or vector. This may be seen explicitly by counting the number of smoothing parameters s that must be specified: $s = 1$ for k -NN (k); $s = 2$ for Breiman et al. (h, k); $s = 2$ for Abramson (h, h_{pilot}); and $s = 2, 3$, or 4 for the transformation approach (h plus 1 to 3 parameters defining the transformation family).

Theoretical and practical aspects of the adaptive problem are investigated below. There is much more known about the former than the latter.

6.6.2 Univariate Adaptive Smoothing

6.6.2.1 Bounds on Improvement

Consider a pointwise adaptive estimator with a p th-order kernel:

$$\hat{f}_1(x) = \frac{1}{nh(x)} \sum_{i=1}^n K\left(\frac{x - x_i}{h(x)}\right) \quad \text{letting } h(x) = h_x.$$

The pointwise AMSE properties of this estimator were summarized in Theorem 6.3 for a p th-order kernel:

$$\text{AV}(x) = f(x)R(K)/[nh(x)] \quad \text{and} \quad \text{ASB}(x) = [h(x)^p f^{(p)}(x)/p!]^2$$

from which the optimal pointwise AMSE(x) may be obtained:

$$\begin{aligned} h^*(x) &= \left[\frac{(p!)^2 f(x)R(K)}{2p f^{(p)}(x)^2} \right]^{1/(2p+1)} n^{-1/(2p+1)} \\ \text{AMSE}^*(x) &= (2p + 1) \left[\frac{f(x)^p f^{(p)}(x)R(K)^p}{(2p)^p p!} \right]^{2/(2p+1)} n^{-2p/(2p+1)}. \end{aligned} \tag{6.82}$$

Recall that the MISE accumulates pointwise errors. Thus accumulating the minimal pointwise errors obtained by using $h^*(x)$ gives the asymptotic lower

bound to the adaptive AMISE:

$$\begin{aligned} \text{AAMISE}^* &= \int_{-\infty}^{\infty} \text{AMSE}^*(x) dx \\ &= (2p+1) \left[\frac{R(K)^p}{(2p)^p p!} \right]^{\frac{2}{2p+1}} \int [f(x)^{2p} f^{(p)}(x)^2]^{\frac{1}{2p+1}} dx \cdot n^{-2p/(2p+1)}. \end{aligned} \quad (6.83)$$

Comparing Equations (6.21) and (6.83), it follows that the bound on the improvement of an adaptive kernel estimator is

$$\frac{\text{AAMISE}^*}{\text{AMISE}^*} = \frac{\int [f(x)^{2p} f^{(p)}(x)^2]^{1/(2p+1)} dx}{[\int f^{(p)}(x)^2 dx]^{1/(2p+1)}}. \quad (6.84)$$

An application of Jensen's inequality to the quantity

$$E[f^{(p)}(X)^2/f(X)]^{1/(2p+1)}$$

shows that the ratio in (6.84) is always ≤ 1 ; see Problem 34. In Table 6.4, this lower bound ratio is computed numerically for the Normal and Cauchy densities. Observe that the adaptivity potential decreases for higher-order kernels if the data are Normal but the opposite holds for Cauchy data. The table gives further evidence of the relative ease when estimating the Normal density. Rosenblatt derived (6.83) in the positive kernel case $p = 2$.

For the case of a positive kernel $p = 2$, the asymptotically optimal adaptive mesh is in fact equally spaced when

$$\frac{f''(x)^2}{f(x)} = c \implies f(x) = \frac{c}{144}(x-a)^4. \quad (6.85)$$

where a is an arbitrary constant; see Section 3.2.8.3. Thus the null space for kernel estimators occurs in intervals where the density is a pure quartic function. Piecing together pure segments of the form (6.85) while ensuring that f and f'

Table 6.4 Ratio of AAMISE* to AMISE* for Two Common Densities as a Function of the Kernel Order

Kernel Order p	Density	
	Normal	Cauchy
1	89.3%	84.0%
2	91.5%	76.7%
4	94.2%	72.0%
6	95.6%	70.0%
8	96.5%	68.9%

are continuous implies that f is monotone; thus there does not exist an entire null adaptive density in C^1 or C^2 unless boundary kernels are introduced.

6.6.2.2 Nearest-Neighbor Estimators

Using the asymptotic value for the adaptive smoothing parameter from Equation (6.80) with a positive kernel ($p = 2$) and $d = 1$,

$$h(x) = \frac{k}{2nf(x)}, \quad (6.86)$$

the adaptive asymptotic integrated squared bias is given by

$$\begin{aligned} \text{AAISB}(k) &= \int_x \text{AISE}(x) dx = \int_x \frac{1}{4} h(x)^4 f''(x)^2 dx \\ &= \frac{k^4}{64n^4} \int_x \frac{f''(x)^2}{f(x)^4} dx. \end{aligned}$$

Surprisingly, the latter integral is easily seen to diverge for such simple densities as the Normal. This divergence does not imply that the bias is ∞ for finite samples, but does indicate that no choice of k can be expected to provide a satisfactory variance/bias trade-off.

The explanation is quite simple: asymptotically, optimal adaptive smoothing depends not only on the density level as in Equation (6.86) but also on the curvature as in Equation (6.82). Thus the simple rule given by Equation (6.86) does not represent an improvement relative to nonadaptive estimation. This phenomenon of performing worse will be observed again and again with many simple ad hoc adaptive procedures, at least asymptotically. Some do provide significant gain for small samples or certain densities. Other approaches, such as transformation, include the fixed bandwidth setting as a special case and hence need not perform significantly worse asymptotically.

6.6.2.3 Sample-Point Adaptive Estimators

Consider the second form for an adaptive estimator with different smoothing parameters at the sample points:

$$\hat{f}_2(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right) \quad \text{letting } h(x_i) = h_i.$$

Terrell and Scott (1992, Appendix) proved under certain conditions that

$$\text{AV}\{\hat{f}_2(x)\} = f(x)R(K)/[nh(x)]$$

and

$$\text{ASB}\{\hat{f}_2(x)\} = \{(p!)^{-1}[h(x)^p f(x)]^{(p)}\}^2.$$

Abramson (1982a) proposed what is obvious from this expression for the bias, namely, that when $p = 2$, the choice

$$h(x) = h/\sqrt{f(x)} \quad (6.87)$$

implies that the second-order bias term in ASB vanishes! The bias is actually

$$\frac{1}{4!} [h(x)^4 f(x)]^{(iv)} = \frac{1}{24} h^4 \left[\frac{1}{f(x)} \right]^{(iv)},$$

as shown by Silverman (1986). This fourth-order bias is usually reserved for negative $p = 4$ th order kernels and apparently contradicts Farrell's (1972) classical result about the best bias rates with positive kernels.

In fact, Terrell and Scott (1992) have provided a simple example that illustrates the actual behavior with Normal data ($f = \phi$):

$$\hat{f}_2(x) = \frac{1}{n} \sum_{i=1}^n \frac{\sqrt{\phi(x_i)}}{h} K\left(\frac{(x - x_i)\sqrt{\phi(x_i)}}{h}\right),$$

from which it follows that

$$E\hat{f}_2(x) = \int \frac{\sqrt{\phi(t)}}{h} K\left(\frac{(x - t)\sqrt{\phi(t)}}{h}\right) \phi(t) dt, \quad (6.88)$$

with a similar expression for the variance. The exact adaptive MISE is difficult to obtain, but may be computed numerically for specific choices of h and n . The authors showed that the exact adaptive MISE was half that of the best fixed bandwidth estimator when $n < 200$; however, the fixed bandwidth estimator was superior for $n > 20,000$. This finding suggests the procedure does not have $O(n^{-8/9})$ MISE.

In fact, it is easy to demonstrate the source of the difference with Silverman (1986) and Hall and Marron (1988). Consider an asymptotic formula for the MSE of the estimator at $x = 0$, without loss of generality. The subtle point is made most clearly by choosing the boxcar kernel $K = U(-1, 1)$ so that Equation (6.88) becomes

$$E\hat{f}(0) = \frac{1}{2h} \int \phi(t)^{3/2} \left\{ I_{[-1, 1]} \left(\frac{t\sqrt{\phi(t)}}{h} \right) \right\} dt. \quad (6.89)$$

Usually, the limits of integration would extend from $-h$ to h . However, a closer examination show that the argument of the kernel is not monotone increasing and, in fact, approaches zero as $|t| \rightarrow \infty$. Thus the integral in Equation (6.89)

covers three intervals, call them $(-\infty, -b)$, $(-a, a)$, and (b, ∞) , where a and b are solutions to the equation

$$\frac{t\sqrt{\phi(t)}}{h} = 1. \quad (6.90)$$

Define $c = (2\pi)^{1/4}h$. Then Equation (6.90) takes two forms that give sufficient approximations to the interval endpoints a and b :

$$\begin{aligned} t e^{-t^2/4} = c &\implies a \approx c + c^3/4 + 5c^5/32 \\ \log t - t^2/4 = \log c &\implies b \approx (-4 \log c + 4 \log \sqrt{-4 \log c})^{1/2}. \end{aligned}$$

Now taking a Taylor's series of the integrand in Equation (6.89) and integrating over $(-a, a)$ gives

$$E\hat{f}(0) = \phi(0) + (2\pi)^{1/2}h^4/40 + O(h^6),$$

which gives the predicted $O(h^4)$ bias. However, the contribution towards the bias from the remaining two intervals totals

$$2 \cdot \frac{1}{2h} \int_{-\infty}^{-b} \phi(x)^{3/2} dx = \left(\frac{2}{9\pi}\right)^{1/4} \frac{1}{h} \Phi(-b\sqrt{3/2}),$$

which, using the approximation for b and the tail approximation $\Phi(x) \approx -\phi(x)/x$ for $x \ll 0$, equals

$$h^2/(24[\log \{(2\pi)^{1/4}h\}]^2).$$

Thus the tails exert an undue influence on the estimate in the middle and destroy the apparent gain in bias. With a smoother kernel, the same effect is observed but not so clearly. Abramson recognized this practical problem and suggested putting an upper bound on h_i by "clipping" the pilot estimator in Equation (6.87) away from zero. Other authors have missed that suggestion. However, the asymptotic inefficiency does not negate the good small-sample properties observed by Abramson (1982a), Silverman (1986), and Worton (1989).

This same analysis may be applied to the original proposal of Breiman, Meisel, and Purcell (1977). The contribution from the tails to the bias turns out to be $O(h/\log h)$. These authors had noted that in spite of excellent empirical bivariate performance, the univariate performance was poor. This slow bias rate helps to explain that observation.

6.6.3 Multivariate Adaptive Procedures

Multivariate adaptive procedures contain some interesting and unique features. These results are available in more detail in Terrell and Scott (1992).

6.6.3.1 Pointwise Adapting

Let $\nabla^2 f(\mathbf{x})$, which is the matrix of second partial derivatives of f at \mathbf{x} , be denoted by $S_{\mathbf{x}}$. From Equation (6.46), the pointwise asymptotic bias is

$$\text{AB}(\mathbf{x}) = \frac{1}{2} h^2 \text{tr}\{A^T S_{\mathbf{x}} A\} = \frac{1}{2} h^2 \text{tr}\{AA^T S_{\mathbf{x}}\}. \quad (6.91)$$

In the univariate setting, the bias is controlled entirely by the scale of the kernel, while in the multivariate setting, the shape of the kernel is also available to control the bias. The importance of the shape in minimizing $\text{tr}\{A^T S_{\mathbf{x}} A\}$ depends upon the properties of the matrix $S_{\mathbf{x}}$.

Case I

$S_{\mathbf{x}}$ is positive or negative definite. As H (and hence A) is full rank by assumption, then $A^T S_{\mathbf{x}} A$ is also positive or negative definite. Because the sum and product of the eigenvalues of a definite matrix equal its trace and determinant, respectively, the matrix with minimum (absolute) trace and determinant equal to 1 has all of its eigenvalues equal. Thus the matrix A should be chosen to satisfy

$$AA^T S_{\mathbf{x}} = |S_{\mathbf{x}}|^{1/d} I_d.$$

Observe that with this choice, the matrix on the right-hand side has the same determinant as $S_{\mathbf{x}}$, and all the eigenvalues of $AA^T S_{\mathbf{x}}$ equal $|S_{\mathbf{x}}|^{1/d}$. Therefore, the best $\text{tr}\{A^T S_{\mathbf{x}} A\} = d |S_{\mathbf{x}}|^{1/d}$. The pointwise asymptotic MSE of $\hat{f}(\mathbf{x})$ follows from Equations (6.47) and (6.91) and may be optimized to yield

$$h^*(\mathbf{x}) = \left[\frac{f(\mathbf{x})R(K)}{n d |S_{\mathbf{x}}|^{2/d}} \right]^{1/(d+4)}$$

so that

$$\text{AMSE}^*(\mathbf{x}) = \left(\frac{d+4}{4d} \right)^{\frac{2(d+2)}{d+4}} \left[f(\mathbf{x})R(K)\sqrt{|S_{\mathbf{x}}|} \right]^{4/(d+4)} n^{-4/(d+4)}.$$

Case II

The density is *saddle-shaped* at \mathbf{x} ; that is, the matrix $S_{\mathbf{x}}$ has both positive and negative eigenvalues. The density is curved upwards in some directions and downwards in others. In this case, it is possible to construct the matrix A so that sum of the eigenvalues of $AA^T S_{\mathbf{x}}$ equals 0; see Terrell and Scott (1992). Thus the order h^2 bias terms vanish with higher-order terms dominating. Therefore, regions where the density is saddle-shaped asymptotically contribute nothing to the AAMISE compared to regions where the density is definite.

How common are saddle-shaped regions? Consider the multivariate Normal density $N(\mathbf{0}, I_d)$. Then the gradient and Hessian of $f(\mathbf{x})$ are

$$\nabla f(\mathbf{x}) = -f(\mathbf{x}) \mathbf{x} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = S_{\mathbf{x}} = f(\mathbf{x})(\mathbf{x}\mathbf{x}^T - I_d).$$

There are $d - 1$ eigenvalues of the matrix $(\mathbf{x}\mathbf{x}^T - I_d)$ equal to -1 and one equal to $\mathbf{x}^T \mathbf{x} - 1$. Therefore, the multivariate Normal density is negative definite when $\|\mathbf{x}\| < 1$ and saddle-shaped *everywhere* outside the unit sphere. Given that the fraction of probability mass inside the unit sphere decreases as the dimension increases, the potential practical significance of this finding seems promising. Finally, observe that exactly on the unit sphere where $\|\mathbf{x}\| = 1$, one of the eigenvalues vanishes. This leads to the final case.

Case III

$S_{\mathbf{x}}$ is semidefinite with at least one zero eigenvalue. The density is flat in certain directions. There is no contribution to the bias in those directions and hence the bias contribution is again of lower order than in Case I. The problem at that point can be reduced to a lower dimension by projection if desired.

6.6.3.2 *Global Adapting*

The problem of selecting the best global adaptive fixed kernel estimator was formulated by Deheuvels (1977a,b), who characterized the solution in differential equation form. The global criterion to be optimized is

$$\text{AAMISE}(h, A) = \frac{R(K)}{nh^d} + \frac{1}{4} h^4 \int_{R^d} \text{tr}\{A^T S_{\mathbf{x}} A\} d\mathbf{x}.$$

Minimizing over h and A separately gives

$$\text{AAMISE}^* = \left[\min_A \int_{R^d} \text{tr}^2\{A^T S_{\mathbf{x}} A\} d\mathbf{x} \right]^{d/(d+4)} \left[\frac{(d+4)R(K)}{4nd} \right]^{4/(d+4)}.$$

For $N(\mathbf{0}, I_d)$ data, the advantage of the adaptive scheme compared to the fixed kernel is great; see Table 6.5 (from Terrell and Scott, 1992). The minimization was done numerically by Terrell. The advantage comes from the large saddle-shaped portion of the density.

The final global comparison is between the fixed kernel estimator and the k -NN multivariate density estimator. The latter estimator will be treated as an adaptive kernel estimator, with the kernel being a uniform density over the unit sphere and the bandwidth being taken as the asymptotic form

$$h \leftarrow [k/(nf(\mathbf{x})V_d)]^{1/d}.$$

Table 6.5 Relative Efficiency of Transformed Adaptive Density Estimate Compared to Fixed Kernel for the Multivariate Normal Density

Dimension d	1	2	3	4	5	6
Adapt/Fixed Efficiency	1	45.5%	30.2%	18.6%	10.6%	5.7%

Table 6.6 Relative Efficiency of k -NN Estimator to Fixed Kernel for the Multivariate Normal Density

Dimension d	1	2	3	4	5	15	100
k -NN/Fixed Efficiency	0	0	0.48	0.87	1.15	1.55	1.49

The AMISE of the fixed kernel estimator using this same kernel follows from the general kernel results earlier (see Terrell and Scott, 1992). It is shown that

$$\frac{\text{AMISE}_h^*}{\text{AAMISE}_k^*} = \left(\frac{d-2}{d} \right)^{\frac{d(d+2)}{2(d+4)}} \left(\frac{4(d^2 - 4)}{d^2 - 6d + 16} \right)^{\frac{d}{d+4}} \rightarrow \frac{4}{e}$$

as $d \rightarrow \infty$; see Table 6.6. Thus the nearest-neighbor estimator, which over-adapts to the tails in the univariate and bivariate cases, is seen to perform better than the fixed kernel estimator when $d \geq 5$, at least for Normal data. This superiority is reassuring since the algorithm has a proven track record in high-dimensional applications such as clustering.

PROBLEMS

1. Compute the IV of estimator (6.2) and compare to the histogram result. What is the kernel if $\hat{f}(x) = [F_n(x + h) - F_n(x)]/h$ is used?
2. Compute the bias and variance directly for estimator (6.3). Demonstrate that the equivalent kernel $K = U(-0.5, 0.5)$.
3. Check that the ASH results are obtained from Theorem 6.1 when the isosceles triangle kernel is specified.
4. Examine the graphs of Wahba's equivalent kernel (6.9) for several choices of the two parameters m and λ .
5. Following the discussion of the rootgram for the histogram, show that the square root of the kernel estimate is variance stabilizing compared to (6.13). Show that the variance of the root-kernel estimate is $R(K)/(4nh)$.
6. Verify Equations (6.19). Find the optimal bandwidths for estimating the first and second derivatives of f . Evaluate these for the $N(\mu, \sigma^2)$ density.
7. Show that 2 kernel estimates with sample sizes in the ratio given in Equation (6.24) have the same AMISE.

8. Show that the IV of the fixed kernel estimate equals (exactly)

$$\text{IV}(h) = \frac{1}{n} \int [\mathbb{E} K_h(x - X)]^2 dy - \frac{1}{n} \int [\mathbb{E} K_h(x - X)]^2 dx.$$

Show that the first term in the IV is *exactly* $R(K)/(nh)$. Show that the next terms are

$$-\frac{R(f)}{n} + \frac{h^2 \mu_2 R(f')}{n} - \frac{h^4}{n} \left(\frac{\mu_2^2}{4} + \frac{\mu_4}{24} \right) R(f'') + \dots,$$

where μ_k is the k th moment of the kernel.

9. Show that the ISB of a fixed kernel estimate is

$$\text{ISB}(h) = \int [\mathbb{E} K_h(x - X) - f(x)]^2 dx.$$

Suppose that the kernel K is symmetric around 0 so that $\int w^i K(w) dw = 0$ for i odd. Show that the first few bias terms equal

$$h^4 \frac{\mu_2^2}{4} R(f'') - h^6 \frac{\mu_2 \mu_4}{24} R(f''') + h^8 \left(\frac{\mu_4^2}{476} + \frac{\mu_2 \mu_6}{720} \right) R(f^{(v)}) .$$

Hint: Watch the cross-product terms and note that $\int f'' f^{(iv)} = - \int (f''')^2$, for example.

10. Verify the bias expressions for the higher-order finite difference estimators in Equation (6.30). Devise your own higher-order boxcar kernels using different spacings than integer multiples of h .
11. Compare the AMSE(x) values for 4 combinations of pointwise kernel estimates — at 0 and 1 with a second-order and fourth-order kernel.
12. Empirically compare the order-4 kernel method to the Terrell–Scott fourth-order ratio estimator for simulated Normal data as well as the snowfall data.
13. Derive the Terrell–Scott fourth-order kernel ratio estimator. Extend the procedure to a sixth-order estimator.
14. Compute the equivalent kernel of the parametric estimator $\phi(x; 0, s^2)$.
15. Compute the “theoretical” k th moment of \hat{f} , treating the kernel estimator as a “true” density.
16. Find the indifferent frequency polygon kernel mentioned in Table 6.2.

17. Consider the class of shifted-Beta kernels, $c_k(1 - t^2)_+^k$. Find their variance and rescale so that each has variance 1. Show that these rescaled kernels converge to a standard Normal kernel as $k \rightarrow \infty$.
18. Derive the product kernel AMISE results from the general multivariate kernel formulas.
19. Verify the equivalent-bandwidth formula for higher-order kernels in Equation (6.29). Check that the last factor is approximately equal to 1 for several kernels.
20. Using a Taylor's series on $\Delta F_n(x, kh)$, verify that the finite difference estimates in Equation (6.30) are of higher order. Derive some additional estimates based on the spacings $h, 2h, 4h, 8h, \dots$.
21. Check by direct integration that the "variance" of the kernel in Equation (6.34) is 0.
22. Find boundary modification kernels based on the Epanechnikov kernel. Compare with kernels supported on $[c, 1]$. Investigate the increase in $R(K_c)$. How much wider (on the right) should the kernel be so that the roughness is the same as for the biweight kernel? Does such an "equivalent roughness" always exist?
23. Verify Equation (6.38).
24. Recall estimator (6.43). Show that it is functionally equivalent to choose K to be $N(\mathbf{0}, \Sigma)$ with $H = I_d$, or to choose K to be $N(\mathbf{0}, I_d)$ with $H = \Sigma^{1/2}$. Thus the linear transformation may be applied to either the data or the kernel as a matter of preference.
25. Finite support kernels need not have only a finite number of derivatives. For example, consider

$$K(t) \propto e^{-\pi(t-t^2)} I_{(-1,1)}(t).$$

Show that the normalizing constant is $\pi/\sqrt{e} = 2\pi\sqrt{\pi e}\Phi(-\sqrt{2})$. Plot the kernel.

26. Show that $c(x - a)^4/144$ is the solution to the null adaptive density differential equation (6.85).
27. Verify the equivalent kernel for the parametric estimator $N(\bar{x}, 1)$. Plot $K(x, t)$. Compute the equivalent kernel for the parametric estimator $N(0, s^2)$ and plot it.

28. Recall that $\sigma_f^2 = s_x^2 + h^2 \sigma_k^2$. From this result, argue that if $s_x > \sigma_f$, then $\hat{h}_{\text{ISE}} > \hat{h}_{\text{MISE}}$ is the likely result. Is the converse true?
29. Try the simple orthogonal series estimator on some Beta data with $0 \leq m \leq 6$. With a larger sample, is there sufficient control on the estimate with m alone?
30. Using the simple estimate for the Fourier coefficients in Equation (6.6), find unbiased estimates of the two unknown terms in Equation (6.59).
31. Show that if a correction factor of the form $(1 + b n^{-\delta})$ is applied to h^* in (6.76), then the best choice is $\delta = 1/5$.
32. Derive Equation (6.80). *Hint:* The fraction of points, k/n , in the ball of radius h centered on \mathbf{x} is approximately equal to $f(\mathbf{x})$ times the volume of the ball.
33. Show that the k -NN estimator using the d_k distance function has infinite integral. *Hint:* In the univariate case, look at the estimator for $x > x_{(n)}$, the largest order statistic.
34. Use Jensen's inequality to show that the ratio in Equation (6.84) is ≤ 1 . *Hint:* The integral in brackets in the denominator is $E[f^{(p)}(X)^2/f(X)]$.
35. (Research). Rather than abandoning UCV in favor of more efficient asymptotic estimators, consider adjusting the data before computing the UCV curve. The noise in UCV is due in part to the fact that the terms $(x_i - x_j)/h$ are very noisy for $|x_i - x_j| < h$. Imagine placing small springs between the data points with the result that the interpoint distances become very smoothly changing. The resulting UCV seems much better behaved. A locally adaptive method replaces $x_i \leftarrow (x_{i-1} + x_{i+1})/2$, assuming that the data have been sorted. Try this modification on simulated data.
36. Prove that the average bias of a kernel estimate is *exactly* zero. *Hint:* The pointwise bias is $\int K(w)f(x - hw)dw - f(x)$.

CHAPTER 7

The Curse of Dimensionality and Dimension Reduction

7.1 INTRODUCTION

The practical focus of most of this book is on density estimation in “several dimensions” rather than in very high dimensions. While this focus may seem misleading at first glance, it is indicative of a different point of view towards counting dimensions. Multivariate data in \mathbb{R}^d are almost never d -dimensional. That is, the *underlying structure* of data in \mathbb{R}^d is almost always of dimension lower than d . Thus, in general, the full space may be usefully partitioned into subspaces of signal and noise. Of course, this partition is not precise, but the goal is to eliminate a significant number of dimensions so as to encourage a parsimonious representation of the underlying structure.

For example, consider a parametric multivariate linear regression problem in \mathbb{R}^{d+1} . The data $\{(x_i, y_i) : 1 \leq i \leq n\}$ are modeled by

$$y_i = \sum_{j=1}^d a_j x_{ij} + \epsilon_i = \mathbf{a}^T \mathbf{x}_i + \epsilon_i.$$

While d may be very large, the relevant structure of the solution data space is precisely 2-dimensional: $\{(w_i, y_i) : 1 \leq i \leq n\}$ where $w_i \equiv \mathbf{a}^T \mathbf{x}_i$. Often the motivation in regression extends beyond the prediction of y and focuses on the structure of that prediction as it is reflected in the relative magnitudes and signs of the coefficients a_j . However, when some of the predictor variables in \mathbf{x} are highly correlated, many choices of weights on those variables can give identical predictions. Collinearity in the \mathbf{x} data (reflected by a nearly singular covariance matrix) adversely affects the interpretability of the coefficients \mathbf{a} . One successful approach is to find a lower-dimensional approximation to the covariance matrix. Factor analysis is a statistical technique for culling out structure in the correlation matrix with only several “factors” being retained

(Johnson and Wichern, 1982). Additional dimensions or factors may provide more information on the margin, but their use requires more careful analysis.

The dimension reduction afforded by these models results in a tremendous increase in the signal-to-noise ratio. The full information is contained in the raw multivariate data themselves, which are sufficient statistics. In more than a few dimensions, noise limits visual interpretability of raw data. Exploratory graphical methods described in Chapter 1 such as brushing work best when the structure in the data is very low-dimensional, two at most. If the structure is too complicated, then the sequence of points highlighted during brushing can only hint at that structure's existence. More often than not, any subtle structure is lost in apparently random behavior.

With ordinary multivariate data, the conclusions are similar. If the structure in 100-dimensional data falls on a 20-dimensional nonlinear manifold, then without prior information, the task of detection is futile. With prior information, appropriate modeling can be brought to bear, but parsimonious representations of nonlinear 20-dimensional manifolds are few. Fortunately, it appears that in practical situations, the dimension of the structure seldom exceeds 4 or 5. It has been suggested that speech recognition, surely a complex example, requires a feature space of 5 dimensions. That is only one expert's opinion, but even if a nonparametric surface in \mathbb{R}^d could be well estimated, the ability to "fully explore" surfaces in dimensions beyond 5 is limited (see Huber, 1985; and Tukey and Tukey, 1981).

Projection of high-dimensional data onto subspaces is unavoidable. The choice remains whether to work with techniques in the full dimension, or to first project and then work in the subspace. The clear preference expressed here is for the latter. The two tasks need not be totally decoupled, but the separation aids in the final interpretation.

The precise cutoff dimension is a subject of intense research. Can a kernel estimate of 20-dimensional data be constructed to aid projection? Almost no one believes so. Even a decade ago, many held the belief that bivariate nonparametric estimation required prohibitively large samples to be sufficiently accurate. In higher dimensions, the limitation focuses on an understanding of the *curse of dimensionality*. The term was first applied by Bellman (1961) to combinatorial optimization over many dimensions, where the computational effort was observed to grow exponentially in d . In statistics, the phrase reflects the sparsity of data in multiple dimensions. Indeed, it was shown in Section 1.5.3.1 that given a uniform sample over the hypercube $[-1, 1]^d$, almost no sample elements will be found in the inscribed hypersphere. Even the "large" histogram bin $[0, 1]^d$ in the first quadrant contains only the fraction 2^{-d} of the data. In 10 dimensions, this fraction is only $1/10,000$. The bin $[-0.5, 1.0]^{10}$, which covers three-quarters of each edge of the hypercube, contains only 5.6% of the data. The conclusion is that in order to include sufficient data the smoothing parameters must be so large that no local behavior of the function can be reasonably approximated without astronomical sample sizes. Epanechnikov

(1969) sought to quantify those sample sizes. His and several other arguments are presented below.

In density estimation, two quite different kinds of structure can have contours that are similar in appearance. The first is data with spherical symmetry, for which many examples have already been presented. The second is of more pedagogical than practical value. Consider a density that is spherically symmetric about the origin but with a “hole” in it. Now the hole should not be a discontinuous feature, but rather a dip in the density in the middle (visualize a volcano). There are 2 notable features about the contours of this density. First, the mode is not a single point, but rather, it is a connected set of points falling on a circle (sphere) in 2 (3) dimensions. Thus the mode is not a 0-dimensional object, but rather a $(d - 1)$ -dimensional object. Second, when “falling off the mode,” contours occur in *nested pairs*; a bivariate example is illustrated in Figure 7.1. Contours in the bimodal setting also occur in pairs, but are not nested. In fact, without any labeling of contours, the nested contours of the “volcano” might be mistaken for the ordinary Normal density. As the result of a finite sample size, the sample mode is not a circle, but is rather a set of

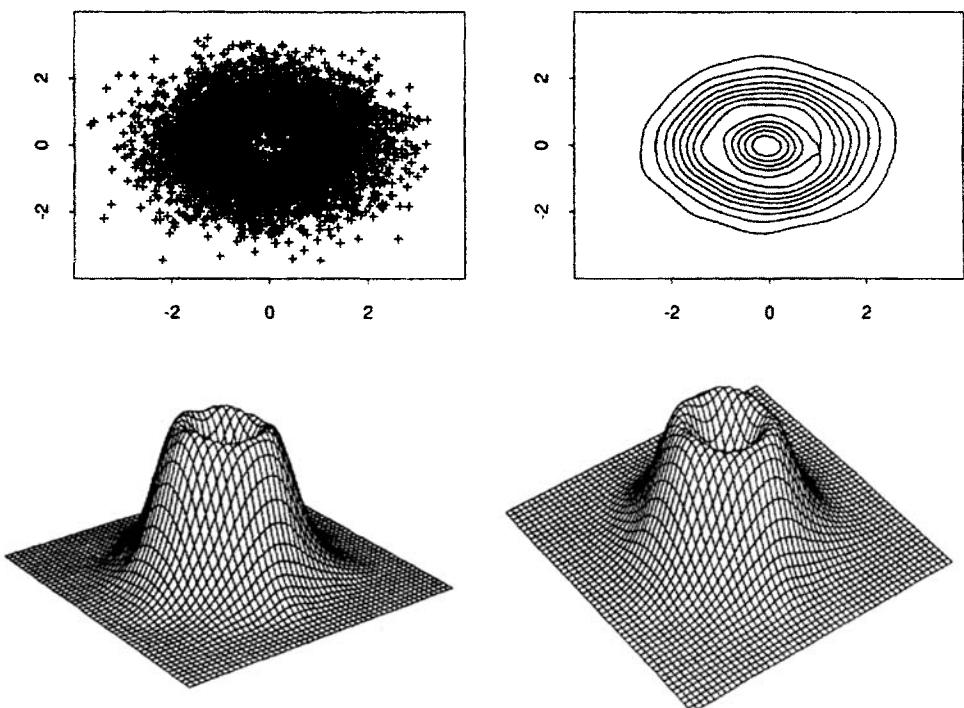


Figure 7.1 Scatter diagram and ASH representations of bivariate data from a density with a hole ($n = 5,000$). Observe how some contours occur in nested pairs.

several points located around the circle. A trivariate example is presented in Section 9.3.4.

Assessing the dimension of structure is not always easy. For example, if the regression model includes a nonlinear term of the k th variable,

$$y_i = \mathbf{a}^T \mathbf{x}_i + g(x_{ik}) + \epsilon_i, \quad (7.1)$$

then it seems reasonable to describe the structure of the model as 3-dimensional. Suppose g was a nonparametric estimator in (7.1). Even though g is usually thought of as being infinite-dimensional, 3-dimensional is still an appropriate description. These *semiparametric regression* or *partial spline* models (Speckman, 1987) are reviewed by Wahba (1990).

The illustration of the histogram in \Re^{10} misses the additional complication of rank deficient data. It was shown in Equation (6.40) that the AMISE of a kernel estimate in two dimensions blows up as the correlation $\rho \rightarrow \pm 1$. Real data in ten dimensions are never independent nor spread uniformly on a cube. Serious density estimation is never performed with an inefficient histogram or uniform kernel estimator. What is in fact feasible? This issue is explored in Section 7.2.

7.2 CURSE OF DIMENSIONALITY

The curse of dimensionality describes the apparent paradox of “neighborhoods” in higher dimensions — if the neighborhoods are “local,” then they are almost surely “empty,” whereas if a neighborhood is not “empty,” then it is not “local.” This description is simply another way of saying that the usual variance–bias trade-off cannot be accomplished very well in higher dimensions without very large samples. If the bandwidth is large enough to include enough data to hold down the variance, the bias is intolerable due to the large neighborhood, and vice versa. Stated in the usual MISE terms, the effects of the curse of dimensionality are investigated for some simple examples.

7.2.1 Equivalent Sample Sizes

In order to demonstrate the progressive deterioration of kernel density estimation as the dimension d increases, it is sufficient to follow the increase of the sample size, $n_d(\epsilon)$, required to attain an equivalent amount of accuracy ϵ . The multivariate MISE has the same units as $(\sigma_1 \sigma_2 \cdots \sigma_d)^{-1}$, which is not dimensionless. Thus direct comparison of MISE across different dimensions is meaningless. Two suggestions for a dimensionless measure of accuracy are

$$(1) \quad [\sigma_1 \sigma_2 \cdots \sigma_d \times \text{MISE}]^{1/2} \quad \text{and} \quad (2) \quad [\text{MISE}/R(f)]^{1/2}. \quad (7.2)$$

The latter suggestion was investigated by Epanechnikov (1969) using the asymptotic MISE. He considered the interesting case where both f and K are

$N(0, I_d)$. His tables were based on comparisons of the AMISE. Fortunately, it is possible to compute the $\text{MISE}(h, n, d)$ exactly in this case by direct integration (Worton, 1989):

$$(4\pi)^{d/2} \text{MISE} = \frac{1}{nh^d} - \frac{(1 + h^2)^{-d/2}}{n} + 1 - 2\left(1 + \frac{h^2}{2}\right)^{-d/2} + (1 + h^2)^{-d/2}.$$

To determine the equivalent sample sizes corresponding to $n = 50$ in \mathbb{R}^1 , the MISE expression in Equation (7.2) was minimized numerically to find the best bandwidth $h^* = 0.520$ and the corresponding $\text{MISE}^* = 0.00869$. Now $R(f) = 2^{-d} \pi^{-d/2}$ and $\sigma_1 \sigma_2 \cdots \sigma_d = 1$; hence, the Epanechnikov criterion in 1-D is 0.176. In \mathbb{R}^2 , the sample size n is found by numerically searching over the corresponding optimal bandwidth so that the criteria match. For example, with the Epanechnikov criterion, $n = 258$ in \mathbb{R}^2 (with $h^* = 0.420$) and $n = 1,126$ in \mathbb{R}^3 (with $h^* = 0.373$). However, the first criterion in (7.2) yields quite a different sequence: $n = 29$ in \mathbb{R}^2 and $n = 6$ in \mathbb{R}^3 . Clearly, not all choices of dimensionless MISE-based criteria are equivalent. Since the estimation problem does not appear to get easier as d increases, the first proposal is ignored.

Epanechnikov's equivalent sample sizes corresponding to 50 points in \mathbb{R}^1 are displayed in Figure 7.2. The graph confirms that the growth in sample size is at least exponential, since the plot is on a log-log scale. Certainly, the sample sizes for dimensions below 6 are manageable with the ASH technology.

Epanechnikov's particular choice of a dimensionless MISE criterion lacks theoretical foundation. The following pointwise criterion (relative root MSE) may also be computed exactly (Fryer, 1976; Silverman, 1986) and used for

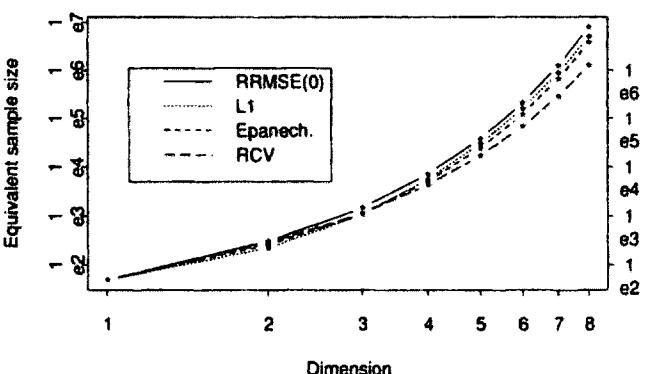


Figure 7.2 Equivalent sample sizes for several criteria that have the same value as in one dimension with 50 sample points. The density and kernel are both $N(0, I_d)$. The criteria values for RRMSE(0), AMIAE, Epanechnikov, and RCV(0) are 0.145, 0.218, 0.176, and 0.127, respectively.

extrapolation:

$$\text{RRMSE}(0) = \sqrt{\text{MSE}\{\hat{f}(0)\}} / f(0).$$

This criterion is the most pessimistic (see Figure 7.2), which is expected as the origin is the most difficult point of the density to estimate. While this criterion is dimensionless, it is only indirectly related to the noise-to-signal ratio.

Scott and Wand (1991) advocate a related criterion (root coefficient variation):

$$\text{RCV}(0) = \sqrt{\text{Var}\{\hat{f}(0)\}} / \text{E}\hat{f}(0).$$

This criterion measures the amount of noise superimposed on the average value of the estimate at the multivariate origin. As a rule of thumb, if $\text{RCV} < 1/3$, then the estimate is unlikely to be totally overwhelmed by noise. In fact, $\text{RCV}(0) = 12.7\%$ in Figure 7.2. The equivalent sample size in 8 dimensions is 33% of the next smallest criterion. Even so, a million points (which the 8-dimensional equivalent requires) is not a routine sample size. If $\text{RCV}(0) = 30\%$ is acceptable, then “only” 20,900 points are required in \mathbb{R}^8 . This prediction is by far the most optimistic.

The pointwise criteria are well-founded but not global. The most attractive criterion for the purpose at hand is the L_1 error. It was noted in Section 2.3.2.2 that the L_1 criterion was a naturally dimensionless criterion suitable for comparisons across dimensions. Fortunately, the univariate asymptotic L_1 formula of Hall and Wand (1988a) can be extended into a multivariate setting. A derivation is given in the next section. Observe that in Figure 7.2, the L_1 equivalent sample sizes are smallest in 2 and 3 dimensions, but are second from the top beyond 3 dimensions. This result should be expected, as the L_1 criterion places more emphasis on errors in the tails. In higher dimensions, virtually all of the probability mass is in the tails. For a much larger error value (corresponding to $n = 4$ in \mathbb{R}^1) given in Scott and Wand (1991), the L_1 equivalent sample sizes were smallest.

7.2.2 Multivariate L_1 Kernel Error

Hall (1984) has shown that the multivariate kernel estimate for fixed \mathbf{x} is asymptotically Normal, so that

$$\hat{f}(\mathbf{x}) - f(\mathbf{x}) \approx b_d(\mathbf{x})h^2 + Z_1 \sqrt{\frac{\sigma_d^2(\mathbf{x})}{nh^d}} = \sqrt{\frac{\sigma_d^2(\mathbf{x})}{nh^d}} \left[Z_1 + b_d(\mathbf{x}) \sqrt{\frac{nh^{d+4}}{\sigma_d^2(\mathbf{x})}} \right], \quad (7.3)$$

where Z_1 is univariate $N(0, 1)$; the pointwise bias and variance quantities were given in Section 6.3.1. Define the function ψ by $\psi(u) = E|Z + u|$. Then the mean absolute error (MAE) may be approximated by

$$\text{MAE}(\mathbf{x}) = E|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \approx \sqrt{\frac{\sigma_d^2(\mathbf{x})}{nh^d}} \psi\left(b_d(\mathbf{x})\sqrt{\frac{nh^{d+4}}{\sigma_d^2(\mathbf{x})}}\right). \quad (7.4)$$

While the absolute value function is not differentiable, it is easy to show that

$$\psi(u) = E|Z + u| = 2u\Phi(u) + 2\phi(u) - u,$$

which is a smooth function. In particular, $\psi'(u) = 2\Phi(u) - 1$. Thus differentiating (7.4) with respect to h and setting it equal to 0, the pointwise asymptotic MAE (AMAE) is minimized when

$$h_1^*(\mathbf{x}) = \left[\frac{\alpha_d^2 \sigma_d^2(\mathbf{x})}{nb_d^2(\mathbf{x})} \right]^{1/(d+4)},$$

where α_d is the unique positive solution to $4\alpha_d\left[\Phi(\alpha_d) - \frac{1}{2}\right] = \phi(\alpha_d)d$. For example, $\alpha_1 = 0.48$ and $\alpha_{10} = 1.22$.

Using the same notation as above, the smoothing parameter which minimizes $\text{MSE}(\mathbf{x})$ is

$$h_2^*(\mathbf{x}) = \left[\frac{\sigma_d^2(\mathbf{x})d}{4nb_d(\mathbf{x})} \right]^{1/(d+4)} \implies \frac{h_1^*(\mathbf{x})}{h_2^*(\mathbf{x})} = \left[\frac{4\alpha_d^2}{d} \right]^{1/(d+4)}.$$

Thus, by observation

Theorem 7.1: *For any kernel, for all qualifying density functions, and for all $\mathbf{x} \in \mathbb{R}^d$, the asymptotically optimal L_1 and L_2 pointwise bandwidths satisfy*

$$0.9635 \leq \frac{h_1^*(\mathbf{x})}{h_2^*(\mathbf{x})} \leq 1. \quad (7.5)$$

Remarkably, this ratio of asymptotically optimal bandwidths does not depend upon the particular choice of kernel, the underlying density function, or the point of estimation; see Figure 7.3. Pointwise, the choice of criterion does not appear to be critical.

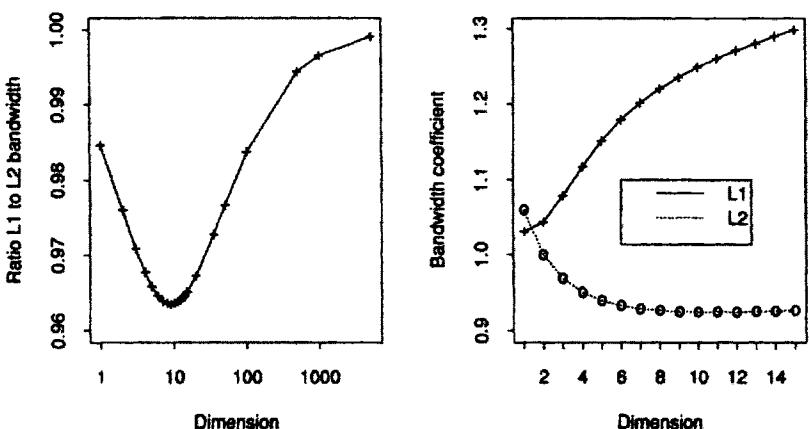


Figure 7.3 Ratio of pointwise optimal L_1 and L_2 bandwidths for all situations. The right frame displays the coefficients of $n^{-1/(d+4)}$ for the global L_1 and L_2 bandwidths for Normal data.

A global multivariate L_1 approximation may be obtained by integrating the pointwise result above. The asymptotic mean integrated absolute error (AMIAE) follows directly from Equation (7.4):

$$\text{AMIAE} = E \int_{\mathbb{R}^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})| \approx \sqrt{\frac{1}{nh^d}} \int_{\mathbb{R}^d} \sigma_d(\mathbf{x}) \psi \left\{ b_d(\mathbf{x}) \sqrt{\frac{nh^{d+4}}{\sigma_d^2(\mathbf{x})}} \right\} d\mathbf{x}.$$

Proceeding as before, the optimal width is $h_1^* = [\nu_d^2/n]^{-1/(d+4)}$, where ν_d is the unique positive solution to

$$\int_{\mathbb{R}^d} \sigma_d(\mathbf{x}) (4\nu_d r_d(\mathbf{x}) [\Phi\{\nu_d r_d(\mathbf{x})\} - 1/2] - d \phi\{\nu_d r_d(\mathbf{x})\}) d\mathbf{x} = 0,$$

letting $r_d(\mathbf{x}) = b_d(\mathbf{x})/\sigma_d(\mathbf{x})$ (Scott and Wand, 1991). By spherical symmetry, this integral may be computed as a univariate integral. The coefficients of $n^{-1/(d+4)}$ for the optimal L_1 and L_2 bandwidths for the case of multivariate Normal data are graphed in Figure 7.3. Thus the difference in the emphasis placed on the tails does lead to somewhat larger bandwidths for the absolute error criterion.

7.2.3 Examples and Discussion

The theoretical arguments above generally suggest that kernel estimation beyond 5 dimensions is fruitless. However, the empirical evidence is actually less pessimistic. Friedman, Stuetzle, and Schroeder (1984) discuss a particular simulation example in \mathbb{R}^{10} . The (x_1, x_2) -space contains the signal, which is an equal mixture of 3 shifted bivariate $N(0, I_2)$ densities with means $(\mu, 0)$,

$(-\mu, 3)$, and $(-\mu, -3)$ with $\mu = 3^{3/2}/2$. The marginal variances are both 7. Next, eight pure Normal noise dimensions are added, each with variance 7. The authors investigated the task of estimating the following slice of $f(\mathbf{x})$:

$$f(x_1, x_2, 0, 0, 0, 0, 0, 0, 0, 0).$$

The authors easily accomplished this task using their projection pursuit density estimation algorithm with only 225 data points.

Given the results of the preceding section, what could be expected of a kernel estimate with only 225 points? Figure 7.4 displays a product kernel estimate of such a sample using a triweight kernel. The same bandwidth $h = 4.0$ was used in each dimension (as suggested by an AMISE computation). Surprisingly, the trimodal structure is quite clearly represented, even though the bandwidth is hardly local. The optimal product kernel bandwidths are 2.0 in the first 2 dimensions and 5.25 in the remaining 8. The structure is even clearer, of course, although the equal bandwidth choice is a fairer test.

The kernel estimate is not as good as the figure suggests. The middle estimate is quite biased at the peaks, only a third of the true modal value. Thus the global error measure is certainly large. However, the important structure in the data is revealed by a nonadaptive kernel estimate with a very modest sample size.

What is to be made of such conflicting theoretical and empirical evidence? The answer lies in the recognition that it may not be unreasonable to accept less global accuracy in 10 dimensions than in 2 dimensions. In some situations, a high-dimensional kernel estimate may be useful, taking into account the known limitations. In other situations, the bias must be examined closely. But one situation where high-dimensional kernel estimation does seem to work is in helping to estimate which subspace on which to focus; see Section 7.3.3.

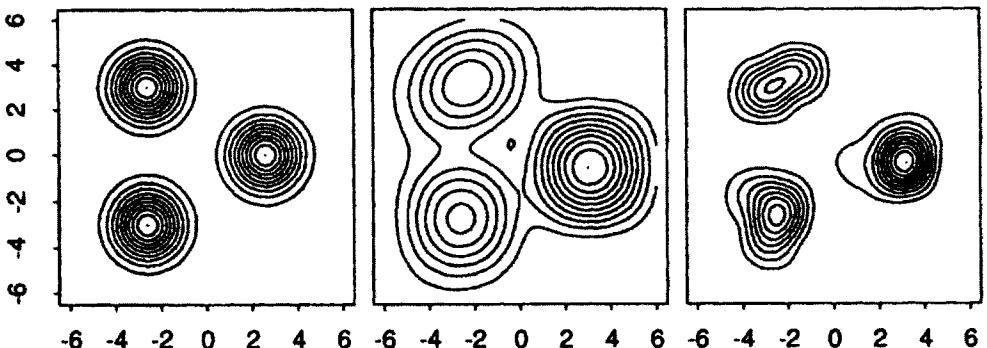


Figure 7.4 Bivariate slice of true density and slices of 2 10-D triweight product kernel estimates. The 9 contour levels are equally spaced up to the mode. The bandwidths for the middle frame were $h_i = 4.0$; for the right frame $h_1 = h_2 = 2.0$ and $h_3 = \dots = h_{10} = 5.25$; see Scott and Wand (1991).

A rather different kind of evidence comes from a simple hill-climbing algorithm investigated by Boswell (1983). He generated samples of size 100 from $N(0, I_d)$ in up to 100 dimensions. Starting at random points, he used a gradient algorithm to find the nearest local peak. Somewhat to his surprise, the algorithm invariably went to the sample mode near the origin and did not find local maxima. The exact role played by the smoothing parameter choice could be looked at more closely, but the result is impressive in any case.

Neither of these two Normal examples explore the other aspect of the curse of dimensionality that results from rank deficiencies in the data. Indeed, these two examples are somewhat unrealistic in that the variables are uncorrelated. The presence of even moderate skewness in several dimensions can adversely affect kernel density estimates. Consider the PRIM4 data set ($n = 500$), one of several sets originating from the Stanford Linear Accelerator Center during development of the original PRIM system (related data sets include PRIM7 and PRIM9). Each of the raw variables is very strongly skewed; see Figure 7.5. By successive application of Tukey's transformation ladder [Equation (3.37)] to each variable, the skewness was reduced. The 4 transformations were

$$\log_{10}(x_1), \sqrt{\log_{10}(1+x_2)}, -\sqrt{\log_{10}(1-x_3)}, -\sqrt{\log_{10}(1-x_4)}.$$

Note that when a marginal variable is bimodal, eliminating the skewness of that variable may *not* be the most desirable result. Rather, the intent is to reduce or possibly eliminate the skewness of each component mixture within that variable. Histograms of the transformed variables are displayed in Figure 7.5.

What is the practical effect on estimation of the multivariate density surface? The same slice $\hat{f}(x_1, x_2, x_3 | x_4)$ of the ASH estimator was computed using the data before and after the transformation. Those slices are shown in Color Plates

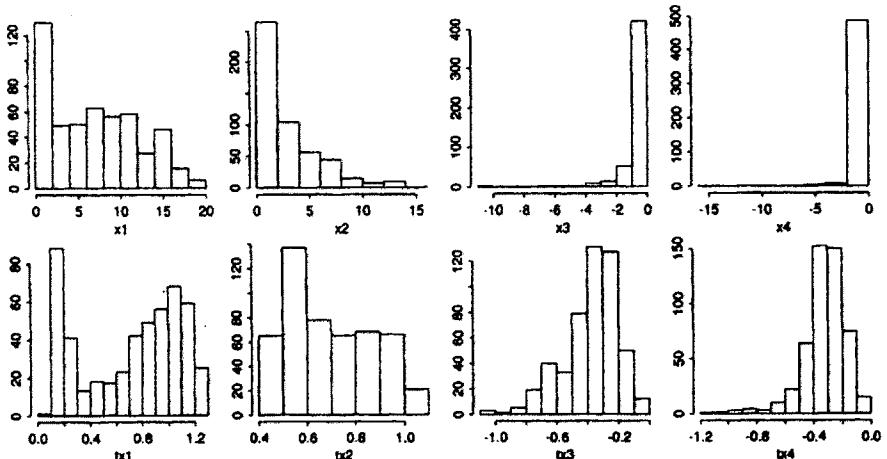


Figure 7.5 Histograms of the original 4 variables in PRIM4 before (top row) and after (bottom row) transformation; $n = 500$.

11 and 12. The original data were crowded into a corner and along a face of the surrounding 4-dimensional hyper-rectangle. After transformation, the data were clustered around the center of the hyper-rectangle. The advantage of the transformation has been to simultaneously reduce the variance and bias in the kernel estimate. The 2 primary clusters are more clearly represented, as are the 2 lesser “bumps” that are suggested.

The role of transformations is of great practical importance. Each marginal variable should be examined and transformed if skewed. The PRIM4 example used simple transformation ideas. It is possible to try to find optimal transformations using a cross-validation criterion. The transformation families need to cope not only with skewness but also with other problems. This task has been accomplished in the univariate setting by Wand, Marron, and Ruppert (1991). They advocate backtransforming to the original scale after applying the kernel estimate in the transformed space. That may or may not be desirable. For example, backtransformation can introduce artifacts such as extra bumps [see Duan (1991) and Scott (1991c)]. For the multivariate setting, visualization is facilitated when the density mass is away from the edges of the support. Back-transformation returns the mass to the edges, although in a smoother manner. It is a trade-off between the ease of examining the contours and the ease of using the scales of the original variables.

The earthquake data preceding the 1982 eruption of Mount St. Helens (Weaver, Zollweg, and Malone, 1983) also illustrates the value of transformation in separating tightly located clusters. A histogram of the depth of the epicenters of the 510 earthquakes is shown in Figure 7.6, together with the transformed variable $-\log_{10}(-z)$. In this case, the recognition of two

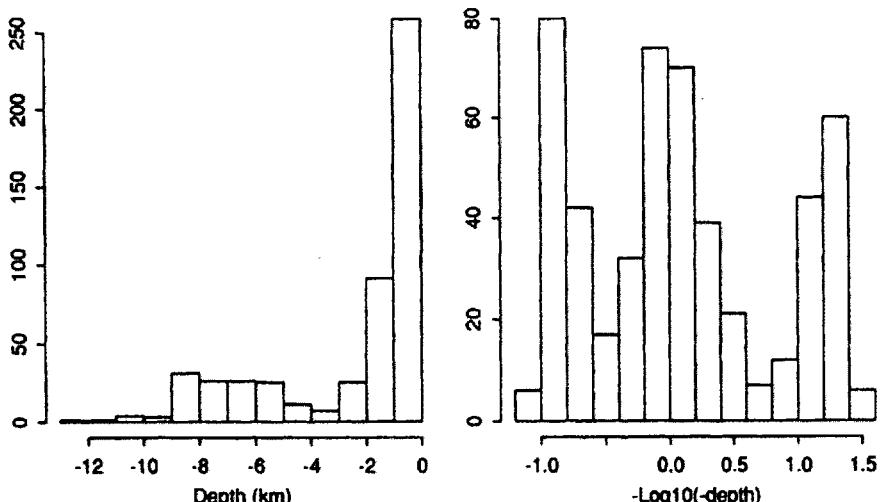


Figure 7.6 Histograms of the depths of 510 earthquake epicenters and the transformed depths.

of the three clusters of earthquake epicenters was almost missed in a histogram of the (untransformed) depth variable. However, the beauty of these data is best captured in the trivariate ASH displayed in Color Plate 8. As this figure is based on 2 months of data preceding the eruption, it is of interest to note that the earlier eruptions are the deep eruptions and the eruptions near the surface concentrated in the last weeks. It would be informative to correlate the contour surfaces with geological structures.

Thus the presence of rank deficiencies in the multivariate data, rather than the fact of high dimensions per se, is the more important component of the curse of dimensionality. The recognition of rank deficiencies and corresponding projection algorithms are the focus of the next section.

7.3 DIMENSION REDUCTION

The strategy advocated here is a two-stage approach: first reduce the dimension of the data, and then perform density estimation and use other graphical techniques in the smaller-dimensional subspace. This section deals with dimension reduction technology. Three steps are outlined for accomplishing the task of reducing the dimension from $\Re^d \rightarrow \Re^{d'}$, where $d' \in [1, 5]$. First, marginal variables are transformed to enhance the shape of the data cloud and move points away from “edges” in the surrounding hyper-rectangle. Second, a linear criterion is used to strip away obviously redundant variables. Finally, a non-linear criterion is applied that retains most of the remaining structure. These steps are discussed below.

7.3.1 Principal Components

If the data cloud takes on the form of a hyper-ellipse, then it may be completely summarized by the mean vector and the covariance matrix. Such is the case with a multivariate Normal density, $\mathbf{X} \sim N(\mu, \Sigma_d)$. Thus the shape and the “dimension” of the data cloud may be determined by examining the eigenvalues of the positive definite covariance matrix Σ_d , assuming Σ_d is of rank d .

The covariance matrix may be computed in the following compact form:

$$\Sigma = \Sigma_d = E\{(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\},$$

dropping the subscript on Σ_d . Suppose that the eigenvalues of Σ are denoted by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$, with corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d$. The eigenvectors are orthonormal; that is, $\mathbf{a}_i^T \mathbf{a}_j = \delta_{ij}$. Then $\Sigma \mathbf{a}_i = \lambda_i \mathbf{a}_i$, $1 \leq i \leq d$, which may be summarized by the matrix equation known as the *spectral representation*:

$$\Sigma \mathbf{A} = \mathbf{A} \Lambda, \quad (\Rightarrow \quad \Sigma = \mathbf{A} \Lambda \mathbf{A}^T) \quad (7.6)$$

where

$$\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_d] \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \lambda_d \end{bmatrix}.$$

The *spherling transformation* of a multivariate random variable is defined by

Spherling transformation :	$\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}),$	(7.7)
----------------------------	--	-------

where $\Sigma^{-1/2} = \mathbf{A}\Lambda^{-1/2}\mathbf{A}^T$. The random variable \mathbf{Z} has mean $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbf{0}$. Its covariance matrix is easily seen to be the identity matrix I_d , recalling that Σ and hence $\Sigma^{-1/2}$ are symmetric matrices:

$$\Sigma_{\mathbf{Z}} = E\{\mathbf{Z}\mathbf{Z}^T\} = E\{\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\Sigma^{-1/2}\} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_d.$$

The spherling transformation destroys all first- and second-order information in the density or data. This transformation is useful when the covariance structure is not desired.

A related transformation is the *principal components transformation*:

Principal components transformation:	$\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu}).$	(7.8)
--------------------------------------	---	-------

This transformation again centers the density, $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{0}$, and produces uncorrelated variables, but without wiping out the variance information:

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= E\{\mathbf{Y}\mathbf{Y}^T\} = E\{\mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{A}\} = \mathbf{A}^T\Sigma_{\mathbf{X}}\mathbf{A} \\ &= \mathbf{A}^T(\mathbf{A}\Lambda\mathbf{A}^T)\mathbf{A} = \Lambda. \end{aligned}$$

Thus the principal components are *uncorrelated* and the variance of the i th principal component, $Y_i = \mathbf{a}_i^T(\mathbf{X} - \boldsymbol{\mu})$, is the i th eigenvalue λ_i . It is shown in any multivariate analysis textbook (e.g., Johnson and Wichern, 1982) that there is no other linear combination with larger variance than Y_1 , and that Y_2 is the linear combination (uncorrelated with Y_1) with the second largest variance, and so on.

The results above hold for data as well as for the density function if Σ is replaced with the maximum likelihood estimate $\hat{\Sigma}$, which is positive definite

with eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_d > 0$, and μ is replaced with \bar{x} . The dimension of the data may be reduced by choosing the smallest $d' < d$ such that

$$\frac{\sum_{i=1}^{d'} \hat{\lambda}_i}{\sum_{i=1}^d \hat{\lambda}_i} = \frac{\sum_{i=1}^{d'} \hat{\lambda}_i}{\text{tr}\{\hat{\Sigma}\}} > 90\%, \quad (7.9)$$

for example. The reduced data are said to have retained at least 90% of the variance ("information") in the original, higher-dimensional data. The formula for transforming data to the principal components subspace in $\mathbb{R}^{d'}$ is given by

PC data transformation:
$$\mathbf{Y}_{n \times d'} = (\mathbf{X}_{n \times d} - \mathbf{1}_{n \times 1} \bar{\mathbf{x}}_{1 \times d}^T) \cdot [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_{d'}]_{d \times d'}$$

The matrix $\mathbf{1} \cdot \bar{\mathbf{x}}^T$ is simply a device used to remove the sample mean from each of the n data vectors, \mathbf{x}_i . In practical situations, d' is typically in the range $1 \leq d' \leq 5$.

In the current setting with non-Normal data, the principal components transformation could be used as follows. Histograms of each of the original d are examined in order to determine the need for marginal transformations. The new marginal variables are then standardized to have mean 0 and variance 1. After this preprocessing, the covariance matrix $\hat{\Sigma}$ is replaced by the correlation matrix $\hat{\mathbf{R}}$. As $r_{ii} = 1$, then $\text{tr}\{\hat{\mathbf{R}}\} = d$ exactly. The goal is to get rid of dimensions that contain no independent linear information, so that $d' < d$ is chosen differently than in (7.9):

$$\frac{\sum_{i=1}^{d'} \hat{\lambda}_i}{\sum_{i=1}^d \hat{\lambda}_i} = \frac{\sum_{i=1}^{d'} \hat{\lambda}_i}{d} > 95\%,$$

or perhaps 99%. The goal is to avoid losing prematurely possible nonlinear information in the data. With this choice, $1 \leq d' \leq 10$ is expected in practice.

7.3.2 Projection Pursuit

With the data "cleaned up" and stripped of obviously redundant information, a final step is proposed that will reduce the data to the final "working dimension" by $\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. To simplify notation, these dimensions are relabeled $\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. If the data fall into clearly defined clumps, then the data may be partitioned at this point, and the further transformations applied separately to each clump or cluster. For the remainder of the discussion, the data will be analyzed as a single unit.

(In general, the data should be “sphered” at this point. As a result, all possible projections will have moments $(\mathbf{0}, I_{d'})$. With this step, no second-order information can be included in the final projection algorithm. Once the small variance dimensions have been removed, the variance criterion is no longer of interest as a measure of structure in the data. Principal components searches for projections with maximum variance, while nonparametric procedures find projections with minimum variance more difficult and therefore more interesting. The need to perform separately principal components and projection pursuit is now evident.)

The idea of a projection criterion other than variance was discussed by Kruskal (1969, 1972) and Switzer (1980). Kruskal discussed an index that would detect clusters. The first successful implementation was reported by Friedman and Tukey (1974), who coined the phrase *projection pursuit* (PP). Projection pursuit is the numerical optimization of a criterion in search of the most interesting low-dimensional linear projection of a high-dimensional data cloud.

Consider a projection from $\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. Denote the projection matrix by $P_{d \times d'}$. Let the density functions of the random variables $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} = P^T \mathbf{X} \in \mathbb{R}^{d'}$ be denoted by $f(\mathbf{x})$ and $g(\mathbf{y} | P)$, respectively. Epanechnikov attempted to find the *smoothest density* in \mathbb{R}^1 by minimizing the dimensionless quantity $\sigma_K R(K)$. The multivariate version of the same quantity may be used as a measure of the *least-smooth density*, which is the most informative in $\mathbb{R}^{d'}$, in the optimization problem

$$\max_P \sigma(\mathbf{y} | P) R[g(\mathbf{y} | P)],$$

where $\sigma(\mathbf{y} | P)$ is the product of the d' marginal standard deviations. A sample version of this problem, letting $\mathbf{X}_{n \times d}$ denote the sample data, would be

$$\max_P \hat{\sigma}(\mathbf{X}P) R\{\hat{f}_h(\mathbf{y} | \mathbf{X}P)\}, \quad (7.10)$$

where the kernel estimate is based on the data matrix $\mathbf{Y} = \mathbf{X}P$ and the smoothing parameter h . The optimization assumes that h is fixed, because the kernel estimate becomes infinitely rough as $h \rightarrow 0$. However, the optimal projection matrix P should be relatively insensitive to the choice of h . The numerical optimization of problem (7.10) is nontrivial. The problem is a constrained nonlinear optimization problem, the constraints being $P^T P = I_{d'}$. Furthermore, the criterion is not strictly concave; it exhibits many local maxima. Taken together, these observations emphasize the necessity of a prior principal components transformation to the original data.

The original univariate Friedman-Tukey (1974) PP index from $\mathbb{R}^d \rightarrow \mathbb{R}^1$ was

$$\hat{\sigma}_a(\mathbf{Y}) \sum_{i,j} (h - |y_i - y_j|)_+, \quad (7.11)$$

where $\hat{\sigma}_\alpha$ denotes the α -trimmed standard deviation (i.e., the standard deviation after $\alpha/2$ of the data are deleted on each end). The criterion is large whenever many points are clustered in a neighborhood of size h . Criterion (7.11) does not obviously resemble criterion (7.10). However, Huber (1985) noted that with the Uniform kernel $U(-0.5, 0.5)$ (see Problem 1),

$$R\{\hat{f}_h(y | \mathbf{Y})\} = \frac{1}{n^2 h^2} \sum_{i,j} (h - |y_i - y_j|)_+ ,$$

so that the two proposals are in fact proportional, except for the modification to $\hat{\sigma}$. The choice of both a discontinuous kernel and a discontinuous measure of scale made the numerical optimization of criterion (7.11) unnecessarily difficult. The use of smoother kernels seems worth further investigation. Smoother kernels could also be used in related PP indices, such as $\sigma^3 R(f')$ and $\sigma^5 R(f'')$. If the data have been spherized, the appearance of σ in these formulas is unnecessary.

Huber (1985) also proposed the use of information criteria as candidates for PP indices. These were formulated in such a manner that any linear transformation of the data left the index unchanged. Two criteria proposed were standardized Fisher and negative Shannon entropy, which are defined to be

$$\sigma^2(y) \int \frac{f'(y)^2}{f(y)} dy - 1 \quad \text{and} \quad \int f \log f + \log [\sigma(y)\sqrt{2\pi e}] ,$$

respectively. These indices are especially interesting because they are each minimized when $f = \phi$. Thus maximizing them provides the *least* Normal projection. This interpretation is significant, as Diaconis and Freedman (1984) proved that most random 1-dimensional projections of multivariate data are approximately Normal, even if the underlying density is multimodal. Projection indices based on higher-order moments were considered by Jones and Sibson (1987).

Jee (1985, 1987) compared the optimal projections for theoretical densities that were mixtures of multivariate Normal pdf's. Observe that the linear projection from $\Re^d \rightarrow \Re^{d'}$ of a mixture of k Normals in \Re^d is also a mixture of k Normals in $\Re^{d'}$ with component mean and covariance matrices given by $\mu_i P$ and $P^T \Sigma_i P$, respectively. Jee also considered two other least-Normal indices. Assume that the random variables have been spherized, so that $(\mu, \sigma) = (0, 1)$ for all possible transformations. Then the L_1 and Hellinger indices are, respectively,

$$\int |f(y) - \phi(y)| dy \quad \text{and} \quad \int \left[\sqrt{f(y)} - \sqrt{\phi(y)} \right]^2 dy .$$

Jee constructed several bivariate mixtures of 2 Normal densities, each with $\Sigma = I_2$. All 4 indices were maximized and each returned the identical "best" projection to \Re^1 : namely, the angle where $f(y)$ displayed the two-component

Normal densities with maximal separation. As expected, the 4 indices were minimized at the angle where the two bivariate densities were superimposed so that $f(y) = \phi(y)$. Note that there are no data associated with these calculations — the 4 indices are calculated exactly as a function of θ over the interval $(0^\circ, 180^\circ)$ degrees, where θ is the angle that the vector P makes with the positive x -axis.

A second PP example from $\mathbb{R}^2 \rightarrow \mathbb{R}^1$ provided different results. Three bivariate Normal densities with identity covariance matrices were placed at 3 of 4 corners of a square $[-4, 4]^2$.² The density is shown at three angles of rotation in Figure 7.7. The last orientation with $\theta = 135^\circ$ clearly displays the trimodal structure after projection. However, only Fisher information takes on its maximum at $\theta = 135^\circ$. The other 3 criteria all take on their maxima at $\theta = 0^\circ, 90^\circ$, and 180° . These angles all correspond to a bimodal projection density with 2 of the densities being superimposed. The angle $\theta = 45^\circ$ is a local maximizer for all 3; it superimposes the diagonal pair of densities. The resulting univariate density is bimodal but not as well separated. From this result, it would appear that only Fisher information can successfully identify multiple clusters. Jee confirmed this finding for the projection $\mathbb{R}^3 \rightarrow \mathbb{R}^2$. Placing 4 trivariate Normals at corners of a tetrahedron, Fisher information was maximized with a projection that was a mixture of 4 bivariate Normals, while bivariate density with optimal Shannon information was only trimodal, with 2 of the original 4 trivariate densities superimposed.

Jee discussed sample-based estimates of these information criteria estimated from histograms. In simulations from the examples above, he found 200 points usually inadequate, 800 points stable, and 3,200 points very reliable. He also applied the criterion based on Fisher information (trace of the standardized Fisher information matrix) to the PRIM7 data (see Figure 7.8) presented by Friedman and Tukey (1974) in the setting $\mathbb{R}^7 \rightarrow \mathbb{R}^2$. The numerical optimization was started and restarted from many random initial guesses. Scatter diagrams of the best and second best projections are displayed in Figure 7.9. The criteria were 9.6 and 7.8, respectively. The second scatter diagram is quite similar to that found by Friedman and Tukey (1974). The special structure in the first seems a remarkable find in this data set. Observe the higher density at the corners of the triangle.

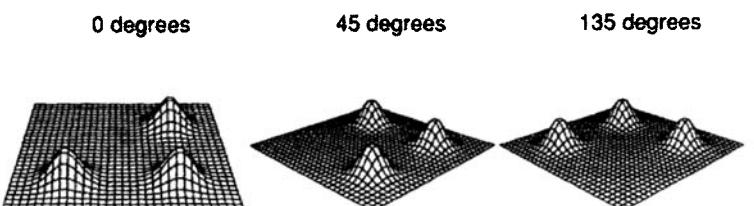


Figure 7.7 Three possible projection angles ($\theta = 0^\circ, 45^\circ, 135^\circ$) for a bivariate mixture of 3 Normal densities.

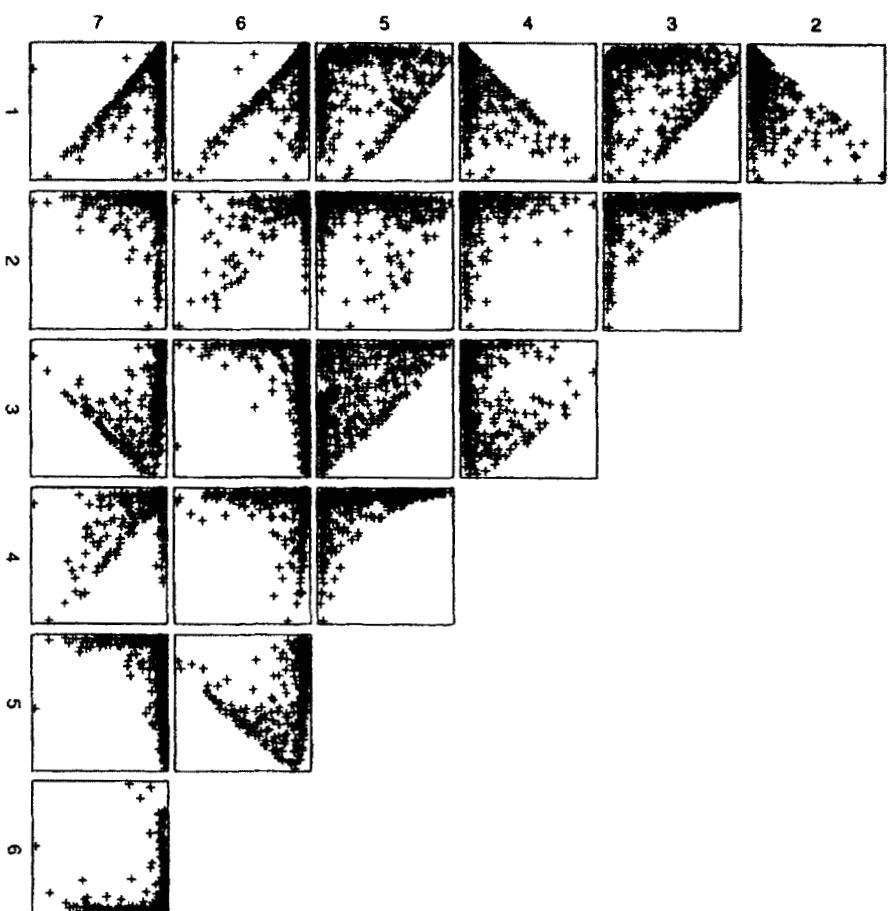


Figure 7.8 Pairs of the 7 variables in the PRIM7 data.

Much work on projection pursuit remains. Many workers skip the PP step and simply rely on the first 2, 3, or 4 principal components. The interesting variables often may be found by looking among all (but not necessarily the first few) of the pairwise scatter plots of the principal components. For example, the third and fourth principal components of the PRIM7 are somewhat close to the optimal Fisher information solution.

7.3.3 Informative Components Analysis

There are noniterative alternatives to projection pursuit besides principal components that are directly related to density estimation. For example, the results in Section 6.6.3.2 concerning optimal adaptive kernel density estimation in terms of a smoothing parameter and rotation matrix (h, A) could lead to possible algorithms. Deheuvels (1977a) characterized the solution (h^*, A^*) by a

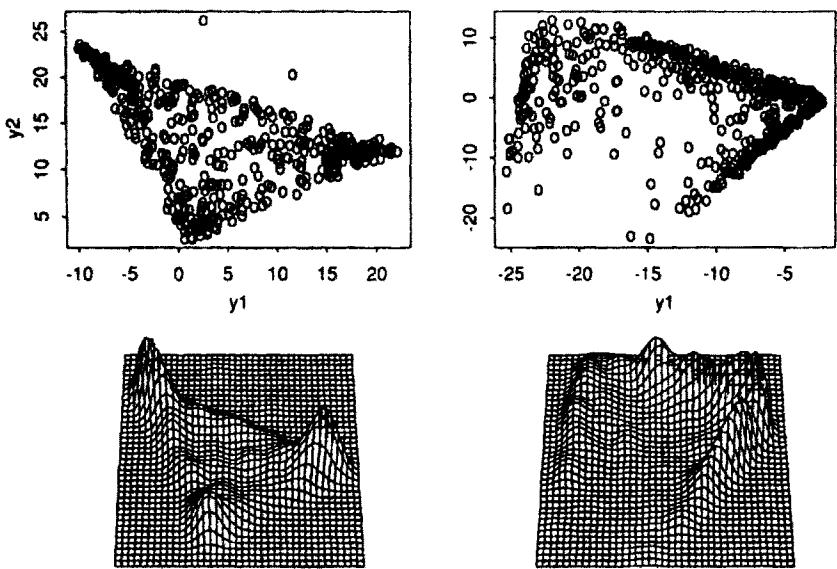


Figure 7.9 Bivariate projections of the PRIM7 data using the global and best local maxima of Fisher information. The ASH estimates for each are shown beneath.

system of nonlinear differential equations. A similar approach for the classical multivariate histogram results in some practical and quite simple projection algorithms. These results are due to Terrell (1985).

Assume that the data have been spherized. As there is no *a priori* knowledge of the scale of the remaining structure in the data, a histogram with square bins will be employed. The pointwise bias of the histogram over bin $B_0 \equiv (-h/2, h/2)^d$ given in Equation (3.57) may be written in vector form as

$$\text{Bias}\{\hat{f}(\mathbf{x})\} = -\mathbf{x}^T \nabla f(\mathbf{0}),$$

where $\nabla f(\mathbf{0})$ is the gradient vector. Now the matrix of integrals $\int_{B_0} \mathbf{x} \mathbf{x}^T d\mathbf{x} = \frac{1}{12} h^{d+2} I_d$, as $\int_{B_0} x_i x_j d\mathbf{x} = 0$ and $\int_{B_0} x_i^2 d\mathbf{x} = h^{d+2}/12$. Therefore, the asymptotic integrated squared bias over bin B_0 equals

$$\int_{B_0} \text{Bias}^2\{\hat{f}(\mathbf{x})\} d\mathbf{x} = \int_{B_0} \nabla f(\mathbf{0})^T \mathbf{x} \mathbf{x}^T \nabla f(\mathbf{0}) d\mathbf{x} = h^d \cdot \left[\frac{h^2}{12} \nabla f(\mathbf{0})^T \nabla f(\mathbf{0}) \right],$$

which is the vector form of Equation (3.58) with $h_i = h \ \forall i$. A similar expression holds in every bin B_k with $\mathbf{0}$ replaced by the bin center ξ_k . Summing over all bins, h^d is the area of the integrating unit, so standard Riemannian approximation yields

$$\text{AISB} = \frac{1}{12} h^2 \int_{\mathbb{R}^d} \nabla f(\mathbf{x})^T \nabla f(\mathbf{x}) d\mathbf{x}.$$

Reviewing the derivation of accumulating individual bin errors, it is apparent that any orientation of the cubical bins would result in the same bias. The only assumption is that the bins have volume h^d . In particular, any rotation of the cubical bin structure would yield the same AISB. As the bin probability $p_k \approx h^d f(\mathbf{x})$ for $\mathbf{x} \in B_k$, it follows from Equation (3.54) that the pointwise variance of the histogram equals $f(\mathbf{x})/(nh^d)$. Obviously, this result is also independent of the orientation of the regular cubical mesh. Therefore, for any rotation of the cubical mesh,

$$\text{AMISE}(h) = \frac{1}{nh^d} + \frac{1}{12} h^2 \int_{\mathbb{R}^d} \nabla f(\mathbf{x})^T \nabla f(\mathbf{x}) d\mathbf{x}.$$

Now by the matrix identity $\text{tr}(\mathbf{y}^T \mathbf{y}) = \text{tr}(\mathbf{y} \mathbf{y}^T)$, it follows that

$$\text{ISB} = \frac{1}{12} h^2 \text{tr} \left\{ \int_{\mathbb{R}^d} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T d\mathbf{x} \right\} = \frac{1}{12} h^2 \text{tr}\{\mathbf{Y}_f\}.$$

The matrix \mathbf{Y}_f was shown by Terrell to be positive definite. Thus the spectral representation (7.6) of \mathbf{Y}_f is

$$\mathbf{Y}_f = \mathbf{B} \Lambda \mathbf{B}^T \quad \Rightarrow \quad \text{tr}\{\mathbf{Y}_f\} = \sum_{i=1}^d \lambda_i.$$

Therefore, the AMISE may be written as

$$\text{AMISE}(h) = \frac{1}{nh^d} + \frac{1}{12} h^2 (\lambda_1 + \lambda_2 + \cdots + \lambda_{d'} + \cdots + \lambda_d).$$

The eigenvalues λ_i of \mathbf{Y}_f reflect the contribution to the overall bias from the particular direction \mathbf{b}_i , because the eigenvectors are orthogonal. Thus the subspace spanned by the first d' eigenvectors is the "most interesting" since that is where the largest contribution to the bias occurs. Terrell calls these directions the "informative components." The choice of d' in informative components is entirely analogous to the choice of d' in principal components. In practice, as the multivariate histogram is constructed with square bins, even along the "singular" dimensions, the eigenvalues in those directions will be overestimated due to the variance inflation of the histogram. The most important directions should still be correctly ranked as the inflation applies equally to all dimensions, but the smallest eigenvalues will not be close to 0 as in principal components.

The practical estimation of the matrix \mathbf{Y}_f is facilitated by the observation that

$$(\mathbf{Y}_f)_{ij} = \int_{\mathbb{R}^d} \frac{\partial f(\mathbf{x})}{\partial x_i} \cdot \frac{\partial f(\mathbf{x})}{\partial x_j} d\mathbf{x} = -E \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

upon integration by parts. $(Y_f)_{ij}$ may be estimated by a leave-one-out estimator

$$-\hat{E} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = -\frac{1}{n} \sum_{k=1}^n \frac{\partial^2 f_{-k}(\mathbf{x}_k)}{\partial x_i \partial x_j}.$$

The informative components analysis (ICA) estimator was used on a sample from Jee's bivariate example shown in Figure 7.7. In Figure 7.10 the raw data with 200 points in each cluster are shown as well as the data on the estimated ICA coordinates. The eigenvalues of 0.377 and 0.283 correctly order the interesting projections. Note that y_1 is trimodal while y_2 is bimodal. Similar reproducible results were obtained with 50 points in each cluster.

The informative components analysis (ICA) estimator was also used on the PRIM7 example shown in Figure 7.9. The pairwise scatter diagrams of the first 3 of 7 ICA coordinates are shown in Figure 7.11. The 7 eigenvalues of \hat{Y}_f were (relative to the largest): 1.0, 0.996, 0.856, 0.831, 0.761, 0.308, and 0.260. Spinning the 3-D scatterplot reveals a data cloud in the shape of a dragonfly with its wings spread open. The ICA coordinates are promising, but more work is required in choosing the bandwidth and coping with outliers.

7.3.4 Model-Based Nonlinear Projection

Statisticians are often justly criticized for using techniques that ignore known scientific models when analyzing data. Dimension-reduction transformations based on prior knowledge are common in applied sciences. Often the transformations reflect an expert's opinion based on years of experience. Objective comparisons of model-based projections and "blind" statistical methodology are rare.

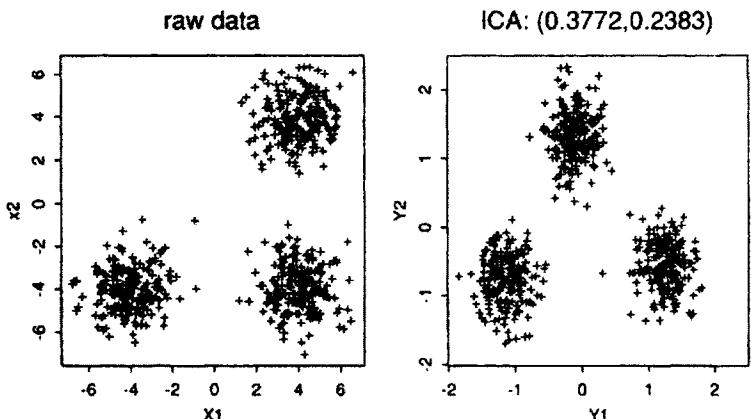


Figure 7.10 ICA of simulated bivariate trimodal data with 200 points in each cluster.

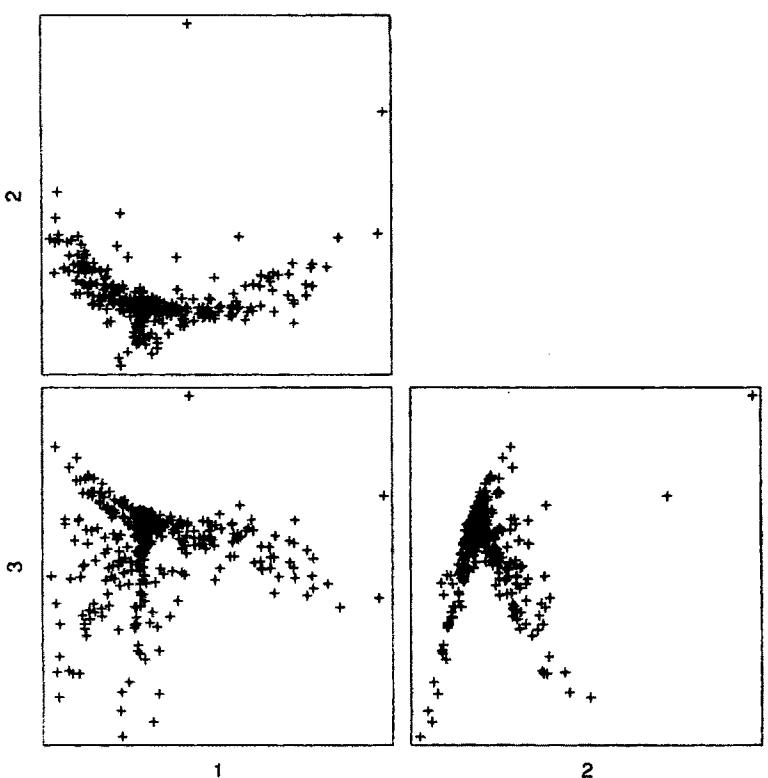


Figure 7.11 Pairwise scatter diagrams of first 3 ICA components for the PRIM7 data.

The LANDSAT data presented earlier fall into the model-based category. The raw 4-dimensional data from the satellite sensors (4 channels) were projected by principal components into a single dimension known as the “greenness” of a pixel. Each pixel was imaged every 3 weeks except when clouds obscured the region. As the region was known to be largely agricultural, the scientists’ idea was to plot the greenness as a function of time and then fit a shape-invariant growth model to those data (Badhwar, Carnes, and Austin, 1982). The model chosen had 3 parameters. Thus the data in the feature space for each pixel are the 3 parameters of the fitted greenness growth model on a pixel-by-pixel basis. The trivariate ASH of these data is displayed in Color Plates 9 and 10. An evaluation of this model-based model compared to a statistical model has not yet been performed.

PROBLEMS

1. Show that the Friedman-Tukey criterion (7.11) is exactly equal to $\hat{\sigma}_\alpha R(\hat{f}_h)$, where \hat{f}_h is a kernel estimate with the $U(-0.5, 0.5)$ kernel.

2. If $Z \sim N(0, 1)$, show that $E|Z + u| = 2u\Phi(y) + 2\phi(u) - u$.
3. Perform an informative components analysis on the iris data. Examine how the choice of smoothing parameter affects your answer.
4. Repeat Problem 3 for the PRIM4 data set found in the S system, both with and without marginal transformations used in Figure 7.5.

CHAPTER 8

Nonparametric Regression and Additive Models

The most commonly used statistical technique, parametric or nonparametric, is regression. The basic assumption in nonparametric regression is the existence of a smooth function $r(\cdot)$ relating the response y and the predictor x :

$$y_i = r(x_i) + \epsilon_i, \quad \text{for } 1 \leq i \leq n, \quad \text{where } \epsilon_i \sim g(\mu, \sigma_\epsilon^2(x)). \quad (8.1)$$

Often g is assumed to be Normal and $\sigma_\epsilon^2(x) = \sigma_\epsilon^2$, a constant.

In this chapter, the relationship of the underlying density estimator to the regression estimator will be explored. Many regression estimators discussed in this chapter are *linear smoothers*, that is, linear combinations of the observed responses. Issues of robustness are relevant, particularly in higher dimensions. Nonlinear nonparametric smoothers will be discussed. For high-dimensional regression problems, additive models of the form $r(y_i) = \sum_{j=1}^d r_j(x_{ij})$ have received much attention; see Stone (1985), Wahba (1990), and Hastie and Tibshirani (1990). An implementation based on ASH estimates will be presented. Other promising techniques, including the little known modal regression method and nonparametric L_1 regression procedure, will be introduced. Several other recent books provide comprehensive treatments of various aspects of this broad field; see Eubank (1988), Müller (1988), and Härdle (1990).

8.1 NONPARAMETRIC KERNEL REGRESSION

8.1.1 The Nadaraya-Watson Estimator

There are two distinct cases of (8.1) that should be considered, depending upon the probabilistic structure in the data $\{(x_i, y_i) : 1 \leq i \leq n\}$. The first case occurs when the $\{x_i\}$ data come from a *fixed design*; that is, the data x_i are not random but chosen by the experimenter. The second case is the *random*

design, which occurs when the data come from a joint probability density function $f(x, y)$. The emphasis in this chapter will be on the latter case, in keeping with the focus of this book.

The theoretical regression function is defined to be

$$r(x) = E(Y|X = x) = \int yf(y|x) dy = \frac{\int yf(x, y) dy}{\int f(x, y) dy}. \quad (8.2)$$

Consider the construction of a nonparametric regression estimator obtained by computing (8.2) using the bivariate product kernel estimator

$$\begin{aligned} \hat{f}(x, y) &= \frac{1}{nh_x h_y} \sum_{i=1}^n K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right) \\ &= \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) K_{h_y}(y - y_i) \end{aligned}$$

in place of the unknown bivariate density $f(x, y)$. The denominator in (8.2) contains the marginal density function, which here becomes

$$\int \hat{f}(x, y) dy = \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) \int K_{h_y}(y - y_i) dy = \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i)$$

as $\int K_h(t) dt = 1$. The restriction for an order-2 kernel that $\int t K_h(t) dt = 0$ implies that $\int y K_h(y - y_i) dy = y_i$; hence, the numerator becomes

$$\int y \hat{f}(x, y) dy = \frac{1}{n} \sum_{i=1}^n y_i K_{h_x}(x - x_i).$$

Therefore, the nonparametric kernel regression estimator corresponding to (8.2) is

$$\hat{r}(x) = \frac{\frac{1}{n} \sum_{i=1}^n y_i K_{h_x}(x - x_i)}{\frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i)} = \sum_{i=1}^n w_{h_x}(x, x_i) y_i, \quad (8.3)$$

where

$$w_{h_x}(x, x_i) = \frac{K_{h_x}(x - x_i)}{\sum_{j=1}^n K_{h_x}(x - x_j)}. \quad (8.4)$$

This estimator was proposed by Nadaraya (1964) and Watson (1964).

There are two observations to be made. First, the kernel regression estimator (8.3) is *linear in the observations* $\{y_i\}$ and is, therefore, a linear smoother. This feature is shared by many other nonparametric regression estimators. Second, the kernel regression estimate is independent of the particular choice

of the smoothing parameter h_y . This feature is not totally unexpected, given the earlier result in Problem 6.15 that $\int x \hat{f}_h(x) dx = \bar{x}$ for any choice of h . However, certain deficiencies in the model such as the nonrobustness of kernel regression follow from this feature.

Finally, suppose that the data are not random but come from a fixed design $\{x_i = i/n, i = 1, \dots, n\}$, and that the regression function is periodic on $[0, 1]$. If the regression estimate is to be computed at the design points, then the denominator in Equation (8.3) is constant, and the weight vector $w_h(x_j, x_i)$ need only be computed once. As evaluating the kernel repeatedly can entail much of the computational burden, the equally spaced fixed design is particularly advantageous. In fact, it is possible to mimic this efficient computation load even with a random design by “rounding” the design points $\{x_i\}$ to an equally spaced fixed mesh. This proposal is reminiscent of the ASH algorithm and has been investigated by Härdle and Scott (1988); see Section 8.4.1.

8.1.2 Local Least-Squares Polynomial Estimators

The use of higher-order polynomials in *parametric regression* to approximate a larger class of possible regression curves deserves some consideration. Here, the degree of the polynomial plays a role corresponding to the smoothing parameter in nonparametric regression. However, parametric polynomial fits can exhibit rapid oscillations as the order of the polynomial grows. Thus the use of higher and higher-order polynomials for “nonparametric” regression does not seem practical. If the true regression curve is smooth, then a low-order polynomial fit should be adequate *locally*; see Stone (1977). This remark is nothing but a restatement of Taylor’s theorem. Two well-known nonparametric estimators emerge from this line of thinking.

Local Constant Fitting

Globally, the *best constant regression fit* (polynomial of degree 0) to a scatter diagram is $\hat{r}(x) = \bar{y}$, which is the estimate that minimizes the least-squares criterion:

$$\bar{y} = \arg \min_a \sum_{i=1}^n (a - y_i)^2; \quad (8.5)$$

here the $\arg \min_a$ notation indicates that the constant $a = \bar{y}$ is the choice (argument) that minimizes the criterion. Consider a *local constant fit* at x to the data. Here, “local” may mean including only those data (x_i, y_i) for which $x_i \in (x - h, x + h)$ in the sum in (8.5), or it may mean including only the k design points nearest to x . It is convenient to introduce the kernel function $K_h(x - x_i)$ to indicate precisely which terms are included, and the weights (if any) on those terms. Thus the best local constant fit is

$$\hat{r}(x) = \arg \min_a \sum_{i=1}^n [K_h(x - x_i) \times (a - y_i)^2]. \quad (8.6)$$

Minimizing the right-hand side of (8.6) with respect to a leads to

$$\hat{r}(x) = \frac{\sum_{i=1}^n y_i K_h(x - x_i)}{\sum_{j=1}^n K_h(x - x_j)}, \quad (8.7)$$

which is precisely the Nadaraya-Watson kernel estimator (8.3). This pointwise result can be extended to the entire regression function by noting that

$$\hat{r}(\cdot) = \arg \min_{a(\cdot)} \int \sum_{i=1}^n K_h(x - x_i) \times [a(x) - y_i]^2 dx,$$

as the integrand is minimized by $\hat{a}(x) = \hat{r}(x)$ in (8.7) for each x .

Local Polynomial Fitting

The local constant fits in the preceding section are zero-order polynomials. The use of local linear or quadratic polynomial fits is intuitively appealing. The use of higher-order polynomials locally results in a different order of bias, as with higher-order kernels. This result has been shown for the fixed design by Härdle (1990) and for the random design by Fan (1990). The local fitting of linear polynomials has been especially advocated by Cleveland (1979) with his LOWESS procedure distributed in the S language (Becker, Chambers, and Wilks, 1988). Cleveland shows that by careful algorithm organization, the work in fitting the many local polynomials may be minimized by adding and deleting design points one at a time to the sum of squares. Wang (1990) has considered local polynomial fitting but with an absolute error criterion applied to the residuals rather than a squared-error criterion.

8.1.3 Pointwise MSE

The MSE properties of the Nadaraya-Watson estimator are rather complicated, because the estimator is the ratio of two correlated random variables. The bias and variance of the Nadaraya-Watson estimator in Equation (8.3) may be obtained by using approximations to the standard errors of functions of random variables (Stuart and Ord, 1987, Section 10.5). If the numerator and denominator in (8.3) each converge to a (positive) constant, then the asymptotic expectation of the ratio is the ratio of the asymptotic expectations of the numerator and denominator, to first order. The properties of the kernel estimator in the denominator have been well-studied [see Equations (6.14) and (6.15)], with the result that

$$E\hat{f}(x) \approx f(x) + h^2 \sigma_K^2 f''(x)/2 \quad \text{and} \quad \text{Var } \hat{f}(x) \approx \frac{R(K)f(x)}{nh}. \quad (8.8)$$

Next, consider the expectation of the numerator:

$$E\left\{ \frac{1}{n} \sum_{i=1}^n y_i K_h(x - x_i) \right\} = \int \int v \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u, v) du dv$$

$$= \int \int v K(s) f(x - hs, v) ds dv, \quad (8.9)$$

after the change of variable $s = (x - u)/h$. Now the conditional density satisfies $f(v|x - hs)f(x - hs) = f(x - hs, v)$. Therefore, the integral over v in (8.9) equals [ignoring $K(s)$]

$$f(x - hs) \int v f(v|x - hs) dv = f(x - hs) r(x - hs),$$

as the integral is the conditional mean and $r(\cdot)$ is the true regression function defined in (8.1). Continuing with (8.9), we have

$$\begin{aligned} &= \int K(s) f(x - hs) r(x - hs) ds \\ &= f(x) r(x) + h^2 \sigma_K^2 [f'(x) r'(x) + f''(x) r(x)/2 + f(x) r''(x)/2] \end{aligned} \quad (8.10)$$

after expanding $f(x - hs)$ and $r(x - hs)$ in Taylor's series up to order h^2 , assuming that K is a second-order kernel. Thus the expectation of the Nadaraya-Watson estimator with a random design is the ratio of the expectations in (8.10) and (8.8), with the density f factored out of each:

$$\begin{aligned} E \hat{r}(x) &\approx \frac{f(x) \cdot [r(x) + h^2 \sigma_K^2 \{f' r'/f + f'' r/(2f) + r''/2\}]}{f(x) \cdot [1 + h^2 \sigma_K^2 f''/(2f)]} \\ &\approx r(x) + \frac{1}{2} h^2 \sigma_K^2 \left\{ r''(x) + 2 r'(x) \frac{f'(x)}{f(x)} \right\} \end{aligned} \quad (8.11)$$

using any symbolic manipulation program, or the approximation $(1 + h^2 c)^{-1} \approx (1 - h^2 c)$ for $h \approx 0$ in the factor in the denominator and multiplying through.

The bias in the fixed design case is $h^2 \sigma_K^2 r''(x)/2$, which appears in Equation (8.11); see Problem 1. The term $2r'(x)f'(x)/f(x)$ in (8.11) is small if there are many data points in the window [i.e., $f(x)$ large]. If more design points are in the interval $(x, x + h)$ than in $(x - h, x)$ [i.e., $f'(x) > 0$], then the local average will be biased as the average will include more responses over $(x, x + h)$ than over $(x - h, x)$. The bias will be positive if $r'(x) > 0$ and negative otherwise. A similar interpretation exists when $f'(x) < 0$. The interaction of the various possibilities for the signs of $f'(x)$ and $r'(x)$ is summarized in that term. Although the expression for the bias with a fixed design matches the random design when $f'(x) = 0$, observe that these two settings are not identical. The random design has zero probability of being equally spaced even when $f = U(0, 1)$.

The variance of $\hat{r}(x)$ may be computed using the following approximation for the variance of the ratio of two random variables (Stuart and Ord, 1987):

$$\text{Var} \frac{U}{V} = \left[\frac{\text{E } U}{\text{E } V} \right]^2 \left[\frac{\text{Var } U}{(\text{E } U)^2} + \frac{\text{Var } V}{(\text{E } V)^2} - \frac{2 \text{Cov}(U, V)}{(\text{E } U)(\text{E } V)} \right]. \quad (8.12)$$

Calculations similar to those following Equation (8.9) show that

$$\begin{aligned} \text{Var} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) y_i \right\} &= \frac{1}{n} \text{E}\{K_h(x - x_i) y_i\}^2 - O(n^{-1}) \\ &\approx \frac{R(K)f(x)}{nh} [\sigma_\epsilon^2 + r(x)^2], \end{aligned} \quad (8.13)$$

using the facts that $\int v^2 f(v|x - hs) = [\sigma_\epsilon^2(x - hs) + r(x - hs)^2]$ and $\sigma_\epsilon^2(x) = \sigma_\epsilon^2$ for all x . The variance of $\hat{f}(x)$ is given in Equation (8.8). Finally,

$$\begin{aligned} \text{Cov} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) y_i, \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) \right\} \\ = \frac{1}{n} \text{E}\{K_h(x - x_i)^2 y_i\} - O(n^{-1}) \approx \frac{R(K)f(x)r(x)}{nh}. \end{aligned} \quad (8.14)$$

Carefully substituting Equations (8.8), (8.10), (8.13), and (8.14) into (8.12) yields

$$\text{Var}\{\hat{r}(x)\} \approx \frac{R(K)\sigma_\epsilon^2}{nhf(x)}.$$

In addition to the familiar factor $R(K)/(nh)$, the variance of $\hat{r}(x)$ includes factors relating to the noise variance σ_ϵ^2 and the (relative) amount of data through $f(x)$.

These results may be collected into the following theorem (Rosenblatt, 1969). A multivariate version is given by Mack and Müller (1989).

Theorem 8.1: *The asymptotic MSE(x) of the Nadaraya-Watson estimator is*

$$\text{AMSE}\{\hat{r}(x)\} = \frac{R(K)\sigma_\epsilon^2}{nhf(x)} + \frac{1}{4} h^4 \sigma_K^4 \left[r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right]^2. \quad (8.15)$$

The fact that the Nadaraya-Watson estimator follows directly from the product kernel estimator in its functional form deserves closer examination.

For the bivariate density case, $h_x^* = O(n^{-1/6})$; however, it happens that $h_x^* = O(n^{-1/5})$ for this particular estimator, which is the rate for univariate density estimation. Thus, the extra smoothing provided by integration of the bivariate density function obviates the need for a wider smoothing parameter.

Several authors have considered weights other than those of Nadaraya-Watson and obtained significantly simpler bias expressions. The scheme of integral or convolution weights devised by Gasser and Müller (1979) has certain advantages (Gasser and Engel, 1990). In addition to a bias expression similar to the fixed design case, the shape of the bias is favorable when $r(x)$ is linear, but the variance is increased. Fan (1990) has recently shown that the local polynomial fitting technique simultaneously achieves good bias and variance properties even with the random design.

8.1.4 Bandwidth Selection

The selection of smoothing parameters by cross-validation in nonparametric regression is easier than in density estimation, if only because the regression curve goes through the *data point cloud*. Two general classes of algorithms have emerged based on simple modifications to the naive average predictive squared error, which is defined by

$$G(h) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{r}(x_i)]^2.$$

Letting $\hat{r}_{-i}(\cdot)$ denote the Nadaraya-Watson estimator of the $n - 1$ points with the point (x_i, y_i) omitted, Allen (1974), Stone (1974), Clark (1975), and Wahba and Wold (1975) proposed choosing

$$\hat{h}_{CV} = \arg \min_h CV(h) \quad \text{where } CV(h) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{r}_{-i}(x_i)]^2.$$

Leaving out the i th data point eliminates the degenerate solution $h = 0$, which corresponds to interpolation of the data points, and the estimate $CV(h) = 0$, a severe underestimate of the true predictive error. Alternatively, underestimation of the predictive error can be avoided by multiplication of a simple factor:

$$\hat{h} = \arg \min_h \left(1 + \frac{2K(0)}{nh} \right) \times G(h). \quad (8.16)$$

It may be shown that this choice provides a consistent estimator of the predictive error. This particular formula in (8.16) is due to Shibata (1981). Rice (1984b) uses the factor $[1 - 2K(0)/(nh)]^{-1}$. Härdle, Hall, and Marron (1988) show how five cross-validation algorithms all have the same Taylor's series as criterion (8.16).

8.1.5 Adaptive Smoothing

A quick review of Theorem 8.1 leads to the conclusion that construction of asymptotically optimal adaptive regression curves is difficult, certainly more difficult than the in density estimation case. The use of the Gasser-Müller weights simplifies matters somewhat since the term involving r' does not appear in the AMSE expansion. Müller (1988) describes several algorithms for adaptive estimation, as do Staniswalis (1989) and Eubank and Schucany (1990).

One of the more innovative adaptive smoothers is Friedman's (1984) *super-smoother*. The algorithm constructs oversmoothed, average, and undersmoothed estimates, referred to as the "woofer," "midrange," and "tweeter," respectively. With nine passes through the data and the use of local cross-validation, an adaptive estimate is pieced together.

8.2 GENERAL LINEAR NONPARAMETRIC ESTIMATION

The kernel regression estimate is a linear combination of the observed responses of the form $\mathbf{w}^T \mathbf{y}$, which is a *linear smoother* by definition. In vector form, the bivariate regression data $\{(x_i, y_i)\}$ will be denoted by

$$\mathbf{x} = (x_1, \dots, x_n)^T \quad \text{and} \quad \mathbf{y} = (y_1, \dots, y_n)^T.$$

The weights for other linear estimators are derived in the next 3 sections.

8.2.1 Local Polynomial Regression

The study of local polynomial regression is only slightly more complicated than the global case. Therefore, the global case is considered first.

Any least-squares polynomial regression is easily shown to be "linear" in the sense described above, even when the polynomial terms are quadratic or cubic. Consider the case of straight-line regression with an intercept term, which may be incorporated into the model by appending a column of one's to the design points:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^T \quad \text{and} \quad \boldsymbol{\beta} = (\beta_0, \beta_1)^T.$$

Then the regression problem may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (8.17)$$

which is the well-known least-squares estimate of $\boldsymbol{\beta}$. Therefore, the best predictor of y at x is

$$\hat{y}(x) = \hat{r}(x) = \mathbf{w}_x^T \mathbf{y}, \quad \text{where } \mathbf{w}_x^T \mathbf{y} = (1 \ x)(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (8.18)$$

Not only is the predictor a linear combination of the responses $\{y_i\}$, but the weights themselves fall on a straight line when plotted in the (x, y) plane. This can be observed empirically in Figure 8.1, but follows from Equation (8.18) by examining the way x enters into the definition of the weight vector w_x ; see Problem 3.

This demonstration is really quite general, as the design matrix X can represent polynomial regression of any order, with multiple predictor variables and with arbitrary powers of those variables. Thus the general linear model in the parametric setting is a linear estimator according to the nonparametric definition as well.

For *local polynomial nonparametric regression* algorithms, such as LOWESS, the demonstration above is easily extended. Rather than a single linear model of the form (8.17) that holds for all x , a local linear model is formed for each point x . Only design points x_i "close" to x are included in the local linear model. The number of points included in the local linear fit plays the role of the smoothing parameter (although the degree of the polynomial also plays a role). The number of points included in the local fit can be designated directly as k , or indirectly as a fraction f of the sample size. For example, LOWESS chooses the $f \cdot n$ points closest to x ; that is, $k = fn$.

8.2.2 Spline Smoothing

The construction of smooth curves passing through design points $\{(x_i, y_i), 1 \leq i \leq n\}$ was a practical problem faced by designers of ship hulls. Their solution was to take a long, thin, and flexible piece of wood and peg it to the design points in such a way that the strain on the wood was minimal. This mechanical device was called a *spline*. Thus, mathematicians considering this interpolation problem started with a mathematical model of the bending strain given by the formula $\int r''(x)^2 dx$. The solution to the interpolation problem

$$\min_r \int_{x_1}^{x_n} r''(x)^2 dx \quad \text{s/t} \quad r(x_i) = y_i, \quad i = 1, \dots, n$$

is called an *interpolating spline*. Schoenberg (1964) showed that the interpolating spline $r(\cdot)$ is a cubic polynomial defined piecewise between adjacent knots (x_i, x_{i+1}) so that r'' is continuous on (x_1, x_n) . The latter statement is denoted by $r \in C^2(x_1, x_n)$.

The statistical version of this problem is to find the smoothest regression function with a fixed residual sum of squares. The *smoothing spline* is the solution to the variational problem

$$\hat{r}_\lambda(\cdot) = \arg \min_r S_\lambda(r) = \sum_{i=1}^n [y_i - r(x_i)]^2 + \lambda \int_{x_1}^{x_n} r''(x)^2 dx. \quad (8.19)$$

Reinsch (1967) demonstrated that the solution to (8.19) is also a cubic spline. The $\lim\{\hat{r}_\lambda\}$ as $\lambda \rightarrow 0$ is the interpolating spline, while the $\lim\{\hat{r}_\lambda\}$ as $\lambda \rightarrow \infty$ is the least-squares linear fit to the data (observing that the bending strain of a line is 0). The parameter λ plays the role of the smoothing parameter. Given λ , the amount of computation required to find \hat{r}_λ is $O(n)$. Packages such as IMSL (1991) provide software for this purpose.

Several properties of the smoothing spline may be derived without an explicit solution. For any function $q \in C^2(x_1, x_n)$ and constant α ,

$$S_\lambda(\hat{r} + \alpha q) \geq S_\lambda(\hat{r}) \quad \forall q \in C^2(x_1, x_n).$$

Therefore, $T(\alpha) = S_\lambda(\hat{r} + \alpha q)$ has a local minimum at $\alpha = 0$ for all q . By the Euler-Lagrange equation, the derivative of T should vanish there. Now

$$\begin{aligned} T'(\alpha) &= \frac{\partial}{\partial \alpha} \left\{ \sum_{i=1}^n [y_i - (\hat{r} + \alpha q)(x_i)]^2 + \lambda \int_{x_1}^{x_n} [(\hat{r} + \alpha q)''(x)]^2 dx \right\} \\ &= \frac{\partial}{\partial \alpha} \left\{ \sum_{i=1}^n [y_i - \hat{r}(x_i) - \alpha q(x_i)]^2 + \lambda \int_{x_1}^{x_n} [\hat{r}''^2 + 2\alpha \hat{r}'' q'' + \alpha^2 q''^2] dx \right\}. \end{aligned}$$

Performing the differentiation, the equation $T'(\alpha) = 0$ when $\alpha = 0$ becomes

$$-2 \sum_{i=1}^n [y_i - \hat{r}(x_i)] q(x_i) + 2\lambda \int_{x_1}^{x_n} \hat{r}''(x) q''(x) dx = 0. \quad (8.20)$$

It is now a short argument to show that \hat{r}_λ is a linear smoother. Suppose that there are two response vectors, $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, each with the same design vector \mathbf{x} . Let the respective smoothing splines be denoted by $\hat{r}_\lambda^{(1)}$ and $\hat{r}_\lambda^{(2)}$. Then, if the smoothing spline is a linear smoother, the smoothing spline for the sum of the two data response vectors should be the sum of the individual smoothing splines; that is,

$$(\mathbf{x}, \mathbf{y}^{(1)} + \mathbf{y}^{(2)}) \implies \hat{r}^{(1+2)} = \hat{r}^{(1)} + \hat{r}^{(2)}. \quad (8.21)$$

As $\{y_i^{(1)}, y_i^{(2)}, r_\lambda^{(1)}, r_\lambda^{(2)}\}$ enter linearly into Equation (8.20), the Euler-Lagrange equation holds for the proposed solution in Equation (8.21) by additivity; see Problem 2.

As a consequence of the linear smoothing property of the smoothing spline, the solution for the original data set (\mathbf{x}, \mathbf{y}) can be thought of as the sum of the solutions to n individual problems, as

$$\mathbf{y} = \sum_{i=1}^n y_i \mathbf{e}_i, \quad \text{where } \mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)^T.$$

If $\hat{r}_\lambda^{(i)}$ is the smoothing spline for the data (x_i, e_i) , then the spline solution for the original data (x, y) is given by

$$\hat{r}_\lambda = \sum_{i=1}^n y_i \hat{r}_\lambda^{(i)}.$$

If the design is periodic (also called circular) and equally spaced, then all of the n component smoothing splines $\hat{r}_\lambda^{(i)}$ are identical in shape. Silverman (1984) demonstrated that the shape converged to the kernel

$$K_s(t) = \frac{1}{2} e^{-|t|/\sqrt{2}} \sin\left(\frac{\pi}{4} + \frac{|t|}{\sqrt{2}}\right) \quad |t| < \infty. \quad (8.22)$$

$K_s(t)$ may be shown to be an order-4 kernel. For data from a random design, Silverman proved that the smoothing spline is an adaptive rather than fixed bandwidth estimator, with weights

$$w_\lambda(x, x_i) \approx \frac{1}{f(x) h(x_i)} K_s\left(\frac{x - x_i}{h(x_i)}\right), \quad \text{where } h(x_i) \approx \left(\frac{\lambda}{nf(x)}\right)^{1/4}.$$

A fourth-order kernel should have bandwidth $h(x_i) = O(n^{-1/9})$, which occurs if $\lambda = O(n^{5/9})$ is chosen. Unfortunately, the spline smoother adapts only to the design density $f(x)$ and not to the curvature of the unknown curve $r(x)$ as required by Theorem 8.1. While the estimator may not provide optimal adaptive estimation, it provides visually appealing estimates in practice. Reinsch's (1967) original proposal included a weight function in (8.19), which could be designed to provide improved adaptive properties.

8.2.3 Equivalent Kernels

A wide range of nonparametric regression algorithms, including LOWESS and smoothing splines, have been shown to be linear smoothers. To compare these estimates to the original kernel estimator in (8.3), it is natural to compare the weight vectors w_x multiplying the response vector y . Following the definition of the kernel weights in (8.4), the weight vectors for other linear smoothers will be called the *equivalent kernel weights* of that estimator. For a particular data set, these weights may be computed exactly for theoretical purposes. In practice it is not necessary to obtain the weights explicitly to compute the estimate.

Consider the gas flow accuracy data of $n = 33$ readings taken when the pressure was set at 74.6 psia in the gas line. The accuracy of the meter is plotted as a function of the logarithm of the rate of the gas flow rate in the pipeline. It is clear from the upper right panel in Figure 8.1 that the relationship is not linear. The first panel shows the collection of equivalent kernel weights for the parametric straight-line fit, and the estimate is shown in the upper right frame. For example, consider the leftmost data point (1.76, 97.25). The 33 weights

on the 33 responses used to compute the regression estimate were calculated according to (8.18). Those weights fall on a straight line plotted in the upper left frame of Figure 8.1, connecting the 2 extreme points $(x, w) = (1.76, 0.145)$ and $(3.59, -0.081)$. To identify the 33 lines, the weight (x_i, w_i) is shown as a small circle on the line representing the i th weight vector. Notice that the weights are not local (i.e., the weights do not vanish outside a neighborhood of x_i).

A local quadratic polynomial fit was also applied to these data. The fit was based on $k = 7$ design points, $\{x_{i-3}, \dots, x_i, \dots, x_{i+3}\}$. At the boundaries, the 7 points closest to the boundary are used for the fit. The 33 local quadratic fits are shown in the bottom right frame of Figure 8.1, and the 33 equivalent weight vectors are shown in the bottom left frame. The small bandwidth allows the estimate to follow the small dip in the middle of the curve. The equivalent

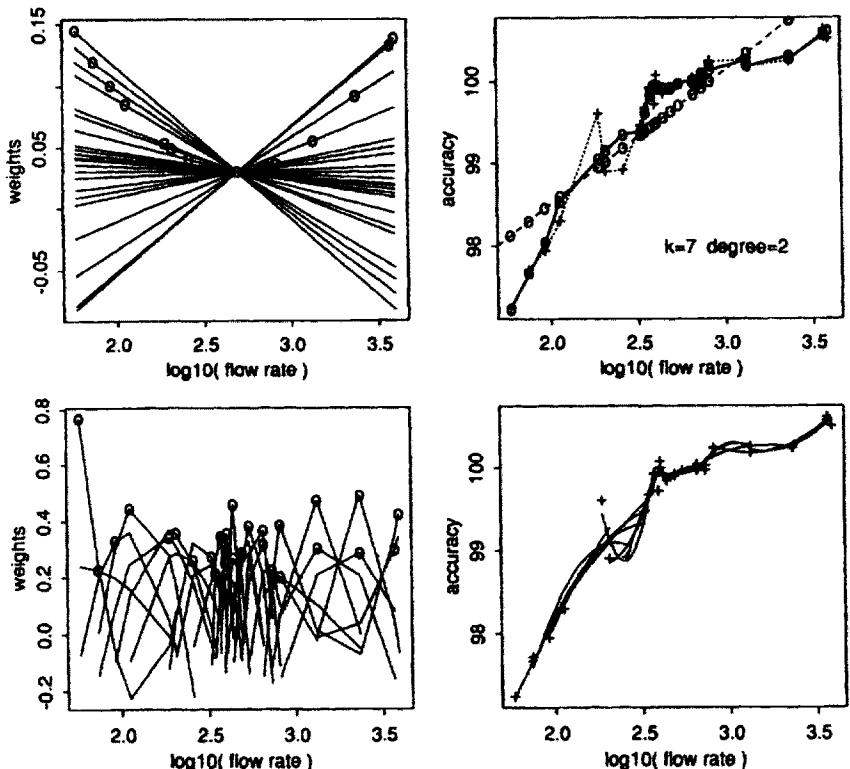


Figure 8.1 Linear smoothers applied to the gas flow data at 74.6 psia. In the first frame, the 33 weight vectors are plotted for a parametric straight-line fit, with the circle showing the weight on the data point itself. In the second frame, the raw data ("+" are shown connected by small dashes; the 33 estimates at $x = x_i$ from frame 1 are shown (the "o" symbols on the straight line connected by big dashes); the 33 estimates "o" connected by a solid line using a local quadratic model with $k = 7$ (3 points on either side of x_i except at the boundaries). In the third frame, the 33 weight vectors for the local quadratic fits are shown with the circle showing the weight on x_i itself. The fourth frame shows the 33 local quadratic fits superimposed on the data.

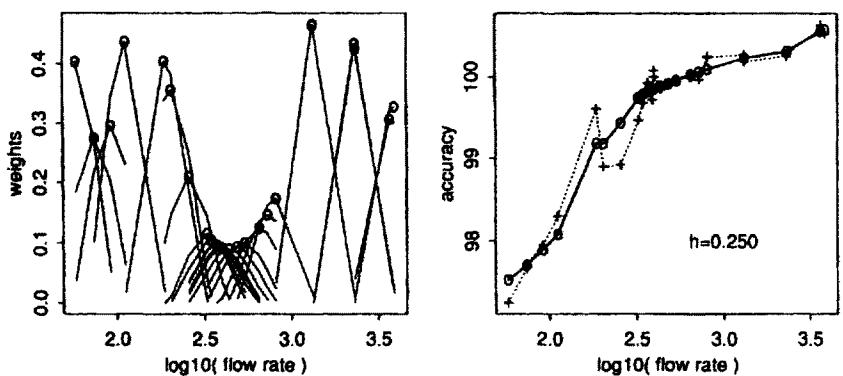


Figure 8.2 Biweight Nadaraya-Watson kernel weights and estimate for the gas flow data.

weight vectors are 0 where they are not explicitly shown. By construction, the weights are local. While the equivalent kernels are fairly rough and take on negative values, a portion of the roughness results from their ability to correctly adapt at the boundaries (Fan, 1990). Compare the boundary kernels here to those shown in Figure 6.10.

The Nadaraya-Watson kernel weights are shown in Figure 8.2 for the biweight kernel with a bandwidth of $h = 0.25$, which provides a smoother estimator than the local polynomial fit. However, a boundary kernel was not employed, as is evident in the estimate at the left boundary and in the picture of the kernel weights.

The smoothing spline fit is shown in Figure 8.3. Observe that the kernel weights are not as local for this small sample because the estimate is much smoother (less local). The boundary behavior is reasonable. For a discussion of the dual of equivalent bandwidths for linear smoothers, see Hastie and Tibshirani (1990).

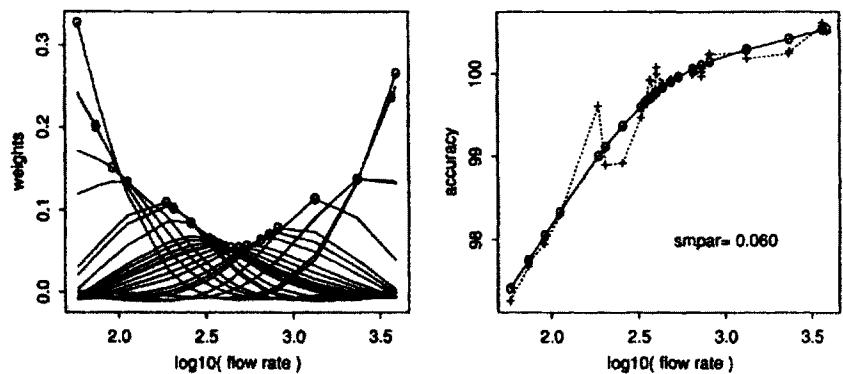


Figure 8.3 Smoothing spline equivalent kernel weights and estimate for gas flow data.

8.3 ROBUSTNESS

In recent years, more effort was required for data entry and verification than for a complete regression and analysis of variance. While changes in computing have been dramatic, the problems of handling bad or influential data points have grown even faster than the growth of the size of data sets. In such cases it is increasingly difficult to identify hundreds of influential points, particularly in higher dimensions. Thus the practical importance of robust procedures that "automatically" handle outliers is ever growing (Rousseeuw and Leroy, 1987).

8.3.1 Resistant Estimators

As linear smoothers are equivalent to a local weighted average of responses, these estimators are not resistant to outliers in the response vector. Following the well-developed literature in robust estimation in the parametric setting (Hampel, 1974; Huber, 1964), two proposals have emerged to improve the resistance of the kernel estimator, focusing on the characterization given in Equation (8.6). The quadratic function of the residuals is replaced by an influence function, $\rho(\cdot)$, that reduces the magnitude of the contribution of large residuals to the optimization criterion. An attractive choice is to replace the unbounded quadratic influence function, $\rho(\epsilon) = \epsilon^2$ with a function that has bounded influence; that is, $|\rho(\epsilon)| < c < \infty$. Given a choice for the robust influence function, the original problem is replaced by

$$\hat{r}_\rho(\cdot) = \operatorname{argmin}_{a(\cdot)} \int \sum_{i=1}^n K_h(x - x_i) \rho(a(x) - y_i) dx.$$

This estimator has been discussed by Härdle (1984) and Härdle and Gasser (1984).

A second algorithm that down-weights the influence of large residuals was proposed by Cleveland (1979). The LOWESS procedure described in Section 8.1.2 can be extended for this purpose. Rather than replacing the quadratic influence function of the residuals $\{\epsilon_i\}$, an additional multiplicative factor is introduced term by term based on the relative magnitude of the sample residuals $\{\hat{\epsilon}_i\}$. The factor is small for large residuals. Cleveland suggests computing a robust scale estimate $\hat{\sigma}$ and weights $\{\delta_i\}$ by

$$\hat{\sigma} = \operatorname{median}\{|\hat{\epsilon}_i|\} \implies \delta_i = [1 - (\hat{\epsilon}_i/(6\hat{\sigma}))^2]_+^2.$$

For example, if fitting a local linear polynomial, the fitted polynomial solves

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{(a,b)} \sum_{i=1}^n \delta_i \cdot w_h(x, x_i) \{y_i - (a + bx)\}^2.$$

The LOWESS estimate is $\hat{a} + \hat{b}x$. The procedure is applied recursively until the estimates remain unchanged. The computational effort is an order of magnitude greater than the nonrobust version.

8.3.2 Modal Regression

It is ironic that while the original kernel density estimator is quite resistant to outliers, the regression function derived from the kernel estimate is not. Is there an alternative regression estimator derived from the kernel estimator that is resistant? The answer comes from a moment's reflection upon the three measures of center discussed in elementary textbooks: mean, median, and mode. The obvious proposal is to go beyond the conditional mean of the kernel estimator and to consider either its conditional median or its conditional mode. Of course, if the errors are Normal or symmetric, all three choices give the same result, asymptotically. A strong argument can be made for summarizing the bivariate data with a trace of the conditional mode. This trace will be called the *modal regression curve* or *trace*. The formal definition is given by

$$\text{Modal regression curve: } \hat{r}(x) = \arg s \max_y \hat{f}(y|x),$$

where the plural on "arg" indicates *all local maxima*. An equivalent definition is

$$\hat{r}(x) = \arg s \max_y \hat{f}(x,y).$$

Some examples reveal the practical advantages of the conditional mode compared to the conditional mean estimators, both robust and nonrobust. Consider the Old Faithful eruption prediction data displayed in Figure 8.4 together with the LOWESS smoother. Because of the (conditional) multimodal behavior for $x > 3$, the resistant kernel estimate LOWESS is actually rougher than a nonresistant LOWESS estimate. But as a scatterplot smoother and data summarizer, any linear smoother is inadequate to summarize the structure. The right frame of the figure shows the modal regression trace, as well as the bivariate ASH from which it was derived. As a measure of center, the mode summarizes the "most likely" conditional values rather than the conditional average. On the other hand, when the conditional density is symmetric, these two criteria match. In this example, they do not.

Observe that the modal trace is much smoother than the usual regression. This smoothness may seem counterintuitive at first. Compare the two regression estimates in the interval (1.5, 2.5), in which the conditional densities are unimodal but not symmetric. The LOWESS estimate faithfully follows the conditional mean, which dips down. In the interval (3.5, 5), the conditional

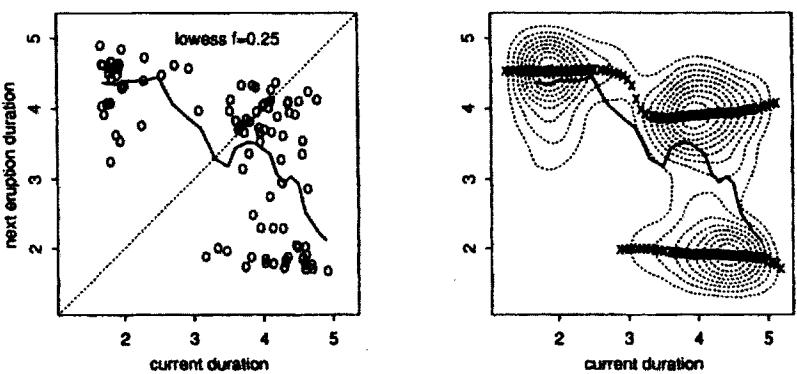


Figure 8.4 Conditional mean (LOWESS) and conditional mode smoothers of the lagged Old Faithful duration data. The conditional mode is displayed with the symbol “o” when above the 25% contour and with an “x” between 5% and 25%. The 45° line is also shown.

densities are bimodal. The bifurcation of the modal regression trace indicates that a single prediction such as the conditional mean misses the structure in the data. Observe that the ordinary regression estimate gives predictions in that interval where there is little probability mass, around $y = 3$ minutes.

A second example provides further evidence for the usefulness of this approach. Figure 8.5 displays the U.S. Lincoln penny thickness data given in Table 4 in Appendix B. Kernel regression, again represented by LOWESS, tracks the changes in penny thickness since World War II. The Treasury restored the penny to its prewar thickness in the 1960s, but reduced its thickness once again in the 1970s. Observe that away from the transition times, the conditional mean and conditional mode estimates are quite similar. However, the conditional mode estimate detects and portrays these abrupt changes. It is interesting to reflect on the common initial perception that the conditional mean estimates appear quite reasonable until compared with the conditional mode estimates.

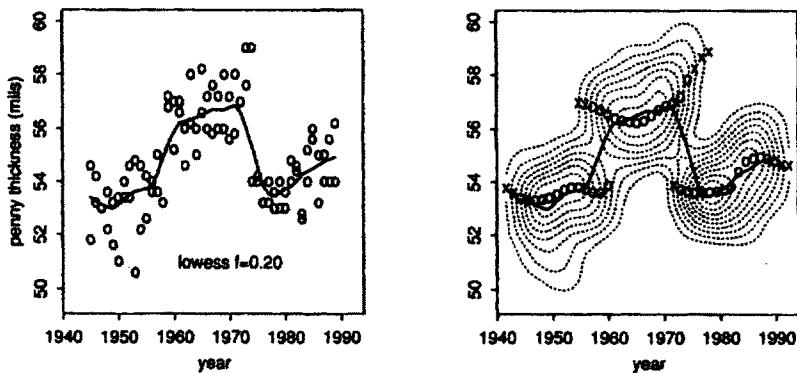


Figure 8.5 Conditional mean and conditional mode smoothers of the U.S. penny thickness data.

The conditional mode traces overlap for several years around 1957 and 1974. This overlapping is the result of the fact that an unknown “boundary” condition exists in the data, and thus the density estimate behaves in a predictable manner. It would be interesting to put internal boundary kernels in this estimation problem (see Problem 4). Note that the conditional mode estimate also detects a smaller rapid change in penny thickness around 1985, but not by a discontinuous jump. The kernel and modal estimates are going in different directions by 1990.

Thus the conditional mode estimate, while introduced as an alternative regression algorithm that is resistant to outliers, has potential beyond that purpose. The problem encountered in the penny data is an example of the “change point” problem where there is a jump in the regression curve; see Siegmund (1988). Displaying the conditional modal trace introduces some interesting design choices. The curves above were computed slice by slice. The estimates are shown with an “x” symbol if the density is less than 25% of the maximum bivariate density value. Other choices could be investigated. In particular, when the data contain many outliers, a consequence is the appearance of an additional modal trace at each of those outliers. Outlier traces will be very short, composed of “x” symbols, and will serve to identify the outliers. The “central” modal traces will be totally unaffected by such outliers; hence, the modal trace is resistant to outliers. Finally, the actual computation of the conditional modes was performed by computing a linear blend of the bivariate ASH estimate using the biweight kernel. Tarter, Lock, and Mellin (1990) have also produced software to compute conditional modes. A related idea is discussed in Sager and Thisted (1982).

8.3.3 L_1 Regression

In the general parametric linear model, the L_1 criterion on the residuals leads to a particularly resistant estimate. Consider data lying on a straight line except for one outlier. As the least-squares fit goes through (\bar{x}, \bar{y}) , it is clear that the fit will be affected by the outlier. On the other hand, the L_1 fit will go through the $n - 1$ points on the line, ignoring the outlier. The reason is simple—any movement of the line incurs a cost for each of the $n - 1$ points and a gain for only the outlying point. This argument makes it clear that the L_1 fit will be resistant to multiple outliers as well.

Wang (1990) investigated the theoretical and practical aspects of the L_1 criterion in nonparametric regression. The idea is essentially equivalent to the criterion in (8.6), with the squared loss replaced by absolute loss. It is well-known that the solution to this problem is the solution of an ordinary linear programming problem (Wagner, 1959). If the i th residual is $\epsilon_i = y_i - x_{(i)}\beta$ and ϵ_i^+ and ϵ_i^- are defined to be the absolute values of the positive and negative parts of ϵ_i , respectively, then

$$\epsilon_i = \epsilon_i^+ - \epsilon_i^- \quad \text{and} \quad |\epsilon_i| = \epsilon_i^+ + \epsilon_i^-.$$

Thus the problem of minimizing the sum of the absolute residuals becomes

$$\begin{aligned} \min_{\epsilon^+, \epsilon^-} & \sum_{i=1}^n (\epsilon_i^+ + \epsilon_i^-) \\ \text{s/t } & \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon^+ - \epsilon^- \quad \text{and} \quad \epsilon^+, \epsilon^- \geq 0. \end{aligned}$$

Modern linear programming (LP) packages can handle tens of thousands of points in reasonable time. To make the procedure nonparametric, the L_1 model is applied locally, with kernel weights on the absolute criterion values. The modified problem remains a LP problem. Wang discusses efficient methods for updating the series of linear programs required to define the L_1 nonparametric regression curve.

The special feature that makes this proposal interesting is its immediate extension to multiple dimensions. Identifying influential points becomes problematic beyond three dimensions (two predictors) and this approach seems quite promising as no interaction (or iteration) is required to eliminate the effects of outliers. An example in Figure 8.6 with a smooth bivariate surface on a 50×50 mesh contaminated by a heavy-tailed Cauchy noise (scale = 0.1) serves to illustrate the power of the approach. The true surface was a portion of a bivariate Normal density over $[0, 1]^2$, where $\mu_i = 1/2$, $\sigma_i = 0.35$, and $\rho = 0$. The local polynomial model was planar and a biweight kernel was introduced in the criterion. The noise is so large that the vertical scale of the raw data surface is truncated, the full vertical range being $(-2.66, 5.91)$. The fit is unaffected by the clusters of outliers. The bias at the corners is eliminated if a local quadratic rather than planar fit is used; see Wang and Scott (1991) for further details and examples.

8.4 REGRESSION IN SEVERAL DIMENSIONS

8.4.1 Kernel Smoothing and WARPing

The extension of the Nadaraya-Watson estimator to several dimensions is straightforward. The multivariate estimator is again a local average of responses, with the product kernel defining the size of the local neighborhood and the specific weights on the responses.

As with the kernel density estimator, the computational burden can be greatly reduced by considering a binning algorithm similar to the ASH. A moment's consideration suggests that binning need only be done in the predictor space, with only the *sum of all the responses* in each bin recorded. For purposes of notation, the sum of the v_k bin responses in the k th bin with average response \bar{y}_k may be computed by $v_k \bar{y}_k$. In the regression (bivariate) binning algorithm (REG-BIN) the quantity $s y_k = v_k \bar{y}_k$ is tabulated, which is the k th bin response sum.

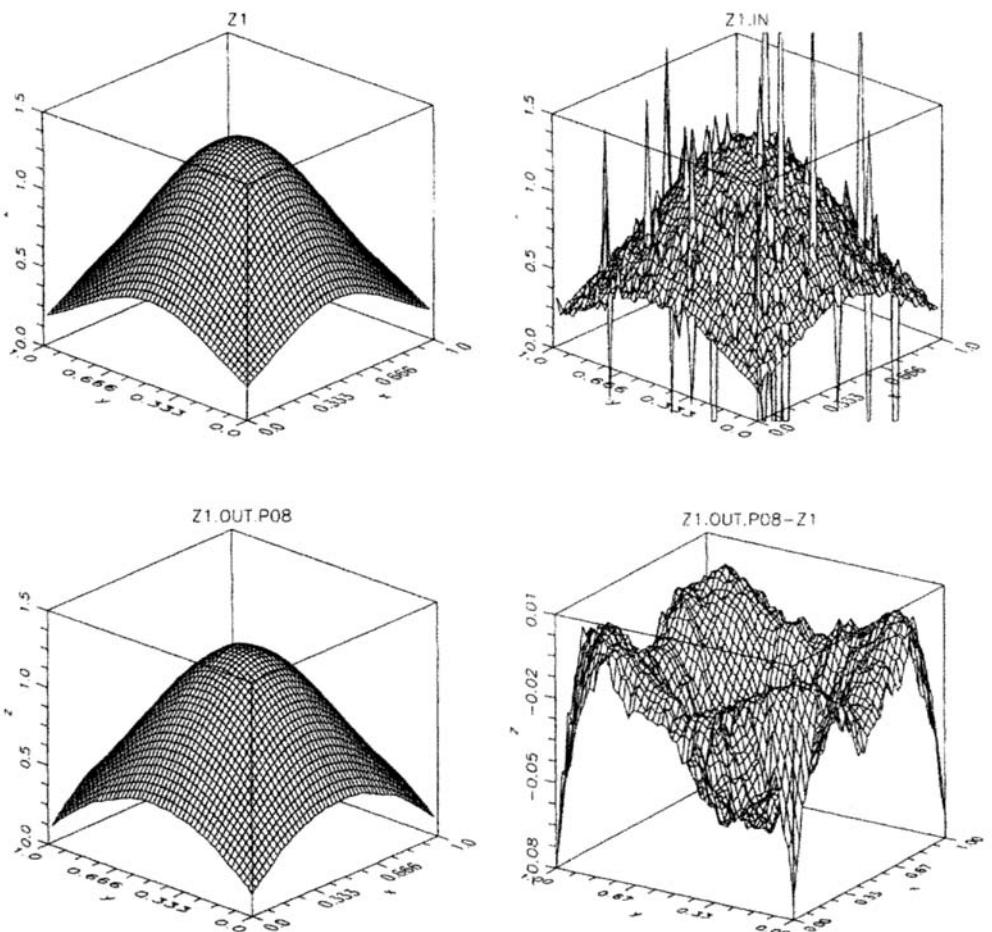


Figure 8.6 Bivariate regression surface, surface contaminated with Cauchy noise, a L_1 nonparametric estimate, and residual surface on a 50×50 mesh.

REG-BIN($x, y, n, a, b, nbin$) Algorithm: (* Bin 2-D regression data *)

```

 $\delta = (b - a)/nbin$ 
for  $k = 1, nbin \{ \nu_k = 0 ; sy_k = 0 \}$ 
for  $i = 1, n \{
    k = (x_i - a)/\delta + 1$  (* integer part *)
    if ( $k \in [1, nbin]$ )  $\nu_k = \nu_k + 1 ; sy_k = sy_k + y_i \}$ 
return ( $\{\nu_k\}, \{sy_k\}$ )

```

The univariate regression version of the ASH (REG-ASH) over the equally spaced mesh $\{t_k\}$ corresponding to the Nadaraya-Watson estimator (8.3) is easily

computed as follows:

$$\hat{r}(x) = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)} \approx \frac{\sum_{k=1}^{nbin} K_h(x - t_k) \nu_k \bar{y}_k}{\sum_{k=1}^{nbin} K_h(x - t_k) \nu_k}.$$

If the regression estimate is computed on the same grid $\{t_k\}$ as the binned data, then the kernel weights need only be computed once. When a finite support kernel is used to compute the weights $w_m(k)$, then the regression estimate in some bins may equal 0/0. The algorithm detects this condition and returns "not-a-number" (*NaN*). The speedup is comparable to that observed with the ASH.

REG-ASH($m, \nu, sy, nbin, a, b, n, w_m$) Algorithm: (* 2-D REG-ASH *)

```

 $\delta = (b - a)/nbin; h = m\delta$ 
for  $k = 1, nbin \{ f_k = 0; r_k = 0; t_k = a + (k - 0.5)\delta \}$ 
for  $k = 1, nbin \{
    \text{if } (\nu_k = 0) \text{ next } k
    \text{for } i = \max(1, k - m + 1), \min(nbin, k + m - 1) \{
        f_i = f_i + \nu_k w_m(i - k)
        r_i = r_i + sy_k w_m(i - k) \}
    \text{for } k = 1, nbin \{ \text{if } (f_k > 0) r_k = r_k/f_k \text{ else } r_k = NaN \}
\}$ 
return (x = { $t_k$ }, y = { $r_k$ }) (* Bin centers & REG heights *)

```

The ASH and REG-ASH are special cases of a general approach to accelerating kernel-like calculations. The basic idea is to round the data to a fixed mesh and perform computations on those data. The idea of using such rounded data was explored by Scott (1981), Scott and Sheather (1985), and Silverman (1982); see also Jones (1989). Härdle and Scott (1988) coined the phrase *weighted averaging of rounded points* or the WARPing method to describe the general approach and applied the WARPing method to other multivariate algorithms.

Silverman (1982) took a different tack with the rounded data, using a fast Fourier transform approach to perform the calculations in the special case of a Normal kernel. The binned data are transformed into the Fourier space, smoothed by a tapering window controlled by a smoothing parameter h , and backtransformed. Such an approach leads naturally to questions of *approximation error* due to rounding in addition to the usual questions of *statistical error*. One advantage of the ASH framework is that it is applied directly on bin counts. A nonzero bin width ($\delta > 0$) limits the statistical accuracy possible. The ASH cannot be applied with a smoothing parameter $h < \delta$. With a kernel approach simply applied to rounded data, $h \rightarrow 0$ may be

computed, and as in the case of unbiased cross-validation provide an artificially attractive choice of smoothing parameter.

8.4.2 Additive Modeling

For more than a few variables, in addition to increased concern about the curse of dimensionality and boundary effects, the kernel method begins to lose its interpretability. There is a need to “constrain” the multivariate kernels in a manner that allows flexible modeling but retains the ease of interpretation of multivariate parametric modeling. *Additive regression models* achieve this goal. The regression surface is modeled by the equation

$$r(\mathbf{x}) = r_0 + \sum_{i=1}^d r_i(x_i) + \epsilon_i .$$

The flexibility is achieved by fitting 1-D linear smoothers for r_i in each dimension. The set of multivariate surfaces that can be generated by this model is not very complex, but it is richer than parametric choices such as x_i or x_i^2 for $r_i(x_i)$. Further flexibility can be achieved by adding in terms of the form $r_{ij}(x_i, x_j)$. However, combinatorial explosion quickly sets in. The level of each additive function is arbitrary, although a constant r_0 can be introduced into the model and each additive function constrained to have average 0.

The fitting of such models takes two extremes. On the one hand, Wahba (1990) has advocated a penalty function approach that estimates the d functions simultaneously as splines. A simpler iterative scheme has been proposed by Friedman and Stuetzle (1981) called *backfitting*. The idea is to begin with an initial set of estimates of r_i , perhaps obtained parametrically, and to iterate over each of the d variables, smoothing the residuals not explained by the $d - 1$ predictors remaining. Any bivariate scatterplot smoother may be used. Simple kernel smoothers may be used as well as the smoothing splines. Friedman and Stuetzle (1981) used *supersmooth* in the more general projection pursuit regression (PPR) algorithm. Hastie and Tibshirani (1990) demonstrate how the backfitting algorithm mimics the Gauss-Siedel algorithm for solving linear equations for certain linear smoothers. In any case, the algorithm usually converges quickly to the neighborhood of a reasonable solution.

The backfitting algorithm is easily implemented using the WARPing regression estimator after removing the mean from $\{y_i\}$:

1. initialize $r_j(x_j) = 0$ (* Backfitting Algorithm *)
2. loop on j until convergence {

$$\epsilon_i \leftarrow y_i - \sum_{j \neq i} r_j(x_j)$$

$$r_j \leftarrow \text{smooth}(\epsilon_i)$$
 }

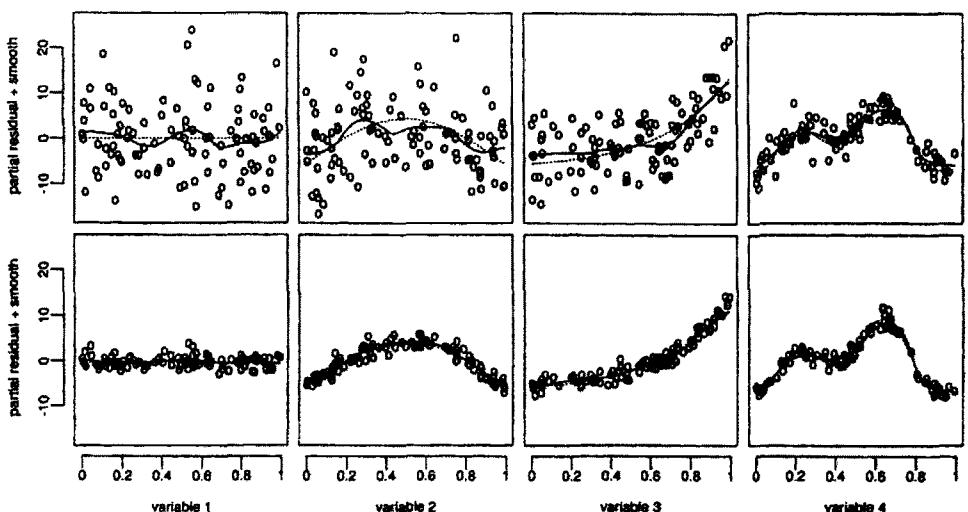


Figure 8.7 Additive model iteration for simulated data from (8.23). The true additive function is shown as a dotted line and the estimated additive function as a solid line. The top row gives the initial loop and the bottom row the final iteration (6 loops).

The REG-ASH additive model (RAM) was applied to a simulation example of Wahba (1990, p. 139) with $\mathbf{x} \in \mathbb{R}^4$, $\epsilon_i \sim N(0, 1)$, $n = 100$ design points sampled uniformly over the hypercube $(0, 1)^4$, and true additive surface

$$r(\mathbf{x}) = 10 \sin \pi x_2 + \exp 3x_3 + 10^6 x_4^{11}(1 - x_4)^6 + 10^4 x_4^3(1 - x_4)^{10}, \quad (8.23)$$

where the true value of the additive function $r_1(x_1) = 0$. In Figure 8.7 the first row displays plots of (x_i, ϵ_i) for the first iteration together with the computed smooth; in the second row, the final iteration is displayed after 5 loops through the algorithm. The initial residual variance was 57.2. Even by the end of the first loop, the residual variation had been significantly reduced to 4.7 and was 1.45 at the final solution.

A second example shown in Figure 8.8 comes from the complete set of gas flow data. The meter was tested at 7 different pressures with between 24 and 51 flows at each pressure. The first variable is log flow rate rescaled to $(0, 1)$. The second variable is log psia also rescaled to $(0, 1)$. The initial residual variance of 0.27 was reduced to 0.12. The relationship with log psia is almost perfectly linear whereas the relationship to log flow rate is more complicated. The accuracy of the meter is affected most by the flow rate but a small effect from the pressure exists as well.

8.4.3 Curse of Dimensionality

An additive model is still a kernel estimator, although with highly constrained kernels, and subject to the same asymptotic problems. The most common

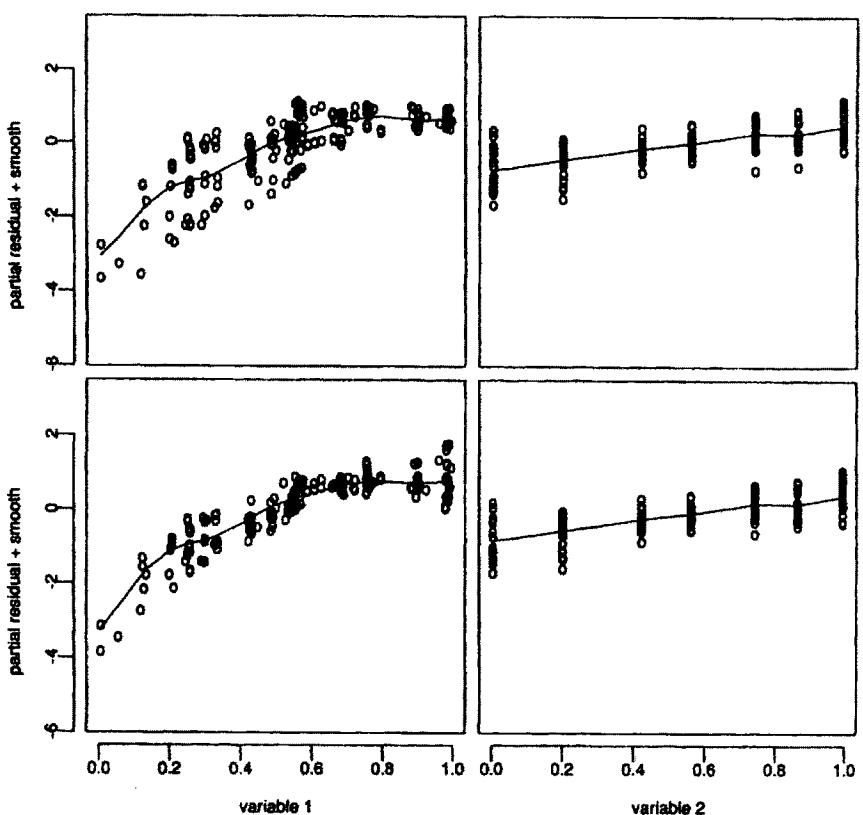


Figure 8.8 Additive model iteration for the gas flow data. The 214 measurements were taken at 7 different pipeline pressures. The additive fits for the initial and final iteration are shown.

problem remains rank deficient predictor data. Some theoretical results require that the data become dense over the hyper-rectangle defining the additive model. This assumption is clearly very strong, and one that presumably is required not in practice but for proving theorems. However, Wahba's example in the preceding section is easily modified to illustrate the effect of the curse of dimensionality. Choose $\mathbf{x} \sim U([0, 1]^4)$ but with pairwise correlations of 0.99. In practice, this is accomplished by applying the inverse of the principal components transformation to points sampled uniformly on the hypercube and then rescaled back to the hypercube. The initial and final additive fits are shown in Figure 8.9. The initial and final residual variances were 6.6 and 2.1. However, as the design points are concentrated along the hyperdiagonal, the individual additive functions are not identifiable since the design space is singular. In fact, only 1 function is required [set $x_2 = x_3 = x_4$ in (8.23)]. In very high dimensions, such singularities are common and the curse of dimensionality applies.

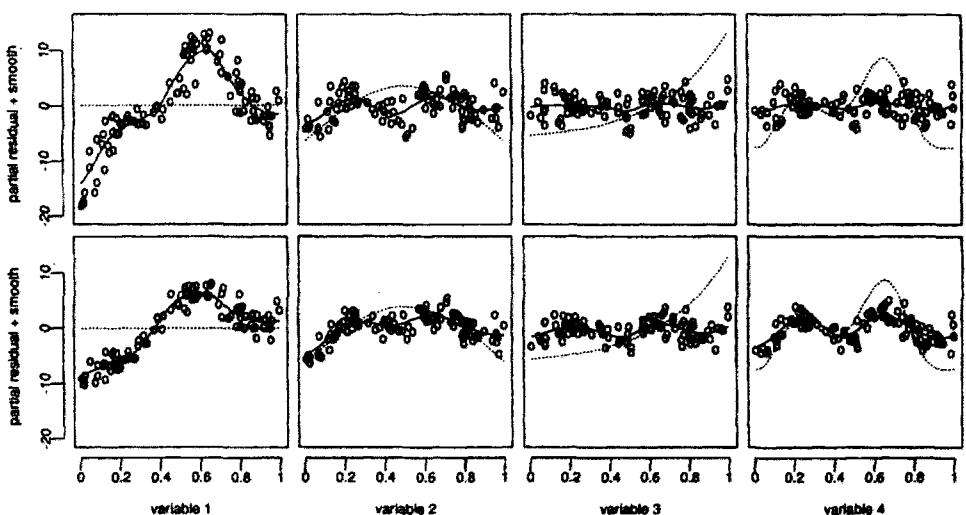


Figure 8.9 Additive model iteration for simulated data from (8.23). The pairwise correlations of the design points are all 0.99. The true additive function is shown as a dotted line and the estimated additive function as a solid line.

An additive model will provide an estimate throughout the hyper-rectangle base, even where there are no data. It might seem straightforward to check if there are data close by, but that leads back to the original density estimation problem and the kernel regression estimator, for which the denominator can be used to determine the “accuracy” of the estimate. Finally, interpreting the shapes of the functions r_i should be performed cautiously and only after careful reflection on the difficulties of estimating the structure in parametric models. Fitting the additive model by permuting the variables and observing whether the order affects the shape of the fits is a good diagnostic tool, as are resampling methods discussed in the next chapter.

As before, if it is possible to project the original data onto a convenient subspace, the nonparametric algorithms should have an enhanced opportunity to do their job. The limitations of additive models are discussed by several authors in the discussion of Friedman (1991). Scott (1991a) also raises some cautionary notes.

Principal components can be applied to the \mathbf{X} matrix before additive modeling. Specifically, the data (\mathbf{X}, y) could be compressed to the form $(\mathbf{X}P, y)$ where P is a projection matrix. However, that may still leave too large a subspace where the response is constant because principal components does not incorporate any of the response information. A new idea, sliced inverse regression (SIR), due to Li (1991) incorporates the response information in an approach designed for Normal data but like principal components has much greater general applicability. The SIR algorithm is a simple extension of the principal components and informative components ideas. The responses y_i are

partitioned into k bins and the mean of the $(d + 1)$ -dimensional vector (\bar{x}_i, \bar{y}_i) computed for each bin to yield the set of k points

$$(\bar{x}_j, \bar{y}_j), \quad j = 1, \dots, k.$$

Principal components is applied to the covariance matrix of the k vectors $\{\bar{x}_j\}$, and those directions with largest eigenvalues are retained as predictors of y .

The basic ideas of SIR are easily illustrated by example. Consider the surface

$$r(x_1, x_2) = 0.5625 - 0.1375x_1 - 0.2875x_2 + 0.0125x_1x_2$$

over $n = 100$ design points sampled from $N(0, 0.4^2 I_2)$ with noise $\epsilon \sim N(0, 0.1^2)$. In Figure 8.10 the outline of the plane defining the true surface (and its projection) is displayed together with the 100 data points. The choice $k = 5$ bins was made so that each bin had 20 points. The 6 tick marks on the y -axis show the locations of the bins. The conditional means for each of the 5 bins were computed and are displayed as diamonds together with a line down to the projection onto the design space. The covariance matrix of the 5 points on the design plane was computed and found to have 2 eigenvalues, given by 0.365 and 0.034. Thus the SIR direction is the eigenvector $(0.667, 0.745)^T$ corresponding to the larger eigenvalue. The ellipse corresponding to the covariance matrix is shown surrounding the 5 means. Notice that the responses are essentially constant in the direction of the other eigenvector. For higher-dimensional data that are not Normal, this simple idea has proven a powerful diagnostic and exploratory tool.

Simple models such as the SIR algorithm have an appeal due solely to the parsimonious solution. A similar simple model is the regression tree model for

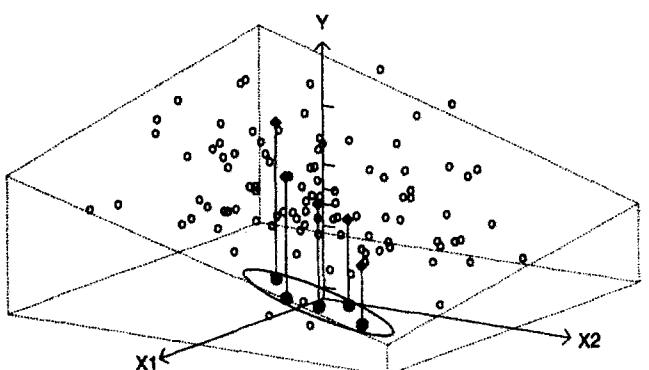


Figure 8.10 Example of the SIR dimension reduction technique; see the text.

the regression surface that is piecewise constant

$$\hat{r}(\mathbf{x}) = \sum_{i=1}^p c_i I\{\mathbf{x} \in N_i\}, \quad \text{where } \bigcup_{i=1}^p N_i = \mathbb{R}^d \quad \text{and} \quad N_i \cap N_j = \emptyset.$$

The so-called CART regression model has had much success in practice for classification problems; see Breiman et al. (1984).

8.5 SUMMARY

The choice of kernel in the regression setting is more flexible than in the density estimation setting. In particular, the use of higher-order kernels is more attractive because the issue of nonnegativity vanishes. Care must be exercised at the boundaries to avoid anomalies in the estimate. Boundary kernels may induce artifacts as often as they correct artifacts.

The direct use of kernel estimates for regression may or may not be the best choice. Certainly, issues of resistance to outliers are of practical importance. Naive kernel regression algorithms provide no protection against influential points. The use of modal regression is to be encouraged. Additive models provide the most easily understood representation of multidimensional regression surfaces, when those surfaces are not too complex. However, the curse of dimensionality suggests that the additive functions can be quite unreliable if the design data are nearly singular. The use of the L_1 norm as a robust measure for local polynomial fitting can be expected to grow dramatically as the availability of good LP codes increases. At the least, the L_1 estimator provides an excellent diagnostic tool for data with many potential outliers and influential points.

PROBLEMS

1. Consider the fixed design with $x_i = i/n$. Show that the MSE of the Nadaraya-Watson estimator in this case is

$$\text{MSE}\{\hat{r}(x)\} = \frac{\sigma_\epsilon^2 R(K)}{nh} + \frac{1}{4} h^4 \sigma_K^2 R(r'').$$

Hint: Use the approximation $\sum r''(x_i)^2 \cdot (1/n) \approx \int_0^1 r''(x)^2 dx$.

2. Verify that the smoothing spline proposed in Equation (8.21) satisfies the Euler-Lagrange equation, assuming that $\hat{r}_\lambda^{(1)}$ and $\hat{r}_\lambda^{(2)}$ do.
3. Prove that the weight vector \mathbf{w}_x in Equation (8.18) is linear in x .

4. Construct an internal boundary modification to the penny data displayed in Figure 8.5.

5. Consider a bivariate additive model for a function $f(x, y)$, where each additive function is a piecewise constant function. Specifically,

$$f_1(x) = \sum a_i I_{A_i}(x) \quad \text{and} \quad f_2(y) = \sum b_j I_{B_j}(y),$$

where $\{A_i\}$ and $\{B_j\}$ are bins along the x and y axes, respectively.

- (a) Given data $\{(x_i, y_i, z_i), i = 1, n\}$, find the equations that correspond to the least-squares estimates of $\{a_i\}$ and $\{b_j\}$.
- (b) Compute and plot this estimator for the gas-flow-accuracy data, for several choices of the mesh.

CHAPTER 9

Other Applications

9.1 CLASSIFICATION, DISCRIMINATION, AND LIKELIHOOD RATIOS

The problems of classification and discrimination are closely related. In each instance, the data, $\mathbf{x}_i \in \mathbb{R}^d$, are assumed to comprise K clusters. If the number of clusters K is known, and a training sample of data is available from each cluster, then the *discrimination problem* is to formulate rules for assigning new unclassified observations to one of the clusters. If the data set is made up of an unknown number of clusters, then the *classification problem* is to stratify the unlabeled data into proposed clusters.

The general principles of discrimination may be understood in the simplest case when $K = 2$. Using a notation common in biostatistical applications, the 2 categories are labeled “positive” or “negative” depending upon the presence or absence of a disease or upon the finding of a diagnostic test. The *prior probabilities* of the categories are denoted by $P(+)$ and $P(-)$, both known and satisfying $P(+) + P(-) = 1$. The ratio $P(+):P(-)$ or $[P(+)/P(-)]$:1 is called the *prior odds* in favor of disease. There is a simple 1:1 relationship between the probability p of an event and the corresponding odds o in favor of the event:

$$o = \frac{p}{1 - p} \quad \text{and} \quad p = \frac{o}{1 + o}. \quad (9.1)$$

When $p = \frac{1}{2}$ the odds are even or 1:1.

Suppose that a discrete covariate such as race or sex is thought to be associated with the risk of disease. Let the event A denote the presence of a covariate. Then Bayes' theorem states that

$$P(+|A) = P(A|+P(+)/P(A)) \quad (9.2)$$

and

$$P(-|A) = P(A|-)P(-)/P(A), \quad (9.3)$$

where $P(+|A)$ and $P(-|A)$ are the *posterior probabilities* given the additional knowledge of the information contained in the event A . The conditional probabilities $P(A|\pm)$ may be estimated through epidemiological studies, or by experiment. Taking the ratio of the 2 equations in (9.2) and (9.3), we have

$$\frac{P(+|A)}{P(-|A)} = \frac{P(A|+)}{P(A|-)} \times \frac{P(+)}{P(-)}, \quad (9.4)$$

where $P(A|+)/P(A|-)$ is the *likelihood ratio* (LR). The ratio on the left-hand side is called the *posterior odds* in favor of the disease. In odds form, (9.4) is called *Bayes' rule*:

$$o(+|A) = \frac{P(A|+)}{P(A|-)} o(+).$$

Depending on whether the likelihood ratio is greater than or less than 1, the posterior odds in favor of disease are increased or decreased, respectively.

If the (multivariate) covariate x is a continuous random variable, then Bayes' rule still holds when the likelihood ratio is replaced with the corresponding ratio of the probability densities at x for the two groups.

Bayes' rule:

$$o(+|x) = \frac{f_+(x)}{f_-(x)} o(+). \quad (9.5)$$

Again the odds change with the likelihood ratio. An unclassified (or unlabeled) point z would be assigned to the "+" group if $o(+|z) > 1$ and "-" otherwise.

In practice, the *shifted-Normal model* is the model most commonly assumed for the pair of densities f_+ and f_- . Specifically, in the univariate case

$$f_+(x) = N(\mu_1, \sigma^2) \quad \text{and} \quad f_-(x) = N(\mu_0, \sigma^2).$$

A little algebra reveals that the logarithm of Bayes' rule is

$$\log [o(+|x)] = \frac{\mu_1 - \mu_0}{\sigma} \left[\frac{1}{\sigma} \left(x - \frac{\mu_0 + \mu_1}{2} \right) \right] + \log [o(+)].$$

That is, the log-odds change linearly in x . The posterior odds equal the prior odds when the covariate x is midway between the two population means; see Figure 9.1. If x is measured in standard units, then the slope of the log-likelihood ratio is $(\mu_1 - \mu_0)/\sigma$, which is called the *coefficient of detection*. The greater (in magnitude) the coefficient of detection, the greater

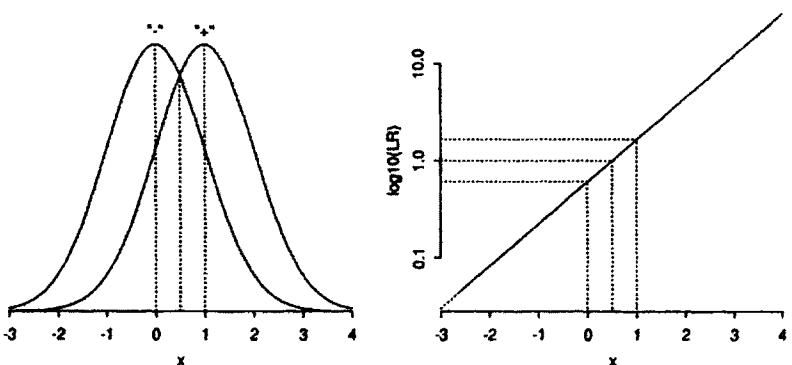


Figure 9.1 Shifted-Normal model of two populations for a single risk factor or covariate. The log-likelihood ratio is linear in x .

the predictive power of Bayes' rule. The multivariate shifted-Normal model is considered in Problem 2.

Nonparametric discrimination simply involves estimation of the unknown densities in the likelihood ratio using ASH estimates. Consider the two populations in the plasma lipid study. By imaging the coronary arteries using angiography, 320 males were found to have significant occlusion (narrowing) while 51 had none. There is no obvious deviation from Normality in the group of 51; see frame (a) in Figure 9.2. The Normal fit to the 320 males shown in frame (b) is not obviously unsatisfactory, although the superposition of the Normal estimate with the ASH estimate in frame (d) indicates the lack-of-fit clearly. Frames (c) and (e) contain the parametric estimates that lead to the contours of the likelihood ratio (on a \log_{10} scale). If the covariance matrices are assumed to be identical, then the \log_{10} -likelihood ratio surface will be a plane (see Problem 2) and the LR contours a series of parallel lines. Looking only at areas with significant probability mass, the logarithmic LR estimates range from about -0.5 to 1.0, which suggests moderate predictive power.

The nonparametric LR shown in frame (f) used the ASH estimate in the numerator and the Normal fit in the denominator. As ASH estimates do not extrapolate well, the LR is plotted only over the region where the ASH estimate was greater than 5% of the modal value. The LR surface is not monotonic increasing, reflecting the multimodal nature of the data. The parametric LR estimates are never grossly incorrect; however, the parametric fit seems to overestimate the importance of high cholesterol and underestimate the importance of high triglyceride concentrations. (Fitting a single covariance matrix to the data results in heavier tails along the cholesterol axis because of the bimodality in the data along that axis.) In the nonparametric LR plot, the relative contributions to risk of the two lipids seem nearly equal. Similar complex interactions were observed in a much larger follow-up study by Scott et al. (1980). That study included age, smoking behavior, and history of hypertension as covariates. For various combinations of levels of the 3 additional covariates, the lipid

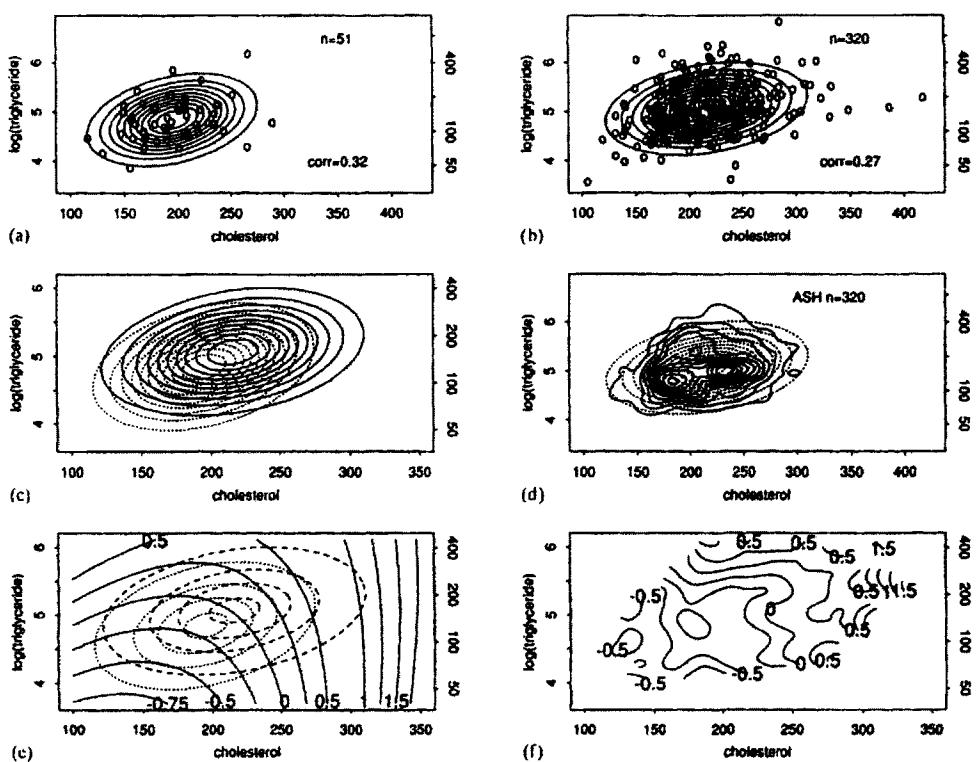


Figure 9.2 Risk analysis of plasma lipid data. (a) Normal fit “-” group. (b) Normal fit “+” group. (c) Overlay of 2 Normal fits. (d) Overlay of ASH of f_+ [biweight kernel $h = (21.7, 0.33)$] and Normal fit to “+” group. (e) Contours of parametric $\log_{10}(LR)$. (f) $\log_{10}(LR)$ nonparametric estimate.

data exhibited bimodality and the risk contours suggested that the predictive power of triglycerides was again underestimated in a shifted-Normal model.

In the author's opinion, bivariate contour plots of surfaces that are not monotone or uncomplicated are much less satisfactory than perspective plots. Perspective plots of the two \log_{10} -likelihood ratio surfaces in Figure 9.2 are shown in Figure 9.3. The parametric LR estimates around the border of the plot are clearly driven by the Normal model, not by the presence of significant data. Caution should be exercised when extrapolating outside the range of the data.

Prediction and classification in \mathbb{R}^d using nonparametric techniques have a long history of success. Habbema, Hermans, and Van Der Broek (1974) reported excellent results using a multivariate Normal kernel estimator for classification. Titterington et al. (1981) used mixed variables. The classification depends only upon the contour where the LR equals 1 [$\log_{10}(LR) = 0$]. As in the univariate case, a new observation is classified as positive if $LR(x) > 1$ and as negative otherwise. Thus classification is a relatively easy task for kernel estimation, as only a relatively small fraction of the data is near the boundary where the

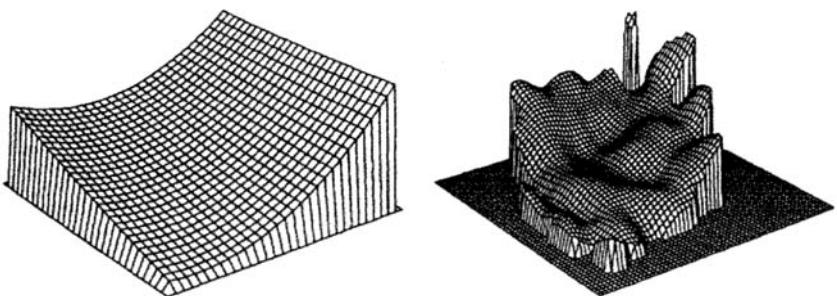


Figure 9.3 Perspective plots of the \log_{10} likelihood ratio surfaces in Figure 9.2. The range of the vertical axes is $(-0.91, 0.91)$ in both frames, corresponding to a range of odds in favor of disease from 0.15:1 to 8.1:1.

likelihood ratio is 1. This ease seems to hold over a wide range of smoothing parameters. For example, Hand (1982, p. 88) makes a connection with the nearest-neighbor classification rule, which assigns a new point x to a category based on the category of the training sample point closest to it in Euclidean norm. This behavior is precisely that of a Gaussian kernel classifier as the bandwidth $h \rightarrow 0$! This equivalence may be seen by noting that if the distance between x and the nearest data point is δ and to the second nearest point $\delta + \epsilon$, then the relative contribution to the kernel estimate $\hat{f}(x)$ is

$$\frac{\phi[-(\delta + \epsilon)^2/(2h^2)]/h}{\phi[-\delta^2/(2h^2)]/h} \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Therefore, for a very small choice of bandwidth, the Gaussian kernel estimates for the 2 training samples are dominated by the nearest sample point in each, with the classification going to the nearest one.

This connection may be extended to very large bandwidths as $h \rightarrow \infty$. As h dwarfs the span of the data, each kernel estimate converges pointwise to a single kernel centered on the sample mean. Therefore, the 2 kernel estimates at the new observation reflect only the distance to the sample mean of each cluster, the classification being made to the closer sample mean. This classification rule is known as average linkage. Thus for extreme values of the bandwidth, the kernel classifier mimics two well-known classification algorithms known to work reasonably well in practice. Of course, for bandwidths in between, the classifier benefits from a good estimate. Therefore, the kernel classifier should perform well in almost all situations.

For the LANDSAT data, the actual agricultural use for each pixel was determined on the ground by trained observers. The primary crops observed were sunflower (1,200 pixels), wheat (1,010), and barley (1,390). Scott (1985c) compared the observed misclassification rates of the trivariate Normal fits and the ASH fits to the training sample data from these 3 clusters. The results are summarized in Table 9.1. Visually, the ASH estimates were egg-shaped,

slightly longer on one end of the egg. The result of the small skewness is a bias in the location of the parametric means, and asymmetric error rates for wheat and barley compared to the ASH estimates. Improved error rates were obtained by considering not only the classification prediction of each pixel but also the predictions for its 8 neighbors. This modification takes advantage of the obvious spatial correlation among neighboring pixels. The final smoothed classification was based on the majority prediction in each block of 9 pixels. The resulting spatial classification appears much smoother visually. The improvement in prediction is significant, as shown in the final column of Table 9.1.

For smaller samples, a jackknife or cross-validation procedure should be used (Efron, 1982). Each sample in the training set is removed from the density estimator before classification. The reuse procedure removes the somewhat overoptimistic error rates obtained when all the data in the training sample are used to form the density estimates and then those same data are used to test the classification performance.

An alternative approach when there are 2 clusters is to treat the problem as a nonparametric regression problem, where the outcome Y is 0 or 1, corresponding to the “-” and “+” clusters. Thus the data are $\{\mathbf{x}_i, y_i\}$. Now,

$$E Y_i = 0 \cdot \Pr \{Y_i = 0\} + 1 \cdot \Pr \{Y_i = 1\} = \Pr (+ | \mathbf{X} = \mathbf{x}_i),$$

which can be estimated from the odds estimate in Equation (9.5) via Equation (9.1). A kernel regression estimate of Y will fall between 0 and 1 if a positive kernel is used. If the sizes of the training samples reflect the relative frequency of the 2 classes and the same smoothing parameter is used for both classes, then

$$\hat{y} = \hat{r}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i K_h(\mathbf{x} - \mathbf{x}_i)}{\sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i)} = \frac{n_1 \hat{f}_+(\mathbf{x})}{n_0 \hat{f}_-(\mathbf{x}) + n_1 \hat{f}_+(\mathbf{x})},$$

which gives identical predictions to the nonparametric likelihood ratio estimate using the same smoothing parameter [as $p = o/(1 + o)$ and $o = \hat{f}_+/\hat{f}_-$].

Table 9.1 Classification Cross-Tabulations Based on Trivariate Gaussian and ASH Fits to the LANDSAT Data^a

		PREDICTED	Sunflwr	Wheat	Barley	% Correct	Smoothed
NORMAL	Sunflwr		1,191	9	0	99.3%	100.0%
	Wheat		10	665	335	65.8%	80.3%
	Barley		10	314	1,066	76.7%	93.7%
ASH	Sunflwr		1,194	5	0	99.5%	100.0%
	Wheat		7	773	230	76.5%	93.7%
	Barley		3	361	1,026	73.8%	89.9%

^aThe first 3 columns summarize the predictions of the classifier using the training data. The last column summarizes the rates using a classification rule based on a majority rule of a pixel and its 8 neighbors.

The obvious disadvantage of the regression approach is the use of the same smoothing parameter for both populations. If $P(+) \neq P(-)$, then the sample sizes will not be equal and h_+ and h_- should not be equal either. Likewise, the same conclusion holds if the shapes of the 2 densities are not identical. For example, with the lipid data, neither condition supports the use of the regression approach. Hall and Wand (1988b) observed that the classification rule $\sigma(+|x) > 1$, together with (9.5), is equivalent to the inequality $P(+)f_+(x) - P(-)f_-(x) > 0$ and investigated a single optimal bandwidth for this weighted density difference.

9.2 MODES AND BUMP HUNTING

9.2.1 Confidence Intervals

The asymptotic distribution of the sample mode provides the basis for finding a confidence region for the true mode. The same approach may be extended to finding confidence regions for the modal regression curves.

Kernel estimates are actually sums of rows in a triangular array of random variables, $\{Y_{n1}, Y_{n2}, \dots, Y_{nn}; n \geq 1\}$, as is demonstrated below. Consider conditions under which sums of triangular arrays converge to Normality. Suppose $Y_{nj} \sim F_{nj}$ (a cdf) with $\mu_{nj} = E Y_{nj}$ and

$$A_n = \sum_{j=1}^n \mu_{nj} \quad \text{and} \quad B_n^2 = \text{Var} \left\{ \sum_{j=1}^n Y_{nj} \right\}.$$

Then Serfling (1980) proves the following proposition: If

$$\sum_{j=1}^n E(Y_{nj} - \mu_{nj})^4 = o(B_n^4) \quad \text{then} \quad \sum_{j=1}^n Y_{nj} \sim AN(A_n, B_n^2). \quad (9.6)$$

From Equation (6.18) the kernel estimate of the derivative is a sum of random variables from a triangular array:

$$\hat{f}'(x) = \sum_{j=1}^n Y_{nj} \quad \text{where } Y_{nj} = \frac{1}{nh^2} K' \left(\frac{x - x_j}{h} \right). \quad (9.7)$$

The asymptotic normality of the mode follows from the asymptotic normality of $\hat{f}'(x)$. The first 4 noncentral moments of Y_{nj} are $f'(x)/n + h^2 \sigma_k^2 f'''(x)/(2n)$, $f(x)R(K')/(n^2 h^3)$, $-f'(x)R((K')^{3/2})/(n^3 h^4)$, and $f(x)R((K')^2)/(n^4 h^7)$; see

Problem 4. Therefore,

$$A_n = \sum_{j=1}^n \mu_{nj} \approx f'(x) + h^2 \sigma_K^2 f'''(x)/2$$

$$B_n^2 = n \operatorname{Var}\{Y_{nj}\} \approx f(x)R(K')/(nh^3).$$

Now the 4th noncentral and central moments are the same to first order, so

$$\sum_{j=1}^n E(Y_{nj} - \mu_{nj})^4 \approx \frac{f(x)R((K')^2)}{n^3 h^7} \quad \text{and} \quad B_n^4 \approx \frac{f(x)^2 R(K')^2}{n^2 h^6};$$

their ratio is $O((nh)^{-1})$, which vanishes if $nh \rightarrow \infty$ as $n \rightarrow \infty$. Thus the conditions of Serfling's proposition hold and

$$\pm\{\hat{f}'(x) - [f'(x) + h^2 \sigma_K^2 f'''(x)/2]\} \sim \operatorname{AN}(0, f(x)R(K')/(nh^3)).$$

Let $\hat{\theta}_n$ be a sequence of sample modes converging to a true mode at $x = \theta$. Then $f'(\theta) = 0$, $\hat{f}'(\hat{\theta}_n) = 0$, $f'(\hat{\theta}_n) \approx f'(\theta) + (\hat{\theta}_n - \theta)f''(\theta) = (\hat{\theta}_n - \theta)f''(\theta)$, and $f'''(\hat{\theta}_n) \approx f'''(\theta)$. Hence,

$$\pm\{0 - [(\hat{\theta}_n - \theta)f''(\theta) + h^2 \sigma_K^2 f'''(\theta)/2]\} \sim \operatorname{AN}(0, f(\theta)R(K')/(nh^3))$$

or

$$(\hat{\theta}_n - \theta) \sim \operatorname{AN}\left(-\frac{h^2 \sigma_K^2 f'''(\theta)}{2f''(\theta)}, \frac{f(\theta)R(K')}{f''(\theta)^2 nh^3}\right),$$

ignoring a $O(1/nh^3)$ term in the bias that comes from the next term in the Taylor's series, $(\hat{\theta}_n - \theta)^2 f'''(\theta)/2$. The variance is $O(1/nh^3)$ and the squared bias is $O(h^4)$; therefore, optimally, $h = O(n^{-1/7})$ with corresponding $\operatorname{MSE} = O(n^{-4/7})$, which are the same rates as for the derivative estimator itself. However, if $h = O(n^{-1/5})$, then the same result holds but the bias is asymptotically negligible and the MSE is dominated by the variance, which is of order $n^{-2/5}$. Hence a $100(1 - \alpha)\%$ confidence interval in this case is

$$\theta \in \hat{\theta}_n \pm z_{\alpha/2} [f(\theta)R(K')/(f''(\theta)^2 nh^3)]^{1/2}.$$

Replacing $f''(\theta)$ with $\hat{f}''(\hat{\theta}_n)$ is often adequate; otherwise, an auxiliary bandwidth must be introduced. Compare this result to the FP approximation obtained in Section 4.1.4.

For the modal regression problem, let $\hat{\theta}_{x,n}$ denote a conditional mode of $\hat{f}(x, y)$ that converges to θ_x . Then defining

$$Y_{nj} = K((x - x_j)/h_x)K'((y - y_j)/h_y)/(nh_x h_y^2),$$

a similar approach shows that Y_{nj} is asymptotically Normal and that

$$\mathbb{E} Y_{nj} \approx f_y(x, y) + \frac{1}{2} h_y^2 \sigma_K^2 f_{yyy}(x, y); \quad \text{Var } Y_{nj} \approx \frac{f(x, y)R(K)R(K')}{nh_x h_y^3}$$

$$\theta_x \in \hat{\theta}_{x,n} \pm z_{\alpha/2} \left[f(x, \theta_x)R(K)R(K') / \left(f_{yy}(x, \theta_x)^2 nh_x h_y^3 \right) \right]^{1/2},$$

where the subscripts on f indicate partial derivatives.

9.2.2 Oversmoothing for Derivatives

Sample modes that are not “real” tend to have confidence intervals that overlap with neighboring sample modes. If a larger bandwidth of optimal order $n^{-1/7}$ is chosen, then the bias becomes significant. Choosing a bandwidth for the derivative could be done by cross-validation (Härdle and Marron, 1988), but this approach is even less reliable than for the density estimate itself. Fortunately, the oversmoothing approach can be extended to f' and f'' as well. The oversmoothed bandwidths can be used as a point of reference for cross-validation or to provide a conservative estimator for modes and bumps. From Theorem 6.2, the functionals to minimize are $R(f'')$ and $R(f'^v)$ for the first and second derivatives, respectively, subject to the constraint $\int x^2 f = 1$. The solutions are

$$f_3^{\text{OS}}(x) = \frac{315}{256} (1 - x^2)_+^4 \quad \text{and} \quad f_4^{\text{OS}}(x) = \frac{693}{512} (1 - x^2)_+^5,$$

leading to the inequalities

$$R(f'') \geq 14,175 \sigma^{-7} 11^{-9/2} \quad \text{and} \quad R(f'^v) \geq 1,091,475 \sigma^{-9} 13^{-11/2}.$$

Substituting into the expressions for the optimal bandwidths in Theorem 6.2 for the choice $K = N(0, 1)$ leads to

Oversmoothed f' :	$h_{f'}^* \leq 1.054 \sigma n^{-1/7} \equiv \text{hos}(f')$	(9.8)
Oversmoothed f'' :	$h_{f''}^* \leq 1.029 \sigma n^{-1/9} \equiv \text{hos}(f'').$	

These formulas are nearly optimal for estimating the derivatives when the data are Normal. For a sample size of $n = 100$, $\text{hos}(f)$, $\text{hos}(f')$, and $\text{hos}(f'')$ equal 0.455, 0.546, and 0.617. When $n = 1,000$ the values are 0.287, 0.393, and 0.478 [which is only 1.66 times $\text{hos}(f)$]. These values may be rescaled for use with other more computationally tractable kernels using an equivalent kernel rule similar to that in Equation (6.27), but with the r th derivative of the kernel (compare h^* in Theorems 6.1 and 6.2); see Problem 5.

9.2.3 Critical Bandwidth Testing

Bump hunting falls precisely into the classification category. Given a density estimate, the modes and bumps are easily located. Some of those modes and bumps may be spurious and should be smoothed away locally. Good and Gaskins (1980) proposed an iterative procedure for computing the odds in favor of the existence of a bump's "reality" over an interval (a_i, a_{i+1}) with their maximum penalized likelihood estimator. Assume that the data are bin counts. Briefly, the counts in the bins covering (a_i, a_{i+1}) were reduced to match the probability mass in those bins computed from a MPL density estimate; then all the bin counts were rescaled so that the sum of the bin counts was again n . (The MPL criterion can easily accommodate the notion of noninteger bin counts.) As all density estimates are biased downward at bumps, the effect was to effectively eliminate the bump by "surgery" in the authors' terminology. Usually, a dozen iterations were sufficient to completely eliminate any hint of the bump. To evaluate the odds in favor of the bump being real, the penalized log-likelihood of the original data for the MPL estimate was computed with and without the bump. Each bump is surgically removed and tested individually. For the LRL data discussed in Section 3.5, Good and Gaskins found that the loss in likelihood when removing bumps 3, 8, 11, and 12 seemed small (numbering the bumps from the left). The authors found that the exact choice of smoothing parameter within a narrow range did not affect the estimated odds very much. As bumps may or may not contain a mode, the evidence in favor of a mode is indirect.

For univariate data, Silverman (1981) suggests a conservative test of the null hypothesis that the unknown density has at most k modes. The role played by the bandwidth is important, because the number of sample modes in the density estimate can be anything between 1 and n , as h ranges over $(0, \infty)$. Surprisingly, the number of modes need not be monotonic increasing as $h \rightarrow 0$. For example, Hart (1985) shows that the property of monotonicity need not hold even if the kernel K is unimodal. Silverman (1981) showed that the property of monotonicity does hold for the Normal kernel.

Starting with a large value of h and a Normal kernel, clearly there exists a bandwidth h_1 such that \hat{f} is unimodal for all $h > h_1$ but not for $h < h_1$. Similarly, \hat{f} is bimodal for $h_1 > h > h_2$ but not if $h < h_2$, and so on. Silverman calls the bandwidths where $h = h_k$ the critical bandwidths, corresponding to a density estimate which has exactly k modes and 1 saddle point. The rationale is simple. If the true density has fewer than k modes, then the density estimate with $h = h_k$ will contain several small spurious modes that will tend not to appear in the bootstrap estimates so that $p \approx 0$ will be observed. If the true density has more than k modes, then the saddle point in the density estimate with $h = h_k$ will in fact be close to a true mode. In the bootstrap samples, perhaps 30–50% of the estimates will show a sample mode there so that the null hypothesis is not rejected. The test is performed by computing an estimate of the p -value by resampling from \hat{f}_{h_k} , which is the same smoothed bootstrap

sample used in cross-validation by Taylor (1989) in Section 6.5.1.4. For each bootstrap sample, the Normal kernel estimate is computed using the bandwidth $h = h_k$ and the number of modes counted. The p -value is the observed fraction of estimates where the number of modes exceeds k .

The test is supported by theory that suggests that the critical bandwidth converges to 0 if k is less than or equal to the true number of modes, while the critical bandwidth does not converge to 0 otherwise. Izenman and Sommer (1988) have used the test to examine differences in thicknesses of new issues of stamps.

Matthews (1983) has experimented with this test and concluded “that when the underlying density is k -modal, the p -values for less than k modes may be large and stable, but that the p -values for k modes or more are often highly variable.” With a distinctly bimodal mixture density $0.5[\phi(x) + \phi(x - 3.2)]$, he computed the number of times the test concluded that the density was exactly bimodal with sample sizes of 40, 200, and 1,000. In 10 simulations each, the test reached the correct conclusion 30%, 40%, and 60% of the time, respectively. Further research directed towards the use of adaptive bandwidths would be interesting.

9.2.4 Other Approaches

For the particular instance of testing the unimodality of the unknown density, Hartigan and Hartigan (1985) propose the “dip test.” The dip test is based on the distance between the sample distribution function and the closest distribution function in the family of unimodal densities. The authors present an algorithm for computing the test statistics and give tables of critical values.

The most powerful technique may simply be the parametric general Normal mixture model (Titterington, Smith, and Makov, 1985). In particular, densities of the form $0.5\phi(-\mu, 1) + 0.5\phi(\mu, 1)$ are not bimodal unless $|\mu| > 1$. Karl Pearson (1894) even considered the estimation of 2 general Normal mixtures based on estimates of the first 5 sample moments and the solution of a ninth-degree polynomial. More recent references include Day (1969), Aitkin and Tunnicliffe Wilson (1980), and Hathaway (1982), who use the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) to fit the model. Roeder (1990) used the EM algorithm with some data pertaining to the distribution of galaxies and reported many practical problems using a mixture model with general parameters, such as delta spikes at isolated data points.

9.3 SPECIALIZED TOPICS

9.3.1 Bootstrapping

Bootstrapping provides a means of estimating statistical error by resampling. The ordinary bootstrap resamples from the empirical density function, while the smoothed bootstrap resamples from the kernel density estimate $\hat{f}_h(\mathbf{x})$ given

in Equation (6.37) or (6.43). In practice, bootstrap samples are generated by treating the kernel estimator as a mixture density; that is, the kernel estimate is viewed as the mixture of n equally probable kernels. To generate a smoothed bootstrap sample of the same size as the original multivariate sample $\{\mathbf{x}_i, i = 1, \dots, n\}$:

1. Generate $\{j_1, j_2, \dots, j_n\} \sim U(1, 2, \dots, n)$ with replacement.
2. Generate $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ from the scaled multivariate kernel.
3. Return $\mathbf{x}_i^* = \mathbf{x}_{j_i} + \mathbf{t}_i, i = 1, \dots, n$.

where $U(1, 2, \dots, n)$ is the discrete uniform density and \mathbf{x}^* is a bootstrap sample. The precise details for step 2 depend upon the form of the multivariate estimator being used. The most common case is the product kernel in (6.37) with univariate kernel K and smoothing parameters $\{h_1, \dots, h_d\}$. Then step 2 becomes

- 2a. Generate $\{u_{i1}, \dots, u_{id}, i = 1, \dots, n\}$ from $K(u)$.
- 2b. $\{\{t_{ij} = h_j u_{ij}, j = 1, \dots, d\}, i = 1, \dots, n\}$.

For the ordinary bootstrap, which corresponds to $h_j = 0$, only the first and third steps are required as $\mathbf{t}_i = \mathbf{0}$. The above algorithm is easily modified for adaptive density estimates.

Generating pseudo-random samples from specific kernels may be accomplished by the probability sampling approach $F_K^{-1}(U)$ where $U \sim U(0, 1)$. Packages are widely available for generating Normal samples. The biweight kernel may be sampled by the following transformation of 6 uniform samples:

$$\log(u_1 u_2 u_3 / u_4 u_5 u_6) \div \log(u_1 u_2 u_3 u_4 u_5 u_6), \quad (9.9)$$

although the computations should not be organized in this fashion due to underflow; see Problem 7.

Generating Monte Carlo samples from \hat{f} is identical to the bootstrap. Certain density estimates are not amenable to resampling. For example, what does it mean to resample from a kernel that is negative? How should a sample be drawn from the nearest-neighbor estimator when its integral is not finite? Ad hoc procedures may be envisioned, but some thought should be given to what the corresponding density estimate is after any modification.

9.3.2 Confidence Intervals

The difficulty with constructing confidence intervals for nonparametric estimates is the presence of bias in the estimates. Eliminating that bias is not possible, but on average the variance dominates the MSE. A practical compromise is to estimate the variance pointwise and construct 2 standard error bars around the estimate; see Problem 3. Such an interval is not a confidence interval for the unknown density function but rather a confidence interval for the nonparametric estimate.

The bootstrap may be used to obtain pointwise error bands for the positive kernel estimate $\hat{f}(x; h, \{x_i\})$. If samples from the original (unsmoothed) bootstrap are used to compute $\hat{f}(x; h, \{x_i^*\})$, then sample percentiles of these bootstrapped kernel density estimates may be superimposed on the original density estimate. In Figure 9.4, error bars (based on the 10th and 90th percentiles) of 200 bootstrap resamples from the silica data ($n = 22$) are shown, with the biweight kernel and $h = 3$. The error bars are shown on the square-root-density scale, which is the variance stabilizing transformation. Except near the boundaries, this technique gives reasonable answers.

The use of the original bootstrap does not address the problem of bias in the kernel estimate and does not help with the question of whether the bumps are spurious. The smoothed bootstrap will result in error bars that reflect this bias. In other words, the bias between $\hat{f}(x; h, \{x_i\})$ and $f(x)$ should be similar to the bias between $\hat{f}(x; h, \{x_i^*\})$ and $\hat{f}(x; h, \{x_i\})$. The right frame in Figure 9.4 displays the error bars of 200 smoothed bootstrap kernel estimates. The downward bias at peaks and upward bias at dips is evident. The trimodal structure still seems plausible based on the smoothed bootstrap error bars. Based upon the analysis by Taylor (1989) discussed in Section 6.5.1.4, the smoothed bootstrap error bars will be somewhat wider than necessary because the smoothed bootstrap inflates the variance about 10%. The temptation is to pivot the error bars around the density estimate to reflect the estimated bias correctly.

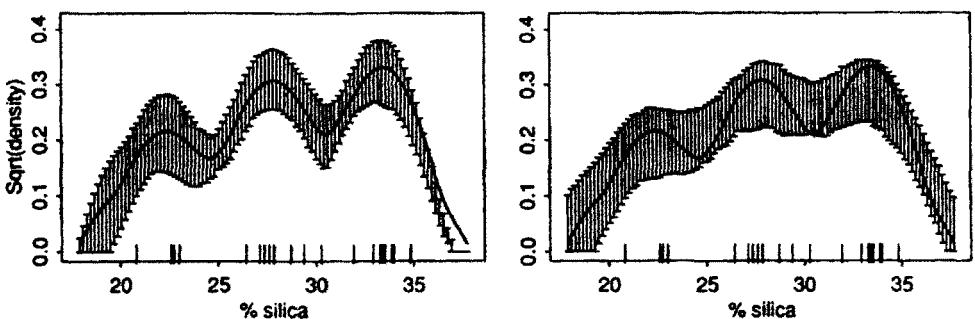


Figure 9.4 Bootstrap error bars for the unsmoothed and smoothed bootstrap resamples for the silica data. The 90th and 10th sample percentiles from 200 bootstrap estimates are shown.

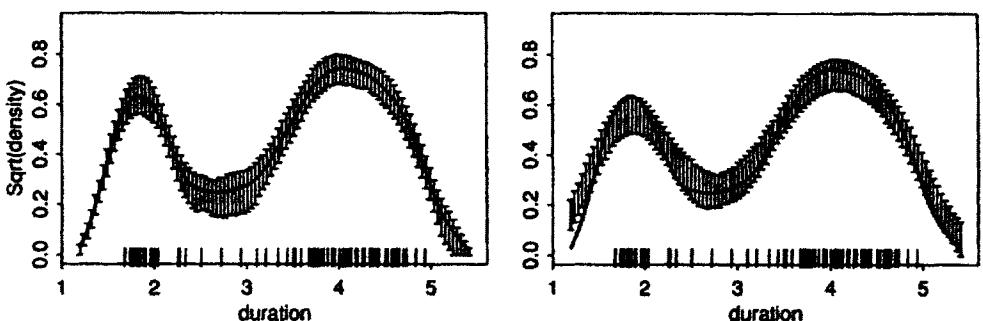


Figure 9.5 Bootstrap error bars for the unsmoothed and smoothed bootstrap resamples for the Old Faithful geyser data. The 10th and 90th sample percentiles from 200 bootstrap estimates are shown.

The same bootstrap procedures were applied to the larger geyser data set ($n = 107$) in Figure 9.5. The biweight kernel with $h = 0.5$ was chosen. The error bars are much narrower and the bimodal structure apparent in all the bootstrap resamples.

For regression problems, either smoothed or original bootstrap samples may be drawn and the regression estimate computed from those data. This idea was demonstrated by McDonald (1982). Alternatively, bootstrap samples can be based on the residuals rather than the data (Härdle and Bowman, 1988). The extension from pointwise to uniform confidence intervals for both density estimation and regression has been discussed recently by Hall and Titterington (1988) and first by Bickel and Rosenblatt (1973). Hall, DiCiccio, and Romano (1989) recommend the use of the smoothed bootstrap asymptotically.

9.3.3 Survival Analysis

Watson and Leadbetter (1964) considered nonparametric estimation of the hazard function

$$h(t) = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log [1 - F(t)],$$

which measures the instantaneous force of mortality given survival to time t . Given a sample of failure times $\{0 \leq t_1 \leq t_2 \leq \dots \leq t_n\}$, the empirical cumulative distribution function may be substituted to provide an estimate of the hazard

$$\begin{aligned}\tilde{h}(t) &= -\frac{d}{dt} \log [1 - F_n(t)] \\ &= -\sum_{i=1}^n \left[\log \left(1 - \frac{i}{n}\right) - \log \left(1 - \frac{i-1}{n}\right) \right] \delta(t - t_i),\end{aligned}$$

which is undefined for $t \geq t_n$ as that term involves $\log(0)$. Usually, the ecdf is multiplied by a factor such as $n/(n + 1)$ to avoid that difficulty. With this modification and convolving the rough estimate with a kernel, a smooth estimate of the hazard function is obtained:

$$\hat{h}(t) = \sum_{i=1}^n \log \left[\frac{n - i + 2}{n - i + 1} \right] K_\lambda(t - t_i),$$

denoting the smoothing parameter by λ rather than h to avoid confusion.

Often survival data are subject to censoring, which occurs when the actual failure time of a subject is not observed either due to loss of follow-up or due to the end of the observation period. In this case, t_i is either the time of death (or failure) or the length of time the subject was observed before being lost (or the survival time at the end of the study). Suppose n_i is the number of individuals still under observation at time t_i , including the individual who died at time t_i . Then the Kaplan-Meier (1958) product-limit estimator replaces the ecdf:

$$\tilde{F}_n(t_i) = 1 - \prod_{j=1}^i \left(\frac{n_j - 1}{n_j} \right).$$

The authors prove that this is a consistent estimate of $F(t)$ assuming that the censoring mechanism is random. The kernel estimate for censored data follows in a similar fashion; see Problem 8.

9.3.4 High-Dimensional Holes

A challenge in higher dimensions is recognizing high-dimensional structure. A simple example is a hole in the data, or more generally, regions of lower density surrounded by regions of higher density. A bivariate example was given in Figure 7.1. A trivariate example of 5,000 data points from a symmetric density with a hole is depicted in Color Plates 13–16. The pairwise scatter diagrams are shown in Figure 9.6. The hole is not visible as was the case with the bivariate example. In fact, applications of interactive tools such as rotation and brushing to the point cloud do not reveal the hole. Color Plate 13 shows a single α -shell with two nested contour surfaces (the inner shell, which is visible through the semitransparent outer shell, has radius about a fourth as long). Again, the appearance of two *nested* contours at the same α level is the *signature* of a hole in the density. (Two *nonnested* contours at the same α level is the signature of a bimodal density.) This figure might easily have resulted from displaying α -shells at two different α levels with ordinary Normal data.

At a slightly higher value of α than used in Color Plate 13, the inner and outer contour surfaces join together, as shown in Color Plate 14. That only one surface is displayed can be demonstrated by slightly rotating the viewpoint and adding outer normal vectors to the surface, as shown in Color Plate 15. Observe that the hole where the contours join is now visible. The outer normal

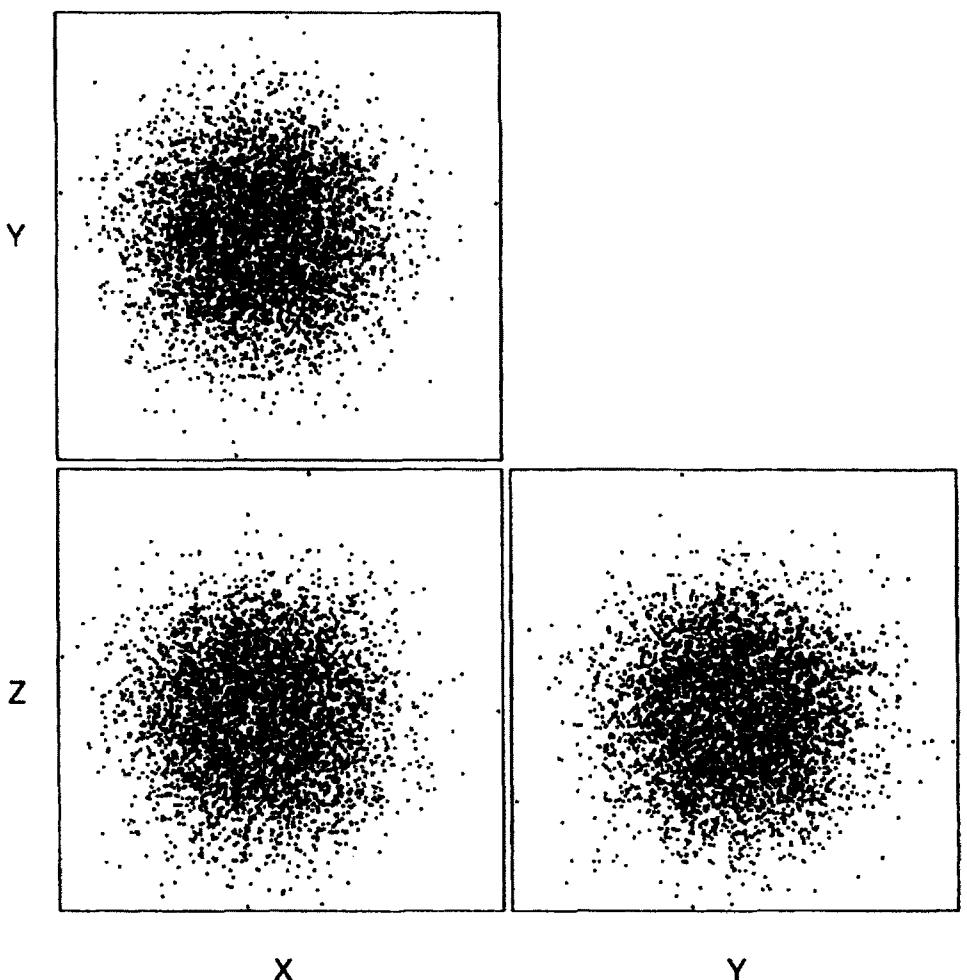


Figure 9.6 Pairwise scatterplots of 5,000 trivariate simulated points with a hole. The hole is actually a region of lower density rather than a region around the origin.

vectors point toward regions of relatively lower probability density, which are located both near and far away from the origin. Finally, at a level of α near 1, the contour surface approaching the sample mode becomes quite complex, as shown in Color Plate 16. These data happen to contain 3 sample modes located near the theoretical mode, which is the surface of a sphere.

In a poster session sponsored by the ASA Statistical Graphics section in Chicago in 1986, David Coleman created a 5-D data set with a true hole carved in it. The presence of the hole was discovered by looking for nested contours in the 3-D slices. However, the density estimate blurred the points inserted by Coleman into the hole that spelled "EUREKA." Nevertheless, multivariate

density estimation is an excellent tool for discovering unusual features in the data.

9.3.5 Image Enhancement

A comprehensive treatment of image enhancement and image processing is beyond the scope of this book. Some references include Ripley (1988) and Wegman and DePriest (1986). A simple example of image enhancement is called histogram equalization. In Figure 9.7, a histogram of the first channel of the LANDSAT image is shown together with a gray scale image. The spatial structure is barely discernible, as the data values are not uniformly spaced over the interval (0, 255). Applying an inverse cdf transformation, a more uniform distribution of these data is obtained (not exactly uniform because the raw data are integers). The corresponding image is much more revealing.

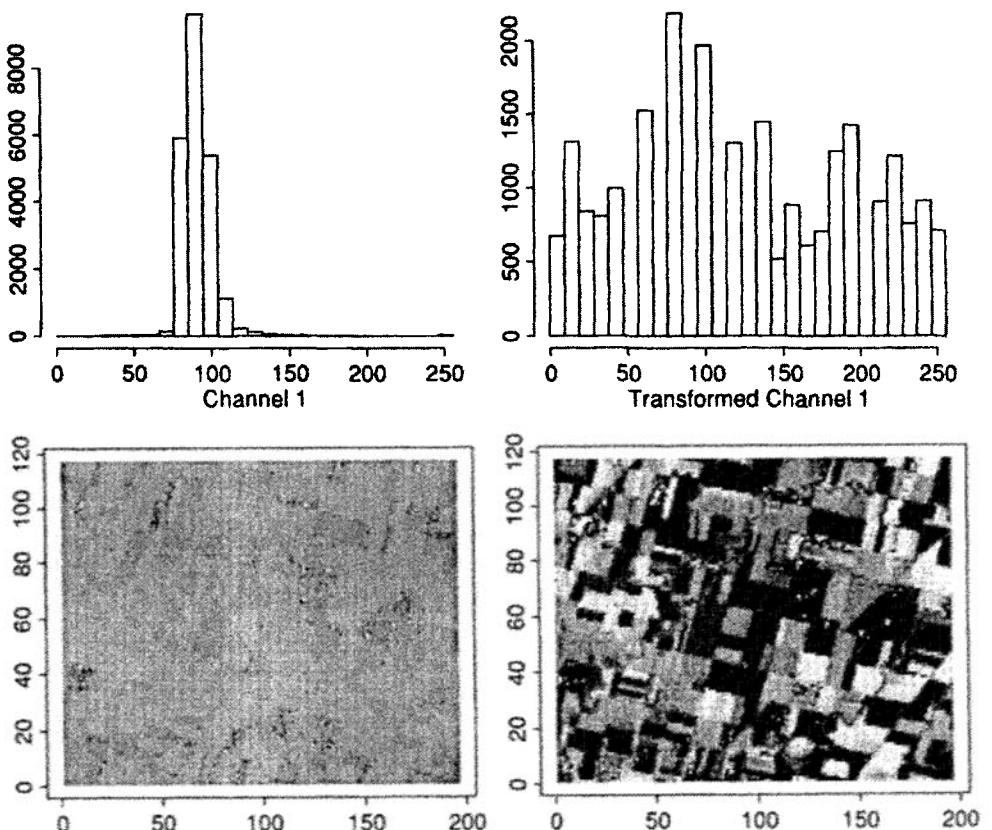


Figure 9.7 Histograms of raw data from LANDSAT scene and transformed data that are more nearly uniform. The increased dynamic range in the gray scale images may be observed.

9.3.6 Nonparametric Inference

Drawing inferences from nonparametric estimates is a growing field. In certain cases, such as nonparametric regression and additive modeling, much progress has occurred because many of these procedures are linear smoothers, which represent a hybrid of parametric and nonparametric ideas. The difficult inference is how many terms should be in the model. Conditional on that information, the fitting often follows other linear parametric inference procedures. A few other examples could be given, but much work remains, particularly in the multivariate case.

Consider Pearson's goodness-of-fit hypothesis test:

$$H_0 : f = f_0 \quad \text{vs.} \quad H_1 : f \neq f_0.$$

Given a sample $\{x_i\}$, a bootstrap test may be constructed by introducing a measure of discrepancy such as

$$d(f_0, \hat{f}_h) = \int [\hat{f}_h(x) - f_0(x)]^2 dx.$$

A p -value can be determined by sampling from f_0 , computing a kernel estimate, and determining the fraction of such samples where the discrepancy exceeds that for the original sample.

While the basis for the test is sound, the discrepancy measures the bias more than anything else because the kernel estimate scale is inflated due to smoothing. Thus the power can be improved by redefining the discrepancy to be

$$d_K(f_0, \hat{f}_h) = \int [\hat{f}_h(x) - [f_0 * K_h](x)]^2 dx.$$

Bowman (1991) reports that the power does not seem to be influenced by the choice of smoothing parameter. If both the null hypothesis and kernel are Normal, then $f_0 * K_h$ is $N(0, 1 + h^2)$, so that the computations are relatively easy.

Other authors have considered tests of the adequacy of parametric fits based on nonparametric alternatives. In the regression context, see Cox, Koh, Wahba, and Yandell (1988) and Eubank and Speckman (1990).

9.3.7 Other Topics

Principal Curves and Density Ridges

When bivariate data are not of the regression type, a lower-dimensional summary may still be desired. Ordinary scatterplot smoothing is not appropriate as there is no response variable. Hastie and Stuetzle (1989) discuss an iterative procedure that moves points towards the "principal curve." The principal

curve can be thought of as a local conditional mean. Alternatively, a similar concept may be developed based on the bivariate density surface. The *density ridge* should be more than just those points for which $\nabla f(x, y) = 0$ and the Hessian $Hf(x, y)$ is negative semidefinite. By the spectral representation, $Hf(x, y) = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T$. Now the curvature of the surface at (x, y) in the direction $\eta = (\eta_1, \eta_2)$ is given by

$$\begin{aligned} f''(x, y)[\eta] &= \eta^T \nabla^2 f(x, y) \eta = \eta^T (\lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T) \\ &= \lambda_1 \cos^2 \theta_1 + \lambda_2 \cos^2 \theta_2 = \lambda_1 \cos^2 \theta_1 + \lambda_2 \sin^2 \theta_1, \end{aligned}$$

where θ_1 and θ_2 are the angles between η and the 2 eigenvectors, respectively, and $\theta_2 = \theta_1 \pm \pi/2$, so that $\cos^2 \theta_2 = \sin^2 \theta_1$. A point is on the density ridge when the maximum negative curvature is perpendicular to the gradient. Thus a point is on the density ridge if $\nabla f(x, y)$ is an eigenvector and the *other* eigenvalue is negative (and is more negative if both eigenvalues are negative). Scott (1991b) investigated the possibility of identifying such structure in the density function.

Time Series Data

Special attention may be paid to time series data when there is substantial autocorrelation. Negative autocorrelation generally is helpful while positive autocorrelation clearly can be difficult in practice. The theory of optimal smoothing and cross-validation changes. For some recent results, see Hart (1984, 1991), Altman (1990), Hart and Vieu (1990), and Chiu (1989) as well as the monograph by Györfi et al. (1989).

Inverse Problems (Deconvolution)

This class of problems is the most difficult in smoothing. Spline methods have been the favorite technique applied in this area. Some recent applications include estimation of yearly infection rates of AIDS (Brookmeyer, 1991). Excellent summaries are available in O'Sullivan (1986) and Wahba (1990).

Densities on the Sphere

See Watson (1985) and Fisher, Lewis, and Embleton (1987).

PROBLEMS

1. Suppose that $\mu_0 = 0$ and $\mu_1 > 0$ in a univariate shifted-Normal model and that the prior odds are 1:1. By varying σ^2 (and hence the coefficient of detection), examine how the posterior odds vary at $x = \mu_0$ and $x = \mu_1$.
2. The multivariate shifted-Normal model is $f_+(\mathbf{x}) = N(\mu_1, \Sigma)$ and $f_-(\mathbf{x}) = N(\mu_0, \Sigma)$. Compute the log-likelihood ratio and describe the level sets

- of constant likelihood ratio. In particular, describe what happens at $\mathbf{x} = (\mu_0 + \mu_1)/2$.
3. Using Serfling's proposition in (9.6), show that the kernel estimate $\hat{f}(x) = AN(f(x) + h^2\sigma_K^2 f''(x)/2, f(x)R(K)/(nh))$.
 4. Compute the asymptotic approximations to the first 4 noncentral moments of the random variables Y_{nj} in Equation (9.7).
 5. Find an equivalent bandwidth rule for the first and second derivative kernel estimators.
 6. Consider a bootstrap sample size n from a data set with n points. Show that the probability that a particular point is *not* in the resample is approximately e^{-1} . Comment on the effect that a single outlier has on the estimated p -values in Silverman's multimodality test.
 7. In order to generate samples X from the biweight kernel, show that $X = 2Y - 1$ where $Y \sim \text{Beta}(3, 3)$; show that $Y = V_1/(V_1 + V_2)$ where $V_i \sim \text{Gamma}(3)$; show that $V_i = W_1 + W_2 + W_3$ where $W_i \sim \text{Gamma}(1)$; and show that $W_i = -\log(U_i)$ where $U_i \sim U(0, 1)$. Prove formula (9.9).
 8. Investigate the Kaplan-Meier product-limit estimator for censored data. Construct a kernel estimator based on it.

APPENDIX A

Computer Graphics in \Re^3

A.1 BIVARIATE AND TRIVARIATE CONTOURING DISPLAY

Bivariate Contouring

Bivariate contouring algorithms on rectangular grids present some subtle and interesting challenges, beyond considerations of speed. The discussion will be limited to a square mesh with only piecewise linear interpolation and piecewise linear contours. The contour is an approximation to the set $\{x \in \Re^2 : f(x) = c\}$. The algorithm should be local; that is, the algorithm is applied bin by bin and is based solely on the 4 function values at the corners. No other information, such as gradient values or values of the function outside the square, is permitted. [An alternative approach is to trace each individual contour (Dobkin et al., 1990); however, as this assumes that a starting point is given, some form of exhaustive search over the entire mesh cannot be avoided.]

A vertex v will be labeled "+" if $f(v) > c$ and "-" if $f(v) \leq c$ [this convention for handling the case $f(v) = c$ is discussed below]. Using linear interpolation along the sides of the square bins and connecting the points of intersection gives approximate contours as shown in Figure A.1. In the left frame, the contour is drawn bin by bin, one line per bin. However, in the next two frames, two lines are drawn in the center bin, as each side of that square

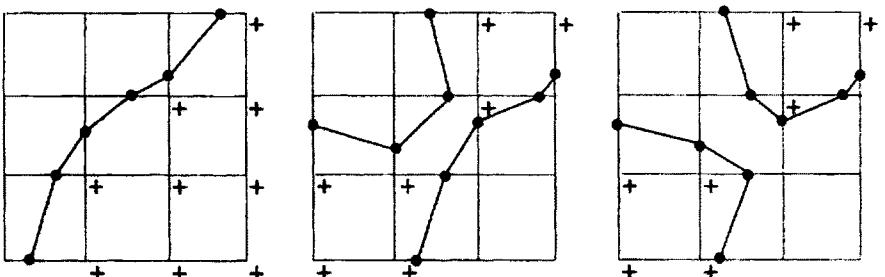


Figure A.1 Three examples of portions of a contour mesh. The unlabeled vertices are "-", that is, $f(v) \leq c$. The right two meshes have identical values on the vertices, but different contours.

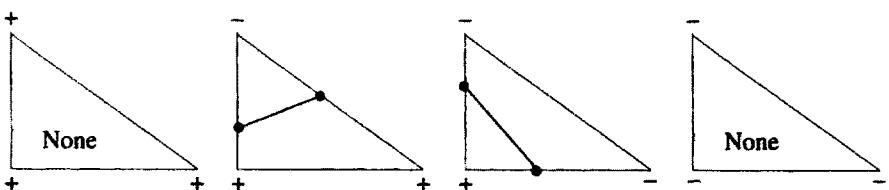


Figure A.2 Examples of function values on triangular mesh elements.

has a point where $f(\mathbf{x}) = c$. Depending on the manner in which the 4 points are connected (assuming that the contour lines cannot cross), one of the shown contour patterns will be chosen.

The ambiguity could be resolved if a finer mesh were available, but by assumption it is not. Alternatively, a triangular mesh can be superimposed on the square mesh, as a triangle can have at most 2 points of intersection on its boundary when using linear interpolation along the 3 sides of the triangle. Save rotations, there are 4 distinct cases, as shown in Figure A.2. Reexamining the center bin in Figure A.1, the preference in a triangular mesh is determined by the orientation of the triangles in that bin. Therefore, the decision is by fiat rather than appeal to any other preference.

The question of how to handle equalities at the vertices is a practical problem. In statistics, very often a contour is attempted at the zero level, with disastrous results if the function has large regions where $f(\mathbf{x}) = 0$. Here, a simple idea is advanced that handles the equality problem neatly. The approach is summarized in Figure A.3. Contrast the second and fourth triangles. A contour line is drawn along the 0–0 edge when the other corner is at a higher level, but a contour line is not drawn if the third corner satisfies $f(\mathbf{x}) \leq c$. The rationale is that such an edge may be drawn in the adjoining triangle, if appropriate. When a large region satisfies $f(\mathbf{x}) = c$, the only contour lines drawn are at the boundary where the function finally begins to increase. This choice seems more correct than drawing all 0–0 edges. Mathematically, the contour is being drawn at the level $f(\mathbf{x}) = c + \epsilon$ as $\epsilon \searrow 0$ and not at the exact level $f(\mathbf{x}) = c$.

A triangular mesh is attractive for its simplicity and its speed. Yet the multiplicity of choices suggests further consideration of the original square meshes. Of the 2^4 cases [depending as $f(\mathbf{v}) > c$ or $f(\mathbf{v}) \leq c$ at each of the 4

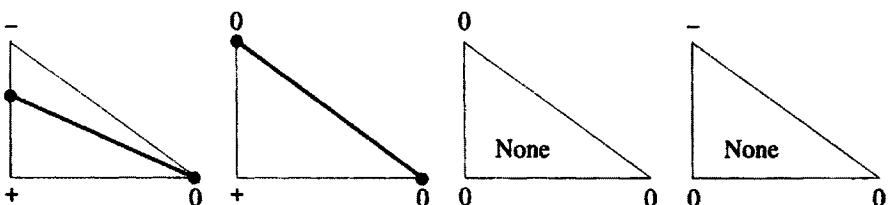


Figure A.3 Examples of function values with equality on triangular mesh elements.

vertices], only the 2 cases illustrated in Figure A.1 result in any ambiguity. The cases occur when the sequence of signs along the 4 corners is $+ - + -$. One approach is to draw the exact contours of an approximation function defined on each square. The simplest function that interpolates the 4 values at the corners is the linear blend, defined by the bilinear form $f(\mathbf{x}) = a + bx_1 + cx_2 + dx_1x_2$. This function is linear along any slice parallel to either axis, and in particular provides linear interpolation along the edges of the square bin (but not along the diagonal as was assumed for the triangular mesh). The linear blend resolves the ambiguity between the 2 alternative contours shown in Figure A.1. Thus the algorithm proposed is to draw piecewise linear contours with the aid of a linear blend surface to resolve any ambiguities.

Consider the 3 square bins in Figure A.4. The dots along each edge indicate where $f(\mathbf{x}) = c$ by linear interpolation. The dotted lines in each square are parallel to the axes; hence, the linear blend approximation linearly interpolates the values along the edges. In the left frame, the contour segments (as drawn) "overlap" along the x -axis but not along the y -axis. It is left to the reader to prove that this configuration is impossible. For example, the vertical dotted line takes on values at its endpoints that are less than c , but takes on values greater than c in the middle. Either there is overlap in both directions or no overlap at all.

The 4 dotted lines in the right pair of square bins take on values less than c as they intersect the edges in the region where $f < c$. Therefore, the linear blend would reject the contours drawn in the middle frame in favor of the contours as drawn in the right frame. There is an easy test to decide when the contour line segments should be drawn with positive slope or negative slope. Define v_{ij} to be the difference of the function value at the vertex less the contour value c . Then the contour line segments with positive slope are drawn if $v_{00}v_{11} < v_{01}v_{10}$ and with negative slope if $v_{00}v_{11} > v_{01}v_{10}$. Equality may occur in 2 distinct cases. The first occurs when $v_{ij} = 0$ and all 4 function values at the corners equal c . By convention, no contour lines are drawn in this square bin. When equality occurs but $v_{ij} \neq 0$, the linear blend offers no guidance. (Verify that the contours of the linear blend in the square bin cross in this rare situation.)

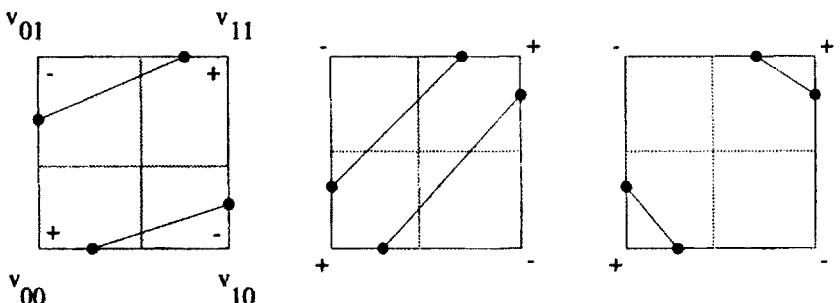


Figure A.4 Using the linear blend to resolve contour ambiguities on a square mesh.

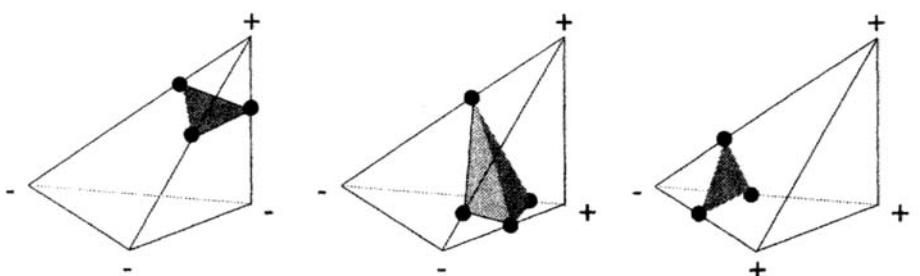


Figure A.5 Trivariate contour patches with tetrahedron bins.

Trivariate Contouring

The detailed description of bivariate contouring carries over to trivariate contouring over a cubical mesh. Assume that the array of function values f_{ijk} is given. In each cube, a collection of triangular patches is computed that defines the contour surface. As in the bivariate case, there are ambiguities when contouring the cubical bins directly. The analog of the triangular mesh is to divide the cube into 6 regions, for example, $x_1 < x_2 < x_3$. The 3 cases of interest with the contour patches are illustrated in Figure A.5. Notice that in the middle frame, a pair of triangular patches are drawn connecting the 4 points on the edges, and that the triangles are not unique. Since the orientation of the 6 tetrahedra in the cube is not unique, different contours can result just as in the bivariate case.

In general, fewer triangular patches are required if the cubical bins are contoured directly. A few of the cases have been illustrated in Chapter 1. As in the bivariate case, saddle points provide the greatest challenge. One special case that requires an unusual solution is illustrated in Figure A.6. The pair of patches leaves a situation on the closest face of the cube with overlap in

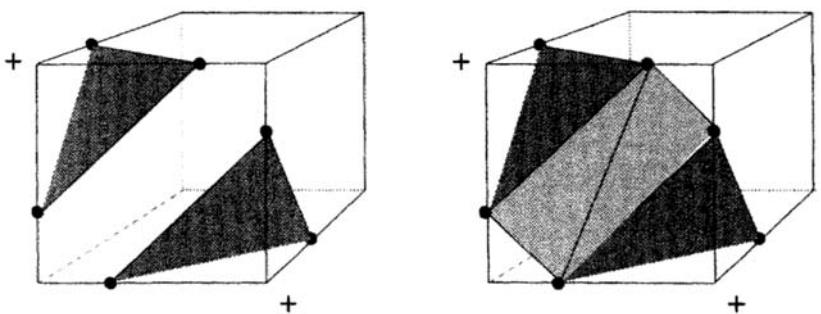


Figure A.6 Special case of cubical contouring.

both directions that would not be permitted with the (bivariate) linear blend fix. In fact, a "hole" in the 3-D contour surface remains and should be filled in during this step.

More visually pleasing surfaces can be drawn through enhancement of the *marching cubes* algorithm (Lorensen and Cline, 1987) by local smoothing over the triangular patches. The difference is illustrated in the Farin (1987), for example. Completely eliminating any noise from the surface is to be discouraged generally, in order to remind the user that these are estimated surfaces and subject to noise. A rough measure of the accuracy of the estimate can be obtained by examining the noise in the contours. These are subjective matters and viewing programs such as MinneView allow toggling between different viewing options. For further details, see Foley and Van Dam (1982), Diamond (1982), Ripley (1981), Farin (1987), and Dobkin et al. (1990).

A.2 DRAWING 3-D OBJECTS ON THE COMPUTER

In this section, a simplified description of the algorithms for presenting 3-D objects on a computer screen is presented, both with and without perspective. Suppose that the object of interest has been positioned and scaled to fit in the cube $[-1, 1]^3$. Let $e \in \mathbb{R}^3$ be the Euclidean coordinates of the viewer and imagine that the computer screen is positioned halfway between the viewer and the origin.

The viewer's position is first transformed to spherical coordinates (r, ϕ, θ) :

$$r = \left[\sum_{i=1}^3 e_i^2 \right]^{1/2}; \quad \phi = \cos^{-1}(e_3/r); \quad \theta = \tan^{-1}(e_2, e_1),$$

where $\tan^{-1}(e_1, e_2)$ is a version of $\tan^{-1}(e_2/e_1)$ that returns a value between $(-\pi, \pi)$ rather than $(-\pi/2, \pi/2)$. Then given $x = (x_1, x_2, x_3)^T$ in the cube, the screen coordinates are x' , where $x' = (x'_1, x'_2, x'_3)^T$ is given by

$$x' = Vx \quad \text{where} \quad V = \begin{pmatrix} -\sin \theta & -\cos \theta \cos \phi & -\cos \theta \sin \phi \\ \cos \theta & -\sin \theta \cos \phi & -\sin \theta \sin \phi \\ 0 & \sin \phi & -\cos \phi \end{pmatrix}. \quad (\text{A.1})$$

The value x'_3 gives the value of the axis orthogonal to screen, but in a left-hand system rather than a right-hand system (so that larger values are farther away from the viewer). If the original object is made up of points or line segments, this transformation is applied to the points or ends of the line segments and then plotted as points or line segments on the plane.

Adding perspective to the equations is not difficult. A "1" is appended to each point \mathbf{x} and the matrix \mathbf{V} becomes the 4×4 matrix

$$\mathbf{V} = \begin{pmatrix} -\sin \theta & -\cos \theta \cos \phi & -\cos \theta \sin \phi & 0 \\ \cos \theta & -\sin \theta \cos \phi & -\sin \theta \sin \phi & 0 \\ 0 & \sin \phi & -\cos \phi & 0 \\ 0 & 0 & r & 1 \end{pmatrix}. \quad (\text{A.2})$$

After computing $\mathbf{x}' = \mathbf{V}[\mathbf{x}^T \ 1]^T$, the screen coordinates (x_1'', x_2'') are given by

$$x_1'' = \frac{x'_1}{2} \cdot \frac{r}{x'_3} \quad \text{and} \quad x_2'' = \frac{x'_2}{2} \cdot \frac{r}{x'_3}. \quad (\text{A.3})$$

Multivariate Density Estimation

DAVID W. SCOTT

Copyright © 1992 by John Wiley & Sons, Inc.

A P P E N D I X B**Data Sets****B.1 UNITED STATES ECONOMIC VARIABLES DATA**

Annual economic variables for the United States between 1925 and 1940. The selection of variables for Chernoff faces in Figure 1.6 is also indicated.

Year	GNP	WPI	CPI	Incom	Banks	Unempl	Fuel	Hous	Suic	Homi
1925	1.794	5.33	5.25	1.274	2.8442	3.2	4.0014	9.37	12.0	8.3
1926	1.900	5.16	5.30	1.274	2.7742	1.8	4.1342	8.49	12.6	8.4
1927	1.898	4.93	5.20	1.274	2.6650	3.3	4.2492	8.10	13.2	8.4
1928	1.909	5.00	5.13	1.274	2.5798	4.2	4.3020	7.53	13.5	8.6
1929	2.036	4.91	5.13	1.274	2.4970	3.2	4.9039	5.09	13.9	8.4
1930	1.835	4.46	5.00	1.167	2.3679	8.7	4.7544	3.30	15.6	8.8
1931	1.693	3.76	4.56	1.108	2.1654	15.9	4.3954	2.54	16.8	9.2
1932	1.442	3.36	4.09	0.949	1.8734	23.6	3.4489	1.34	17.4	9.0
1933	1.415	3.40	3.88	0.921	1.4207	24.9	3.5274	0.93	15.9	9.7
1934	1.543	3.86	4.01	0.981	1.5348	21.7	3.9367	1.26	14.9	9.5
1935	1.695	4.13	4.11	1.068	1.5488	20.1	4.0797	2.21	14.3	8.3
1936	1.930	4.17	4.15	1.198	1.5329	16.9	5.0144	3.19	14.3	8.0
1937	2.032	4.45	4.30	1.236	1.5094	14.3	5.3560	3.36	15.0	7.6
1938	1.929	4.05	4.22	1.153	1.4867	19.0	4.8560	4.06	15.3	6.8
1939	2.094	3.98	4.16	1.232	1.4667	17.2	5.7958	5.15	14.1	6.4
1940	2.272	4.05	4.20	1.303	1.4534	14.6	6.2942	6.03	14.4	6.3

GNP	- 1958 prices; \$\$ 10**8	area of face
WPI	- Wholesale Price Index	shape of face
CPI	- Consumer Price Index	length of nose
Income	- Personal 1958 prices	location of mouth
Banks	- Number commercial banks	curve of smile
Unemployment	- Percent civilian	width of mouth
Fuel	- Electric Utilities Cost	location of eyes
House	- New starts (000's)	separation of eyes
Suicides	- Rate/100,000	angle of eyes
Homicide	- Rate/100,000	shape of eyes

Source: U.S. Department of Commerce, Bureau of the Census, "Historical Statistics of the United States: Colonial Times to 1970," Washington, D.C., 1975.

B.2 UNIVERSITY DATA

Characteristics of 28 selected universities around 1984.

COLLEGE	\$/F	S/F	G/U	Tuit	Bks	\$\$/F	NMP	\$R&D	Bk/F
Amherst	0.59	9.7	0.00	81.5	2.4	1.2	0.8	3.	3.8
Brown	0.27	13.6	0.22	82.0	4.2	0.9	0.7	34.	3.8
Cal Tech	0.70	5.8	1.05	75.0	2.0	0.4	3.8	110.	1.3
Carnegie-M	0.29	12.9	0.38	63.0	2.2	0.6	0.4	51.	1.2
Chicago	0.38	7.5	1.73	70.7	6.8	0.5	3.3	51.	4.4
Columbia	0.52	12.3	2.49	78.9	7.2	0.4	0.3	56.	3.5
Dartmouth	0.89	17.1	0.16	81.9	3.9	3.1	0.8	36.	5.1
Duke	0.11	7.5	0.43	62.1	5.6	0.2	0.7	31.	2.1
Emory	0.90	21.2	1.34	62.0	4.1	0.9	1.2	53.	5.2
Harvard	0.70	6.6	1.46	81.9	10.2	0.5	4.1	36.	4.3
J. Hopkins	0.91	11.0	0.47	67.0	4.9	1.6	1.2	1260.	8.3
MIT	0.63	9.4	1.04	87.0	4.4	0.8	1.9	146.	1.9
Northwest.	0.27	8.5	0.66	80.8	5.4	0.4	1.4	24.	2.1
Notre Dame	0.29	12.4	0.26	59.5	3.7	3.2	0.4	19.	2.0
Oberlin	0.44	16.7	0.38	75.1	3.2	0.6	0.3	1.	4.4
U. Penn.	0.06	4.7	0.97	80.0	4.8	0.2	0.6	22.	0.7
Princeton	1.28	8.5	0.32	83.8	5.8	1.5	3.3	34.	4.9
Rice	0.95	9.3	0.31	35.0	3.3	1.7	5.5	20.	3.0
USC	0.11	16.6	1.13	68.4	4.7	0.5	0.3	30.	1.4
SMU	0.03	20.3	0.74	50.0	4.4	1.2	0.3	4.	4.3
Southwest.	0.66	17.5	0.00	35.0	1.0	3.3	0.2	0.	1.8
Stanford	0.66	10.1	0.79	82.2	7.0	0.8	1.2	90.	4.2
Swarthmore	0.56	7.6	0.02	71.3	2.4	0.7	1.5	2.	3.4
Texas A&M	0.02	10.6	0.26	12.0	3.7	0.1	0.6	12.	0.5
U. Texas	0.96	24.5	0.35	12.0	6.9	0.2	0.3	23.	2.5
Tulane	0.10	9.7	0.35	59.5	3.7	0.5	0.7	13.	2.1
Vanderbilt	0.14	7.0	0.63	61.0	3.9	0.2	0.3	21.	1.2
Yale	0.47	6.4	0.97	81.9	8.8	0.7	2.7	46.	4.8

1. Endowment per faculty (\$millions)
2. Total students per faculty
3. Ratio graduate/undergraduate students
4. Tuition (\$100s)
5. Square root (# Library Books/100,000)
6. Fund drive per faculty (\$10,000s)
7. Percentage National Merit undergraduates
8. Federal R&D funds per faculty (\$1,000s)
9. Library books (1,000s)/faculty

Source: Rice University Self-Study, 1985.

B.3 BLOOD FAT CONCENTRATION DATA

Concentration of plasma cholesterol and plasma triglycerides (mg/dl) in 371 patients evaluated for chest pain. The data are listed sequentially for each patient $\{x_1, y_1, x_2, y_2, \dots, x_{51}, y_{51}\}$ for the first group and similarly for the second group.

Data for 51 males with no evidence of heart disease:

195	348	237	174	205	158	201	171	190	85	180	82	193	210	170	90	150	167	200	154
228	119	169	86	178	166	251	211	234	143	222	284	116	87	157	134	194	121	130	64
206	99	158	87	167	177	217	114	234	116	190	132	178	157	265	73	219	98	266	486
190	108	156	126	187	109	149	146	147	95	155	48	207	195	238	172	168	71	210	91
208	139	160	116	243	101	209	97	221	156	178	116	289	120	201	72	168	100	162	227
207	160																		

Data for 320 males with narrowing of the arteries:

184	145	263	142	185	115	271	128	173	56	230	304	222	151	215	168	233	340	212	171
221	140	239	97	168	131	231	145	221	432	131	137	211	124	232	258	313	256	240	221
176	166	210	92	251	189	175	148	185	256	184	222	198	149	198	333	208	112	284	245
231	181	171	165	258	210	164	76	230	492	197	87	216	112	230	90	265	156	197	158
230	146	233	142	250	118	243	50	175	489	200	68	240	196	185	116	213	130	180	80
208	220	386	162	236	152	230	162	188	220	200	101	212	130	193	188	230	158	169	112
181	104	189	84	180	202	297	232	232	328	150	426	239	154	178	100	242	144	323	196
168	208	197	291	417	198	172	140	240	441	191	115	217	327	208	262	220	75	191	115
119	84	171	170	179	126	208	149	180	102	254	153	191	136	176	217	283	424	253	222
220	172	268	154	248	312	245	120	171	108	239	92	196	141	247	137	219	454	159	125
200	152	233	127	232	131	189	135	237	400	319	418	171	78	194	183	244	108	236	148
260	144	254	170	250	161	196	130	298	143	306	408	175	153	251	117	256	271	285	930
184	255	228	142	171	120	229	242	195	137	214	223	221	268	204	150	276	199	165	121
211	91	264	259	245	446	227	146	197	265	196	103	193	170	211	122	185	120	157	59
224	124	209	82	223	80	278	152	251	152	140	164	197	101	172	106	174	117	192	101
221	179	283	199	178	109	185	168	181	119	191	233	185	130	206	133	210	217	226	72
219	267	215	325	228	130	245	257	186	273	242	85	201	297	239	137	179	126	218	123
279	317	234	135	264	269	237	88	162	91	245	166	191	90	207	316	248	142	139	173
246	87	247	91	193	290	332	250	194	116	195	363	243	112	271	89	197	347	242	179
175	246	138	91	244	177	206	201	191	149	223	154	172	207	190	120	144	125	194	125
105	36	201	92	193	259	262	88	211	304	178	84	331	134	235	144	267	199	227	202
243	126	261	174	185	100	171	90	222	229	231	161	258	328	211	306	249	256	209	89
177	133	165	151	299	93	274	323	219	163	233	101	220	153	348	154	194	400	230	137
250	160	173	300	260	127	258	151	131	61	168	91	208	77	287	209	308	260	227	172
168	126	178	101	164	80	151	73	165	155	249	146	258	145	194	196	140	99	187	390
171	135	221	156	294	135	167	80	208	201	208	148	185	231	159	82	222	108	266	164
217	227	249	200	218	207	245	322	242	180	262	169	169	158	204	84	184	182	206	148
198	124	242	248	189	176	260	98	199	153	207	150	206	107	210	95	229	296	232	583
267	192	228	149	187	115	304	149	140	102	209	376	198	105	270	110	188	148	160	125
218	96	257	402	259	240	139	54	213	261	178	125	172	146	198	103	222	348	238	156
273	146	131	96	233	141	269	84	170	284	149	237	194	272	142	111	218	567	194	278
252	233	184	184	203	170	239	38	232	161	225	240	280	218	185	110	163	156	216	101

Source: Scott, Gotto, Cole, and Gorry (1978), "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease: A Study of 370 Males with Chest Pain," *Journal of Chronic Diseases* 31:337-345.

B.4 PENNY THICKNESS DATA

Thickness in *mils* of a sample of 90 U.S. Lincoln pennies dated from 1945 to 1989. Two pennies were measured for each year.

1945	1946	1947	1948	1949	1950	1951	1952	1953	1954
51.8	53.2	53.0	53.6	53.2	53.4	53.4	53.4	50.6	52.2
54.6	54.2	53.0	52.2	51.6	51.0	54.0	54.6	54.8	54.6
1955	1956	1957	1958	1959	1960	1961	1962	1963	1964
52.6	54.0	53.6	53.2	56.8	57.0	56.6	54.6	58.0	55.0
54.2	53.6	55.0	53.2	57.2	55.2	57.0	56.0	56.2	56.0
1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
56.6	57.2	57.6	57.2	58.0	57.2	58.0	57.0	59.0	59.0
58.2	56.0	55.8	56.0	56.0	55.6	55.8	57.0	57.6	54.0
1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
54.0	53.2	54.0	53.0	54.0	53.6	54.8	54.4	52.8	55.2
54.2	53.2	53.2	53.6	53.0	53.0	54.0	54.6	52.6	54.0
1985	1986	1987	1988	1989					
55.6	55.0	55.0	54.0	56.2					
56.0	53.2	54.0	55.6	54.0					

Source: Bradford S. Brown, Registered Professional Engineer, Houston, Texas.

B.5 GAS METER ACCURACY DATA

Accuracy of a gas meter as a function of flow rate and pressure. Alternating rows give the flow rate and the accuracy (100% is perfect).

44.217 psia:

35.00	54.00	69.0	71.00	88.0	107.00	179.00	252.00	314.00	341.00
96.36	97.33	97.7	97.86	97.7	98.17	98.25	98.54	98.83	99.05
360.00	360.0	376.00	395.00	721.00	722.00	1076.00	1077.00		
99.16	99.1	99.21	99.25	99.87	99.99	100.41	100.32		
2243.00	2354.00	3265.00	3614.00	3625.00	3772.00				
100.35	100.38	100.45	100.73	100.85	100.87				

74.609 psia:

58.00	74.00	74.00	92.00	112.0	187.00	205.0	258.00	327.00	346.00
97.25	97.67	97.72	97.95	98.3	99.61	98.9	98.92	99.47	99.68
369.00	372.00	392.00	402.00	402	438.00	438.00	486.00	486.00	
99.93	99.92	99.72	100.08	100	99.88	99.85	99.93	99.91	
538.00	650.00	650.00	728.00	729.00	813.00	814.00	1328.00		
99.97	100.05	99.99	100.03	99.97	100.25	100.24	100.27		
1331.00	2326.00	2348.00	3682.00	3686.00	3899.00				
100.19	100.25	100.28	100.59	100.63	100.52				

134.59 psia:

54.00	72.00	73.00	92.00	111.00	181.00	186.00	257.00	271.00	
97.93	98.54	98.71	98.82	98.75	99.31	99.12	99.51	99.72	
303.00	335.00	344.00	348.00	360.00	361.00	405.00	741.00	748.00	
100.41	99.69	99.92	99.91	100.05	100.08	100.32	100.51	100.46	
1079.0	1079.00	2343.00	2368.00	3614.00	3614.0	3815.0			
100.8	100.78	100.47	100.46	100.58	100.5	100.3			

194.6 psia:

37.00	55.00	73.00	74.00	92.00	111.00	183.00	183.00	188.00	
97.68	98.75	98.88	98.69	99.03	98.98	99.37	99.27	99.19	
189.00	265.00	328.00	342.00	358.00	360.00	385.00	408.00	658.00	
99.44	100.14	99.97	100.17	100.34	100.36	100.64	100.63	100.53	
681.00	1094.0	1101.0	2386.00	2391.00	3596.00	3694.00	3755.00		
100.49	100.8	100.7	100.49	100.51	100.64	100.69	100.57		

314.58 psia:

25.00	38.00	56.00	73.00	74.00	92.00	109.00	183.00	184.00	251.00
96.65	98.33	99.22	99.53	99.48	99.76	99.79	99.89	99.76	100.05
251.00	324.00	330.00	330.00	345.00	351.0	355.00	356.00	372.00	
99.88	100.04	100.14	100.04	100.19	100.3	100.29	100.17	100.71	
380.00	380.0	383.00	401.00	402.00	482.0	539.00	637.00	638.00	
100.72	100.8	100.83	100.85	100.75	100.8	100.92	100.69	100.75	
640.0	712.00	725.00	739.00	746.00	887.00	901.00	902.00	1066.00	
100.7	100.71	100.59	100.74	100.74	100.68	100.63	100.91	100.95	
1069.00	1082.00	1093.00	1117.00	1126.00	1194.00	2149.00	2323.00		
100.87	100.94	100.77	100.88	100.89	100.85	100.66	100.74		
2348.00	2403.00	2669.00	3608.00	3632.00	3781.00				
100.86	100.68	100.64	100.46	100.37	100.51				

434.6 psia:

19.00 36.00 56.00 74.00 74.00 90.00 108.00 178.00 183.00
 96.27 98.78 99.31 99.58 99.58 99.85 99.94 99.85 99.74
 187.00 257.00 257.00 328.00 345.00 362.00 363.00 364.00 407.00
 99.89 100.02 99.88 100.15 100.22 100.99 100.35 100.34 100.81
 726.00 726.00 1085.0 1116.00 1131.00 2291.00 2293.00 3627.00
 100.62 100.64 100.7 100.68 100.72 100.46 100.44 100.43
 3653.00 3767.00
 100.38 100.59

614.7 psia:

19.00 36.00 56.00 71.00 71 73.0 74.0 95.00 110.00 180.00
 97.17 98.75 99.32 100.05 100 99.8 99.7 100.01 100.17 100.06
 181.00 251.0 327.00 342.00 362.00 362.00 377.00 400.00 750.00
 100.03 100.5 100.37 100.38 100.33 100.32 101.04 100.95 100.75
 757.00 1073.00 1082.00 2118.00 2177.00 3557.00 3577.00 3742.00
 100.72 100.74 100.79 100.92 100.88 100.39 100.34 100.57

B.6 OLD FAITHFUL DATA

Duration in minutes of 107 nearly consecutive eruptions of the Old Faithful geyser. A dash indicates missing observations in the sequence.

4.37 3.87 4.00 4.03 3.50 4.08 2.25 4.70 1.73 4.93
 1.73 4.62 3.43 ---- 4.25 1.68 3.92 3.68 3.10 4.03
 1.77 4.08 1.75 3.20 1.85 4.62 1.97 ---- 4.50 3.92
 4.35 2.33 3.83 1.88 4.60 1.80 4.73 1.77 4.57 1.85
 3.52 ---- 4.00 3.70 3.72 4.25 3.58 3.80 3.77 3.75
 2.50 4.50 4.10 3.70 3.80 3.43 ---- 4.00 2.27 4.40
 4.05 4.25 3.33 2.00 4.33 2.93 4.58 1.90 3.58 3.73
 3.73 ---- 1.82 4.63 3.50 4.00 3.67 1.67 4.60 1.67
 4.00 1.80 4.42 1.90 4.63 2.93 ---- 3.50 1.97 4.28
 1.83 4.13 1.83 4.65 4.20 3.93 4.33 1.83 4.53 2.03
 ---- 4.18 4.43 4.07 4.13 3.95 4.10 2.72 4.58 1.90
 4.50 1.95 4.83 4.12

Source: S. Weisberg (1985), *Applied Linear Regression*, John Wiley, New York, Table 9.1, p. 213.

B.7 SILICA DATA

Percentage of silica in 22 chondrites meteors.

20.77	22.56	22.71	22.99	26.39	27.08	27.32	27.33	27.57	27.81
28.69	29.36	30.25	31.89	32.88	33.23	33.28	33.40	33.52	33.83
33.95	34.82								

Source: Ahrens (1965) and Good and Gaskins (1980).

B.8 LRL DATA

Bin counts from a particle physics experiment. There are 172 bins with centers ranging from 285 to 1,995 with bin width of 10 MeV. The sample size is 25,752.

5	11	17	21	15	17	23	25	30	22	36	29	33
43	54	55	59	44	58	66	59	55	67	75	82	98
94	85	92	102	113	122	153	155	193	197	207	258	305
332	318	378	457	540	592	646	773	787	783	695	774	759
692	559	557	499	431	421	353	315	343	306	262	265	254
225	246	225	196	150	118	114	99	121	106	112	122	120
126	126	141	122	122	115	119	166	135	154	120	162	156
175	193	162	178	201	214	230	216	229	214	197	170	181
183	144	114	120	132	109	108	97	102	89	71	92	58
65	55	53	40	42	46	47	37	49	38	29	34	42
45	42	40	59	42	35	41	35	48	41	47	49	37
40	33	33	37	29	26	38	22	27	27	13	18	25
24	21	16	24	14	23	21	17	17	21	10	14	18
16	21	6										

Source: Good and Gaskins (1980).

B.9 BUFFALO SNOWFALL DATA

Annual snowfall in Buffalo, NY, 1910–1972, in inches.

126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5	25.0
69.3	53.5	39.8	63.6	46.7	72.9	79.6	83.6	80.7	60.3
79.0	74.4	49.6	54.7	71.8	49.1	103.9	51.6	82.4	83.6
77.8	79.3	89.6	85.5	58.0	120.7	110.5	65.4	39.9	40.1
88.7	71.4	83.0	55.9	89.9	84.8	105.2	113.7	124.7	114.5
115.6	102.4	101.4	89.8	71.5	70.9	98.3	55.5	66.1	78.4
120.5	97.0	110.0							

Source: Carmichael (1976) and Parzen (1979).

Multivariate Density Estimation

DAVID W. SCOTT

Copyright © 1992 by John Wiley & Sons, Inc.

APPENDIX C

Notation

General Mathematical and Probability Notation

I_A	indicator function for the set A	34
$E(X)$	expectation of X	34
$\text{Var}(X)$	variance of X	34
$\text{Cov}(X, Y)$	covariance of X and Y	97
$X \sim f$	X has density f	30
$C.V.$	coefficient of variation	168
Φ	cdf of the $N(0, 1)$ density	201
ϕ	pdf of the $N(0, 1)$ density	40
$\chi^2(n)$	pdf of the chi-squared density	30
$\Pr(\cdot)$	probability of an event	30
$B(n, p)$	binomial pdf	47
$U(a, b)$	uniform pdf	39
$P(\lambda)$	Poisson pdf	87
$B(\mu, \nu)$	Beta function [= $\Gamma(\mu)\Gamma(\nu)/\Gamma(\mu + \nu)$]	28
$\Gamma(x)$	Gamma function	28
$\text{Beta}(\mu, \nu)$	Beta pdf	68
$N(\mu, \Sigma)$	multivariate Normal pdf	22
IQ	interquartile range	55
$R(g)$	$\int g(x)^2 dx$ (roughness of g)	53
$\arg \min$	argument that minimizes a function	177
$\arg \max$	argument that maximizes a function	88
s/t	subject to (some constraints)	72
$f * g$	convolution of f and g	127
$\delta(t)$	Dirac delta function	35
$\delta_{\mu\nu}$	Kronecker delta function	128
$\text{tr}(A)$	trace of the matrix A	153
I_d	$d \times d$ identity matrix	22
$ H $	determinant of the matrix H	42
∇f	gradient of f ($\partial f / \partial x_i$)	153
$\nabla^2 f$	Hessian of $[\partial^2 f / (\partial x_i \partial x_j)]$	153

$(1 - x^2)_+$	$= \max(0, 1 - x^2)$	68
$a_n = O(b_n)$	$\iff a_n/b_n \rightarrow c$ as $n \rightarrow \infty$	39
$a_n = o(b_n)$	$\iff a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$	52
AN	asymptotically Normal	167
$V_d(a)$	volume of sphere in \mathbb{R}^d of radius a	28

Density Abbreviations

cdf	cumulative distribution function F	34
ecdf	cumulative distribution function F_n	34
pdf	probability density function f	34
ASFP	averaged shifted frequency polygon	113
ASH	averaged shifted histogram	113
FP	frequency polygon	95
FP-ASH	frequency polygon interpolant of the ASH	113
LBFP	linear blend of a frequency polygon	106
WARP	weighting average of rounded points	238
\hat{f}	probability density estimator	22
\hat{f}_{-i}	probability density estimator omitting x_i	76
k -NN	k th nearest neighbor (density estimator)	182
B_k	k th bin	49
$K_h(t)$	$= K(t/h)/h$	125
S_α	α -level contour surface	22

Error Measure Abbreviations

AAISB	asymptotic adaptive integrated squared bias	185
AAMISE	asymptotic adaptive MISE	67
AAMSE	asymptotic adaptive mean squared error	67
AB	asymptotic bias	188
ASB	asymptotic squared bias	183
AV	asymptotic variance	183
AISB	asymptotic integrated squared bias	54
AIV	asymptotic integrated variance	82
AMAIE	asymptotic mean absolute integrated error	202
AMISE	asymptotic mean integrated squared error	40
AMSE	asymptotic mean squared error	103
Bias	bias of an estimator	37
BMISE	bootstrap mean integrated squared error	170
BMSE	bootstrap mean squared error	170
IMSE	integrated mean squared error	38
ISB	integrated squared bias	54
ISE	integrated squared error	38
IV	integrated variance	52
MAE	mean absolute error	201

MISE	mean integrated squared error	38
MSE	mean squared error	37
RCV	root coefficient of variation	200
RRMSE	relative root mean squared error	200
$(\cdot)^*$	the optimal value of (\cdot) , for example, MISE*	54

Smoothing Parameter Abbreviations

CV	cross-validation	75
BCV	biased cross-validation	76
OS	oversmoothed	73
PI	plug-in	174
UCV	unbiased cross-validation	77
h	smoothing parameter (bin or kernel width)	49
h^*	optimal smoothing parameter	54
h_{MISE}	optimal smoothing parameter w.r.t. MISE	161
\hat{h}	data-based choice for smoothing parameter	55
\hat{h}_{ISE}	data-based optimal choice w.r.t. ISE	161
\hat{h}_{BCV}	data-based optimal choice w.r.t. BCV	167
\hat{h}_{UCV}	data-based optimal choice w.r.t. UCV	168
\hat{h}_{OS}	data-based optimal choice w.r.t. OS	161
\hat{h}_{PI}	data-based optimal choice w.r.t. PI	174

References

- Abramson, I.S. (1982a). "On Bandwidth Variation in Kernel Estimates—A Square Root Law" *Ann. Statist.* **10** 1217–1223.
- Abramson, I.S. (1982b). "Arbitrariness of the Pilot Estimator in Adaptive Kernel Methods" *J. Multivariate Analysis* **12** 562–567.
- Ahrens, L.H. (1965). "Observations on the Fe–Si–Mg Relationship in Chondrites" *Geochimica et Cosmochimica Acta* **29** 801–806.
- Aitkin, M. and Tunnicliffe Wilson, G. (1980). "Mixture Models, Outliers, and the EM Algorithm" *Technometrics* **22** 325–331.
- Allen, D.M. (1974). "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction" *Technometrics* **16** 125–127.
- Altman, N.S. (1990). "Kernel Smoothing of Data with Correlated Errors" *J. Amer. Statist. Assoc.* **85** 749–759.
- Andrews, D.F. (1972). "Plots of High Dimensional Data" *Biometrics* **28** 125–136.
- Asimov, D. (1985). "The Grand Tour: A Tool for Viewing Multidimensional Data" *SIAM J. Sci. Statist. Comp.* **6** 128–143.
- Badhwar, G.D., Carnes, J.G., and Austin, W.W. (1982). "Use of Landsat-Derived Temporal Profiles for Corn-Soybean Feature Extraction and Classification" *Remote Sensing of Environment* **12** 57–79.
- Banchoff, R.F. (1986). "Visualizing Two-Dimensional Phenomena in Four-Dimensional Space: A Computer Graphics Approach." In *Statistical Image Processing and Graphics*, E.J. Wegman and D.J. Depriest (eds.), pp. 187–202. Marcel Dekker, New York.
- Bartlett, M.S. (1963). "Statistical Estimation of Density Functions" *Sankhyā Ser. A* **25** 245–254.
- Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). *The New S Language*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Becker, R.A. and Cleveland, W.S. (1987). "Brushing Scatterplots" *Technometrics* **29** 127–142.
- Bellman, R.E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

- Bickel, P.J. and Rosenblatt, M. (1973). "On Some Global Measures of the Deviations of Density Function Estimates" *Ann. Statist.* **1** 1071–1095.
- Boswell, S.B. (1983). "Nonparametric Mode Estimation for Higher Dimensional Densities." Ph.D. thesis, Department of Mathematical Sciences, Rice University.
- Bowman, A.W. (1984). "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates" *Biometrika* **71** 353–360.
- Bowman, A.W. (1985). "A Comparative Study of Some Kernel-Based Nonparametric Density Estimators" *J. Statist. Comput. Simul.* **21** 313–327.
- Bowman, A.W. (1991). "Density Based Tests for Goodness-of-Fit." Manuscript.
- Bowyer, A. (1980). "Experiments and Computer Modelling in Stick-Slip." Ph.D. thesis, University of London, England.
- Box, G.E.P. and Cox, D.R. (1964). "An Analysis of Transformations" *J. Roy. Statist. Soc. B* **26** 211–243.
- Breiman, L., Friedman, J.H., Olshen, A., and Stone, C.J. (1984). *CART: Classification and Regression Trees*. Wadsworth, Belmont, CA.
- Breiman, L., Meisel, W., and Purcell, E. (1977). "Variable Kernel Estimates of Multivariate Densities and Their Calibration" *Technometrics* **19** 135–144.
- Brookmeyer, R. (1991). "Reconstruction and Future Trends of the AIDS Epidemic in the United States" *Science* **253** 37–42.
- Cacoullos, T. (1966). "Estimation of a Multivariate Density" *Ann. Inst. Statist. Math.* **18** 178–189.
- Carmichael, J.-P. (1976). "The Autoregressive Method: A Method for Approximating and Estimating Positive Functions." Ph.D. thesis, SUNY, Buffalo, NY.
- Carr, D.B., Littlefield, R.J., Nicholson, W.L., and Littlefield, J.S. (1987). "Scatterplot Matrix Techniques for Large N" *J. Amer. Statist. Assoc.* **83** 596–610.
- Carr, D.B. and Nicholson, W.L. (1988). "EXPLOR4: A Program for Exploring Four-Dimensional Data Using Stereo-Ray Glyphs, Dimensional Constraints, Rotation, and Masking." In *Dynamic Graphics for Statistics*, W.S. Cleveland and M.E. McGill (eds.), pp. 309–329. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Carr, D.B., Nicholson, W.L., Littlefield, R.J., and Hall, D.L. (1986). "Interactive Color Display Methods for Multivariate Data." In *Statistical Image Processing and Graphics*, E.J. Wegman and D.J. Depriest (eds.), pp. 215–250. Marcel Dekker, New York.
- Cencov, N.N. (1962). "Evaluation of an Unknown Density from Observations" *Soviet Mathematics* **3** 1559–1562.
- Chamayou, J.M.F. (1980). "Averaging Shifted Histograms" *Computer Physics Communications* **21** 145–161.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA.
- Chernoff, H. (1973). "The Use of Faces to Represent Points in k -Dimensional Space Graphically" *J. Amer. Statist. Assoc.* **68** 361–368.
- Chiu, S.T. (1989). "Bandwidth Selection for Kernel Estimates with Correlated Noise" *Statist. Prob. Letters* **8** 347–354.

- Chiu, S.T. (1991). "Bandwidth Selection for Kernel Density Estimation" *Ann. Statist.* **19** 1883–1905.
- Chow, Y.S., Geman, S. and Wu, L.D. (1983). "Consistent Cross-Validated Density Estimation" *Ann. Statist.* **11** 25–38.
- Clark, R.M. (1975). "A Calibration Curve for Radiocarbon Dates" *Antiquity* **49** 251–266.
- Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatter-plots" *J. Amer. Statist. Assoc.* **74** 829–836.
- Cleveland, W.S. and McGill, M.E. (eds.) (1988). *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Cox, D.D., Koh, E., Wahba, G., and Yandell, B. (1988). "Testing the (Parametric) Null Model Hypothesis in (Semiparametric) Partial and Generalized Spline Models" *Ann. Statist.* **16** 113–119.
- Davis, K.B. (1975). "Mean Square Error Properties of Density Estimates" *Ann. Statist.* **3** 1025–1030.
- Day, N.E. (1969). "Estimating the Components of a Mixture of Normal Distributions" *Biometrika* **56** 463–474.
- Deheuvels, P. (1977a). "Estimation non paramétrique de la densité par histogrammes généralisés" *Revue de Statistique Appliquée*, v.15 25/3 5–42.
- Deheuvels, P. (1977b). "Estimation non paramétrique de la densité par histogrammes généralisés II" *Publications de l'Institut Statistique de l'Université Paris* **22** 1–23.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). "Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm" *J. Roy. Statist. Soc. B* **39** 1–38.
- Devroye, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.
- Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York.
- Diaconis, P. and Freedman, D. (1984). "Asymptotics of Graphical Projection Pursuit" *Ann. Statist.* **12** 793–815.
- Diaconis, P. and Friedman, J.H. (1983). "M and N Plots." In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, M.H. Rizvi, J.S. Rustagi, D. Siegmund (eds.), pp. 425–447. Academic Press, New York.
- Diamond, R. (1982). "Two Contouring Algorithms." In *Computational Crystallography*, D. Sayre (ed.), pp. 266–272. Clarendon Press, Oxford, England.
- DIW (1983). "Das Sozio-Ökonomische Panel." Deutsches Institut für Wirtschaftsforschung, Berlin.
- Doane, D.P. (1976). "Aesthetic Frequency Classifications" *Amer. Statist.* **30** 181–183.
- Dobkin, D.P., Levy, S.V.F., Thurston, W.P., and Wilks, A.R. (1990). "Contour Tracing by Piecewise Linear Approximations" *ACM Transactions on Graphics* **9** 389–423.
- Donoho, D.L. (1988). "One-Sided Inference About Functionals of a Density" *Ann. Statist.* **16** 1390–1420.
- Donoho, A.W., Donoho, D.L., and Gasko, M. (1988). "MACSPIN: Dynamic Graphics on a Desktop Computer." In *Dynamic Graphics for Statistics*, W.S. Cleveland and M.E. McGill (eds.), pp. 331–351. Wadsworth & Brooks/Cole, Pacific Grove, CA.

- Duan, N. (1991). "Comment on 'Transformations in Density Estimation (with Discussion)" by Wand, M.P., Marron, J.S., and Ruppert, D." *J. Amer. Statist. Assoc.* **86** 355–356.
- Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.
- Duin, R.P.W. (1976). "On the Choice of Smoothing Parameter for Parzen Estimators of Probability Density Functions" *IEEE Trans. Comput.* **C-25** 1175–1179.
- Efron, B. (1982). *The Jackknife, Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia.
- Emerson, J.D. and Hoaglin, D.C. (1983). "Stem-and-Leaf Displays." In *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller, and J.W. Tukey (eds.), pp. 7–32. John Wiley, New York.
- Epanechnikov, V.K. (1969). "Non-Parametric Estimation of a Multivariate Probability Density" *Theory Probab. Appl.* **14** 153–158.
- Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.
- Eubank, R.L. and Schucany, W.R. (1990). "Adaptive Bandwidth Choice for Kernel Regression." Technical Report.
- Eubank, R.L. and Speckman, P. (1990). "Curve Fitting by Polynomial-Trigonometric Regression" *Biometrika* **77** 1–10.
- Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.
- Fan, J.Q. (1990). "Design-Adaptive Nonparametric Regression" *J. Amer. Statist. Assoc.* In press.
- Farin, G.E., ed. (1987). *Geometric Modeling: Algorithms and New Trends*. SIAM, Philadelphia.
- Farrell, R.H. (1972). "On the Best Obtainable Asymptotic Rates of Convergence in Estimation of the Density Function at a Point" *Ann. Math. Statist.* **43** 170–180.
- Fienberg, S.E. (1979). "Graphical Methods in Statistics" *Amer. Statist.* **33** 165–178.
- Fisher, N.I., Lewis, T., and Embleton, J.J. (1987). *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge, England.
- Fisher, R.A. (1922). "On the Mathematical Foundations of Theoretical Statistics" *Philosophical Trans. Royal Society London (A)* **222** 309–368.
- Fisher, R.A. (1932). *Statistical Methods for Research Workers*, Fourth Edition. Oliver and Boyd, Edinburgh.
- Fisherkeller, M.A., Friedman, J.H., and Tukey, J.W. (1974). "PRIM-9: An Interactive Multidimensional Data Display and Analysis System." SLAC-PUB-1408, Stanford Linear Accelerator Center, Stanford, CA.
- Fix, E. and Hodges, J.L., Jr. (1951). "Nonparametric Discrimination: Consistency Properties." Report Number 4, USAF School of Aviation Medicine, Randolph Field, Texas.
- Flury, B. and Riedwyl, H. (1981). "Graphical Representation of Multivariate Data by Means of Asymmetrical Faces" *J. Amer. Statist. Assoc.* **76** 757–765.

- Foley, J.D. and Van Dam, A. (1982). *Fundamentals of Interactive Computer Graphics*. Addison-Wesley, Reading, MA.
- Freedman, D. and Diaconis, P. (1981). "On the Histogram as a Density Estimator: L_2 Theory" *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57** 453–476.
- Friedman, J.H. (1984). "A Variable Span Smoother." Technical Report 5, Department of Statistics, Stanford University.
- Friedman, J.H. (1991). "Multivariate Adaptive Regression Splines (with Discussion)" *Ann. Statist.* **19** 1–141.
- Friedman, J.H. and Stuetzle, W. (1981). "Projection Pursuit Regression" *J. Amer. Statist. Assoc.* **76** 817–23.
- Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984). "Projection Pursuit Density Estimation" *J. Amer. Statist. Assoc.* **79** 599–608.
- Friedman, J.H. and Tukey, J.W. (1974). "A Projection Pursuit Algorithm for Exploratory Data Analysis" *IEEE Transactions in Computers* **C-23** 881–890.
- Fryer, M.J. (1976). "Some Errors Associated with the Non-Parametric Estimation of Density Functions" *J. Inst. Maths. Applies.* **18** 371–380.
- Galton, F. (1886). "Regression Towards Mediocrity in Hereditary Stature" *J. Anthropological Institute* **15** 246–263.
- Gasser, T. and Engel, J. (1990). "The Choice of Weights in Kernel Regression Estimation" *Biometrika* **77** 377–381.
- Gasser, T. and Müller, H.G. (1979). "Kernel Estimation of Regression Functions." In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics 757, pp. 23–68. Springer-Verlag, Berlin.
- Gasser, T., Müller, H.G., and Mammitzsch, V. (1985). "Kernels for Nonparametric Curve Estimation" *J. Roy. Statist. Soc. B* **47** 238–252.
- Good, I.J. and Gaskins, R.A. (1972). "Global Nonparametric Estimation of Probability Densities" *Virginia Journal of Science* **23** 171–193.
- Good, I.J. and Gaskins, R.A. (1980). "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by the Scattering and Meteorite Data (with Discussion)" *J. Amer. Statist. Assoc.* **75** 42–73.
- Graunt, J. (1662). *Natural and Political Observations Made upon the Bills of Mortality*. Martyn, London.
- Gross, A.J. and Clark, V.A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. John Wiley, New York.
- Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer-Verlag, Berlin.
- Habbema, J.D.F., Hermans, J., and Van Der Broek, K. (1974). "A Stepwise Discriminant Analysis Program Using Density Estimation," *COMPSTAT 1974, Proceedings in Computational Statistics*, G. Bruckman (ed.), pp. 101–110, Physica-Verlag, Vienna.
- Hald, A. (1990). *A History of Probability and Statistics and Their Application Before 1750*. John Wiley, New York.
- Hall, P. (1984). "Central Limit Theorem for Integrated Square Error of Multivariate Density Estimators" *J. Multivariate Analysis* **14** 1–16.

- Hall, P., DiCiccio, T.J., and Romano, J.P. (1989). "On Smoothing and the Bootstrap" *Ann. Statist.* **17** 692–704.
- Hall, P. and Marron, J.S. (1987a). "On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator" *Ann. Statist.* **15** 163–181.
- Hall, P. and Marron, J.S. (1987b). "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation" *Probability Theory and Related Fields* **74** 567–581.
- Hall, P. and Marron, J.S. (1987c). "Estimation of Integrated Squared Density Derivatives" *Statist. Probab. Lett.* **6** 109–115.
- Hall, P. and Marron, J.S. (1988). "Variable Window Width Kernel Estimates of Probability Densities" *Probability Theory and Related Fields* **80** 37–49.
- Hall, P. and Marron, J.S. (1989). "Lower Bounds for Bandwidth Selection in Density Estimation" *Probab. Theory Related Fields* To appear.
- Hall, P., Sheather, S.J., Jones, M.C., and Marron, J.S. (1991). "On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation" *Biometrika* **78** 263–270.
- Hall, P. and Titterington, D.M. (1988). "On Confidence Bands in Nonparametric Density Estimation and Regression" *J. Multivariate Analysis* **27** 228–254.
- Hall, P. and Wand, M.P. (1988a). "Minimizing L_1 Distance in Nonparametric Density Estimation" *J. Multivariate Analysis* **26** 59–88.
- Hall, P. and Wand, M.P. (1988b). "On Nonparametric Discrimination Using Density Differences" *Biometrika* **75** 541–547.
- Hampel, F.R. (1974). "The Influence Curve and Its Role in Robust Estimation" *J. Amer. Statist. Assoc.* **69** 383–393.
- Hand, D.J. (1982). *Kernel Discriminant Analysis*. Research Studies Press, Chichester, England.
- Härdle, W. (1984). "Robust Regression Function Estimation" *J. Multivariate Analysis* **14** 169–180.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, England.
- Härdle, W. and Bowman, A.W. (1988). "Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands" *J. Amer. Statist. Assoc.* **83** 102–110.
- Härdle, W. and Gasser, T. (1984). "Robust Nonparametric Function Fitting" *J. Roy. Statist. Soc. B* **46** 42–51.
- Härdle, W., Hall, P., and Marron, J.S. (1988). "How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum? (with Discussion)" *J. Amer. Statist. Assoc.* **83** 86–101.
- Härdle, W. and Scott, D.W. (1988). "Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points." Technical Report 88–16, Rice University.
- Hart, J.D. (1984). "Efficiency of a Kernel Density Estimator Under an Autoregressive Dependence Model" *J. Amer. Statist. Assoc.* **79** 110–117.
- Hart, J.D. (1985). "A Counterexample to a Claim Concerning the Convolution of Multimodal Distributions." *Comm. Statist.-Theor. Meth.* **14** 2943–2945.

- Hart, J.D. (1991). "Kernel Regression Estimation with Time Series Errors" *J. Roy. Statist. Soc. B* **53** 173–187.
- Hart, J.D. and Vieu, P. (1990). "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data" *Ann. Statist.* **18** 873–890.
- Hartigan, J.A. and Hartigan, P.M. (1985). "The Dip Test of Unimodality" *Ann. Statist.* **13** 70–84.
- Hastie, T.J. and Stuetzle, W. (1989). "Principal Curves" *J. Amer. Statist. Assoc.* **84** 502–516.
- Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London.
- Hathaway, R.J. (1982). "The EM Algorithm for the Maximum-Likelihood Estimation of Gaussian Mixtures." Ph.D. thesis, Department of Mathematical Sciences, Rice University.
- Hiebert-Dodd, K.L. (1982). "An Evaluation of Mathematical Software That Solves Systems of Nonlinear Equations" *ACM Trans. Math. Soft.* **8** 5–20.
- Hjort, N.L. (1986). "On Frequency Polygons and Averaged Shifted Histograms in Higher Dimensions." Technical Report 22, Stanford University.
- Hodges, J.L. and Lehmann, E.L. (1956). "The Efficiency of Some Nonparametric Competitors of the t -test" *Ann. Math. Statist.* **27** 324–335.
- Hoerl, A.E. and Kennard, R.W. (1970). "Ridge Regression: Biased Estimation for Non-Orthogonal Problems" *Technometrics* **12** 55–67.
- Huber, P.J. (1964). "Robust Estimation of a Location Parameter" *Ann. Math. Statist.* **33** 73–101.
- Huber, P.J. (1985). "Projection Pursuit (with Discussion)" *Ann. Statist.* **13** 435–525.
- Hüsemann, J.A. and Terrell, G.R. (1991). "Optimal Parameter Choice for Error Minimization in Bivariate Histograms" *J. Multivariate Analysis* **37** 85–103.
- IMSL (1991). "Fortran Subroutine Library and Exponent Graphics Package Manuals." Houston, TX.
- Inselberg, A. (1985). "The Plane with Parallel Coordinates" *The Visual Computer* **1** 69–91.
- Izenman, A.J. and Sommer, C.J. (1988). "Philatelic Mixtures and Multimodal Densities" *J. Amer. Statist. Assoc.* **83** 941–953.
- Jee, J.R. (1985). "A Study of Projection Pursuit Methods." Ph.D. thesis, Department of Mathematical Sciences, Rice University.
- Jee, J.R. (1987). "Exploratory Projection Pursuit Using Nonparametric Density Estimation," *Proceeding of the Statistical Computing Section*, pp. 335–339, American Statistical Association, Alexandria, VA.
- Johnson, N.L. (1949). "Systems of Frequency Curves Generated by Methods of Translation" *Biometrika* **36** 149–176.
- Johnson, R.A. and Wichern, D.W. (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Jones, M.C. (1989). "Discretized and Interpolated Kernel Density Estimates" *J. Amer. Statist. Assoc.* **84** 733–741.
- Jones, M.C. (1990). "Variable Kernel Density Estimates" *Austral. J. Statist.* **32** 361–371.

- Jones, M.C. and Kappenman, R.F. (1992). "On a Class of Kernel Density Estimate Bandwidth Selectors" *Scandinavian Journal of Statistics*. In press.
- Jones, M.C. and Sibson, R. (1987). "What Is Projection Pursuit? (with Discussion)" *J. Roy. Statist. Soc. A* **150** 1–36.
- Kaplan, E.L. and Meier, P. (1958). "Nonparametric Estimation from Incomplete Observations" *J. Amer. Statist. Assoc.* **53** 457–481.
- Kendall, M.G. (1961). *A Course in the Geometry of n Dimensions*. Griffin's Statistical Monographs and Courses.
- Klonias, V.K. (1982). "Consistency of Two Nonparametric Maximum Penalized Likelihood Estimators of the Probability Density Function" *Ann. Statist.* **10** 811–824.
- Kogure, A. (1987). "Asymptotically Optimal Cells for a Histogram" *Ann. Statist.* **15** 1023–1030.
- Kronmal, R.A. and Tarter, M.E. (1968). "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods" *J. Amer. Statist. Assoc.* **63** 925–952.
- Kruskal, J.B. (1969). "Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation That Optimizes a New Index of Condensation." In *Statistical Computation*, R.C. Milton and J.A. Nelder (eds.), pp. 427–440. Academic Press, New York.
- Kruskal, J.B. (1972). "Linear Transformation of Multivariate Data to Reveal Clustering." In *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Vol. I, Theory*, R.N. Shepard, A.K. Romney, and S.B. Nerlove (eds.), pp. 179–191. Seminar Press, London.
- Li, K.-C. (1991). "Sliced Inverse Regression for Dimension Reduction (with Discussion)" *J. Amer. Statist. Assoc.* **86** 316–342.
- Loftsgaarden, D.O. and Quesenberry, C.P. (1965). "A Nonparametric Estimate of a Multivariate Density Function" *Ann. Math. Statist.* **36** 1049–1051.
- Lorensen, W.E. and Cline, H.E. (1987). "Marching Cubes: A High Resolution 3D Surface Construction Algorithm" *Computer Graphics* **21** 163–169.
- Mack, Y.P. and Müller, H.G. (1989). "Convolution Type Estimators for Nonparametric Regression" *Statist. Prob. Letters* **7** 229–239.
- Marron, J.S. and Nolan, D. (1988). "Canonical Kernels for Density Estimation" *Statist. Prob. Letters* **7** 195–199.
- Marron, J.S. and Wand, M.P. (1991). "Exact Mean Integrated Squared Error" *Ann. Statist.* In press.
- Marshall, A.W. and Olkin, I. (1985). "A Family of Bivariate Distributions Generated by the Bivariate Bernoulli Distribution" *J. Amer. Statist. Assoc.* **80** 332–338.
- Mathematica (1991). "A System for Doing Mathematics by Computer." Champaign, IL.
- Matthews, M.V. (1983). "On Silverman's Test for the Number of Modes in a Univariate Density Function." Honors Bachelor's thesis, Harvard University.
- McDonald, J.A. (1982). "Interactive Graphics for Data Analysis." ORION Technical Report 011, Stanford University.
- de Montricher, G.F., Tapia, R.A., and Thompson, J.R. (1975). "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods" *Ann. Statist.* **3** 1329–1348.

- Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, Berlin.
- Nadaraya, E.A. (1964). "On Estimating Regression" *Theory Probab. Applic.* **15** 134–137.
- Nezames, D. (1980). "Some Results for Estimating Bivariate Densities Using Kernel, Orthogonal Series, and Penalized-Likelihood Procedures." Ph.D. thesis, Department of Mathematical Sciences, Rice University.
- O'Sullivan, F. (1986). "A Statistical Perspective on Ill-Posed Inverse Problems" *Statist. Sci.* **1** 502–527.
- Park, B.U. and Marron, J.S. (1990). "Comparison of Data-Driven Bandwidth Selectors" *J. Amer. Statist. Assoc.* **85** 66–72.
- Parzen, E. (1962). "On Estimation of Probability Density Function and Mode" *Annals Math. Statist.* **33** 1065–1076.
- Parzen, E. (1979). "Nonparametric Statistical Data Modeling" *J. Amer. Statist. Assoc.* **74** 105–131.
- Pearson, E.S. (1938). *Karl Pearson: An Appreciation of Some Aspects of His Life and Work*. Cambridge University Press, Cambridge, England.
- Pearson, K. (1894). "Contributions to the Mathematical Theory of Evolution" *Philosophical Trans. Royal Society London (A)* **185** 71–110.
- Pearson, K. (1902a). "On the Systematic Fitting of Curves to Observations and Measurements, I" *Biometrika* **1** 265–303.
- Pearson, K. (1902b). "On the Systematic Fitting of Curves to Observations and Measurements, II" *Biometrika* **2** 1–23.
- Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Academic Press, Orlando, FL.
- Redner, R.A. and Walker, H.F. (1984). "Mixture Densities, Maximum Likelihood and the EM Algorithm" *SIAM Review* **26** 195–202.
- Reinsch, C.H. (1967). "Smoothing by Spline Functions" *Numerische Mathematik* **10** 177–183.
- Rice, J.A. (1984a). "Boundary Modification for Kernel Regression" *Commun. Statist.* **13** 893–900.
- Rice, J.A. (1984b). "Bandwidth Choice for Nonparametric Kernel Regression" *Ann. Statist.* **12** 1215–1230.
- Ripley, B.D. (1981). *Spatial Statistics*. John Wiley, New York.
- Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge, England.
- Roeder, K. (1990). "Density Estimation with Confidence Sets Exemplified by Super-clusters and Voids in the Galaxies" *J. Amer. Statist. Assoc.* **85** 617–624.
- Romano, J.P. and Siegel, A.F. (1986). *Counterexamples in Probability and Statistics*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function" *Ann. Math. Statist.* **27** 832–837.

- Rosenblatt, M. (1969). "Conditional Probability Density and Regression Estimates." In *Multivariate Analysis II*, P.R. Krishnaiah (ed.), pp. 25–31. Academic Press, New York.
- Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.
- Rudemo, M. (1982). "Empirical Choice of Histograms and Kernel Density Estimators" *Scandinavian Journal of Statistics* **9** 65–78.
- S-PLUS (1990). "User's Manual." StatSci, Inc., Seattle, WA..
- Sager, T.W. and Thisted, R.A. (1982). "Maximum Likelihood Estimation of Isotonic Modal Regression" *Ann. Statist.* **10** 690–707.
- Sain, S., Baggerly, K., and Scott, D.W. (1992). "Bootstrap Cross-Validation of Multivariate Densities." Technical Report, Rice University.
- Schoenberg, I. (1964). "On Interpolation by Spline Functions and Its Minimum Properties" *Internat. Ser. Numer. Analysis* **5** 109–129.
- Schucany, W.R. (1989). "Locally Optimal Window Widths for Kernel Density Estimation with Large Samples" *Statist. Prob. Letters* **7** 401–405.
- Schucany, W.R. and Sommers, J.P. (1977). "Improvement of Kernel Density Estimators" *J. Amer. Statist. Assoc.* **72** 420–423.
- Schuster, E.F. and Gregory, G.G. (1981). "On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimators," *Proceedings of the Thirteenth Interface of Computer Science and Statistics*, W.F. Eddy (ed.) pp. 295–298. Springer-Verlag, New York.
- Schwartz, S.C. (1967). "Estimation of a Probability Density by an Orthogonal Series" *Ann. Math. Statist.* **38** 1262–1265.
- Scott, D.W. (1976). "Nonparametric Probability Density Estimation by Optimization Theoretic Techniques." Ph.D. thesis, Department of Mathematical Sciences, Rice University.
- Scott, D.W. (1979). "On Optimal and Data-Based Histograms" *Biometrika* **66** 605–610.
- Scott, D.W. (1980). "Comment on a Paper by Good and Gaskins" *J. Amer. Statist. Assoc.* **75** 61–62.
- Scott, D.W. (1981). "Using Computer-Binned Data for Density Estimation," *Proceedings of the Thirteenth Interface of Computer Science and Statistics*, W.F. Eddy (ed.), pp. 292–294. Springer-Verlag, New York.
- Scott, D.W. (1983). "Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions," *Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing*, pp. 387–397.
- Scott, D.W. (1984). "Multivariate Density Function Representation," *Proceedings of the Sixth Annual National Computer Graphics Association Conference*, Volume II, pp. 794–801.
- Scott, D.W. (1985a). "Frequency Polygons" *J. Amer. Statist. Assoc.* **80** 348–354.
- Scott, D.W. (1985b). "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions" *Ann. Statist.* **13** 1024–1040.
- Scott, D.W. (1985c). "Classification Using Multivariate Nonparametric Density Estimation," *Proceedings of the Sixth Annual National Computer Graphics Association Conference*, Volume III, pp. 715–718.

- Scott, D.W. (1986). "Data Analysis in 3 and 4 Dimensions With Nonparametric Density Estimation." In *Statistical Image Processing and Graphics*, E.J. Wegman and D.J. Deprist (eds.), pp. 291–305. Marcel Dekker, New York.
- Scott, D.W. (1988a). "A Note on Choice of Bivariate Histogram Bin Shape" *J. Official Statistics* **4** 47–51.
- Scott, D.W. (1988b). "Comment on "How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum" by W. Härdle, P. Hall, and J.S. Marron" *J. Amer. Statist. Assoc.* **83** 96–98.
- Scott, D.W. (1990). "Statistics in Motion: Where Is It Going?" *Proceedings of the Statistical Graphics Section*, pp. 17–22, American Statistical Association, Alexandria, VA.
- Scott, D.W. (1991a). "On Estimation and Visualization of Higher Dimensional Surfaces." In *IMA Computing and Graphics in Statistics*, Volume 36 in IMA Volumes in Mathematics and its Applications, P. Tukey and A. Buja (eds.), pp. 187–205. Springer-Verlag, New York.
- Scott, D.W. (1991b). "On Density Ridges." Technical Report, Rice University.
- Scott, D.W. (1991c). "Comment on "Transformations in Density Estimation (with Discussion)" by Wand, M.P., Marron, J.S., and Ruppert, D." *J. Amer. Statist. Assoc.* **86** 359.
- Scott, D.W. and Factor, L.E. (1981). "Monte Carlo Study of the Three Data-Based Nonparametric Density Estimators" *J. Amer. Statist. Assoc.* **76** 9–15.
- Scott, D.W., Gorry, G.A., Hoffmann, R.G., Barboriak, J.J., and Gotto, A.M. (1980). "A New Approach for Evaluating Risk Factors in Coronary Artery Disease: A Study of Lipid Concentrations and Severity of Disease in 1847 Males" *Circulation* **62** 477–484.
- Scott, D.W., Gotto, A.M., Cole, J.S., and Gorry, G.A. (1978). "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease: A Study of 371 Males with Chest Pain" *Journal of Chronic Diseases* **31** 337–345.
- Scott, D.W. and Hall, M.R. (1989). "Interactive Multivariate Density Estimation in the S Language," *Proceedings of the Twentieth Interface of Computer Science and Statistics*, E.J. Wegman (ed.), pp. 241–245, American Statistical Association, Alexandria, VA.
- Scott, D.W. and Jee, J.R. (1984). "Nonparametric Analysis of Minnesota Spruce and Aspen Tree Data and Landsat Data," *Proceedings of the Second Symposium on Mathematical Pattern Recognition and Image Analysis*, pp. 27–49, NASA & Texas A&M University.
- Scott, D.W. and Schmidt, H.-P. (1988). "Calibrating Histograms with Applications to Economic Data" *Empirical Economics* **13** 155–168.
- Scott, D.W. and Schmidt, H.-P. (1989). "Calibrating Histograms with Applications to Economic Data." Reprinted in *Semiparametric and Nonparametric Economics*, A. Ullah (ed.), pp. 33–46. Physica-Verlag, Heidelberg.
- Scott, D.W. and Sheather, S.J. (1985). "Kernel Density Estimation with Binned Data" *Comm. Statist.* **14** 1353–1359.
- Scott, D.W., Tapia, R.A., and Thompson, J.R. (1977). "Kernel Density Estimation Revisited" *J. Nonlinear Analysis Theory Meth. Applic.* **1** 339–372.

- Scott, D.W. and Terrell, G.R. (1987). "Biased and Unbiased Cross-Validation in Density Estimation" *J. Amer. Statist. Assoc.* **82** 1131–1146.
- Scott, D.W. and Thompson, J.R. (1983). "Probability Density Estimation in Higher Dimensions," *Proceedings of the Fifteenth Interface of Computer Science and Statistics*, J.E. Gentle (ed.), pp. 173–179, North-Holland, Amsterdam.
- Scott, D.W. and Wand, M.P. (1991). "Feasibility of Multivariate Density Estimates" *Biometrika* **78** 197–206.
- Scott, D.W. and Wilks, A.R. (1990). "Animation of 3 and 4 Dimensional ASH Surfaces." Videotape, U. Minnesota Geometry Project.
- Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.
- Sheather, S.J. (1983). "A Data-Based Algorithm for Choosing the Window Width When Estimating the Density at a Point" *Computational Statist. Data Analysis* **1** 229–238.
- Sheather, S.J. and Jones, M.C. (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation" *J. Roy. Statist. Soc. B* **53** 683–690.
- Shibata, R. (1981). "An Optimal Selection of Regression Variables" *Biometrika* **68** 45–54.
- Siegmund, D. (1988). "Confidence Sets in Change-Point Problems" *Int. St. Review* **56** 31–48.
- Silverman, B.W. (1978a). "Choosing the Window Width When Estimating a Density" *Biometrika* **65** 1–11.
- Silverman, B.W. (1978b). "Density Ratios, Empirical Likelihood and Cot Death" *Applied Statistics* **27** 26–33.
- Silverman, B.W. (1981). "Using Kernel Density Estimates to Investigate Multimodality" *J. Roy. Statist. Soc. B* **43** 97–99.
- Silverman, B.W. (1982). "Algorithm AS176. Kernel Density Estimation Using the Fast Fourier Transform" *Appl. Statist.* **31** 93–99.
- Silverman, B.W. (1984). "Spline Smoothing: The Equivalent Variable Kernel Method" *Ann. Statist.* **12** 898–916.
- Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- Speckman, P. (1987). "Kernel Smoothing in Partial Linear Models" *J. Roy. Statist. Soc. B* **50** 413–436.
- Staniswalis, J.G. (1989). "Local Bandwidth Selection for Kernel Estimates" *J. Amer. Statist. Assoc.* **84** 276–283.
- Staniswalis, J.G., Messer, K., and Finston, D.R. (1990). "Kernel Estimators for Multivariate Smoothing." Technical Report 90-01, Biostatistics, Virginia Commonwealth University.
- Stein, C.M. (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium Math. Statist. Prob.*, Vol. 1, pp. 197–206, U. California Press, Berkeley, CA.
- Stigler, S.M. (1986). *The History of Statistics*. Harvard University Press, Cambridge, MA.

- Stone, C.J. (1977). "Consistent Nonparametric Regression (with Discussion)" *Ann. Statist.* **5** 595–645.
- Stone, C.J. (1985). "Additive Regression and Other Nonparametric Models" *Ann. Statist.* **13** 689–705.
- Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions (with Discussion)" *J. Roy. Statist. Soc. B* **36** 111–147.
- Stuart, A. and Ord, J.K. (1987). *Kendall's Advanced Theory of Statistics, Volume 1*. Oxford University Press, New York.
- Sturges, H.A. (1926). "The Choice of a Class Interval" *J. Amer. Statist. Assoc.* **21** 65–66.
- Switzer, P. (1980). "Extension of Linear Discriminant Analysis for Statistical Classification of Remotely Sensed Satellite Imagery" *Mathematical Geology* **12** 367–376.
- Tapia, R.A. (1971). "The Differentiation and Integration of Nonlinear Operators." In *Nonlinear Functional Analysis and Applications*, L.B. Rall (ed.), pp. 45–101. Academic Press, New York.
- Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation*. John Hopkins University Press, Baltimore.
- Tarter, M.E. and Kronmal, R.A. (1970). "On Multivariate Density Estimates Based on Orthogonal Expansions" *Ann. Math. Statist.* **41** 718–722.
- Tarter, M.E. and Kronmal, R.A. (1976). "An Introduction to the Implementation and Theory of Nonparametric Density Estimation" *Amer. Statist.* **30** 105–112.
- Tarter, M.E., Lock, M.D., and Mellin, C.C. (1990). *Curves: Background and Program Description*. Precision Data Group, Berkeley, CA.
- Taylor, C.C. (1989). "Bootstrap Choice of the Smoothing Parameter in Kernel Density Estimation" *Biometrika* **76** 705–712.
- Terrell, G.R. (1983). "The Multilinear Frequency Spline." Technical Report, Department of Math Sciences, Rice University.
- Terrell, G.R. (1984). "Efficiency of Nonparametric Density Estimators." Technical Report, Department of Math Sciences, Rice University.
- Terrell, G.R. (1985). "Projection Pursuit Via Multivariate Histograms." Technical Report 85-7, Department of Math Sciences, Rice University.
- Terrell, G.R. (1990). "The Maximal Smoothing Principle in Density Estimation" *J. Amer. Statist. Assoc.* **85** 470–477.
- Terrell, G.R. and Scott, D.W. (1980). "On Improving Convergence Rates for Nonnegative Kernel Density Estimators" *Ann. Statist.* **8** 1160–1163.
- Terrell, G.R. and Scott, D.W. (1983). "Variable Window Density Estimates." Technical report presented at ASA meetings in Toronto.
- Terrell, G.R., and Scott, D.W. (1985). "Oversmoothed Nonparametric Density Estimates" *J. Amer. Statist. Assoc.* **80** 209–214.
- Terrell, G.R. and Scott, D.W. (1992). "Variable Kernel Density Estimation" *Ann. Statist.* In press.
- Thompson, J.R. and Tapia, R.A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM, Philadelphia.
- Tierney, L. (1990). *LISP-STAT*. John Wiley, New York.

- Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., and Gelpke, G.J. (1981). "Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients" *J. Roy. Statist. Soc. A* **144** 145–175.
- Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.
- Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
- Tukey, P.A. and Tukey, J.W. (1981). "Graphical Display of Data Sets in 3 or More Dimensions." In *Interpreting Multivariate Data*, V. Barnett (ed.), pp. 187–275. John Wiley, Chichester, England.
- Turner, D.W. and Tidmore, F.E. (1980). "FACES-A FORTRAN Program for Generating Chernoff-Type Faces on a Line Printer" *Amer. Statist.* **34** 187–187.
- U.S. Bureau of the Census (1987). "U.S. Decennial Life Tables for 1979-81; Some Trends and Comparisons of United States Life Table Data: 1900-1981." Volume 1, Number 4, DHHS Pub No. PHS 87-1150-4.
- Van Ryzin, J. (1973). "A Histogram Method of Density Estimation" *Communications in Statistics* **2** 493–506.
- Wagner, H.M. (1959). "Linear Programming Techniques for Regression Analysis" *J. Amer. Statist. Assoc.* **54** 206–212.
- Wahba, G. (1971). "A Polynomial Algorithm for Density Estimation" *Ann. Math. Statist.* **42** 1870–1886.
- Wahba, G. (1977). "Optimal Smoothing of Density Estimates." In *Classification and Clustering*, J. Van Ryzin (ed.), pp. 423–458. Academic Press, New York.
- Wahba, G. (1981). "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates" *Ann. Statist.* **9** 146–156.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
- Wahba, G. and Wold, S. (1975). "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation" *Comm. Statist.* **4** 1–17.
- Walter, G. and Blum, J.R. (1979). "Probability Density Estimation Using Delta Sequences" *Ann. Statist.* **7** 328–340.
- Wand, M.P., Marron, J.S., and Ruppert, D. (1991). "Transformations in Density Estimation (with Discussion)" *J. Amer. Statist. Assoc.* **86** 343–352.
- Wand, M.P. and Jones, M.C. (1991). "Comparison of Smoothing Parameterizations in Bivariate Density Estimation." Technical Report, Rice University.
- Wang, F.T. (1990). "A New Method for Robust Nonparametric Regression." Ph.D. thesis, Department of Statistics, Rice University.
- Wang, F.T. and Scott, D.W. (1991). "The L_1 Method for Robust Nonparametric Regression." Technical Report, Rice University.
- Wang, P.C.C. (Ed.) (1978). *Graphical Representation of Multivariate Data*. Academic Press, New York.
- Watson, G.S. (1964). "Smooth Regression Analysis" *Sankhyā Ser. A* **26** 359–372.

- Watson, G.S. (1969). "Density Estimation by Orthogonal Series" *Ann. Math. Statist.* **40** 1496–1498.
- Watson, G.S. (1985). *Statistics on Spheres*. John Wiley, New York.
- Watson, G.S. and Leadbetter, M.R. (1963). "On the Estimation of the Probability Density I" *Ann. Math. Statist.* **34** 480–491.
- Watson, G.S. and Leadbetter, M.R. (1964). "Hazard Analysis I" *Biometrika* **51** 175–184.
- Weaver, C.S., Zollweg, J.E., and Malone, S.D. (1983). "Deep Earthquakes Beneath Mount St. Helens: Evidence for Magmatic Gas Transport?" *Science* **221** 1391–1394.
- Wegman, E.J. (1970). "Maximum Likelihood Estimation of a Unimodal Density Function" *Ann. Statist.* **41** 457–471.
- Wegman, E.J. (1990). "Hyperdimensional Data Analysis Using Parallel Coordinates" *J. Amer. Statist. Assoc.* **85** 664–675.
- Wegman, E.J. and DePriest, D.J. (eds.) (1986). *Statistical Image Processing and Graphics*. Marcel Dekker, New York.
- Weisberg, S. (1985). *Applied Linear Regression*. John Wiley, New York.
- Wertz, W. (1978). *Statistical Density Estimation: A Survey*. Vandenhoeck and Ruprecht, Göttingen.
- Wharton, S.W. (1983). "A Generalized Histogram Clustering Scheme for Multidimensional Image Data" *Pattern Recognition* **16** 193–199.
- Woodroffe, M. (1970). "On Choosing a Delta-Sequence" *Ann. Math. Statist.* **41** 1665–1671.
- Worton, B.J. (1989). "Optimal Smoothing Parameters for Multivariate Fixed and Adaptive Kernel Methods" *J. Statist. Comp. Simul.* **32** 45–57.

Multivariate Density Estimation

DAVID W. SCOTT

Copyright © 1992 by John Wiley & Sons, Inc.

Author Index

- Abramson, I.S., 182, 183, 186, 187, 285
 Ahrens, L.H., 279, 285
 Aitkin, M., 257, 285
 Allen, D.M., 225, 285
 Altman, N.S., 171, 265, 285
 Andrews, D.F., 13, 285
 Asimov, D., 11, 285
 Austin, W.W., 7, 216, 285
 Badhwar, G.D., 7, 216, 285
 Baggerly, K., 180, 294
 Banchoff, R.F., 3, 285
 Barboriak, J.J., 295
 Bartlett, M.S., 133, 139, 285
 Becker, R.A., 6, 8, 222, 285
 Bellman, R.E., 82, 196, 285
 Bickel, P.J., 260, 286
 Blum, J.R., 125, 155, 156, 158, 298
 Boswell, S.B., 204, 286
 Bowman, A.W., 162, 166, 260, 264, 286, 290
 Bowyer, A., 136, 286
 Box, G.E.P., 71, 183, 286
 Breiman, L., 182, 183, 187, 244, 286
 Brookmeyer, R., 265, 286
 Brown, B.S., 276
 Cacoullos, T., 125, 286
 Carmichael, J.-P., 279, 286
 Carnes, J.G., 7, 216, 285
 Carr, D.B., 9, 11, 85, 286
 Cencov, N.N., 125, 128, 286
 Chamayou, J.M.F., 123, 286
 Chambers, J.M., 6, 7, 222, 285, 286
 Chernoff, H., 12, 15, 286
 Chiu, S.T., 171, 173, 265, 286, 287
 Chow, Y.S., 163, 287
 Clark, R.M., 225, 287
 Clark, V.A., 17, 289
 Cleveland, W.S., 3, 8, 11, 12, 222, 232,
 285–287
 Cline, H.E., 26, 271, 292
 Cole, J.S., 275, 295
 Coleman, D.E., 262
 Cox, D.D., 264, 287
 Cox, D.R., 71, 183, 286
 Davis, K.B., 138, 287
 Day, N.E., 87, 257, 287
 Deheuvels, P., 189, 212, 287
 Dempster, A.P., 257, 287
 DePriest, D.J., 263, 299
 Devroye, L., 3, 41, 287
 Diaconis, P., 8, 31, 54–56, 210, 287, 289
 Diamond, R., 271, 287
 DiCiccio, T.J., 260, 290
 Doane, D.P., 48, 91, 287
 Dobkin, D.P., 267, 271, 287
 Donoho, A.W., 9, 287
 Donoho, D.L., 9, 64, 287
 Duan, N., 205, 288
 Duda, R.O., 51, 288
 Duin, R.P.W., 163, 288
 Efron, B., 36, 252, 288
 Embretson, J.J., 265, 288
 Emerson, J.D., 47, 288
 Engel, J., 225, 289
 Epanechnikov, V.K., 83, 125, 134, 139–
 141, 192, 196, 198, 199, 209, 288
 Eubank, R.L., 3, 219, 226, 264, 288
 Everitt, B.S., 87, 288
 Factor, L.E., 161–164, 295
 Fan, J.Q., 222, 225, 231, 288
 Farin, G.E., 271, 288
 Farrell, R.H., 186, 288
 Fienberg, S.E., 13, 288

- Finston, D.R., 155, 296
 Fisher, N.I., 265, 288
 Fisher, R.A., 4, 33, 95, 213, 288
 Fisherkeller, M.A., 12, 288
 Fix, E., 122, 125, 288
 Flury, B., 12, 288
 Foley, J.D., 271, 289
 Freedman, D., 54–56, 210, 287, 289
 Friedman, J.H., 8, 9, 12, 31, 202, 209, 211,
 216, 226, 239, 242, 286–289
 Fryer, M.J., 137, 199, 289
- Galilei, Galileo, 17
 Galton, F., 4, 289
 Gaskins, R.A., 18, 86–88, 155, 256, 279, 289
 Gasko, M., 9, 287
 Gasser, T., 141, 146, 164, 225, 226, 232,
 289, 290
 Gelpke, G.J., 298
 Geman, S., 163, 287
 Good, I.J., 18, 86–88, 155, 256, 279, 289
 Gorry, G.A., 275, 295
 Gotto, A.M., 275, 295
 Graunt, J., 17, 289
 Gregory, G.G., 163, 294
 Gross, A.J., 17, 289
 Györfi, L., 3, 41, 265, 287, 289
- Habbema, J.D.F., 163, 250, 289, 298
 Hald, A., 17, 289
 Hall, D.L., 286
 Hall, M.R., 26, 295
 Hall, P., 41, 77, 90, 91, 161, 167, 173,
 186, 200, 225, 253, 260, 289, 290
 Hampel, F.R., 232, 290
 Hand, D.J., 87, 251, 288, 290
 Härdle, W., 3, 26, 123, 219, 221, 222, 225,
 232, 238, 255, 260, 289, 290
 Hart, J.D., 256, 265, 290, 291
 Hart, P.E., 51, 288
 Hartigan, J.A., 257, 291
 Hartigan, P.M., 257, 291
 Hastie, T.J., 3, 219, 231, 239, 264, 291
 Hathaway, R.J., 87, 257, 291
 Hermans, J., 163, 250, 289
 Hieber-Dodd, K.L., 12, 291
 Hjort, N.L., 90, 106, 107, 120, 291
 Hoaglin, D.C., 47, 288
 Hodges, J.L., 72, 122, 125, 139, 288, 291
 Hoerl, A.E., 37, 291
 Hoffmann, R.G., 295
 Huber, P.J., 11, 196, 210, 232, 291
 Hüsemann, J.A., 80, 291
- Inselberg, A., 13, 291
 Izenman, A.J., 86, 257, 291
- Jee, J.R., 7, 210, 211, 215, 291, 295
 Johnson, N.L., 44, 291
 Johnson, R.A., 196, 207, 291
 Jones, M.C., 162, 164, 172, 173, 180, 182,
 210, 238, 290–292, 296, 298
- Kaplan, E.L., 261, 266, 292
 Kappenman, R.F., 162, 172, 292
 Kendall, M.G., 27, 292
 Kennard, R.W., 37, 291
 Kleiner, B., 286
 Klonias, V.K., 155, 292
 Kogure, A., 67, 292
 Koh, E., 264, 287
 Kronmal, R.A., 125, 127–129, 162, 163,
 292, 297
- Kruskal, J.B., 209, 292
- Laird, N.M., 257, 287
 Leadbetter, M.R., 125, 157, 158, 260, 299
 Lehmann, E.L., 72, 139, 291
 Leroy, A.M., 232, 294
 Levy, S.V.F., 287
 Lewis, T., 265, 288
 Li, K.-C., 242, 292
 Littlefield, J.S., 286
 Littlefield, R.J., 286
 Lock, M.D., 235, 297
 Loftsgaarden, D.O., 125, 182, 292
 Lorenzen, W.E., 26, 271, 292
- Mack, Y.P., 224, 292
 Makov, U.E., 257, 298
 Malone, S.D., 15, 205, 299
 Mammitzsch, V., 141, 289
 Marron, J.S., 71, 77, 138, 142, 161, 162,
 173, 183, 186, 205, 225, 255, 290, 292,
 293, 298
- Marshall, A.W., 44, 292
 Matthews, M.V., 257, 292
- McDonald, J.A., 8, 260, 292
 McGill, M.E., 3, 8, 11, 12, 287
 Meier, P., 261, 266, 292
 Meisel, W., 182, 183, 187, 286
 Mellin, C.C., 235, 297
 Messer, K., 155, 296
 de Montricher, G.F., 155, 292
 Müller, H.G., 3, 141, 146, 148, 219, 224–
 226, 289, 292, 293
- Murray, G.D., 298
 Murray, L.S., 298

- Nadaraya, E.A., 220, 222–225, 231, 236, 237, 244, 293
- Nezames, D., 180, 293
- Nicholson, W.L., 9, 11, 286
- Nolan, D., 142, 292
- O'Sullivan, F., 265, 293
- Olkin, I., 44, 292
- Olshen, A., 286
- Ord, J.K., 222, 224, 297
- Park, B.U., 162, 293
- Parzen, E., 109, 122, 125, 131, 279, 293
- Pearson, E.S., 4, 5, 293
- Pearson, K., 4, 5, 33, 43–45, 257, 264, 293
- Prakasa Rao, B.L.S., 3, 293
- Purcell, E., 182, 183, 187, 286
- Quesenberry, C.P., 125, 182, 292
- Redner, R.A., 87, 293
- Reinsch, C.H., 228, 229, 293
- Rice, J.A., 146, 148, 225, 293
- Riedwyl, H., 12, 288
- Ripley, B.D., 263, 271, 293
- Roeder, K., 257, 293
- Romano, J.P., 39, 260, 290, 293
- Rosenblatt, M., 36, 122, 125, 126, 138, 143, 184, 224, 260, 286, 293, 294
- Rousseeuw, P.J., 232, 294
- Rubin, D.B., 257, 287
- Rudemo, M., 76, 77, 166, 294
- Ruppert, D., 71, 183, 205, 298
- Sager, T.W., 235, 294
- Sain, S., 180, 294
- Sarda, P., 289
- Schmidt, H.-P., 79, 295
- Schoenberg, I., 227, 294
- Schroeder, A., 202, 289
- Schucany, W.R., 137, 138, 226, 288, 294
- Schuster, E.F., 163, 294
- Schwartz, S.C., 125, 294
- Scott, D.W., 5–7, 16, 24, 26, 41, 44, 54–56, 67, 72, 73, 76, 77, 79, 80, 84, 85, 93, 95, 98, 101, 106, 109, 113, 119, 120, 123–125, 131, 138, 141, 156, 161–170, 180, 185–191, 200, 202, 203, 205, 221, 236, 238, 242, 249, 251, 265, 275, 290, 294–298
- Serfling, R.J., 253, 254, 266, 296
- Sheather, S.J., 164, 173, 238, 290, 295, 296
- Shibata, R., 225, 296
- Sibson, R., 210, 292
- Siegel, A.F., 39, 293
- Siegmund, D., 235, 296
- Silverman, B.W., 3, 30, 44, 125, 136, 161, 164, 186, 187, 199, 229, 238, 256, 266, 296
- Skene, A.M., 298
- Smith, A.F.M., 257, 298
- Sommer, C.J., 86, 257, 291
- Sommers, J.P., 138, 294
- Speckman, P., 198, 264, 288, 296
- Spiegelhalter, D.J., 298
- Staniswalis, J.G., 155, 226, 296
- Stein, C.M., 37, 296
- Stigler, S.M., 4, 115, 296
- Stone, C.J., 219, 221, 286, 297
- Stone, M., 225, 297
- Stuart, A., 222, 224, 297
- Stuetzle, W., 202, 239, 264, 289, 291
- Sturges, H.A., 47, 48, 51, 55, 75, 79, 91, 297
- Switzer, P., 209, 297
- Tapia, R.A., 3, 44, 155, 157, 164, 292, 295, 297
- Tarter, M.E., 125, 127–129, 162, 163, 235, 292, 297
- Taylor, C.C., 170, 171, 257, 259, 297
- Terrell, G.R., 44, 67, 72–77, 80, 93, 101, 106, 125, 138, 156, 159, 161, 162, 165–170, 180, 181, 185–191, 213, 214, 291, 296, 297
- Thisted, R.A., 235, 294
- Thompson, J.R., 3, 7, 16, 26, 44, 84, 155, 164, 292, 295–297
- Thurston, W.P., 287
- Tibshirani, R.J., 3, 219, 231, 239, 291
- Tidmore, F.E., 12, 298
- Tierney, L., 11, 161, 297
- Titterington, D.M., 250, 257, 260, 290, 298
- Tufte, E.R., 6, 16, 298
- Tukey, J.W., 3, 4, 8, 9, 12, 20, 47, 52, 71, 196, 204, 209, 211, 216, 288, 289, 298
- Tukey, P.A., 3, 8, 9, 196, 286, 298
- Tunnicliffe Wilson, G., 257, 285
- Turner, D.W., 12, 298
- Van Dam, A., 271, 289
- Van Der Broek, K., 163, 250, 289
- Van Ryzin, J., 71, 298
- Vieu, P., 265, 289, 291
- Wagner, H.M., 235, 298
- Wahba, G., 3, 125, 129, 161, 163, 164, 190, 198, 219, 225, 239–241, 264, 265, 287, 298
- Walker, H.F., 87, 293
- Walter, G., 125, 155, 156, 158, 298
- Wand, M.P., 41, 71, 90, 91, 138, 180, 183, 200, 202, 203, 205, 253, 290, 292, 296, 298

- Wang, F.T., 222, 235, 236, 298
Wang, P.C.C., 3, 6, 12, 298
Watson, G.S., 125, 128, 129, 157, 158, 220,
 222–225, 231, 236, 237, 244, 260, 265,
 298, 299
Weaver, C.S., 15, 205, 299
Wegman, E.J., 13, 29, 32, 71, 263, 299
Weisberg, S., 18, 278, 299
Wertz, W., 3, 299
Wharton, S.W., 83, 299
Wichern, D.W., 196, 207, 291
Wilks, A.R., 6, 24, 26, 222, 285, 287, 296
Wold, S., 225, 298
Woodroffe, M., 125, 164, 299
Worton, B.J., 187, 199, 299
Wu, L.D., 163, 287
Yandell, B., 264, 287
Zollweg, J.E., 15, 205, 299

Subject Index

- Adaptive estimators:**
 ad hoc vs. optimal, 71
 frequency polygons, *which see*
 heuristic, 69
 histograms, *which see*
 kernel estimators, *which see*
 multivariate, 181, 187–190, 212
 nearest-neighbor estimator, *which see*
 null space, 68–69, 184–185, 192
 regression, *which see*
- Additive models, *see* Regression**
- Algorithms, *see also* Cross-validation**
 ASH1, 118
 ASH2, 121
 backfitting, 239
 BIN1, 117
 BIN2, 121
 bootstrap, 258
 contouring, 267–271
 marching cubes, *which see*
 RAM, 239–240
 REG-ASH, 238
 REG-BIN, 237
 SIR, 242
- Andrews' curve, *see* Glyphs**
- Animation, 5, 11, 26, 161**
- Asymptotic Normality, 167–169, 253–255**
- Asymptotic relative efficiency, 139–140**
- Averaged shifted histograms, 113–124, 199, 238, 240**
 asymptotic relative efficiency, 139–140
 averaged shifted frequency polygons, 113, 124, 139–140
 bin origin problems, 109, 113–115
 bin weights, 116–118
 bumps, 114
- construction, 113–115**
- cross-validation algorithm, 170**
- in discrimination, 249–251**
- equivalent sample sizes, 119–120**
- frequency polygon of, 113**
- historical note, 115–116**
- indifferent histogram, 122, 131, 139, 190**
- limiting form, 121–123, 150**
- MISE, 119–120**
- multivariate, 114–115**
 linear blend, 113, 120, 235
 as product triangle kernel, 123, 150
 rate of convergence, 120–121
 smoothing parameter, 120
- Normal reference rule, 119**
- regression, 236–238**
- WARPing, *which see***
- Backfitting, 239**
- Bandwidth:**
 critical, *see* Bump hunting
 parameter, *see* Smoothing parameters
 selection, *see* Cross-validation
- Basis functions, 128–129, 162**
- Bayes' rule, 42, 247–249**
- Beta distribution, 43, 53, 68, 70, 92–93, 104, 112, 147–148, 193**
- Bias, 37**
 component, 37
 downward at peaks, 256, 259–260
 inherent property, 36–37, 258
 of parametric estimator, 33
 variance-bias trade-off, 50
- Bilinear form, 269**
- Bills of Mortality, 17**

- Binning:
 averaged shifted histogram, 114, 117, 121
 data, 17, 114
 frequency polygon, *which see*
 histogram, *which see*
 multivariate:
 equilateral triangle, 85
 hexagonal, 85–86
 square, 85
 regression, 236–239
- Binomial distribution:
 empirical cdf, 34
 frequency polygon error, 97
 histogram error, 49, 87, 91
 Sturges' rule, 47–48, 55–56
- Blood fats, *see* Data sets, lipid
- Blurring or jittering:
 in cross-validation, 79–80
 scatter diagram, 7, 19, 179
- Bootstrap:
 algorithms, 258
 confidence intervals, 257–260
 critical bandwidth, *see* Bump hunting
 cross-validation, *which see*
 empirical distribution, 36
 goodness-of-fit test, 264
 MISE and MSE, 170–171
 ordinary, 170, 258, 266
 smoothed, 170–171, 256–258
- Boundary effects:
 frequency polygon, *which see*
 histogram, *which see*
 kernel estimator, *which see*
 scatter diagrams, *which see*
- Brushing, *see* Scatter diagrams
- Buffalo snowfall data, *see* Data sets, snowfall
- Bump hunting, *see also* Modes
 critical bandwidth testing, 256–257
 definition, 86
 dip test, 257
 iterative surgery, 256
 Normal mixture model, 86, 94, 257
- Bumps:
 in data, 17–18, 87–88, 114, 149, 179
 definition, 86
 derivative, *see* Density derivative estimators
 multivariate, 205–206
 oversmoothing, 255
 spurious, 25, 87, 89, 105–106, 149, 256
- Cauchy distribution, 68, 75, 92, 104, 106, 112, 184, 236–237
- Censored data, 261, 266
- Change-point problem, 234–235
- Chernoff faces, *see* Glyphs
- Chi-squared distribution, 84
 goodness-of-fit, 4, 33, 58, 67, 70, 93, 264
- Cholesterol, *see* Data sets, lipid
- Classification, 247
 average linkage, 251
 bump hunting, *which see*
 CART, 244
 clusters, 247, 251
 brushing, 9
 elliptical, 3
 multiple, 6, 15–16, 23–24, 205–206, 208–211, 215, 251
 of outliers, 236
 parallel coordinates, 31–32
 density function, 35
 jackknife, 252
 likelihood ratio, 248–251
 majority prediction, 252
 misclassification error rates, 42–43, 252
 nearest-neighbor, 251
 prediction, 250–253
 by regression, 252–253
- Clusters, *see* Classification
- Coefficient of detection, 248, 265
- Coefficient of variation, 168
- Collinearity, 195. *See also* Rank deficiency
- Complete statistic, 34, 36
- Computer rendering of 3-D objects, 271–272
- Conditional density, 23, 223, 233. *See also*
 Slicing
- Conditional mode, *see* Regression
- Confidence intervals:
 density estimates, 259–260
 modes:
 conditional, 254–255
 sample, 105–106, 253–255
 regression estimates, 260
- Consistency of:
 empirical distribution function, 34
 Fourier coefficients, 128
 nonparametric estimators, 45, 51, 157–159
- Constrained oversmoothing, *see* Cross-validation
- Contour:
 algorithm:
 bivariate, 25, 267–269
 localness, 25, 267, 270
 marching cubes, 26–27, 271
 trivariate, 25, 270–271
 α -level, 22, 24–26, 32
 bivariate, 21, 267–269

- clusters, 196, 261
- fraction of dimension, 21, 23
- holes, 196, 262, Color Plates 13–16
- likelihood ratio, 249–251
- nested, 22, 197–198
 - as signature of hole, 261
- nonnested, 261
- Normal, 22, 30, 32
- parallel planes, 25
- vs. perspective, 250
- regression, 24
- of rootgram, 52
- shell, 25–27
- smoothness, 26
- stereo, 22
- surfaces, 22, 25–27
- transparency, 26, Color Plates 6, 8, 10–12, 15–16
- trivariate, 22–24, 26, 270–271
- quadrivariate, 24
- Convolution, 123, 127–129, 171, 225, 265
- Covariance matrix, 1, 30, 154, 181, 195, 206–208, 210–211, 243, 249
- Cramer-Rao lower bound, 51
- Criteria:
 - choice of, 37
 - error, 37. *See also* Mean squared error
 - approximation, 238
 - statistical, 238
 - global, 38, 52, 70, 189, 200, 202–203
 - Hellinger distance, 39, 210
 - imperfection, 2, 37
 - integrated mean squared error, 38
 - integrated squared bias, 54
 - integrated squared error, 38, 60–61
 - asymptotic, 54
 - exact, 60, 162
 - integrated variance, 52
 - asymptotic, 54
 - exact, 60, 191
 - Kullback-Leibler, 38, 45
 - L_1 , 38, 41–43
 - asymptotics, 41
 - bounds, 41
 - dimensionless, 41–42, 199–202
 - histogram, 90–91
 - interpretability, 42–43
 - kernel, 200–202
 - vs. L_2 , 38, 41, 45
 - as misclassification probability, 43
 - transformation invariance, 42
 - L_2 , 38, 41
 - L_4 , L_6 , L_8 , 91
 - L_p , 39
 - L_∞ , 38
 - mean integrated squared error, 38. *See also* each estimator
 - asymptotic, 41, 45, 54, 150
 - asymptotically optimal, 55, 63–64, 151–152
 - exact, 39, 45, 60–64, 135, 199
 - minimax, 2
 - optimality, 2, 37
 - pointwise, 37–38
 - sensitivity to smoothing parameter, 58–60, 63, 92, 99, 136–137, 161
 - stability, 2
 - subjectivity, 37
 - Critical bandwidth testing, *see* Bump hunting
 - Cross-validation, 72–80, 101–103, 160–181
 - autocorrelation effect, 171, 265
 - averaged shifted histogram, 170
 - biased cross-validation:
 - asymptotic theory, 168–169
 - examples, 78–80, 170–173, 175–176, 179
 - frequency polygon, 101–102, 112
 - histogram, 75–80, 93, 101–102
 - horizontal variance, 168
 - kernel, 160–162, 166–173
 - local minimizer, 167, 171–173, 175
 - Normal kernel, 167
 - orthogonal series, 163
 - problem with large h , 77–78
 - slow convergence, 162
 - vertical variance, 168
 - bootstrap, 170–172, 176
 - example, 171–173
 - multivariate, 179–180
 - Normal kernel, 171
 - constrained oversmoothing, 177–179
 - fixed-point algorithm:
 - Gasser's modification, 164
 - Scott-Tapia-Thompson, 164
 - generalized, 161, 163
 - interactive search, 161, 177
 - least-squares, *see* Cross-validation, unbiased
 - leave-one-out, 163, 214
 - maximum likelihood, 32, 163
 - Normal reference rules:
 - averaged shifted histogram, 119
 - compared to oversmoothed, 166
 - compared to plug-in, 175
 - frequency polygon, 99–100, 103, 109
 - histogram, 55, 82

- Cross-validation (*Continued*)
 kernel, 131, 152
 robust kernel version, 174–175
 use in fixed-point algorithm 164
- orthogonal series, 44
 exact MISE, 162
 generalized cross-validation, 163
 unbiased estimate, 162–163
- oversmoothed:
 constrained, 177–179
 derivative functions, 255
 examples, 170–173
 frequency polygon, 101–103
 histogram, 72–75, 79
 kernel, 165–166
 multivariate, 180–181
 upper bounds, 72
- plug-in, 172–177
 auxiliary bandwidth, 162, 174–176
 divergence, 174
 examples, 176
fast convergence, 172
 improved AMISE, 173
 lack of local minimum, 175–176
 local minimum, 174
 Normal kernel, 174
 origin, 164
 vs. oversmoothed, 175–176
 sensitivity, 175–176
- rate of convergence, 76–77, 162, 168, 172, 175, 187
- regression, 225
- simulation results, 162, 168
- target bandwidth:
 adequate accuracy, 162
 AMISE, 162
 correlation to sample variance, 161
 ISE, 137, 161–162
 MISE, 161–162
 slow convergence, 162
- target function, 161
- test graph method, 161, 164–165
- unbiased cross-validation:
 adjusting data, 193
 blurring, 79
 examples, 78–80, 170–173, 175–176, 179
 frequency polygon, 101, 112
 histogram, 75–80
 horizontal variance, 168
 kernel, 160–162, 166–173
 orthogonal series, 162–163
 problem with small h , 78–79
 slow convergence, 162
- vertical variance, 168
- Cumulative distribution function, 34
 empirical:
 bivariate, 35, 36, 45
 univariate, 34–35, 157, 260–261
 use in finite difference, 126
 unbiased estimation of, 34, 36
- Curse of dimensionality, 3, 5, 29, 71, 82–85, 152, 193, 196–206, 239–241, 244
 additive models, *see* Regression
 Bellman, 196
 due to dimension deficiency, 85, 198
 effect of data transformation, 71, 204–206
 empty space phenomenon, 84, 198
 hill-climbing, 204
 illustration, 196
 rank deficiency, 198, 204
 vs. high dimensions, 206
- Curse of optimality, 2, 33
- Data-based algorithms, *see* Cross-validation
- Data sets:
 DIW, 79–80, 95, 101–102, 118
 earthquake, 15, 26, 205–206, Color Plate 8
 gas flow, 20–21, 229–231, 240–241, 245, 276–277
 geyser, 18, 34–35, 86, 171, 173, 176, 233–234, 260, 277–278
 hole, bivariate, 197, 261
 hole, trivariate, 31, 261–262, Color Plates 13–16
 iris, 6, 9, 11, 15, 24, 31, 217, Color Plates 1–2
 LANDSAT, 7–8, 19–20, 23, 216, 251–252, 263, Color Plates 9–10
 life table, 16–17
 lipid, 6–7, 86, 115, 143–145, 180, 249–251, 253, 275
 LRL, 18, 73–74, 87–88, 90, 93, 95, 102, 112, 256, 278–279
 marathon times, 149
 Normal, trivariate, 23–24, Color Plates 3–7, Book jacket
 penny, 37, 87, 234–235, 245, 276
 PRIM4, 9–10, 204–205, 214, 217, Color Plates 11–12
 PRIM7, 204, 211–213, 215–216
 silica, 123, 259, 279
 snowfall, 73–74, 93, 109–110, 114, 118, 127, 137, 171–172, 176, 179, 191, 279
 steel surface, 136, 170, 176, 179
 U.S. economy, 6, 12–14, 31, 273
 universities, 6, 12, 31, 274

- Data transformation:**
 alternative to adaptive estimation, 71, 104, 153
Box-Cox family, 71
 adaptive estimation, 183
 optimal choice, 71
 equivalence to kernel choice, 153, 192
 examples, 7, 79, 204–206, Color Plates 11–12
 marginal variables, 1, 71, 180, 185, 204–206, 208, 217
 model-based, 215–216
 principal components, 7, 206–209, 212, 214, 216, 241–243
 probability, 262
 sphering, 207, 209, 213
 Tukey ladder, 71, 180, 204
 variance-stabilizing, which *see*
- Delta function, Kronecker**, 128
Delta methods, 155–159
Delta sequences, 125, 157
Density estimators:
 adaptive, which *see*
 averaged shifted histograms, which *see*
 finite difference estimators, which *see*
 frequency polygons, which *see*
 histograms, which *see*
 kernel estimators, which *see*
 maximum penalized-likelihood, 155–156
 nearest-neighbor estimators, which *see*
 orthogonal series estimators, which *see*
 parametric, which *see*
 rootgram, *see* Histograms
 on sphere, 265
 WARPing, which *see*
- Density derivative estimators:**
 difficulties:
 nonconsistency of second derivative, 132
 slower rates of convergence, 132
 finite difference estimator, 75, 126–127, 143–145, 190
 higher-order, 138, 173
 kernel estimator, 131
 sensitivity to smoothing parameter, 58–59, 131
MISE:
 bias, 132
 optimal, 132
 rates of convergence, 132
 variance, 132
 oversmoothed, 255
 smoothing parameter, 132
Density ridge, 264–265
- Derivative**, *see* Density derivative estimators
Designer kernels, 133, 141, 143, 146–149
Dimensional analysis:
 density function, 41
 density functionals, 91
 L_1 criterion dimensionless, 41, 200
 L_2 criterion, 41
Dimension reduction, 152, 193, 196, 206–216, 243
Dip test, *see* Bump hunting
Dirac delta function, 35, 92, 126, 157, 257
Discrimination, 6, 125, 247
 coefficient of detection, 248
 likelihood ratio, 248–251
 posterior odds, 248
 prior odds, 237
 shifted-Normal model, 248–250
DIW, *see* Data sets
Dot plot, 36
- Earthquake**, *see* Data sets
Economic data, *see* Data sets, U.S. economy and DIW
Eigenstructure, 188–189, 206–208, 214–215, 243, 265
Empirical cdf, *see* Cumulative distribution function
Empirical pdf, *see* Probability density function
Equivalent kernel smoothing, 118, 141–144, 156
 nonparametric regression, *see* Regression Normal example, 160
 orthogonal series, 129, 190
 for parametric estimator, 159–160, 191–192
table for conversion, 142
Equivalent sample sizes:
 averaged shifted histogram, 119–120
 across dimension:
 histogram, 83
 kernel, 198–205
 frequency polygon, 99, 139–140
 histogram vs. parametric, 57–58
 kernel, 198–205
Euler-Lagrange equation, 228, 244
Exploratory data analysis, 1, 3–4
- Factor analysis**, 195
Fast Fourier transform, 238
Filters, *see also* Kernel functions
 boxcar, 128–129, 186
 convolution, 127

- Filters (Continued)**
- half-power point, 128
 - high-pass, 126
 - leakage, 129
 - low-pass, 126
 - orthogonal series equivalent kernels, 129, 190
 - sinc function, 129, 138
 - tapering window, 129, 238
- Finite difference estimators:**
- density derivative, which *see*
 - higher-order, 143–145, 191–192
 - one-sided, 126–127, 190
 - two-sided, 126–127, 190
- Fisher information**, 210–213
- Frequency polygons**, 95–112
- adaptive meshes:
 - definition, 103, 111
 - examples, 104
 - vs. histogram, 104
 - nearest-neighbor meshes, 112
 - theoretical improvement, 103–104, 112
 - asymptotic relative efficiency, 139–140
 - averaged shifted, 113, 124, 139–140
 - bias, 97
 - bin construction:
 - adaptive, 103–104
 - bin origin effect, 109–111
 - edge problems, 109–111
 - linear blend, 106, 108–109
 - multivariate, 106–109
 - number of bins, 102
 - simplex, 106–107
 - triangular, 106–107
 - bin width:
 - adaptive mesh, 103–104
 - asymptotically optimal, 98
 - bivariate Normal, 109
 - kurtosis modification, 57, 100–101, 152
 - multivariate, 107–109
 - Normal reference rule, 99–100, 103, 109
 - oversmoothed, 101–103
 - sensitivity, 99
 - skewness modification, 57, 100–101, 152
 - upper bound, 101–103
 - boundary effects, 98–99
 - bumps, 105–106
 - comparison to histogram, 98, 100
 - construction, 95–96
 - cross-validation, which *see*
 - equivalent sample sizes, 99, 139–140
 - Fisher's caution, 95
 - indifferent frequency polygon, 124, 139–140, 191
- MISE:**
- adaptive mesh, 103
 - asymptotically optimal, 98, 100
 - effect of boundary, 98–99, 111
 - multivariate, 106–109
 - sensitivity, 99
 - modes, 105–106
 - confidence interval, 105–106
 - effect of bin edge, 109–111
 - false sample modes, 99
 - multivariate, 106–109
 - rate of convergence, 107–108
 - rate of convergence, 98
 - variance, 97
- Fubini's theorem**, 36, 38
- Gas flow**, *see* Data sets
- Gâteaux derivative**, 156–157, 165
- Gauss–Siedel algorithm**, 239
- Geometry:**
- data representation, 4
 - hypercube, 29, 32
 - diagonals, 30
 - hypersphere, 28–29, 32
 - polar coordinates, 27–28
 - volume hypersphere, 28
- German income data**, *see* Data sets, DIW
- Geyser**, *see* Data sets
- Glyphs**, *see also* Scatter diagrams
- Andrews' curve, 13, 16, 31–32
 - asymmetric treatment of variables, 12
 - Chernoff faces, 12–13, 15, 31, 273
 - asymmetrical, 12
 - definition, 11
 - Fourier series, 13
 - hexagonal bivariate histogram, 85–86, 115
 - histogram as glyph, 16
 - star diagram, 13–14, 32
 - stem-and-leaf plot, 47
 - stick pin plot, 20–21
- Goodness-of-fit test**, *see* Chi-squared distribution
- Grand tour**, *see* Scatter diagrams
- Hanning filter**, 20
- Hazard function**, 260–261
- Hellinger distance**, *see* Criteria
- Hessian:**
- definite, 188, 265
 - flat, 189
 - saddle-shaped, 188–189

- Histograms, 47–94, 149
 adaptive meshes:
 examples, 68–69
 vs. fixed, 48, 71
 vs. frequency polygon, 104
 nearest-neighbor estimator, 156
 null space, 68–69
 percentile meshes, 69–70
 reliability, 52
 theoretical improvement, 67, 90, 93
 vs. transformation, 71
- bias, 50, 53–54, 213
- bin construction:
 adaptive, 52
 bin origin choice, 47, 54
 bin origin effect, 65–66, 109–111
 bivariate regular meshes, 85–86
 equilateral triangle, 85
 hexagonal, 85–86
 intervals, 47
 multivariate, 80
 number of bins, 73
 origin, 47, 54
 square, 85
 bin width, 47
 adaptive mesh, 67
 asymptotically optimal, 54, 62–64, 83
 bivariate Normal, 84
 Doane's rule, 48
 Freedman-Diaconis rule, 55–56
 kurtosis modification, 57
 multivariate, 82–83
 Normal reference rule, 55–56, 72, 82
 oversmoothed, 72–75, 79
 Scott's rule, 56
 sensitivity, 58–59, 63, 135
 skewness modification, 56–57
 smoothing parameter, 50
 Sturges' rule, 48, 55–56, 75
 upper bound, 74–75
 boundary effects, 65–66
 bumps, 86–89
 consistency, 49–55
 cross-validation, which *see*
 curse of dimensionality, 82–84, 152
 density estimator, 2, 49
 equalization in images, 263
 equivalent sample sizes, 57–58
 frequency form, 48
 frequency polygon construction, 95–96
 glyph, which *see*
 graphical interpretation, 122–123
 higher-order bias at midpoint, 53
 indifferent histogram, 122, 131, 139, 190
- $L_1, L_2, L_4, L_6, L_8, L_\infty$, 90–91
- MISE, 52–55
 adaptive mesh, 66–67
 asymptotically optimal, 54, 62–64
 decomposition, 214
 effect of boundary, 65–66
 exact, 60–64
 multiple minima, 63–64
 multivariate, 82, 214
 sensitivity, 58–59, 63
 modes, 86–89
 false sample modes, 89–90
- mortality, 17
- MSE, 49–51
- multivariate, 18–19, 80–86, 213–214
 informative components, *see* Projection
 MISE decomposition, 214
 rate of convergence, 83
- origin of term, 43
- rate of convergence, 51
- rootogram, 52
- scatterplot smoother, 16, 264
- trivariate, 19
- variance, 49–50
 variance-stabilized, 51–52
- Holes in data, *see* Data sets
- Hypergeometric distribution, 43
- Image enhancement, 263
- Inference, 264
- Influence:
 of data point, 44, 122, 129, 144, 155–157,
 232
 function, 232
- Informative components, *see* Projection
- Integrated squared bias, *see* Criteria
- Integrated squared error, *see* Criteria
- Integrated variance, *see* Criteria
- Interpolation, 25
 linear, frequency polygon, 79, 95–96, 113
 linear blend, 106–109, 235, 267–271
 piecewise triangular, 106–107
 spline, which *see*
- Interquartile range, 55, 75, 93, 100, 175, 177
- Inverse problem, 265
- Iris flower data, *see* Data sets
- Jackknife, 138, 252
- Jensen's inequality, 67, 69, 104, 184, 193
- Jittering, *see* Blurring
- Kaplan-Meier product-limit estimate, 261, 266

- Kernel estimators, 122–123, 125–193
 adaptive estimators:
 Abramson's, 182–183, 186–187
 backtransformation, 183, 205
 Breiman et al., 182–183, 187
 exact MSE, 137, 186
 vs. frequency polygon, 137
 global, 189–190
 vs. higher-order, 137, 184, 186–187
 multivariate, 181, 187–190, 212
 nearest-neighbor, 156, 182, 185, 189–190
 vs. nonadaptive, 184–185, 188–189
 nonuniqueness, 137
 null space, 184–185, 192
 optimal smoothing parameters, 137, 183, 185, 188
 pointwise, 181, 183, 188, 203, 257
 sample-point, 185
 theoretical improvement, 183–184, 189–190
 two versions, 181–182
 variable kernel, 181
 as arithmetic mean, 130
 backtransformation, 183, 205
 bias, 126, 130, 133, 150
 boundary effects, 146–149
 artifacts, 148–149
 crucial kernel condition, 146
 floating boundary, 146–149
 irregular regions, 155
 negative kernels, 147–148
 reflection technique, 149
 zero boundary, 147–149
 bumps, *which see*
 choice of kernel, 133–149. *See also Kernel functions*
 higher-order, 133–138
 relative importance, 133
 clipping, 136, 187
 construction, 122–123
 convolution smoothing, 127, 129
 cross-validation, *which see*
 data transformation, 153, 204–206
 definition:
 multivariate, general kernel, 153–154
 multivariate, product kernel, 150
 univariate, 125
 derivative, *see Density derivative estimators*
 equivalent kernels, 141–144, 191–192
 canonical, 142
 rescaling, 142, 144
 table, 142
 equivalent sample sizes, 198–205
 filtering, *see Filters*
 finite difference, 126–127, 143–145, 190, 192
 as Gâteaux derivative, 156–157
 general kernel, 155–160
 asymptotic equivalent kernel, 156
 bias, 158
 conditions, 159
 consistency, 159
 of parametric estimator, 159–160
 variance, 159
 generalized, 44, 156
 graphical interpretation, 122–123
 higher-order, 133–138, 140–146
 asymptotic benefit, 134, 184
 bias, 133, 143
 clipping, 136
 derivatives, 138, 173
 equivalent smoothing, 141–144, 192
 exact MISE, 135
 exact MSE, 137
 family of order-4 kernels, 145
 finite differences, 127, 143–145, 191–192
 4-, 6-, 8-point kernels, 143–144
 kernels, 134
 negative kernels, 129, 136
 optimal kernels, lack of, 140–141
 optimal MISE, 133, 138, 140
 rates of convergence, 133–134
 ratio estimator, 138
 by removing bias, 145–146
 rescaling, 144
 sensitivity, 135–136
 sinc function, 129, 138
 smoothing parameter, 133
 Terrell–Scott estimator, 138, 191
 indifferent histogram, 122, 139
 kernel choice, *see Kernel functions*
 L_1 criterion asymptotics, 200–202
 least-smooth density, 209
 MISE,
 adaptive, 183–184, 189–190
 asymptotically optimal, 131, 164
 boundary, 147–149
 data-based estimate, 164
 decomposition at optimum, 131
 effect of boundary, 146–149, 155
 higher-order, 133–134
 improved approximation, 173, 191
 multivariate, 150–152
 multivariate, general, 154–155
 sensitivity, 127, 135–136

- Kernel estimators (Continued)**
- modes, *which see*
 - moments, 191
 - multivariate, 149–155, 197–206
 - bias, 150, 153–154
 - exact MISE, 199
 - L_1 criterion, 200–202
 - MISE, 150, 152, 155
 - Normal reference rule, 152
 - product kernel, *see* Kernel estimators
 - Scott's rule, 152
 - variance, 150, 154
 - as Normal mixture, 258
 - orthogonal series estimator, *which see*
 - product kernel, 123, 149–153, 181, 192
 - bootstrap, 258
 - regression, 220, 224, 236
 - 10-D example, 203
 - rate of convergence,
 - higher-order, 133–134
 - smoothest density, 209
 - smoothing parameter:
 - adaptive, 137, 156, 181–182
 - adaptive smoothing function, 182–183
 - adaptive smoothing vector, 182–183
 - asymptotically optimal, 131, 164
 - auxiliary bandwidth, 145–146, 162, 174–176
 - bivariate Normal, 151–152
 - boundary, 147–149
 - equivalent smoothing, 141–143
 - exact MSE, 137
 - higher-order, 133
 - kurtosis modification, 152
 - L_1 criterion, 201
 - L_1 vs. L_2 , 201
 - multivariate, adaptive, 188–189
 - multivariate, product, 151–152
 - multivariate, general, 153–154
 - Normal reference rule, 131, 144, 160, 164, 166, 174–175, 181
 - oversmoothed, 165–166
 - relative importance, 133
 - sensitivity, 58–59, 135–137
 - skewness modification, 152
 - upper bound,
 - variance, 126, 130, 150, 190–191
 - variance-stabilized, 190, 259–260
 - Kernel functions, *see also* Filters
 - asymptotic relative efficiency, 139–140
 - Beta family, 140, 168, 192, 266
 - biweight, 117–118, 140, 142, 146–147, 166, 192, 231, 235–236, 250, 258–260, 266
 - boundary, 146–149
 - crucial kernel condition, 146
 - designer kernels, 148–149, 192
 - equivalent smoothing failure, 146
 - floating boundary, 146–149
 - irregular regions, 155
 - zero boundary, 147–149
 - boxcar, 128–129, 144, 186, 191
 - canonical, 142
 - Cauchy, 138
 - conditions, 130–131, 159
 - cosine arch, 140
 - definition, 122, 130
 - in derivative estimators, 132
 - designer kernels, *which see*
 - double Epanechnikov, 92, 140
 - double exponential, 140
 - Epanechnikov kernel, 134, 139–141, 192
 - equivalent kernel smoothing, *which see*
 - filters, *which see*
 - finite difference kernels, 126–127, 143–145, 190
 - Gasser-Müller kernels, 141
 - higher-order, 133–138
 - examples, 134
 - family of order-4 kernels, 145
 - finite difference, 143–145, 190, 192
 - lack of optimality, 140–141
 - negative lobes, 129, 136
 - optimal, lack of, 140–141
 - polynomial kernels, 134
 - regression, 221
 - sensitivity to smoothing parameter, 58–59
 - sinc function, 138
 - indifferent frequency polygon, 139–140, 191
 - indifferent histogram, 139–140
 - isosceles triangle, 116, 118–119, 122–123, 131, 139–140, 190
 - multivariate:
 - general kernels, 153–154
 - linear transformation, 153, 192
 - marginal kernels, 153
 - moment conditions, 153
 - product kernels, 149–151
 - spherical support, 153, 155–156, 189
 - Normal, 131, 140
 - number of modes, 256
 - optimal Epanechnikov kernel, 139
 - order- p , 133
 - orthogonal series equivalent kernels, 129, 190
 - product kernels, 149–151

- Kernel functions (Continued)**
- quartic, 117
 - rectangular or uniform, 116, 122, 126–127, 140, 156, 189, 198, 210
 - in regression, 244
 - shifted exponential, 140
 - sinc function, 129, 138
 - skewed, 140
 - spherical support, 153, 155–156, 189
 - symmetric, 127
 - table of kernel conversion factors, 142
 - triweight, 140, 142, 170, 176, 203
- Kinematic display, see Animation**
- Kronecker delta function, 128**
- L_1, L_2, L_∞ , see Criteria**
- L_1 regression, see Regression**
- Lagrangian, 165**
- LANDSAT, see Data sets**
- Least-squares cross-validation, see Cross-validation, unbiased**
- Likelihood ratio, see Classification**
- Lincoln penny data, see Data sets, penny**
- Linear blend, 106–109, 113, 120, 129, 235, 269–271**
- Linear programming, 235–236**
- Linear smoothers, see Regression**
- Lipid, see Data sets**
- Lipschitz continuity, 50–52**
- Lognormal distribution, 68**
- histogram bin width, 63–64, 72, 75
 - non-Normal reference, 56, 92–93, 100
 - very rough, 53, 65–66
- LOWESS, see Regression**
- LRL data, see Data sets, LRL**
- Macspin, 9**
- Marathon times, see Data sets**
- Marching cubes, 26–27, 271**
- Martingales, 167**
- Maximum penalized-likelihood estimators, 155–156**
- Mean integrated squared error, see Criteria**
- Mean squared error, 37**
- consistency, definition, 51
 - decomposition, 37
- Mean value theorem, 50**
- generalized, 53
- Minkowski's inequality, 73**
- MinneView, 271**
- Mixture density, 86, 94, 122, 178, 202, 204, 210–211, 257–258**
- Modal regression, see Regression**
- Modal trace, 233–235**
- Modes, see also Bump hunting**
- asymptotic Normality, 253
 - confidence intervals, 105–106, 253–255
 - definition, 86
 - false sample modes, 89–90, 105–106, 109–111, 255
 - nonnested contours, 261
- Mortality tables, see Data sets, life table**
- Mount St. Helens, see Data sets, earthquake**
- Multivariate estimators, see particular estimator**
- Nearest-neighbor estimators:**
- as adaptive estimator, 156, 182, 185, 189–190, 258
 - in classification, 251
 - definition, 156
 - efficiency, 190
 - overadapting, 190
 - percentile mesh, see Histograms, adaptive meshes
 - poor bias property, 185, 190
- Negative exponential, 92, 111, 147–148**
- Nonparametric, see also Density estimators**
- conditions on equivalent kernel, 159
 - definition, 44
 - as general kernel estimator, 156–159
 - localness, 157
 - when nonparametric, 157–159
- Normal mixture model, see Bump hunting**
- Normal vectors, 261–262**
- Not-a-number, 238**
- Null space, see Adaptive estimators**
- Old Faithful geyser, see Data sets, geyser**
- Order statistics, 34, 39, 164, 193**
- Orthogonal series estimator:**
- cross-validation, which see
 - of empirical pdf, 44, 128, 157
 - equivalence to kernel estimators, 129, 155
 - equivalent kernels, 129, 190
 - Fourier series basis, 128, 157, 193
- Orthonormal functions, 128, 162, 206**
- Outer normal vectors, 261–262**
- Oversmoothing:**
- constrained, see Cross-validation
 - of derivative estimates, 255
 - frequency polygon bin width, 102–103, 119
 - general use, 161
 - histogram bin width, 47, 55, 72–79, 93
 - interquartile range, 93, 100, 177

- kernel smoothing parameter, 136, 139, 160–161, 165–166
- multivariate, 180–181
- number of bins, 73, 102
- range constraint, 72–73, 102
- as starting value for search, 161
- variance constraint, 74–75, 102–103
- Parallel coordinates, 13–16, 31–32
 - clusters, *see* Classification, clusters
 - grand tour, 15
 - use of ASH with, 123
- Parametric:
 - estimation, 33
 - equivalent kernel, 159–160, 191–192
 - MISE, 39–41
 - problem, 33
 - modeling, 1–2
 - sample size advantage, 57–58
 - specification problem, 33, 57
 - Parametric kernel estimator, 160
- Particle physics data, *see* Data sets, LRL, PRIM4, PRIM7
- Pearson distribution family, 43–45
- Perspective:
 - vs. contour, 250
 - plots, 21
 - projection on computer, 271–272
- Poisson distribution, 51, 87
- Polar coordinates in \Re^d , 27, 32
- Posterior odds, 248, 265
- PRIM4, *see* Data sets
- PRIM7, *see* Data sets
- PRIM-9, 12, 204
- Principal components, 206–209
 - data transformation, which *see* LANDSAT data, 7, 216
 - in regression, 241–243
- Principal curves, 264
- Prior odds, 247–248, 265
- Probability cdf, *see* Cumulative distribution function
- Probability density function, 34
 - empirical, 35–36, 160
 - bootstrap, 170
 - filtered, 127
 - Fourier coefficients of, 128
 - limit in cross-validation, 163
 - limit of histogram, 50
 - penalized-likelihood solution, 155
 - smoothing of, 127–128
 - too noisy, 126–127
 - Johnson family, 44
 - Pearson family, 43–44
- smoothest, *see* Oversmoothing
- Product kernel, *see* Kernel estimators
- Projection:
 - informative components, 212–215, 217, 242
 - model-based nonlinear, 215–216
 - pursuit, 208–212
 - Fisher index, 210–211
 - Friedman-Tukey index, 209–210, 216
 - Hellinger index, 210
 - information indices, 210–211
 - invariance, 209
 - L_1 index, 210
 - moment index, 210
 - regression, 239
 - Shannon entropy, 210–211
 - subspace, 5, 29, 195–196, 203, 206, 208, 214, 242
 - of 3-D objects on computer, 271–272
- Rank deficiency, 198, 204–206, 241
- Rate of convergence:
 - cross-validation, 76
 - frequency polygon, 98
 - higher-order kernel, 133–138
 - histogram MSE, 51
 - kernel, 133–134
 - multivariate, 83, 107–108
 - parametric estimator, 39
- Regression:
 - adaptive, 226, 229, 231
 - additive models, 3, 219, 239–242, 244–245, 264
 - ASH, 236–238
 - bootstrap, 260
 - CART, 244
 - in classification, 252–253
 - conditional:
 - mean, 220, 233
 - mode, 233–235, 254–255
 - confidence intervals, 260
 - cross-validation, 225
 - curse of dimensionality, 239–244
 - definition, 220
 - design:
 - fixed, 219, 221
 - random, 219–220
 - equivalent kernel weights, 229–231
 - higher-order kernel, 221
 - integral weights, 225
 - kernel, 219–225
 - L_1 , 235–237, 244
 - linear, 195
 - linear smoothers, 219–220, 226–231

- Regression (Continued)**
- local polynomial fits, 221–222, 226–227, 230, 235
 - LOWESS, 222, 227, 229, 232–234
 - modal, 233–235, 254–255
 - Nadaraya-Watson, 219–225, 231, 236
 - from kernel density estimate, 220, 224, 236
 - MSE, 222–224, 244
 - nonparametric, 31, 219–245
 - partial spline, 198
 - projection pursuit, 239
 - regressogram, 20
 - resistant estimators, 232–233, 235, 244
 - ridge, 37
 - scatterplot smoother, 19, 264
 - semiparametric, 197
 - sliced inverse, 242–243
 - spline, 3
 - cubic, 228
 - equivalent kernel, 229
 - interpolating, 227
 - smoothing, 227–229, 231
 - supersmooth, 226, 239
 - 3R smoother, 20
 - visualization, 24–25
 - WARPing, 236–239
- Remote sensing data, *see* Data sets, LAND-SAT**
- Ridge, density, 264–265**
- Riemannian integral approximation, 52, 54, 81, 92–94, 97, 213**
- Robust:**
- bin width rule, 55
 - estimation, 33
- Rootgram, *see* Histograms**
- Roughness:**
- bivariate Normal, 151–152
 - cross-validation, 167, 172
 - definition, 53
 - estimation of, 75–76, 93, 164, 167, 173–174
 - histogram, 54, 93
 - kernel estimate, 97, 164
 - lognormal density, 56, 63, 100
 - mathematical, 53
 - Normal density, 55, 131, 151–152, 175
 - penalty, 155
 - statistical, 53
 - t-distribution, 57
- S language, 6, 9, 217, 222**
- S-plus, 9–10**
- Scatter diagrams:**
- blurring, 7, 19, 179
 - boundary emphasis, 8, 29
 - brushing, 8–10, 16, 196, 261
 - clusters, *see* Classification, clusters
 - data cloud, 8, 24, 71, 180, 206, 209, 215, 225, 261
 - dot plot, 36
 - glyphs, 11–12, 15
 - grand tour, 11, 15
 - jittering, 7, 19, 179
 - linked, 8–9, 16, 31
 - multivariate, 6–12
 - overplotting, 7–9
 - pairwise, 6–10, 12, 16, 30–31, 212, 215–216, 241–242, 261–262
 - rotation, 9–11, 15–16, 23, 261
 - scatterplot matrix, 8–9, 16
 - scatterplot smoothing, 16, 18–19, 233, 239, 264
 - sensitivity, *see* Smoothing parameters
 - stem-and-leaf plot, 47
 - stick pin plot, 20–21
 - tail emphasis, 8
 - thinning, 16
 - too many points, 7
 - too much ink, 16, 60, 123
 - trivariate, 9, 23
- Shannon entropy, 210–211**
- Shell, *see* Contour**
- Shifted-Normal model, 248–250**
- Shrinkage estimator, 37**
- Silica, *see* Data sets**
- Simulation:**
- large sample, 59–64, 99
 - Monte Carlo, 41, 258
 - small sample, 59–64
- SIR, *see* Regression, sliced inverse**
- Slicing:**
- contours, 23
 - density function, 3, 22, 24, 203–204
- Smoothing parameters:**
- bin width, *see* Histograms; Frequency polygons
 - kernel estimators, *which see* oversmoothed, *see* Oversmoothing
 - selection, *see* Cross-validation
 - sensitivity to choice, 58–59
 - variance-bias trade-off, 50
 - wrongness, 2
- Snowfall, *see* Data sets**
- Spectral representation, 206, 214, 265**
- Spherical coordinates, 271**
- Spherling transformation, 207, 209, 213**

- Splines, 14, 26, 177, 227, 265. *See also* Regression
- Star diagram, *see* Glyphs
- Statistical data analysis goal, 5
- Steel surface data, *see* Data sets
- Stem-and-leaf plot, *see* Glyphs
- Stereography:
- contour surfaces, 22, 26
 - scatter diagrams, 9, 11
 - stereograms, 4–5
- Stick pin plot, *see* Glyphs
- Structure of data, 195–196
- Sufficient statistic, 34, 36
- Surfaces, *see* Contour
- Survival analysis, 260–261
- Table of kernel scale conversion factors, 142
- Tail probabilities, 29–30
- Taylor series, 40, 93, 96, 111, 126, 138, 158, 187, 192, 221, 223, 254
- Terrell–Scott higher-order estimator, 138, 191
- Test graph method, 161, 164–165
- Time series, 12, 18, 126, 171, 265
- Total variation, 54, 92
- Transformation:
- data, see* Data transformation
 - variance-stabilizing, which see*
- Triangular array, 253
- Triglyceride, *see* Data sets, lipid
- U*-statistics, 167, 172
- Unbiasedness, 34
- of nonparametric estimator, 36*
- Uniform density:
- maximum likelihood, 39*
 - optimal mise, 39*
- U.S. economy data, *see* Data sets
- Universities, *see* Data sets
- Unparametric method, 1
- Variable kernel, *see* Kernel estimators, adaptive
- Variance:
- component, 37*
 - histogram, 49, 80*
 - variance-bias trade-off, 50*
- Variance-stabilizing transformation, 51–52, 190, 259
- Visualization:
- data, 2*
 - empirical distribution, 35–36*
 - functions, 3, 16, 20–26*
 - linear blend, 106–108, 235, 269–271*
 - marching cubes, 26–27, 271*
 - scientific, 3*
- WARPing, 123, 236–239
- Weight function, *see* Kernel function

Multivariate Density Estimation

DAVID W. SCOTT

Copyright © 1992 by John Wiley & Sons, Inc.

**WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS**

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors

*Vic Barnett, Ralph A. Bradley, Nicholas I. Fisher, J. Stuart Hunter,
J. B. Kadane, David G. Kendall, Adrian F. M. Smith,
Stephen M. Stigler, Jozef L. Teugels, Geoffrey S. Watson**Probability and Mathematical Statistics*ANDERSON • An Introduction to Multivariate Statistical Analysis,
*Second Edition*BARNETT • Comparative Statistical Inference, *Second Edition*

BERNARDO and SMITH • Bayesian Statistical Concepts and Theory

BHATTACHARYYA and JOHNSON • Statistical Concepts and Methods

BILLINGSLEY • Probability and Measure, *Second Edition*

BOROVKOV • Asymptotic Methods in Queueing Theory

BRANDT, FRANKEN, and LISEK • Stationary Stochastic Models

CAINES • Linear Stochastic Systems

CHEN • Recursive Estimation and Control for Stochastic Systems

CONSTANTINE • Combinatorial Theory and Statistical Design

COVER and THOMAS • Elements of Information Theory

*DOOB • Stochastic Processes

DUDEWICZ and MISHRA • Modern Mathematical Statistics

ETHIER and KURTZ • Markov Processes: Characterization and Convergence

FELLER • An Introduction to Probability Theory and Its Applications, Volume I,
Third Edition, Revised; Volume II, Second Edition

FULLER • Introduction to Statistical Time Series

FULLER • Measurement Error Models

GIFI • Nonlinear Multivariate Analysis

GUTTORP • Statistical Inference for Branching Processes

HALD • A History of Probability and Statistics and Their Applications before 1750

HALL • Introduction to the Theory of Coverage Processes

HANNAN and DEISTLER • The Statistical Theory of Linear Systems

HEDAYAT and SINHA • Design and Inference in Finite Population Sampling

HOEL • Introduction to Mathematical Statistics, *Fifth Edition*

HUBER • Robust Statistics

IMAN and CONOVER • A Modern Approach to Statistics

KAUFMAN and ROUSSEEUW • Finding Groups in Data: An Introduction to Cluster
AnalysisLARSON • Introduction to Probability Theory and Statistical Inference,
Third Edition

LESSLER and KALSBEEK • Nonsampling Error in Surveys

MORGENTHALER and TUKEY • Configural Polysampling: A Route to Practical
Robustness

MUIRHEAD • Aspects of Multivariate Statistical Theory

OLIVER and SMITH • Influence Diagrams, Belief Nets and Decision Analysis

*PARZEN • Modern Probability Theory and Its Applications

PILZ • Bayesian Estimation and Experimental Design in Linear Regression Models

PRESS • Bayesian Statistics: Principles, Models, and Applications

PURI and SEN • Nonparametric Methods in General Linear Models

PURI, VILAPLANA, and WERTZ • New Perspectives in Theoretical and Applied
Statistics

RAO • Asymptotic Theory of Statistical Inference

RAO • Linear Statistical Inference and Its Applications, *Second Edition*

ROBERTSON, WRIGHT, and DYKSTRA • Order Restricted Statistical Inference

ROGERS and WILLIAMS • Diffusions, Markov Processes, and Martingales, Volume
II: Ito Calculus

ROHATGI • A Introduction to Probability Theory and Mathematical Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Probability and Mathematical Statistics (Continued)

- ROSS • Stochastic Processes
 RUBINSTEIN • Simulation and the Monte Carlo Method
 RUZSA and SZEKELY • Algebraic Probability Theory
 SCHEFFE • The Analysis of Variance
 SEBER • Linear Regression Analysis
 SEBER • Multivariate Observations
 SEBER and WILD • Nonlinear Regression
 SERFLING • Approximation Theorems of Mathematical Statistics
 SHORACK and WELLNER • Empirical Processes with Applications to Statistics
 STAUDTE and SHEATHER • Robust Estimation and Testing
 STOYANOV • Counterexamples in Probability
 STYAN • The Collected Papers of T. W. Anderson: 1943–1985
 WHITTAKER • Graphical Models in Applied Multivariate Statistics
 YANG • The Construction Theory of Denumerable Markov Processes

Applied Probability and Statistics

- ABRAHAM and LEDOLTER • Statistical Methods for Forecasting
 AGRESTI • Analysis of Ordinal Categorical Data
 AGRESTI • Categorical Data Analysis
 ANDERSON and LOYNES • The Teaching of Practical Statistics
 ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG • Statistical Methods for Comparative Studies
 ASMUSSEN • Applied Probability and Queues
 *BAILEY • The Elements of Stochastic Processes with Applications to the Natural Sciences
 BARNETT • Interpreting Multivariate Data
 BARNETT and LEWIS • Outliers in Statistical Data, *Second Edition*
 BARTHOLOMEW, FORBES, and MCLEAN • Statistical Techniques for Manpower Planning, *Second Edition*
 BATES and WATTS • Nonlinear Regression Analysis and Its Applications
 BELSLEY • Conditioning Diagnostics: Collinearity and Weak Data in Regression
 BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
 BHAT • Elements of Applied Stochastic Processes, *Second Edition*
 BHATTACHARYA and WAYMIRE • Stochastic Processes with Applications
 BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN • Measurement Errors in Surveys
 BLOOMFIELD • Fourier Analysis of Time Series: An Introduction
 BOLLEN • Structural Equations with Latent Variables
 BOX • R. A. Fisher, the Life of a Scientist
 BOX and DRAPER • Empirical Model-Building and Response Surfaces
 BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
 BOX, HUNTER, and HUNTER • Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
 BROWN and HOLLANDER • Statistics: A Biomedical Introduction
 BUCKLEW • Large Deviation Techniques in Decision, Simulation, and Estimation
 BUNKE and BUNKE • Nonlinear Regression, Functional Relations and Robust Methods: Statistical Methods of Model Building
 CHATTERJEE and HADI • Sensitivity Analysis in Linear Regression
 CHATTERJEE and PRICE • Regression Analysis by Example, *Second Edition*
 CLARKE and DISNEY • Probability and Random Processes: A First Course with Applications, *Second Edition*
 COCHRAN • Sampling Techniques, *Third Edition*
 *COCHRAN and COX • Experimental Designs, *Second Edition*
 CONOVER • Practical Nonparametric Statistics, *Second Edition*
 CONOVER and IMAN • Introduction to Modern Business Statistics
 CORNELL • Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Continued on back end papers

Applied Probability and Statistics (Continued)

- COX • A Handbook of Introductory Statistical Methods
- *COX • Planning of Experiments
- CRESSIE • Statistics for Spatial Data
- DANIEL • Applications of Statistics to Industrial Experimentation
- DANIEL • Biostatistics: A Foundation for Analysis in the Health Sciences, *Fifth Edition*
- DAVID • Order Statistics, *Second Edition*
- DEGROOT, FIENBERG, and KADANE • Statistics and the Law
- *DEMING • Sample Design in Business Research
- DILLON and GOLDSTEIN • Multivariate Analysis: Methods and Applications
- DODGE and ROMIG • Sampling Inspection Tables, *Second Edition*
- DOWDY and WEARDEN • Statistics for Research, *Second Edition*
- DRAPER and SMITH • Applied Regression Analysis, *Second Edition*
- DUNN • Basic Statistics: A Primer for the Biomedical Sciences, *Second Edition*
- DUNN and CLARK • Applied Statistics: Analysis of Variance and Regression, *Second Edition*
- ELANDT-JOHNSON and JOHNSON • Survival Models and Data Analysis
- FLEISS • The Design and Analysis of Clinical Experiments
- FLEISS • Statistical Methods for Rates and Proportions, *Second Edition*
- FLEMING and HARRINGTON • Counting Processes and Survival Analysis
- FLURY • Common Principal Components and Related Multivariate Models
- GALLANT • Nonlinear Statistical Models
- GROSS and HARRIS • Fundamentals of Queueing Theory, *Second Edition*
- GROVES • Survey Errors and Survey Costs
- GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, and WAKSBERG • Telephone Survey Methodology
- HAHN and MEEKER • Statistical Intervals: A Guide for Practitioners
- HAND • Discrimination and Classification
- HEIBERGER • Computation for the Analysis of Designed Experiments
- HELLER • MACSYMA for Statisticians
- HOAGLIN, MOSTELLER, and TUKEY • Exploratory Approach to Analysis of Variance
- HOAGLIN, MOSTELLER, and TUKEY • Exploring Data Tables, Trends and Shapes
- HOAGLIN, MOSTELLER, and TUKEY • Understanding Robust and Exploratory Data Analysis
- HOCHBERG and TAMHANE • Multiple Comparison Procedures
- HOEL • Elementary Statistics, *Fifth Edition*
- HOGG and KLUGMAN • Loss Distributions
- HOLLANDER and WOLFE • Nonparametric Statistical Methods
- HOSMER and LEMESHOW • Applied Logistic Regression
- IMAN and CONOVER • Modern Business Statistics
- JACKSON • A User's Guide to Principle Components
- JOHN • Statistical Methods in Engineering and Quality Assurance
- JOHNSON • Multivariate Statistical Simulation
- JOHNSON and KOTZ • Distributions in Statistics
 - Discrete Distributions
 - Continuous Univariate Distributions—1
 - Continuous Univariate Distributions—2
 - Continuous Multivariate Distributions
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE • The Theory and Practice of Econometrics, *Second Edition*
- JUDGE, HILL, GRIFFITHS, LÜTKEPOHL, and LEE • Introduction to the Theory and Practice of Econometrics, *Second Edition*
- KALBFLEISCH and PRENTICE • The Statistical Analysis of Failure Time Data
- KASPRZYK, DUNCAN, KALTON, and SINGH • Panel Surveys
- KISH • Statistical Design for Research
- KISH • Survey Sampling
- LAWLESS • Statistical Models and Methods for Lifetime Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

- LEBART, MORINEAU, and WARWICK • Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices
- LEE • Statistical Methods for Survival Data Analysis, *Second Edition*
- LePAGE and BILLARD • Exploring the Limits of Bootstrap
- LEVY and LEMESHOW • Sampling of Populations: Methods and Applications
- LINHART and ZUCCHINI • Model Selection
- LITTLE and RUBIN • Statistical Analysis with Missing Data
- MAGNUS and NEUDECKER • Matrix Differential Calculus with Applications in Statistics and Econometrics
- MAINDONALD • Statistical Computation
- MALLows • Design, Data, and Analysis by Some Friends of Cuthbert Daniel
- MANN, SCHAFER, and SINGPURWALLA • Methods for Statistical Analysis of Reliability and Life Data
- MASON, GUNST, and HESS • Statistical Design and Analysis of Experiments with Applications to Engineering and Science
- McLACHLAN • Discriminant Analysis and Statistical Pattern Recognition
- MILLER • Survival Analysis
- MONTGOMERY and PECK • Introduction to Linear Regression Analysis, *Second Edition*
- NELSON • Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- NELSON • Applied Life Data Analysis
- OCHI • Applied Probability and Stochastic Processes in Engineering and Physical Sciences
- OSBORNE • Finite Algorithms in Optimization and Data Analysis
- PANKRATZ • Forecasting with Dynamic Regression Models
- PANKRATZ • Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- RACHEV • Probability Metrics and the Stability of Stochastic Models
- RÉNYI • A Diary on Information Theory
- RIPLEY • Spatial Statistics
- RIPLEY • Stochastic Simulation
- ROSS • Introduction to Probability and Statistics for Engineers and Scientists
- ROUSSEEUW and LEROY • Robust Regression and Outlier Detection
- RUBIN • Multiple Imputation for Nonresponse in Surveys
- RYAN • Statistical Methods for Quality Improvement
- SCHUSS • Theory and Applications of Stochastic Differential Equations
- SCOTT • Multivariate Density Estimation: Theory, Practice, and Visualization
- SEARLE • Linear Models
- SEARLE • Linear Models for Unbalanced Data
- SEARLE • Matrix Algebra Useful for Statistics
- SEARLE, CASELLA, and McCULLOCH • Variance Components
- SKINNER, HOLT, and SMITH • Analysis of Complex Surveys
- STOYAN • Comparison Methods for Queues and Other Stochastic Models
- STOYAN, KENDALL, and MECKE • Stochastic Geometry and Its Applications
- THOMPSON • Empirical Model Building
- TIERNEY • LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
- TIJMS • Stochastic Modeling and Analysis: A Computational Approach
- TITTERINGTON, SMITH, and MAKOV • Statistical Analysis of Finite Mixture Distributions
- UPTON and FINGLETON • Spatial Data Analysis by Example, Volume I: Point Pattern and Quantitative Data
- UPTON and FINGLETON • Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
- VAN RIJCKEVORSEL and DE LEEUW • Component and Correspondence Analysis
- WEISBERG • Applied Linear Regression, *Second Edition*
- WHITTLE • Optimization Over Time: Dynamic Programming and Stochastic Control, Volume I and Volume II

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

WHITTLE • Systems in Stochastic Equilibrium
WONNACOTT and WONNACOTT • Econometrics, *Second Edition*
WONNACOTT and WONNACOTT • Introductory Statistics, *Fifth Edition*
WONNACOTT and WONNACOTT • Introductory Statistics for Business and
Economics, *Fourth Edition*
WOOLSON • Statistical Methods for the Analysis of Biomedical Data

Tracts on Probability and Statistics

BILLINGSLEY • Convergence of Probability Measures
TOUTENBURG • Prior Information in Linear Models



9 780471 547709