Recent Developments in Nonparametric Density Estimation

ALAN JULIAN IZENMAN*

Advances in computation and the fast and cheap computational facilities now available to statisticians have had a significant impact upon statistical research, and especially the development of nonparametric data analysis procedures. In particular, theoretical and applied research on nonparametric density estimation has had a noticeable influence on related topics, such as nonparametric regression, nonparametric discrimination, and nonparametric pattern recognition. This article reviews recent developments in nonparametric density estimation and includes topics that have been omitted from review articles and books on the subject. The early density estimation methods, such as the histogram, kernel estimators, and orthogonal series estimators are still very popular, and recent research on them is described. Different types of restricted maximum likelihood density estimators, including order-restricted estimators, maximum penalized likelihood estimators, and sieve estimators, are discussed, where restrictions are imposed upon the class of densities or on the form of the likelihood function. Nonparametric density estimators that are data-adaptive and lead to locally smoothed estimators are also discussed; these include variable partition histograms, estimators based on statistically equivalent blocks, nearest-neighbor estimators, variable kernel estimators, and adaptive kernel estimators. For the multivariate case, extensions of methods of univariate density estimation are usually straightforward but can be computationally expensive. A method of multivariate density estimation that did not spring from a univariate generalization is described, namely, projection pursuit density estimation, in which both dimensionality reduction and density estimation can be pursued at the same time. Finally, some areas of related research are mentioned, such as nonparametric estimation of functionals of a density, robust parametric estimation, semiparametric models, and density estimation for censored and incomplete data, directional and spherical data, and density estimation for dependent sequences of observations.

KEY WORDS: Adaptive estimators; Censored data; Delta sequences; Directional data; Histograms; Kernel estimators; Maximum penalized likelihood; Method of sieves; Multivariate density estimation; Nearest neighbor methods; Order-restricted maximum likelihood methods; Orthogonal series; Projection pursuit density estimation; Statistically equivalent blocks.

1. INTRODUCTION

The field of nonparametrics has broadened its appeal in recent years with an array of new tools for statistical analysis. These new tools offer sophisticated alternatives to traditional parametric models for exploring large amounts of univariate or multivariate data without making specific distributional assumptions. As one of those tools, *nonparametric density estimation* has become a prominent statistical research topic. If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a random d-dimensional sample from a continuous probability density function f, where

$$f(\mathbf{x}) \geq 0, \qquad \int_{\mathbf{R}^d} f(\mathbf{x})\, dx = 1, \tag{1.1}$$

the general problem is to estimate f when no formal parametric structure is specified. In other words, f is taken to belong to a large enough family of densities so that it cannot be represented through a finite number of parameters. "Smoothness" conditions are usually imposed on f and its derivatives, although there are applications (e.g., X-ray transmission tomography) in which discontinuities in f (tissue density) are natural (see Johnstone and Silverman 1990).

Perhaps the earliest nonparametric estimator of a univariate density f was the histogram. Further breakthroughs—initially, with the kernel, orthogonal series, and nearest-neighbor methods—were inspired by application to nonparametric discrimination and developments in spectral density estimation for stationary time series. Later, methods such as penalized likelihood, polynomial spline, variable kernel, sieves, and projection pursuit were introduced with other objectives in mind. What has helped make nonparametric density estimation (and related methods) popular today can be traced to a combination of circumstances: the growing importance of computers in statistical research, the public availability of quality statistical software, and a general awareness of the advantages of high-level graphics.

For example, in comparing data from two independent samples, nonparametric density estimates can be very helpful. In a study by Kasser and Bruce (1969) of coronary heart disease patients and age-matched "normals," a number of variables were recorded on 117 men in each group. These variables included heart rates recorded at rest and at their maximum following exercise. Figure 1 shows kernel density estimates of resting heart rate and maximum heart rate for both groups. Notice that the maximum heart rate density estimate for the patient group appears to be bimodal, while for the normal group, the density estimate is essentially unimodal. The opposite appears to be the case for resting heart rate. Figures 2 and 3 show a contour plot and a perspective plot, respectively, of the bivariate density estimate of resting and maximum heart rates for both groups. The shapes of both bivariate density estimates, especially the direction and extent of bimodality, could be used to classify future males into one of the two diagnostic groups.

Researchers have thus found nonparametric density es-

* Alan Julian Izenman is Associate Professor and Director of the Statistical Computing Laboratory, Department of Statistics, Temple University, Philadelphia, PA 19122. The author thanks Luc Devroye, M. C. Jones, David Scott, Simon Sheather, Bernard Silverman, Mike Steele, Michael Tarter, and Ed Wegman for their detailed comments, suggestions, preprints, reprints, and references, and Partha Bagchi and Albert Cheung for assistance with the graphs. Thanks also go to the coordinating editor and referees, whose comments helped greatly in revising this article.

Journal of the American Statistical Association
March 1991, Vol. 86, No. 413, Review Paper

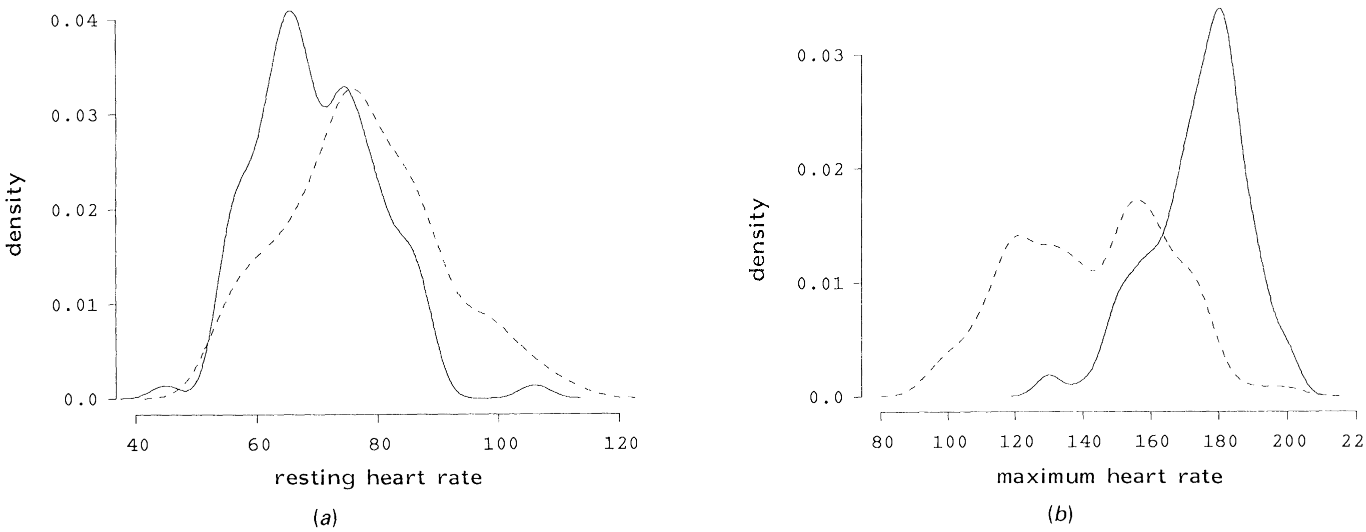

Figure 1. Gaussian Kernel Density Estimates of (a) Resting Heart Rate and (b) Maximum Heart Rate Following Exercise for a Group of 117 Male Heart Patients (Dotted Lines) and for a Group of 117 Age-Matched Male "Normals" (Solid Lines) in a Study of Coronary Heart Disease (Kasser and Bruce 1969). For each density estimate, the window-width was taken to reflect sample variation. Note especially the bimodal density estimate for maximum heart rate for the patient group and the bimodal density estimate for resting heart rate for the normal group. Source of data: Kronmal and Tarter (1973).

timates effective in the following situations: (a) In *exploratory* analysis, descriptive features of the density estimate, such as multimodality, tail behavior, and skewness, are of special interest, and a nonparametric approach may be more

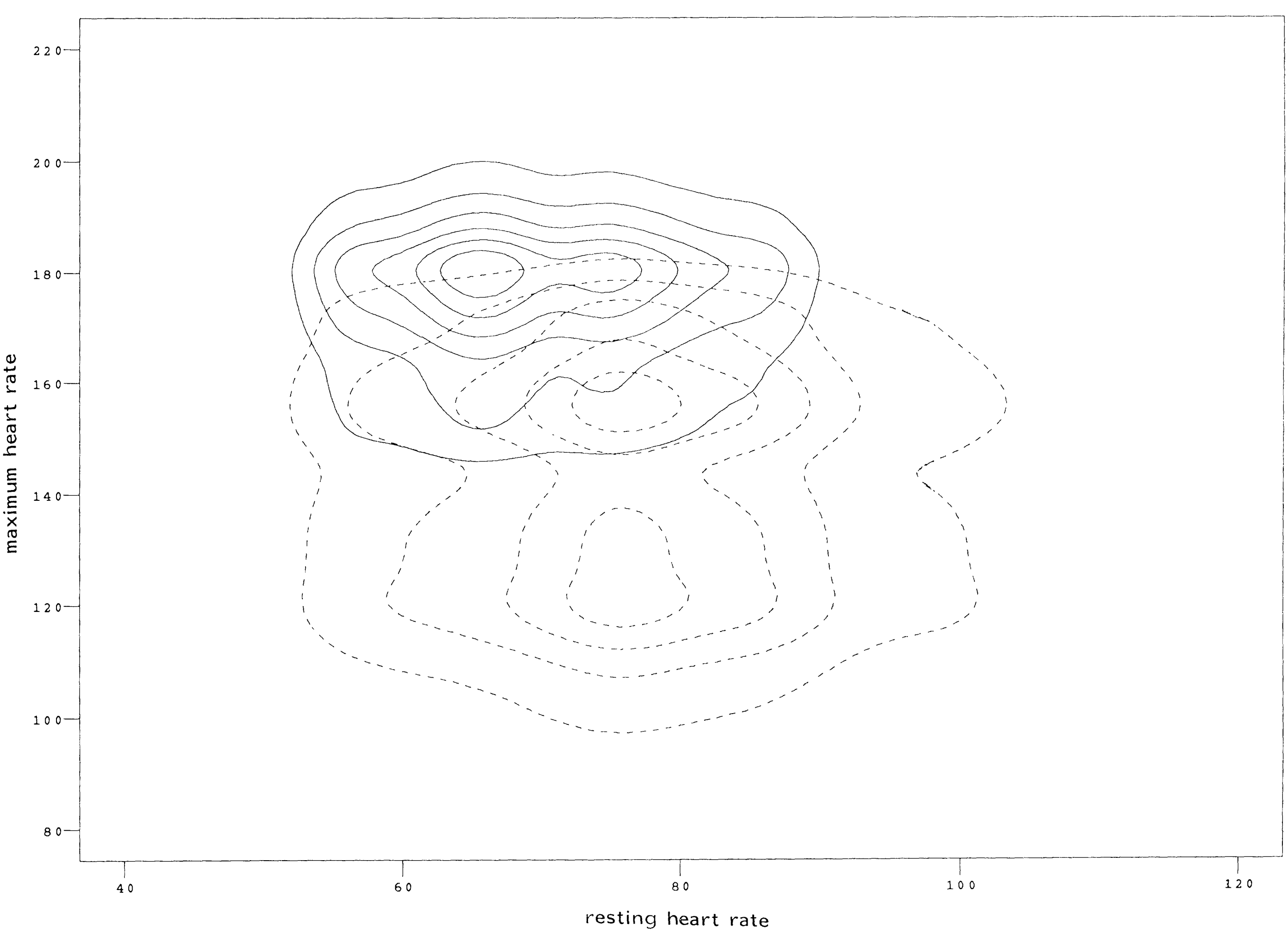

Figure 2. Equal Probability Contours of Bivariate Gaussian Kernel Density Estimates of Resting Heart Rate and Maximum Heart Rate From Figure 1. The normals-group density contours are shown as solid lines and the patient-group density contours are shown as dotted lines. Notice that the bimodal orientations of the density contours of the two groups appear orthogonal to each other.

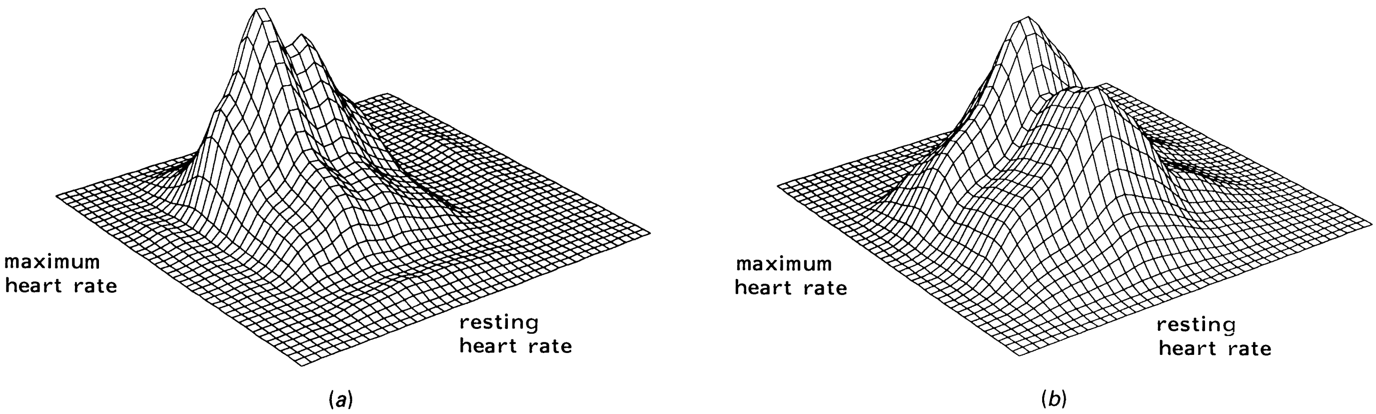

Figure 3. Three-Dimensional Perspective Plots of Bivariate Gaussian Kernel Density Estimates of Resting Heart Rate and Maximum Heart Rate From Figure 1. The normals group is displayed in (a) and the patient group in (b).

flexible than the traditional parametric methods; (b) in *confirmatory* analysis, nonparametric density estimates are used in decision making, such as nonparametric discrimination and classification analysis, testing for modes, and random variate testing; and (c) for *presentational* purposes, statistical peculiarities of the data often can be readily explained to clients through simple graphical displays of estimated density curves (See Silverman 1981a). There is a very revealing example of (a) by Park and Marron (1990) where they display a sequence of annual lognormal density estimates for net income data that indicated unimodal densities hardly changing from year to year, while nonparametric density estimates indicated at least two modes and significant changes in shape over time. Further published applications of nonparametric density estimation can be found listed and briefly described in Table 1.

The last two decades have seen a consolidation and a critical assessment of nonparametric density estimation methods. Several review articles (Bean and Tsokos 1980; Fryer 1977; Leonard 1978; Rosenblatt 1971; Tarter and Kronmal 1976; and Wegman 1972, 1982) and an extensive bibliography (Wertz and Schneider 1979) were published, as well as nine books (Devroye 1987; Devroye and Gyorfi 1985; Hand 1982; Nadarya 1989; Prakasa Rao 1983; Silverman 1986; Tapia and Thompson 1978; Van Es 1990; and Wertz 1978); certain books emphasized density estimation methods preferred by the authors, while others were more comprehensive in their treatment of the diverse material. As with most statistical research, much of what has been written on the subject of nonparametric density estimation, including most of these books, has been completely theoretical; some books (such as Silverman 1986), however, contain discussions of real-data examples, simulation studies, and computational issues. References to *JASA* reviews of some of these books are listed in Table 2. See also the book review by Silverman (1985). The successful development of nonparametric density estimation techniques led, in turn, to the formulation of *nonparametric regression* (Eubank 1988; Muller 1988; Nadarya 1989), including the nonparametric analysis of growth curves, and *nonparametric statistical pattern recognition* (Devijver and Kittler 1982; Fukunaga 1972, chap. 6).

This article surveys recent developments in nonparametric density estimation, as well as topics that were omitted from previous review articles and books. Section 2 discusses desirable statistical properties of nonparametric den-

Table 1. Case Studies Involving Nonparametric Density Estimation

Reference	*Topic*	*Method*	*Remarks*
Silverman (1978c)	Identifying the causes of "cot death"	MPL	Univariate data; assessing bimodality
Scott, Gotto, Cole, and Gorry (1978)	Coronary heart disease	Kernel	Bivariate data; classification problem
Good and Gaskins (1980)	High-energy physics and "bump-hunting"	MPL	Univariate grouped data; assessing a bump in a mass spectrum histogram
Dubuisson and Lavison (1982)	Surveillance of a nuclear reactor	Kernel	Multivariate data; classification problem
Scott and Thompson (1983)	Remote sensing of satellite agricultural crop data	ASH	Trivariate data; exploratory analysis
Aitchison and Lauder (1985)	Compositional data for geology and consumer demand analysis	Kernel	Multivariate data vectors of proportions summing to unity
De Jager, Swanepoel, and Raubenheimer (1986)	Gamma-ray astronomy for estimating light curves and identifying periodic sources	Kernel	Univariate data; assessing whether light curve differs from uniform density
Izenman and Sommer (1988)	Identifying the components of a philatelic mixture	Kernel	Univariate data; assessing multimodality; comparison with parametric mixture

Table 2. Citations of Reviews in JASA of Books on Nonparametric Density Estimation

Author	Source of review	Reviewer	General comments
Wertz (1978)	*JASA*, 75 (1980), 241	K.-S. Lii	Emphasizes kernel methods; theoretical
Tapia and Thompson (1978)	no *JASA* review	—	Emphasizes MPL method; theoretical; Monte Carlo simulations
Hand (1982)	*JASA*, 78 (1983), 990–991	J. D. Knoke	Kernel methods only; some applications; univariate and multivariate approaches
Prakasa Rao (1983)	*JASA*, 81 (1986), 264	V. Surarla	Comprehensive; theoretical; applications to different topics
Devroye and Gyorfi (1985)	*JASA*, 82 (1987), 344	J. R. Thompson	Comprehensive; theoretical; L_1 viewpoint
Silverman (1986)	*JASA*, 83 (1988), 269–270	A. J. Izenman	Comprehensive; numerous real-data applications; univariate and multivariate approaches; computational details
Devroye (1987)	no *JASA* review	—	Emphasizes kernel methods; theoretical; L_1 viewpoint
Nadarya (1989)	*JASA*, 85 (1990), 598	D. W. Scott	Emphasizes kernel methods; theoretical

sity estimates, followed in Sections 3–9 by reviews of the various estimation methods. Finally, in Section 10, some remarks are made about related research areas. Note that the references, though numerous, should not be regarded as exhaustive.

2. STATISTICAL PROPERTIES OF DENSITY ESTIMATORS

Like any statistical procedure, nonparametric density estimators are recommended only if they possess desirable properties. Finite-sample properties of nonparametric density estimators are available for special situations (Deheuvels 1977; Fryer 1976), but, in general, research emphasis has settled on developing large-sample properties.

2.1 Unbiasedness

Consider, for example, unbiasedness. An estimator $\hat{f}$ of a probability density function f is *unbiased* for f if, for all $x \in \mathbf{R}^d$, $E_f[\hat{f}(x)] = f(x)$. Although unbiased estimators of parametric densities, such as the normal, Poisson, exponential, and geometric, do exist (Ghurye and Olkin 1969), no bona fide density estimator [that is, satisfying (1.1)] can exist that is unbiased for all continuous densities (Rosenblatt 1956). Hence attention has since focused on sequences $\{\hat{f}_n\}$ of nonparametric density estimators that are *asymptotically unbiased* for f; that is, for all $x \in \mathbf{R}^d$, $E_f[\hat{f}_n(x)] \to f(x)$ as $n \to \infty$.

2.2 Consistency

A more important property is consistency. The simplest notion of *consistency* of a density estimator is where $\hat{f}$ is (*weakly*) *pointwise consistent* for a univariate f if $\hat{f}(x) \to f(x)$ in probability for every $x \in \mathbf{R}$, and is *strongly pointwise consistent for* f if convergence holds almost surely. Other types of consistency depend upon the error criterion (L_1 or L_2, in general); see Hall (1989b).

The L_2 Approach. If f is assumed square integrable, then the performance of $\hat{f}$ at $x \in \mathbf{R}$ is measured by the mean squared error,

$$\begin{aligned}\mathrm{MSE}(x) &= E_f[\hat{f}(x) - f(x)]^2 \qquad (2.1)\\ &= \mathrm{var}[\hat{f}(x)] + \{\mathrm{bias}[\hat{f}(x)]\}^2,\end{aligned}$$

where $\mathrm{var}[\hat{f}(x)] = E_f\{\hat{f}(x) - E_f[\hat{f}(x)]\}^2$ and $\mathrm{bias}[\hat{f}(x)] = E_f[\hat{f}(x)] - f(x)$. If $\mathrm{MSE}(x) \to 0$ for all $x \in \mathbf{R}$ as $n \to \infty$, then $\hat{f}$ is said to be a *pointwise consistent estimator of f in quadratic mean*. A more important performance criterion relates to how well the entire curve $\hat{f}$ estimates f. One such measure of goodness of fit is found by integrating (2.1) over all values of x, yielding the integrated mean squared error,

$$\mathrm{IMSE} = \int_{-\infty}^{\infty} E_f[\hat{f}(x) - f(x)]^2\, dx. \qquad (2.2)$$

Another measure commonly used is integrated squared error (or L_2 norm),

$$\mathrm{ISE} = \int_{-\infty}^{\infty} [\hat{f}(x) - f(x)]^2\, dx. \qquad (2.3)$$

Taking expectations over f in (2.3) gives the mean integrated squared error, MISE = E_f(ISE). Note that MISE = IMSE. ISE is often preferred as a criterion, rather than its expected value MISE, since ISE determines how closely $\hat{f}$ approximates f for a given data set, whereas MISE is concerned with the average over all possible data sets. Under mild conditions, ISE has been shown to be a reasonably random approximation to MISE (Marron and Hardle 1986), while, in certain situations, MISE may actually be a better performance criterion than ISE (Hall and Marron 1988). Farrell (1972) showed that for bona fide density estimates, the best possible asymptotic rate of convergence for MISE is $O(n^{-4/5})$, and Boyd and Steele (1978) proved that no $\hat{f}$ can exist with a MISE better than $O(n^{-1})$, even if f is a normal density.

The L_1 Approach. One problem with the L_2 approach to nonparametric density estimation is that the tail behavior of a density becomes less important, possibly resulting in peculiarities in the tails of the density estimate. Further objections to the L_2 approach can be found in Donoho and Johnstone (1989). In two books (Devroye 1987; Devroye and Gyorfi 1985), and in a host of articles, an alternative L_1 theory of nonparametric density estimation was vigorously pursued by Devroye and his colleagues. Specifically, Devroye and Gyorfi (1985, p. 1) claimed that L_1 is "the natural space for densities," and showed that the integrated absolute error (also known as the total variation or the L_1 norm),

$$\text{IAE} = \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)|\, dx, \tag{2.4}$$

is always well defined as a norm on that space, is invariant under monotone transformations, and $0 \leq \text{IAE} \leq 2$. If IAE $\to 0$ in probability as $n \to \infty$, then $\hat{f}$ is said to be a *consistent estimator of* f; strong consistency of $\hat{f}$ occurs when convergence holds almost surely. The distance IAE is related to Kullback–Leibler relative entropy and Hellinger distance; see Devroye and Gyorfi (1985, chap. 8) for details. The expectation of (2.4) over all densities f yields the mean integrated absolute error, MIAE $= E_f[\text{IAE}]$. Some quite remarkable results were proved by Devroye and his colleagues concerning the asymptotic behavior of IAE and MIAE under little or no assumptions on f. Hall and Wand (1988) derived a general asymptotic expression for MIAE and showed that its minimization reduced to numerically solving a particular equation. One thing, however, is clear: The technical labor needed to get L_1 results is substantially more difficult than that needed to obtain analogous L_2 results.

2.3 Bona Fide Density Estimates

Of the density estimation methods currently available, some always yield bona fide density estimates, while others generally yield density estimates that contain negative ordinates (especially in the tails) or have an infinite integral. Negativity can occur naturally, as a result of data sparseness in certain regions (Boneva, Kendall, and Stefanov 1971; Kronmal and Tarter 1968), or it can be caused by relaxing the nonnegativity constraint in (1.1) in order to improve the rate of convergence of an estimator of f. Moreover, in the quest for faster convergence rates of estimators, some researchers have chosen to relax the integral constraint in (1.1) rather than the nonnegativity constraint; see Terrell and Scott (1980). There are several ways to alleviate such problems. The density estimate may be truncated to its positive part and renormalized; alternatively, one might estimate a transformed version of f, say $\log f$ or $f^{1/2}$, and then transform back to get a nonnegative estimate of f. Gajek (1986) proposed a simple improvement scheme by which any density estimator that was not a bona fide density could be made to converge to a bona fide density.

3. THE HISTOGRAM

Traditionally, the histogram has been used to provide a visual clue to the general shape of f. Suppose f has support $\Omega = [a, b]$, where a and b are usually taken to encompass the observed data. Partition $[a, b]$ into a grid (or mesh) or m nonoverlapping bins (or cells) $T_i = [t_{n,i}, t_{n,i+1})$ $(i = 1, 2, \ldots, m)$, where $a = t_{n,1} < t_{n,2} < \ldots < t_{n,m+1} = b$, and the bin edges $\{t_{n,i}\}$ are shown depending on the sample size n. This is generally termed a *fixed partition* of Ω. Let I_{T_i} be the indicator function of the ith bin and let N_i be the number of sample values falling into T_i $(i = 1, 2, \ldots, m)$, where $\Sigma_{i=1}^{m} N_i = n$. Then, the *histogram*, defined by

$$\hat{f}(x) = \sum_{i=1}^{m} \frac{N_i/n}{(t_{n,i+1} - t_{n,i})} I_{T_i}(x), \tag{3.1}$$

satisfies (1.1). If $h_n = t_{n,i+1} - t_{n,i}$ $(i = 1, 2, \ldots, m)$, is a common bin width, then (3.1) reduces to

$$\hat{f}_{h_n}(x) = \frac{1}{nh_n} \sum_{i=1}^{m} N_i I_{T_i}(x). \tag{3.2}$$

As a density estimator, however, the histogram leaves much to be desired, with defects that include "the fixed nature of the cell structure, the discontinuities at cell boundaries, and the fact that it is zero outside a certain range" (Hand 1982, p. 15). A much more serious defect relates to the sensitivity of histogram shapes to the choice of origin; see Silverman (1986, sec. 2.2) for an example.

3.1 The Histogram As a Maximum Likelihood Estimator

Let $H(\Omega)$ be a specified class of real-valued functions defined on Ω. The maximum likelihood (ML) problem is to find an f to maximize the likelihood function $L(f) = \Pi_{i=1}^{n} f(X_i)$, or its logarithm, subject to $f \in H(\Omega)$, $\int_\Omega f(t)\, dt = 1$, and $f(t) \geq 0$ $(\forall t \in \Omega)$. If $H(\Omega)$ is finite dimensional, then a (not necessarily unique) solution to this problem exists and is called an *ML estimator of* f. The uniqueness of the solution depends upon the specification of $H(\Omega)$. The histogram is the unique ML estimator based on the random sample $X_1, \ldots, X_n$, where H consists of functions of the form $\Sigma_{i=1}^{m} y_i I_{T_i}$ $(y_i \in \mathbf{R})$. See de Montricher, Tapia, and Thompson (1975), where the histogram was also described as a polynomial spline of degree 0 (functions which are piecewise constant) with knots at the points $t_{n,1}, \ldots, t_{n,m+1}$. More generalized versions of the histogram using polynomial splines of higher degree appear in Tapia and Thompson (1978, chap. 3).

3.2 Statistical Properties

Under different sets of conditions on f and (3.2), Scott (1979) and Freedman and Diaconis (1981b) showed that if $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then IMSE $\to 0$, and that IMSE is asymptotically minimized if $h_n^* = [6/R(f')]^{1/3} \times n^{-1/3}$, where $R(g) = \int_{-\infty}^{\infty} [g(x)]^2\, dx$. For Gaussian data with variance σ^2, for example, $h_n^* = 3.49\sigma n^{-1/3}$. The optimal IMSE convergence rate of $O(n^{-2/3})$ is substantially slower than most other kinds of density estimators, such as kernel estimators, and gives a more technical reason why histograms should not be used as density estimators. Devroye and Gyorfi (1985, secs. 3.3 and 5.4) showed that the histogram (3.2) was strongly consistent for all f and that MIAE was of order $O(n^{-1/3})$. See also Freedman and Diaconis (1981a).

3.3 Choice of Bin Width

Since h_n^* depends upon the unknown f through $R(f')$, an estimate $\hat{f}$ of f can be "plugged into" h_n^*. For example, Scott (1979) found that the approximate optimal bin width $\hat{h}_n^* = 3.49sn^{-1/3}$, where s is the sample standard deviation, worked well for Gaussian samples, while it led to overly large bin widths and hence oversmoothing otherwise. Freedman and Diaconis (1981b) suggested a "simple, robust rule [that] often gives quite reasonable results," namely, $\hat{h}_n^* = 2(\mathrm{IQR})n^{-1/3}$, where IQR is the interquartile range of the data. Numerical comparisons by Emerson and Hoaglin (1983) of the Scott and Freedman–Diaconis rules showed the Freedman–Diaconis rule led to narrower bin widths, although "in practical applications the two rules will often lead to the same choice of interval width." Terrell and Scott (1985) and Terrell (1990) argued that h_n should be chosen conservatively by restricting the choice of bin width to the value that yields the smoothest density, subject to a given measure of spread (such as the standard deviation or range). Information-based methods for the histogram were studied by Taylor (1987), who used Akaike's information criterion for determining an optimal histogram bin width, and by Rodriguez and van Ryzin (1985), who defined maximum entropy histograms. Scott (1988) studied hexagonal and square *bin shapes* for bivariate histograms.

3.4 Related Estimators

By modifying the block-like shape of the histogram, a faster rate of IMSE convergence of $O(n^{-4/5})$ (or close to it) can be attained by the following estimators.

The *averaged shifted histogram* (ASH) of Scott and Thompson (1983) and Scott (1985a) is constructed by averaging several histograms with equal bin widths but different bin locations and was motivated by the need to resolve the problem of a choice of bin origin; its computational efficiency in the multivariate case has made the ASH popular among many researchers.

The classical *frequency polygon* (FP), studied by Scott (1985b), is constructed by connecting the mid-bin values of the histogram with straight lines. The FP was especially recommended for interpolating the ASH, leading to the ASH-FP. Jones (1989) studied discretization and interpolation problems related to the ASH and ASH-FP.

The *histospline* of Boneva, Kendall, and Stefanov (1971) is a cardinal quadratic spline fitted to the histogram and is obtained by interpolating the knots of the sample distribution function $\hat{F}_n = n^{-1} \sum_{i=1}^n I_{[X_i \le x]}$ and then differentiating the cubic spline estimator of the distribution function F.

A *weighted histogram* estimator of f, also referred to as a *Bernstein polynomial-type approximation,* was proposed by Vitale (1975) and Gawronski and Stadtmuller (1980), where the bin counts were weighted by empirical Poisson probabilities.

4. KERNEL DENSITY ESTIMATION

The *multivariate kernel density estimator of f* has the form

$$\hat{f}_h(\mathbf{x}) = (nh^d)^{-1} \sum_{j=1}^n K\left(\frac{\mathbf{x} - \mathbf{X}_j}{h}\right), \qquad \mathbf{x} \in \mathbf{R}^d, \quad (4.1)$$

where the choice of *kernel function K* and the *window width* $h = h_n > 0$ determine the performance of $\hat{f}_h$ as an estimator of f. It is interesting to note that Cacoullos (1966) appears to have been the first to call K in (4.1) a *kernel function;* previously, K was referred to as a *weight* function. Note that the same amount of smoothing is used in (4.1) for each of the d dimensions. The fast Fourier transform is recommended for computing (4.1) in the univariate case ($d = 1$); see Silverman (1982a) and Jones and Lotwick (1984). Since (4.1) shows that $\hat{f}_h$ inherits whatever properties the kernel K possesses, it is important that K have desirable properties.

The simplest class of kernels consists of probability density functions that satisfy

$$K(\mathbf{x}) \ge 0, \qquad \int_{\mathbf{R}^d} K(\mathbf{x})\, d\mathbf{x} = 1. \quad (4.2)$$

If a kernel K from this class is used in (4.1), then $\hat{f}_h$ will always be a bona fide probability density. Popular choices of univariate kernels include the Gaussian kernel with unbounded support,

$$K(x) = (2\pi)^{-1/2} e^{-x^2/2}, \qquad x \in \mathbf{R}, \quad (4.3)$$

and the compactly supported "polynomial" kernels,

$$K(x) = \kappa_{rs}(1 - |x|^r)^s I_{[|x| \le 1]},$$

$$\kappa_{rs} = \frac{r}{2\mathrm{Beta}(s + 1, 1/r)}, \qquad r > 0, s \ge 0. \quad (4.4)$$

The rectangular kernel obtains in (4.4) if $s = 0$ ($\kappa_{r0} = 1/2$); the triangular kernel if $r = 1$, $s = 1$ ($\kappa_{11} = 1$); the Bartlett–Epanechnikov kernel if $r = 2$, $s = 1$ ($\kappa_{21} = 3/4$); the biweight kernel if $r = 2$, $s = 2$ ($\kappa_{22} = 15/16$); the triweight kernel if $r = 2$, $s = 3$ ($\kappa_{23} = 35/32$); and, after a suitable rescaling, the Gaussian kernel if $r = 2$, $s = \infty$. The triangular kernel density estimate is asymptotically related to the ASH since the former is obtained as a limit of the latter as the number of shifted histograms becomes infinite. For $\mathbf{x} \in \mathbf{R}^d$, multivariate kernels are usually radially symmetric unimodal densities such as the Gaussian $K(x) = (2\pi)^{-d/2} e^{-(1/2)\mathbf{x}^\tau\mathbf{x}}$, and the Bartlett–Epanechnikov, $K(\mathbf{x}) = ((d + 2)/2c_d)(1 - \mathbf{x}^\tau\mathbf{x})I_{[\mathbf{x}^\tau\mathbf{x} \le 1]}$, $c_d = \pi^{d/2}/\Gamma((d/2) + 1)$.

In certain situations (Cacoullos 1966), *product kernels* may be appropriate, where $K(\mathbf{x}) = \prod_{i=1}^d K(x_i)$ is a product of univariate kernel functions. For example, Figures 2 and 3 were computed using bivariate product Gaussian kernel density estimates. In a similar study, Scott, Gotto, Cole, and Gorry (1978) used bivariate product biweight kernel density estimates.

4.1 Statistical Properties

Deriving asymptotic properties of kernel density estimates depends on the particular viewpoint considered. Devroye (1983), using the L_1 approach, proved the remarkably simple result that if K satisfies (4.2), then the kernel estimator (4.1) will be a strongly consistent estimator of f if and only if $h_n \to 0$ and $nh_n^d \to \infty$, as $n \to \infty$, without any conditions on f. Devroye and Penrod (1984) also showed that, for the univariate case, MIAE was of order $O(n^{-2/5})$, better than the L_1 rate for histograms. Explicit formulas for minimum MIAE and asymptotically optimal smoothing parameters for kernel estimators were obtained by Hall and Wand (1988).

For the L_2 approach, under regularity conditions on K and f, Parzen (1962) showed that if $h_n \to 0$ as $n \to \infty$, then the univariate kernel estimator was both asymptotically unbiased and asymptotically normal. Cacoullos (1966) showed that the asymptotic expression for IMSE for the d-dimensional case was minimized over all h satisfying the above conditions by $h_n^{\mathrm{IMSE}} = \alpha(K)\beta(f)n^{-1/(d+4)}$, where $\alpha(K)$ depends only on the kernel K and $\beta(f)$ depends only on f; furthermore, IMSE $\to 0$ at rate $O(n^{-4/(d+4)})$. The results show clearly the dimensionality effect, since these convergence rates become slower as d increases. In the univariate case, if K is the standard Gaussian kernel (4.3) and f is a Gaussian density with variance σ^2, then $h_n^{\mathrm{IMSE}} = 1.06\sigma n^{-1/5}$ would be the optimal window width. Additional consistency results were obtained by Hall and Hannan (1988).

4.2 Choice of Kernel

It has been known for some time that although the Bartlett–Epanechnikov kernel minimizes the optimal asymptotic IMSE with respect to K, IMSE is quite insensitive to the shape of the kernel. Marron and Nolan (1987) gave further results in this direction. As a result, more exotic types of kernels are now being studied. The most important of these developments concerns a hierarchy of classes of kernels defined by the existence of certain moments of K. In this scheme, those univariate symmetric kernels K that integrate to unity are called *order 0* kernels, while *order s* kernels, for some positive integer s, are those order 0 kernels whose first $s - 1$ moments vanish but whose sth moment is finite. Thus second-order kernels have zero mean and finite variance and include all compactly supported kernels. Order s kernels, for $s \geq 3$, have zero variance, which can be achieved only if K takes on negative values. Such kernels are important for bias reduction and improving the IMSE convergence rate. For example, if K is an order s kernel, then the fastest asymptotic rate of MSE convergence of $\hat{f}$ to f is $O(n^{-2s/(2s+1)})$; thus, for a fourth-order kernel, which cannot be nonnegative, the minimum asymptotic MSE convergence rate of $\hat{f}$ to f is of order $O(n^{-8/9})$, which is faster than the best such rate, $O(n^{-4/5})$, for nonnegative kernels (see Gasser, Muller, and Mammitzsch 1985). Hall and Marron (1988) considered optimal selection of the order s. Cline (1988) defined the admissibility of kernel estimators and showed that while the Bartlett–Epanechnikov kernel is not admissible among all kernels, it is admissible among all nonnegative kernels.

4.3 Choice of Window Width

Early work on the kernel method emphasized asymptotic results, whereas determining an optimal h is the main research focus today. Since the optimal window width, h_n^{IMSE}, depends explicitly on the unknown f through $\beta(f)$, it cannot be computed exactly. Several "plug-in" procedures were proposed whereby $\beta(\hat{f})$ was used to estimate $\beta(f)$, but these were generally unsatisfactory (e.g., see Scott and Terrell 1987).

An automatic method for determining the optimal window width is cross-validation (CV). The basic algorithm involves removing a single value, say X_i, from the sample, computing the appropriate density estimate at that X_i from the remaining $n - 1$ sample values,

$$\hat{f}_{h,i}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right), \qquad (4.5)$$

and then choosing h to optimize some given criterion involving all values of $\hat{f}_{h,i}(X_i)$ $(i = 1, 2, \ldots, n)$. Two different versions of CV have been used in density estimation: likelihood cross-validation and least squares cross-validation. For *likelihood cross-validation,* h^{LCV} is that h that *maximizes* the "pseudo-likelihood" $L(h) = \prod_{i=1}^{n} \hat{f}_{h,i}(X_i)$. For *least squares cross-validation,* h^{LSCV} is that h that *minimizes* $LS(h) = R(\hat{f}_h) - (2/n) \sum_{i=1}^{n} \hat{f}_{h,i}(X_i)$, which is exactly unbiased for MISE $- R(f)$. Marron (1987b) provided an excellent survey of these and other automatic smoothing parameter methods.

Mixed results have been obtained for CV methods in kernel density estimation. It has been shown, for example, that when using compactly supported kernels [such as (4.4)], likelihood CV produces consistent estimates of compactly supported densities (Chow, Geman, and Wu 1983) but does not necessarily do so for estimating infinitely supported densities (Schuster and Gregory 1981). The complex influence that the tails of both K and f have on likelihood CV was studied by Hall (1987a) in terms of the Kullback–Leibler norm. Broniatowski, Deheuvels, and Devroye (1989) related such convergence problems to the stability of the extreme order statistics. Simulation studies by Scott and Factor (1981) indicated that, depending upon the type of kernel employed, likelihood CV could lead to either a severely undersmoothed or oversmoothed density estimate. Furthermore, the criterion $L(h)$ was found to be very sensitive to outliers. Obvious modifications of $L(h)$, including truncating f, have been considered; see Hall (1982) and Marron (1985).

Least squares CV does not seem to display the peculiar behavior exhibited by likelihood CV. Indeed, very mild tail conditions on f and K are needed to prove asymptotic optimality results for least squares CV. See, for example, Hall (1983a) and Stone (1984), who showed that h^{LSCV} asymptotically minimized ISE. Bowman (1984) also showed, via simulation, that least squares CV achieved satisfactory results for long-tailed f. Hall and Marron (1987a, b) proved that h^{LSCV} performed asymptotically as well as the optimal (but unattainable) window width h^{IMSE}; they then went on to show that although h^{LSCV} converged very slowly, the least squares CV choice of window width could not be improved upon asymptotically. Scott and Terrell (1987) introduced a version of the criterion $LS(h)$ that was *biased* for MISE and showed that although large asymptotic performance gains could be obtained from such a biased CV procedure, no currently available (biased or unbiased) CV procedure could be considered highly reliable for very small samples.

The high sampling variability of CV estimates led Terrell (1990) to propose that the smoothest density estimate be chosen that is compatible with the estimated scale of the density. Taylor (1989) and Hall (1990) showed that the bootstrap also works well for selecting h in large samples and if resampling is carried out with a reduced sample size.

4.4 Related Estimators

Applying the ideas of sequential analysis to kernel density estimation led to the development of *sequential density estimators* by Deheuvels (1973), Davies and Wegman (1975), and Carroll (1976); for this type of estimator, sequential sampling is carried out, and the kernel estimator is computed at each sample size until the conditions of a given stopping rule are satisfied, so that sample size is random. A related estimator is the *recursive density estimator,* where the kernel density estimator is calculated recursively, $\hat{f}_n$ from $\hat{f}_{n-1}$; this estimator was introduced independently by Wolverton and Wagner (1969) and Yamato (1971), and further studied by Devroye (1979) and Wegman and Davies (1979). See Prakasa Rao (1983, chap. 5).

5. LOCAL ADAPTIVE SMOOTHING

The methods for nonparametric density estimation so far described are quite insensitive to local peculiarities in the data, such as data clumping in certain regions and data sparseness in others, particularly the tails. In this section, we describe attempts at constructing nonparametric density estimators that are more sensitive to the clustering of sample values.

5.1 Variable Partition Histograms

The results described in Section 3 were restricted to the fixed partition case. Some work has appeared in which the histogram concept has been made more data-sensitive as an estimator of f. This development, which led to the *variable partition histogram,* was originally suggested by Wegman (1969, 1975). Variable partition histograms are constructed in a similar manner as fixed partition histograms, but in this case the partition depends upon the gaps between the order statistics $X_{(1)}, \ldots, X_{(n)}$. Choose an integer $m \in [2, n]$ to be the number of bins of the histogram and then set $k = [n/m]$. A partition $\mathbf{P} = \{P_{in}\}$ can be obtained by defining $P_{1n} = [X_{(1)}, X_{(k)}]$, $P_{2n} = (X_{(k)}, X_{(2k)}]$, $\ldots$, $P_{mn} = (X_{((m-1)k)}, X_{(n)}]$, so that each interval contains about k sample values. Then, for any $x \in [X_{(1)}, X_{(n)}]$, estimate f by

$$\hat{f}(x) = \sum_{i=1}^{m} \frac{k/n}{(X_{(ik)} - X_{((i-1)k+1)})} I_{P_{in}}(x). \tag{5.1}$$

Clearly, $\hat{f}$ is constant on the intervals $\{P_{in}\}$ and is, therefore, a histogram-type estimator of f. Wahba (1971) and Van Ryzin (1973) indicated that variable partition histograms were related to polynomial spline estimators. In the L_1 approach, Devroye and Gyorfi (1983, sec. 7.5) showed that if $k = k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, then $\hat{f}$ in (5.1) is a strongly consistent estimator of f. Similar results for the L_2 case can be found in Prakasa Rao (1983, sec. 2.4), Lecoutre (1986), and Kogure (1987). Note that the results of Lecoutre are not valid when f is Gaussian. The rate of convergence for MISE of the estimator (5.1) is $O(n^{-2/3})$, the same order as for the fixed partition case. Kanazawa's (1988) results used the Hellinger distance approach and a dynamic programming algorithm, but gave no asymptotic rate of convergence for the estimator.

5.2 Estimators Based on Statistically Equivalent Blocks

A multivariate version of the variable partition histogram was constructed by Gessaman (1970) and applied to nonparametric discrimination in Gessaman and Gessaman (1972). See also Quesenberry and Gessaman (1968). This estimator was defined over a partitioning of the sample space into statistically equivalent blocks (a term introduced by Tukey and abbreviated 'se-blocks'). An se-block is a multivariate analog of the gap between two adjacent order statistics, and was originally used for constructing nonparametric tolerance regions (Anderson 1966; Fraser 1951, 1953, 1957, sec. 4.3; Fraser and Guttman 1956; Tukey 1947, 1948; Wald 1943; and Wilks 1962, sec. 8.7). Since this estimator does not appear in any book or review of nonparametric density estimation, some detail is provided here.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample on $\mathbf{X} \in \mathbf{R}^d$. The procedure for constructing se-blocks depends on a sequence, $h_1(\mathbf{x}), \ldots, h_n(\mathbf{x})$, of n real-valued functions of $\mathbf{X}$, not necessarily different, and a set of integers, $(j_1, j_2, \ldots, j_n)$, that forms a permutation of $(1, 2, \ldots, n)$. Typically, $h_\alpha(\mathbf{x}) = x_k$, the kth coordinate of $\mathbf{x}$. At the first step, $h_{j_1}(\mathbf{x})$ is used to order the $\{\mathbf{X}_\alpha\}$. Define $\mathbf{X}^{(j_1)}$ as that $\mathbf{X}_\alpha$ for which $h_{j_1}(\mathbf{x}^{(j_1)})$ is the j_1st smallest of the $h_{j_1}(\mathbf{x}_\alpha)$ values. The *cut* $h_{j_1}(\mathbf{x}) = h_{j_1}(\mathbf{x}^{(j_1)})$ creates two disjoint *blocks* $B_{1\ldots j_1} = \{\mathbf{x}: h_{j_1}(\mathbf{x}) \le h_{j_1}(\mathbf{x}^{(j_1)})\}$ and $B_{j_1+1\ldots n+1} = \{\mathbf{x} : h_{j_1}(\mathbf{x}) > h_{j_1}(\mathbf{x}^{(j_1)})\}$. Thus, there are exactly $j_1 - 1$ $\mathbf{X}_\alpha$ in $B_{1\ldots j_1}$ and exactly $n - j_1$ in $B_{j_1+1\ldots n+1}$. At the second step, if $j_2 < j_1$, then $h_{j_2}(\mathbf{x})$ is used to order the $j_1 - 1$ $\mathbf{X}_\alpha$'s in $B_{1\ldots j_1}$. Define $\mathbf{X}^{(j_2)}$ as that $\mathbf{X}_\alpha$ for which $j_2 - 1$ $\mathbf{X}_\alpha$'s satisfy $h_{j_2}(\mathbf{x}_\alpha) < h_{j_2}(\mathbf{x}^{(j_2)})$ and $h_{j_1}(\mathbf{x}_\alpha) < h_{j_1}(\mathbf{x}^{(j_1)})$ and $j_1 - j_2 - 1$ $\mathbf{X}_\alpha$'s satisfy $h_{j_2}(\mathbf{x}_\alpha) > h_{j_2}(\mathbf{x}^{(j_2)})$ and $h_{j_1}(\mathbf{x}_\alpha) < h_{j_1}(\mathbf{x}^{(j_1)})$. The cut $h_{j_2}(\mathbf{x}) = h_{j_2}(\mathbf{x}^{(j_2)})$ divides the block $B_{1\ldots j_1}$ into subblocks $B_{1\ldots j_2} = B_{1\ldots j_1} \cap \{\mathbf{x} : h_{j_2}(\mathbf{x}) \le h_{j_2}(\mathbf{x}^{(j_2)})\}$ and $B_{j_2+1\ldots j_1} = B_{1\ldots j_1} \cap \{\mathbf{x} : h_{j_2}(\mathbf{x}) > h_{j_2}(\mathbf{x}^{(j_2)})\}$. If, on the other hand, $j_1 < j_2$, then the block $B_{j_1+1\ldots n+1}$ is divided into subblocks $B_{j_1+1\ldots j_2} = B_{j_1+1\ldots n+1} \cap \{\mathbf{x} : h_{j_2}(\mathbf{x}) \le h_{j_2}(\mathbf{x}^{(j_2)})\}$ and $B_{j_2+1\ldots n+1} = B_{j_1+1\ldots n+1} \cap \{\mathbf{x} : h_{j_2}(\mathbf{x}) > h_{j_2}(\mathbf{x}^{(j_2)})\}$. This is done by ranking the $n - j_1$ $\mathbf{X}_\alpha$'s in $B_{j_1+1\ldots n+1}$ according to $h_{j_2}(\mathbf{x})$ and letting $\mathbf{X}^{(j_2)}$ be the $(j_2 - j_1)$ smallest in the ranking. This procedure is continued. At the mth step, the block that is divided is the one having j_m in its index set, and the $\mathbf{X}_\alpha$ in that block are ordered by $h_{j_m}(\mathbf{x})$ and the $(j_m - j_{m_0})$ smallest value chosen to represent the cut, where j_{m_0} is the largest of the $j_1, \ldots, j_{m-1}$ that are less than j_m. After n steps there will be $n + 1$ *se-blocks,* $B_1, B_2, \ldots, B_{n+1}$. The map of se-blocks is completely determined by the functions $\{h_\alpha\}$ and the permutation used.

To construct the density estimator, consider the bivariate case [$d = 2$, where $\mathbf{X} = (X_1, X_2)$]. Let $k_n > 0$ be an integer (Gessaman suggested $k_n = [n^{1/3}]$). Superimposed over the map of se-blocks, make $[(n/k_n)^{1/2}] - 1$ evenly spaced *vertical* line cuts at the ordered X_1-observations. After deleting the observations used to make the cuts, make a further $[(n/k_n)^{1/2}] - 1$ evenly spaced *horizontal* line cuts at the ordered X_2-observations. The plane will now be partitioned into $[(nk_n)^{1/2}]$ subblocks or *probability squares* (Gessaman and

Gessaman 1972). Each probability square will be the union of about k_n se-blocks and, therefore, will contain about k_n observations. If B_n is a *bounded* probability square and $\mathbf{x} \in B_n$, set

$$\hat{f}(\mathbf{x}) = \frac{k_n/(n+1)}{\text{area}(B_n)}. \tag{5.2}$$

On *unbounded* probability squares, estimate f as 0. Gessaman (1970) showed that if $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, then the estimator (5.2) was weakly consistent for f. Convergence rates and some optimal choice for k_n in (5.2) have yet to be determined, however.

5.3 Nearest Neighbor Methods

Fix and Hodges (1951) proposed the nearest neighbor estimator in the context of nonparametric discrimination. See Silverman and Jones (1988) for a modern interpretation. At a fixed point $\mathbf{x}$ and for fixed integer k, let $D_k(\mathbf{x})$ be the Euclidean distance from $\mathbf{x}$ to its kth nearest neighbor among the $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, and let $\text{vol}_k(\mathbf{x}) = c_d[D_k(\mathbf{x})]^d$ be the volume of the d-dimensional sphere of radius $D_k(\mathbf{x})$, where c_d is the volume of the unit d-dimensional sphere. The *kth nearest neighbor* (k-NN) *density estimator* is then given by

$$\hat{f}(\mathbf{x}) = \frac{k/n}{\text{vol}_k(\mathbf{x})}. \tag{5.3}$$

Tukey and Tukey (1981, sec. 11.3.2) called (5.3) the *balloon density estimate* of f. An advantage of the k-NN estimator is that it is always positive, even in regions of sparce data. Loftsgaarden and Quesenberry (1965) proved (5.3) was consistent if $k = k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Abramson (1984) proposed that in the d-dimensional case, k_n should be chosen proportional to $n^{4/(d+4)}$, the constant of proportionality depending on $\mathbf{x}$. The k-NN estimator (5.3) can be written as an kernel density estimator by setting

$$\hat{f}(\mathbf{x}) = \frac{1}{n[D_k(\mathbf{x})]^d} \sum_{j=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_j}{D_k(\mathbf{x})}\right), \tag{5.4}$$

where the smoothing parameter is now k and the kernel K is the rectangular kernel. Moore and Yackel (1977) and Mack and Rosenblatt (1979) analyzed the bias and variance of (5.3). Rosenblatt (1979) studied the global behavior of generalized nearest neighbor estimates of f. See also Mack (1980) and Abramson (1984). Although the k-NN estimator appeared reasonable for estimating a density at a point, it was not particularly successful for estimating the entire density function f. Indeed, the estimator was not a bona fide density since (5.3) was discontinuous and had an infinite integral due to very heavy tails. Devroye and Gyorfi (1985, p. 21) noted that, because of these difficulties, "it is impossible to study its properties in L_1."

5.4 Variable Kernel Estimators

The variable kernel estimator, which was an attempt to avoid the problems associated with the k-NN estimator, was defined by setting

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{H_{jk}^d} K\left(\frac{\mathbf{x} - \mathbf{X}_j}{H_{jk}}\right), \tag{5.5}$$

where the variable window width $H_{jk} = hD_k(\mathbf{X}_j)$ does not depend on $\mathbf{x}$ as did (5.4), h is a smoothing parameter, and k controls the local behavior of H_{jk}. The estimator (5.5) is a bona fide density if the kernel K satisfies (4.2). It was apparently first considered by Meisel in 1973 in the context of pattern recognition and then studied empirically by Breiman, Meisel, and Purcell (1977), who listed its advantages as having the smoothness properties of kernel estimators, the data-adaptive character of the k-NN approach, and very little computational penalty. In their simulation studies, the estimator (5.5) performed very poorly unless k was large, on the order of $.10n$. Conditions for consistency of the variable kernel estimator were obtained by Wagner (1975) and Devroye (1985); Devroye and Penrod (1986) proved the strong uniform consistency of (5.5).

5.5 Adaptive Kernel Estimators

The variable kernel estimator (5.5) led, in turn, to the *adaptive kernel estimator*. Abramson (1982a,b), who was concerned with estimating f at a point, proposed a two-step algorithm for computing a data-adaptive window width. First, a clipped (or winsorized) version $\tilde{f}_h^0$ is constructed from a pilot kernel density estimate $\hat{f}_h^0$ with fixed window width h and then the adaptive kernel estimator is defined as

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_j^d} K\left(\frac{\mathbf{x} - \mathbf{X}_j}{h_j}\right), \tag{5.6}$$

where $h_j = h[\tilde{f}_h^0(\mathbf{X}_j)]^{-1/2}$. Two modifications of Abramson's h_j have been suggested. Silverman (1986, sec. 5.3) set $h_j = h[(1/g)\,\hat{f}_h^0(\mathbf{X}_j)]^{-\alpha}$, where g is a scale factor [such as the geometric mean of the $\hat{f}_h^0(X_i)$, $i = 1, 2, \ldots, n$] and $0 \le \alpha \le 1$ reflects the sensitivity of the window width to variations in the pilot estimate; examples of Silverman's adaptive window widths and $\alpha = 1/2$ were also given that demonstrated better tail behavior than the corresponding fixed window width kernel estimator. Hall and Marron (1988) set $h_j = h_F[\hat{f}_{h_P}^0(\mathbf{X}_j)]^{-1/2}$ in (5.6), where h_P was the smoothing parameter of the pilot estimate and h_F was the smoothing parameter of the final estimate; they showed that their modification had a very fast rate of MSE convergence.

6. ORTHOGONAL SERIES ESTIMATORS

Orthogonal series density estimators were introduced by Cencov (1962) and have since been applied to several different areas, especially pattern recognition and discrimination and classification; see Greblicki and Pawlak (1981). The method has been used to estimate multivariate densities for dichotomous (Ott and Kronmal 1976), polychotomous (Butler and Kronmal 1985), and mixed continuous and discrete variables (Hall 1983b).

6.1 Arbitrary Orthogonal Expansions

This method assumes that a square-integrable f can be represented as a convergent orthogonal series expansion,

$$f(x) = \sum_{k=-\infty}^{\infty} a_k \varphi_k(x), \qquad x \in \Omega, \tag{6.1}$$

where $\{\varphi_k\}$ is a complete orthonormal system of functions on a set Ω of the real line [that is, satisfying $\int_\Omega \varphi_j(x)\varphi_k(x)\,dx = \delta_{jk}$, where δ_{jk} is the Kronecker delta] and $\{a_k\}$ are coefficients defined by $a_k = E_f[\varphi_k^*(X)]$, where φ_k^* is the complex conjugate of φ_k. This formulation allows for systems of real- or complex-valued orthonormal functions. Orthonormal systems proposed for $\{\varphi_k\}$ are those with compact support (such as the Fourier, trigonometric, and Haar systems on [0, 1], and Legendre system on [−1, 1]) and those with unbounded support [such as the Hermite system on **R** and Laguerre system on [0, ∞)].

Given an independent sample, $X_1, X_2, \ldots, X_n$, from f and a system $\{\varphi_k\}$, the $\{a_k\}$ can be estimated unbiasedly by

$$\hat{a}_k = \frac{1}{n}\sum_{j=1}^{n} \varphi_k^*(X_j). \tag{6.2}$$

The obvious estimator of f, obtained by plugging (6.2) into (6.1) in place of a_k, may not be well defined: It has infinite variance and is not consistent in the ISE sense. Tapered estimators of the form

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} b_k \hat{a}_k \varphi_k(x), \qquad x \in \Omega, \tag{6.3}$$

have been studied, where $0 < b_k < 1$ is a symmetric weight ($b_{-k} = b_k$) that shrinks $\hat{a}_k$ towards the origin, and $\Sigma|b_k| < \infty$ is needed for pointwise convergence of (6.3). See, for example, Watson (1969), Rosenblatt (1971), Brunk (1978), and Hall (1986). Tapered orthogonal series estimators were used by Johnstone and Silverman (1990) to estimate bivariate glucose density within the brain. The choice $b_k = 1$ for $-r \le k \le r$ and 0 otherwise leads to the partial sums of (6.1) being approximated by

$$\hat{f}_r(x) = \sum_{k=-r}^{r} \hat{a}_k \varphi_k(x), \qquad x \in \Omega, \tag{6.4}$$

where $\{\hat{a}_k\}$ are given by (6.2). Wahba (1981) considered a two-parameter system of weights, $b_k = (1 + \lambda(2\pi k)^{2m})^{-1}$ for $-r \le k \le r$, where $\lambda > 0$ is a smoothing parameter and $m > 1/2$ is a shape parameter. Other systems of weights were discussed by Hall (1987) and Lock (1990). To estimate the $\{b_k\}$, likelihood cross-validation was proposed by Wahba (1981) and least squares cross-validation by Hall (1987b). In related work, Anderson and de Figueiredo (1980) developed an adaptive orthogonal series estimator.

6.2 Statistical Properties

The most popular orthogonal series estimator for densities with unbounded support, usually **R** or [0, ∞), is the *Hermite series estimator*. The normalized Hermite functions given by $\varphi_k(x) = c_k(x)H_k(x)$ ($k = 0, 1, 2, \ldots$), where $c_k = e^{-x^2/2}/(2^k k!\pi^{1/2})^{1/2}$ and $H_k(x) = (-1)^k e^{-x^2/2}(d^k/dx^k)(e^{-x^2})$ is the kth Hermite polynomial, form an orthonormal basis for an L_2 approach. They are heavily weighted in the tails by $e^{-x^2/2}$ and provide sufficient protection against unusual tail behavior of X; see Hall (1987b). Schwartz (1967) showed that if $r = r_n$ in (6.4) satisfies $r_n/n \to 0$ as $r_n \to \infty$, then IMSE → 0 as $n \to \infty$; moreover, if $r_n = O(n^{1/q})$ for $q \ge 2$, then IMSE $= O(n^{-(1-1/q)})$. Walter (1977) improved this last result slightly. Note that the IMSE convergence rate is independent of the dimension of the data, which gives the Hermite series estimator an advantage over the kernel estimator for multivariate density estimation. The Hermite system does not form a basis for the L_1 approach, however, and the Hermite series estimator is neither translation invariant nor consistent in the L_1 sense.

If f has compact support [0, 1], say, the popular *Fourier (or trigonometric) series estimate,* which is the real part of (6.4), is formed from the system of discrete Fourier functions, defined by $\varphi_k(x) = e^{2\pi ikx}$ [$i = (-1)^{1/2}$, $k = 0, 1, 2, \ldots$]. See Wahba (1975a, 1975b, 1981) and Hall (1981) for details and comments about the influence of periodicity and the Gibbs phenomenon on Fourier series density estimates. Devroye and Gyorfi (1985, sec. 12.4) proved that for the Fourier series estimator, under suitable conditions on f and if $r_n/n \to 0$ as $r_n \to \infty$, then MIAE → 0 as $n \to \infty$.

Arguments about the relative merits of the Hermite system versus the Fourier system can be found in Walter (1977) and Good and Gaskins (1980). Wahba (1981) suggested that "in many applications it might be preferable to assume the true density has compact support and to scale the data to the interior of [0, 1]."

6.3 Choice of Number of Terms

The performance and smoothness of the orthogonal series density estimate (6.4) depend on r, the number of terms in the expansion. Kronmal and Tarter (1968) proposed a term-by-term optimal stopping rule for choosing r by minimizing an estimated MISE criterion. Disadvantages of that rule were pointed out by Crain (1973), who suggested that it might not yield the optimal r; by Hart (1985), who noted from simulation studies that the rule tended to stop too soon, thus yielding oversmoothed density estimates; and by Diggle and Hall (1986), who warned about the possible poor performance and inconsistency of the rule in multimodal situations. Improvements were suggested by Hart (1985) and Diggle and Hall (1986), and Lock (1990) combined choice of the number of terms with a tapered estimator and showed its advantages in a simulation study.

7. DELTA SEQUENCE DENSITY ESTIMATORS

Many of the different methods described so far for nonparametric density estimation are special cases of the following general class of density estimators. Let $\delta_\lambda(x, y)$ ($x, y \in \mathbf{R}$), be a bounded function indexed by a smoothing parameter $\lambda > 0$. The sequence $\{\delta_\lambda(x, y)\}$ is called a *delta sequence on* **R** if $\int_{-\infty}^{\infty} \delta_\lambda(x, y)\phi(y)\,dy \to \phi(x)$ as $\lambda \to \infty$ for every infinitely differentiable function ϕ on **R**. Any estimator that can be written in the form

$$\hat{f}_\lambda(x) = \frac{1}{n}\sum_{j=1}^{n} \delta_\lambda(x, X_j), \qquad x \in \mathbf{R}, \tag{7.1}$$

where $\{\delta_\lambda(x, y)\}$ is a delta sequence, is called a *delta sequence density estimator*. Thus histograms, kernel estimators, and orthogonal series estimators can each be written in the form (7.1):

histograms: $\delta_m(x, X_j) = \Sigma_{i=1}^{m} (t_{i+1} - t_i)^{-1} I_{T_i}(x) I_{T_i}(X_j)$ [see (3.1)]

kernels: $\delta_h(x, X_j) = \dfrac{1}{h} K((x - X_j)/h)$ [see (4.1)]

orthogonal series: $\delta_r(x, X_j) = \Sigma_{k=-r}^{r} \varphi_k(x)\varphi_k^*(X_j)$ [see (6.2), (6.4)]

In some cases (such as histograms and orthogonal series estimators), λ will be integer-valued as in the number of terms in an expansion, while in others (such as kernel estimators), λ will be real-valued. Such general density estimators were first studied by Whittle (1958). Watson and Leadbetter (1964) called them *δ-function sequences* and showed that they were asymptotically unbiased as density estimators. Further work along the same lines was carried out by Foldes and Revesz (1974). Walter and Blum (1979) and Prakasa Rao (1983, sec. 2.8) gave a long list of special cases and established MSE rates of convergence; but, see Hall (1981) for a cautionary note. Silverman (1986, sec. 2.9) referred to (7.1) as a *general weight function estimator*. Marron (1987a) used delta sequence estimators as a means of comparing different density estimators.

8. RESTRICTED MAXIMUM LIKELIHOOD ESTIMATORS

The ML method of Section 3.1 fails miserably when the class of densities H over which the likelihood L is to be maximized is otherwise unrestricted. For that case, the likelihood is maximized by a linear combination of Dirac delta functions (or "spikes") at the n sample values, resulting in a value of $+\infty$ for the likelihood. In this section, approaches to the ML problem are described in which restrictions are placed either on H or L.

8.1 Order-Restricted Methods

Consider, first, an order restriction on H. For example, densities that are *monotone decreasing* over the range $[0, \infty)$ are especially important in survival analysis; see Denby and Vardi (1986). Grenander (1956) showed that the ML estimator for a nonincreasing density on $[0, \infty)$ was a step function with jumps at the order statistics $\{X_{(i)}\}$. Specifically, if $\hat{F}_n$ is the sample distribution function, then the ML estimator of a nonincreasing density is the slope of the least concave majorant of $\hat{F}_n$, namely,

$$\hat{f}(x) = \min_{s \leq t-1} \max_{t \geq i} \frac{\hat{F}_n(X_{(t)}) - \hat{F}_n(X_{(s)})}{X_{(t)} - X_{(s)}}, \qquad X_{(i-1)} < x < X_{(i)}, \tag{8.1}$$

and 0 for $x < 0$ and $x < X_{(x)}$. Figure 4 displays the least concave majorant for a sample of size $n = 15$. The *Grenander estimator* (8.1) is strongly consistent for monotone decreasing f (Groeneboom 1983) with an MIAE convergence rate of $O(n^{-1/3})$ (Devroye 1987, chap. 8). It is also reasonably well behaved when f is close to decreasing (Birge 1986, 1989). Some modifications have been suggested to improve the performance of (8.1), including smoothing in the neighborhood of zero. For different approaches to computing (8.1), see Barlow, Bartholomew, Bremner, and Brunk (1972, chap. 5) and Denby and Vardi (1986). Alternative approaches to estimating a decreasing density were given by Birge (1987a,b).

A related order restriction concerns *unimodal* densities. First, without loss of generality, assume that the mode $M = 0$ is known. Since a unimodal density f is nondecreasing in x prior to the mode and nonincreasing thereafter, it suffices to consider only ML estimation of f_+, the conditional density on $[0, \infty)$, since a similar argument holds for f_-, the conditional density on $(-\infty, 0)$. The ML estimate of f is then given by $\hat{f} = \hat{\alpha}\hat{f}_+ + (1 - \hat{\alpha})\hat{f}_-$, where $\hat{f}_+$ is the slope of the least concave majorant of $\hat{F}_n$, $\hat{f}_-$ is the slope of the greatest convex minorant of $\hat{F}_n$, and $0 \leq \hat{\alpha} \leq 1$ is the proportion of sample values that fall into $[0, \infty)$. See, for example, Robertson, Wright, and Dykstra (1988, chap. 7). Robertson (1967) showed that the ML estimate for a univariate, unimodal density with known mode can also be expressed as a conditional expectation given the σ lattice of all intervals that contained the mode, together with the empty set, and demonstrated that isotonic regression algorithms can efficiently compute the ML estimate. When the mode is unknown, Wegman (1969) obtained the appropriate ML estimator and showed consistency; in this case the σ lattice was defined in terms of all intervals that contained a consistent estimate of the mode. Sager (1982) generalized the results of Robertson and Wegman and illustrated his results by estimating the contours of a bivariate density applied to a problem in cartography. See also Sager (1986). A related minimum-distance estimator for unimodal densities was studied by Reiss (1976).

8.2 Method of Sieves

The method of sieves is another restricted ML density estimation method in which H is restricted. It is different, however, in that the choice of "sieve" determines the density estimation method. The essence of the method of sieves is the following: For each $h > 0$, select a subset S_h of densities for which a ML estimator does exist; next, find the restricted ML density estimator $\hat{f}_h$ by maximizing the likelihood function

$$L_h(f) = \prod_{i=1}^{n} f(X_i), \qquad f \in S_h; \tag{8.2}$$

and, finally, let the subset S_h grow (in some sense) with the sample size n, while allowing $h = h_n \to 0$ as $n \to \infty$ in such a way as to ensure that the ML estimator converges to a density function. The sequence $\{S_h\}$ of these subsets is called a *sieve*, *h* is called the *sieve parameter* or *mesh size*, and the estimation procedure is called the *method of sieves*. For specific sieves, this method produced the histogram, MPL, and orthogonal series estimators, but, surprisingly, not the Gaussian kernel estimator.

The method was introduced by Grenander (1981, part III), motivated by his work in pattern analysis and "based on an idea of Wald refined by Bahadur." It was further developed by Geman and Hwang (1982) and Walter and Blum (1984). See also Wegman (1975). As with density estimators in

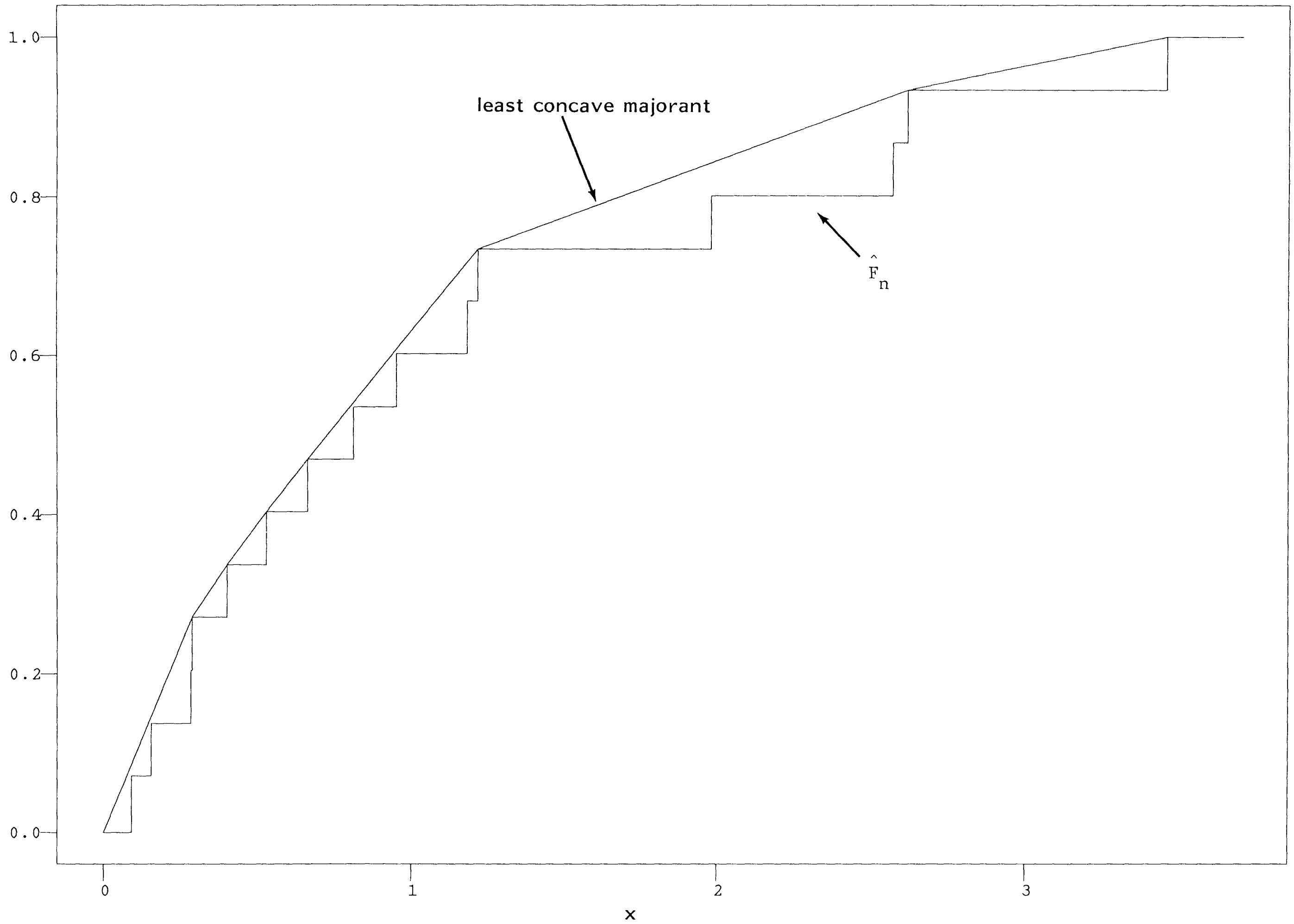

Figure 4. The Empirical Distribution Function $\hat{F}_n$ and Its Least Concave Majorant for a Sample of Size $n = 15$.

general which depend upon a smoothing parameter, the performance of the method of sieves estimator depends particularly upon the sequence of sieve parameters which should decrease to zero "at a sufficiently slow rate" (Grenander 1981, p. 426). It has been shown that this method leads to consistent estimators in the L_1 sense, although exact rates of convergence have not yet been determined. To date, the method has been studied only theoretically.

8.3 Maximum Penalized Likelihood Method

The most popular method for restricted ML density estimation, however, involves penalizing the likelihood function L for producing density estimates that are "too rough." See Good and Gaskins (1971). Thus, if Φ is a given nonnegative (*roughness*) *penalty functional* defined on H, then the Φ-*penalized likelihood* of f is defined to be

$$\tilde{L}(f) = \prod_{i=1}^{n} f(X_i)e^{-\Phi(f)}. \tag{8.3}$$

The optimization problem calls for $\tilde{L}(f)$ in (8.3), or its logarithm, to be maximized subject to $f \in H(\Omega)$, $\int_\Omega f(t)\, dt = 1$, and $f(t) \geq 0$ ($\forall t \in \Omega$). If it exists, a solution, $\hat{f}$, of that problem is called a *maximum penalized likelihood* (MPL) *estimate* of f corresponding to the penalty function Φ and class of functions H. For example, $\Phi(f) = \alpha \int_{-\infty}^{\infty} [f''(x)]^2\, dx$ is used in the International Mathematical and Statistical Libraries, Inc. (1987) routine DESPL, where $\alpha > 0$ is a *smoothing parameter*. Based on this penalty function, Figure 5 shows MPL density estimates with different α using $n = 63$ observations of Buffalo snowfall recorded during 1910–1972. Good and Gaskins observed that the MPL method could, for certain types of problems, be interpreted as "quasi-Bayesian" since (8.3) resembles a posterior density for a parametric estimation problem. Furthermore, the MPL method is closely related to Tikhonov's *method of regularization* used for solving ill-posed inverse problems (O'Sullivan 1986).

De Montricher, Tapia, and Thompson (1975) rigorously established the existence and uniqueness of MPL density estimates, and showed that the MPL method was intimately related to spline methods. For example, if f has finite support Ω and $H(\Omega)$ is a suitable class of smooth functions on Ω, then the MPL estimate $\hat{f}$ exists, is unique, and is a polynomial spline with join points (or "knots") only at the sample values.

The case when f has infinite support is more complicated. Good and Gaskins (1971) proposed penalty functionals designed to estimate the root-density, $\gamma = f^{1/2}$, so that $\hat{f} = \hat{\gamma}^2$ would be a nonnegative (and *bona fide*) estimator of f.

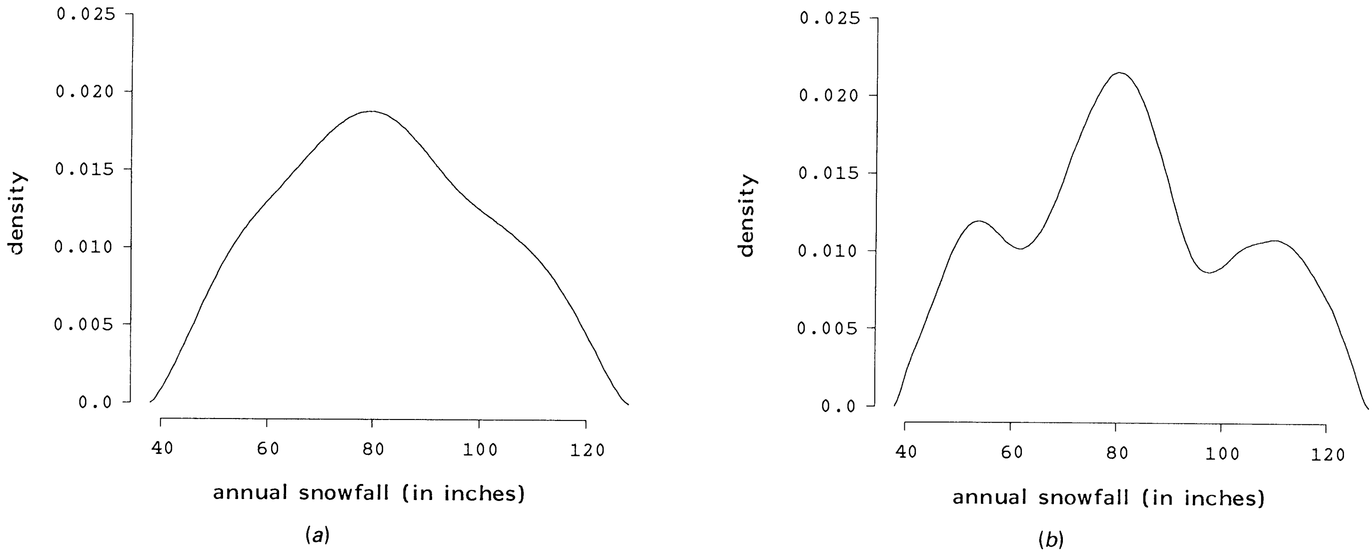

Figure 5. Maximum Penalized Likelihood Density Estimates of the 63 Annual Observations on Buffalo Snowfall, 1910–1972. The data are given in Scott (1985a). The penalty function used was $\Phi(f) = \alpha\int[f''(x)]^2\,dx$, and the smoothing-parameter values were (a) $\alpha = 10^7$, and (b) $\alpha = 10^6$. The trimodal shape [see (b)] is generally regarded as the most reasonable density estimate for these data.

The penalty functionals were

$$\Phi_1(f) = 4\alpha \int_{-\infty}^{\infty} [\gamma'(x)]^2\,dx, \qquad \alpha > 0, \tag{8.4}$$

$$\Phi_2(f) = 4\alpha \int_{-\infty}^{\infty} [\gamma'(x)]^2\,dx + \beta \int_{-\infty}^{\infty} [\gamma''(x)]^2\,dx, \qquad \alpha \geq 0, \quad \beta \geq 0, \tag{8.5}$$

where the *hyperparameters* α and β, with $\alpha + \beta > 0$ in (8.5), control the amount of smoothing. Motivation for Φ_1 and Φ_2 rested on how best to represent the "roughness" of f. Good and Gaskins preferred (8.5) to (8.4), arguing that curvature as well as slope of the density estimate should be penalized. In follow-up papers, Good and Gaskins (1980) and Good and Deaton (1981) set $\alpha = 0$ in (8.5) and used $\beta \int[\gamma''(x)]^2\,dx$ as the measure of roughness of f, where β was to be determined from the data. Klonias and Nash (1983) and Klonias (1984) investigated a very general class of penalty functionals [that included (8.4) and (8.5) as special cases] whose primary motivation was to improve estimation of peaks and valleys of f.

For the penalty function (8.4) and a given value of α, De Montricher et al. (1975) showed that, if the optimization problem is set up correctly, then the resulting estimator $\hat{\gamma}_\alpha$, say, exists, is unique, and is a positive exponential spline with knots only at the sample values. An exponential spline rather than a polynomial spline is the price to be paid for requiring nonnegativity of the density estimate. The MPL estimator is then given by $\hat{f}_\alpha = \hat{\gamma}_\alpha^2$. Klonias (1982) demonstrated consistency of $\hat{f}_\alpha$ in a number of different norms, including L_1 and L_2. As for determining the value of α, Silverman (1978c) suggested, in a slightly different setup, that α be chosen informally using graphical methods. If the penalty function is (8.5) and given values of α and β, then, provided the optimization problem is set up correctly, the resulting estimate $\hat{\gamma}_{\alpha,\beta}$ exists and is unique. The MPL estimate of f is given by $\hat{f}_{\alpha,\beta} = \hat{\gamma}_{\alpha,\beta}^2$. Good and Gaskins also gave some recommendations for (α, β) that performed well in their examples.

Another way of guaranteeing a *bona fide* density estimate using the MPL method was devised by Silverman (1982b), who used a roughness penalty based on $g = \log f$, and showed that this approach led to a wide range of possible density estimates. Solving the appropriate optimization problem yielded an estimator $\hat{g}$ of g, so that a nonnegative MPL estimate for f was given by $\hat{f} = e^{\hat{g}}$. Silverman developed a very general theory of penalty functionals based on $\log f$, and then proved the existence, consistency, and asymptotic normality of the resulting estimators. This approach was studied further by Silverman (1984).

Implementation of the MPL method depends upon the quality of the numerical solutions to the restricted optimization problems. Since $\gamma = f^{1/2}$ is square-integrable, Good and Gaskins (1980) suggested using mixtures of orthonormal expansions for γ, terminating the expansions at some finite number of terms. Scott, Tapia, and Thompson (1980) studied a discrete approximation to the spline solutions of the MPL problems, and proved that the resulting *discrete MPL estimator* exists, is unique, converges to the spline MPL estimator, and is a strongly pointwise consistent estimator of f. Further computational work on the discrete MPL estimator was carried out by Good and Deaton (1981).

9. PROJECTION PURSUIT DENSITY ESTIMATION

Multivariate kernel density estimators tend to be poor performers when it comes to dealing with high-dimensional data since extremely large sample sizes are needed to match the sort of numerical accuracy that is possible in low dimensions. In light of this, Friedman and Stuetzle (1982) and Friedman, Stuetzle, and Schroeder (1984) developed *projection pursuit density estimation* (PPDE). The PPDE method has been shown in simulations to possess excellent properties, and several quite striking applications of PPDE to real data have also been published.

9.1 The PPDE Paradigm

When dealing with small samples of high-dimensional data, the PPDE procedure may be jump-started by restricting attention to the subspace spanned by the first few significant principal components; see Friedman (1987) and Jee (1987) for examples. A PPDE of f is then formed using the following stepwise procedure. First, transform the data to have center the origin and covariance matrix the identity. Second, choose $\hat{f}^{(0)}$ to be an initial multivariate density estimate of f, usually taken to be standard multivariate Gaussian. Third, find the direction $\mathbf{a}_1 \in \mathbf{R}^d$ for which the (model) marginal $f_{\mathbf{a}_1}$ along $\mathbf{a}_1$ differs most from the current estimated (data) marginal $\hat{f}_{\mathbf{a}_1}$ along $\mathbf{a}_1$. Choice of direction $\mathbf{a}_1$ will not generally be unique. Fourth, given $\mathbf{a}_1$, define a univariate "augmenting function" $g_1(\mathbf{a}_1^T\mathbf{x})$ as the ratio of the two marginals, namely, $g_1(\mathbf{a}_1^T\mathbf{x}) = f_{\mathbf{a}_1}(\mathbf{a}_1^T\mathbf{x})/\hat{f}_{\mathbf{a}_1}(\mathbf{a}_1^T\mathbf{x})$, and update the initial estimate so that $\hat{f}^{(1)}(\mathbf{x}) = \hat{f}^{(0)}(\mathbf{x})g_1(\mathbf{a}_1^T\mathbf{x})$. Repeat this procedure on the modified density $\hat{f}^{(1)}$ so that a second direction $\mathbf{a}_2 \in \mathbf{R}^d$ and augmenting function $g_2(\mathbf{a}_2^T\mathbf{x}) = f_{\mathbf{a}_2}(\mathbf{a}_2^T\mathbf{x})/\hat{f}_{\mathbf{a}_2}(\mathbf{a}_2^T\mathbf{x})$ are found, and the density is again modified to be $\hat{f}^{(2)}(\mathbf{x}) = \hat{f}^{(1)}(\mathbf{x})g_2(\mathbf{a}_2^T\mathbf{x})$. Repeat the procedure as many times as necessary so that, at the kth iteration,

$$\hat{f}^{(k)}(\mathbf{x}) = \hat{f}^{(0)}(\mathbf{x}) \prod_{j=1}^{k} g_j(\mathbf{a}_j^T\mathbf{x}) = \hat{f}^{(k-1)}(\mathbf{x})g_k(\mathbf{a}_k^T\mathbf{x}) \qquad (9.1)$$

will be the current multivariate density estimate, where

$$g_j(\mathbf{a}_j^T\mathbf{x}) = \frac{f_{\mathbf{a}_j}(\mathbf{a}_j^T\mathbf{x})}{\hat{f}_{\mathbf{a}_j}(\mathbf{a}_j^T\mathbf{x})}, \qquad j = 1, 2, \ldots, k. \qquad (9.2)$$

In (9.1), the vectors $\{\mathbf{a}_j\}$ are unit length directions in $\mathbf{R}^d$, and the augmenting (or ridge) functions $\{g_j\}$ are used to build up the structure of $\hat{f}^{(0)}$ so that $\hat{f}^{(k)}$ converges to f in some appropriate sense as $k \to \infty$. The number of iterations k operates as a smoothing parameter and a stopping rule is determined by balancing bias against the variance of the estimate. Friedman et al. (1984) suggested graphical inspection of the augmenting functions [plotting $g_j(\mathbf{a}_j^T\mathbf{x})$ against $\mathbf{a}_j^T\mathbf{x}$ for $j = 1, 2, \ldots, k$] as a termination criterion for the iterative procedure.

Computation of the augmenting functions (9.2) has been discussed by Friedman et al. (1984), Huber (1985, sec. 15) and discussants Buja and Stuetzle (especially pp. 487–489), and Jones and Sibson (1987, sec. 3). Given $\mathbf{a}_j$, estimate $f_{\mathbf{a}_j}$ by first projecting the sample data along the direction $\mathbf{a}_j$, thus obtaining $z_i = \mathbf{a}_j^T\mathbf{x}_i$ ($i = 1, 2, \ldots, n$) and then compute a kernel density estimate from the $\{z_i\}$. Monte Carlo sampling is used to compute $\hat{f}_{\mathbf{a}_j}$, followed by kernel density estimation. Alternatives to kernel smoothing include cubic spline functions (Friedman et al. 1984) and average shifted histograms (Jee 1987).

9.2 Projection Indexes

PPDE is driven by a projection index usually of the form

$$I(f) = \int J(f(z))f(z)\,dz = E_f[J(f)], \qquad (9.3)$$

where J is a smooth real-valued functional and z is a one-dimensional projected version of $\mathbf{x}$. As a functional on f, $I(f)$ should be absolutely continuous with easily computable first derivatives. "Interesting" projections should correspond to large values of $I(f)$, while small values of $I(f)$ should correspond to random or unstructured projections.

Estimates of $I(f)$ should be amenable to fast computation, unaffected by the overall covariance structure of the data and by outliers or heavy tails; see Huber (1985, sec. 4). Friedman (1987) stressed that a very reliable and thorough numerical optimizer was absolutely essential for finding "substantive" maxima of $I(f)$, since sampling fluctuations tend to trap ineffective optimizers within a multitude of local maxima. If $\{z_i\}$ are the projected data, then (9.3) is estimated by $\hat{I}(f) = \int J(\hat{f}(z))\,d\hat{F}_n(z) = (1/n)\sum_{i=1}^n J(\hat{f}(z_i))$. Thus if $J(f(z)) = f(z)$, then $I(f) = \int [f(z)]^2\,dz$ can be estimated by $\hat{I}(f) = (1/n)\sum_{i=1}^n \hat{f}_h(z_i)$, where $\hat{f}_h$ is a kernel estimate with window width h; see Friedman and Tukey (1974) and Tukey and Tukey (1981). Another choice is to take $J(f(z)) = \log f(z)$, so that $I(f) = \int f(z)\log f(z)\,dz$, which is (negative) cross-entropy, and (9.3) can be estimated at the kth iteration by $(1/n)\sum_{i=1}^n \log \hat{f}^{(k)}(z_i)$; see Friedman et al. (1984). Joe (1987) discussed kernel estimation of functionals such as (9.3) and showed that, for moderate-sized samples, statistical properties of $\hat{I}$ were improved either through bias corrections or by using a rescaled kernel.

Other projection indexes that have also been used include a moment index based on the sum of squares of the third and fourth sample cumulants of the projected data (Jones and Sibson 1987), and the ISE criterion (Friedman 1987; Hall 1989a). The latter approaches, though related, differed on whether or not to first transform the projected data. Friedman used ISE between the transformed projected data density and the uniform density, while Hall's version used ISE between the untransformed projected data density and the standard normal. Both Friedman and Hall used orthogonal series density estimators (Legendre polynomials and Hermite functions, respectively) to study their projection indexes.

Each of these indexes was designed to search for deviations from "uninterestingness," whose definition depended on the application in question. Thus, the Friedman–Tukey index searched for evidence of "clottedness" as well as departures from a parabolic density; the entropy index searched for departures of the projected data from normality since the normal distribution maximizes entropy; and the moment index and ISE criteria also set up normality as the least interesting data feature. Other indexes are also being studied for specific applications.

10. RELATED TOPICS

Functionals of a Density. Examples of functionals, $\alpha(F)$, say, of the distribution function F associated with a density f include the quantile function F^{-1}, the hazard function $\lambda = f/(1 - F)$, any L_p-norm of the derivatives of f, Shannon negative entropy $\int f \log f$, and Fisher information $\int (f')^2/f$. Certain of these are used as projection indexes in PPDE. Typically, "plug-in" estimators of the form $\alpha(\hat{F})$ are used, where $\hat{F}$ is taken to be a smoothed version of $\hat{F}_n$. Note that estimating F using the kernel method requires less smoothing than that best suited for estimating f. Kernel estimation of the hazard rate was discussed by Singpurwalla and Wong (1983) and Hassani, Sarda, and Vieu (1986), and that of the quantile function $\xi_p = F^{-1}(p)$, $0 < p < 1$, by Parzen (1979), Falk (1984), and Sheather and Marron (1988). The bootstrap and its smoothed versions have been used to estimate $\alpha(F)$ directly, especially for kernel quantile estimation. See Silverman and Young (1987), Yang (1985), Hall, Diciccio, and Romano (1989), and Hall (1990). Note, however, that bootstrap smoothing using a non-bona fide kernel density estimator of a nonnegative quantity, such as a probability or a variance, can make a nonnegative estimate negative.

Assessing Multimodality. Integer-valued nonlinear functionals of f, such as the number of mixture components needed to represent f, and the number of modes of f, are also of interest, and different nonparametric approaches to determining the values of such functionals have been considered. Donoho (1988) developed a general theory for determining nonparametric lower bounds on such functionals. Good and Gaskins (1980) used the MPL method together with certain "bump hunting" surgical techniques to assess the existence of any "real" dips and bumps in mass spectra obtained from scattering experiments. Silverman (1981b, 1983) used the kernel method together with the smoothed bootstrap procedure to develop a confirmatory test of the most probable number of modes in a density; see Silverman (1986, sec. 6.6) and Izenman and Sommer (1988).

Robust Estimation. Nonparametric density estimation has been used to obtain robust estimators for parametric inference. The main tool has been the use of Hellinger distance between two probability densities f and g, namely,

$$HD(f, g) = \frac{1}{2}\int_{-\infty}^{\infty} ([f(x)]^{1/2} - [g(x)]^{1/2})^2 \, dx. \quad (10.1)$$

The minimum Hellinger distance (MHD) estimator is that value $\hat{\theta}$ of θ that minimizes $HD(\hat{f}, f_\theta)$, where $\hat{f}$ is a nonparametric density estimator of f and f_θ, $\theta \in \Theta$, is a member of some parametric family. The distance HD is always finite and is invariant under strictly monotone transformations. Beran (1977a,b) Birge (1986), Tamura and Boos (1986), and Simpson (1987, 1989) proved asymptotic results and established impressive robustness properties of MHD location estimators based on the kernel density estimator. For related work on minimum distance estimators of densities, see Reiss (1976) and Birge (1983).

Semiparametric Models. Olkin and Spiegelman (1987) developed an approach to density estimation that combined parametric and nonparametric approaches. Their density estimator was given by

$$\tilde{f}_\pi(x) = \pi f_{\hat{\theta}}(x) + (1 - \pi)\hat{f}(x), \quad (10.2)$$

where $f_{\hat{\theta}}$ is a ML parametric estimator of f, $\hat{f}$ is a kernel estimator of f, and $0 \le \pi \le 1$ is unknown. The parameter π was chosen to minimize the Hellinger distance, $HD(\tilde{f}_\pi, f)$, and asymptotic results were obtained under regularity conditions on f. Figure 6 shows the semiparametric density estimate constructed from annual wind speed measurements from Olkin and Spiegelman. For that example, the parametric model appeared to be appropriate.

Directional Data. In astronomy, geology, and studies of animal behavior, it is often of interest to estimate the

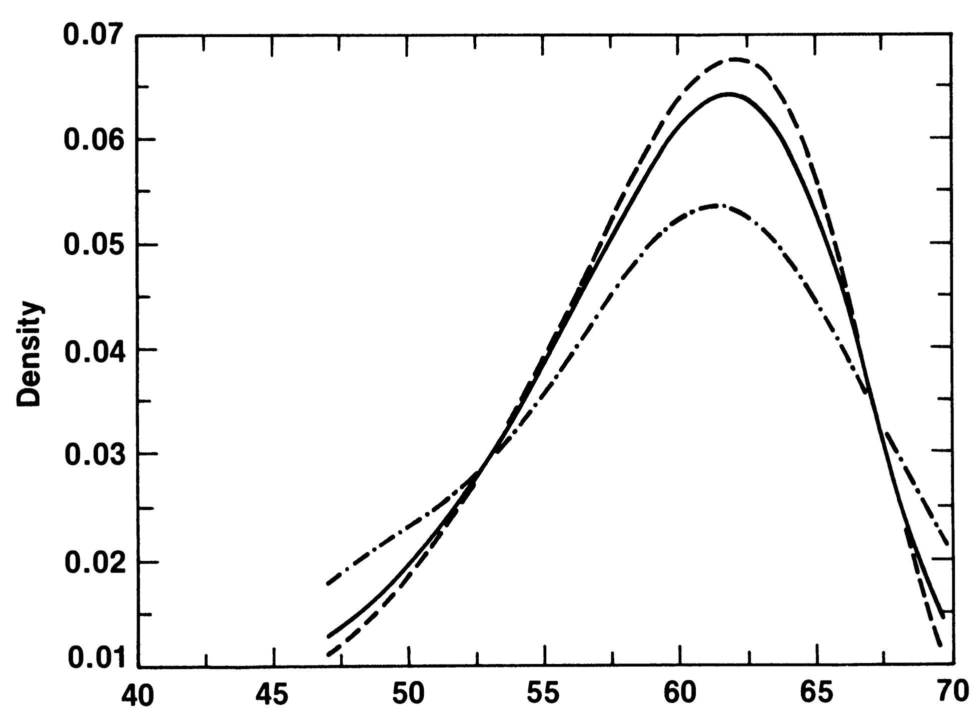

Figure 6. Density Estimates for 20 Measurements on Annual Maximum Wind Speeds in the N. Direction Taken in Sheridan, Wyoming, During 1958–1977. Reproduced from Olkin and Spiegelman (1987). The dotted-and-dashed line shows the kernel density estimate with smoothing parameter h = .7s, where s is the sample standard deviation; the dashed line shows the parametric density estimate; and the solid line shows the semiparametric density estimate with estimated weight $\hat{\pi}$ = .8.

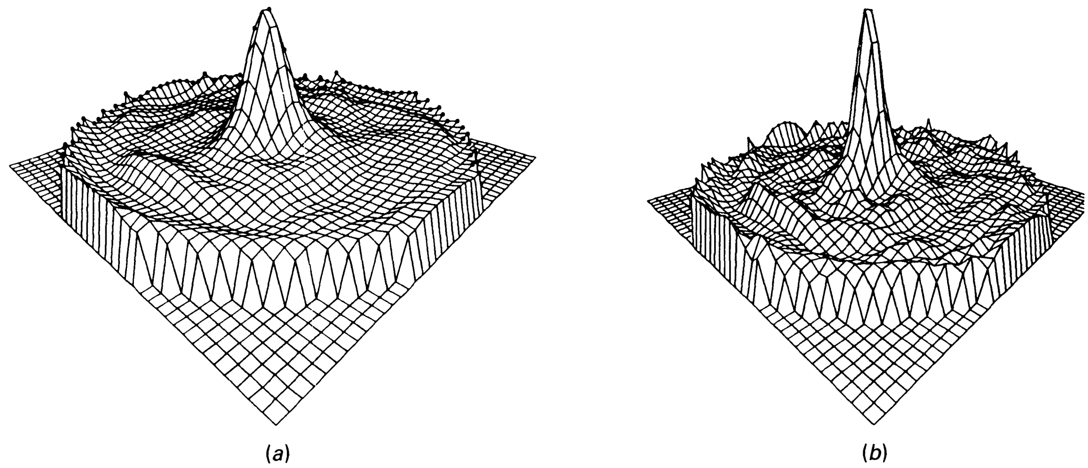

Figure 7. Perspective Plots for 685 Measurements on the Orbits of all Known Comets. Reproduced from Hall, Watson, and Cabrera (1987). Smoothing was obtained by (a) likelihood cross-validation, and (b) least squares cross-validation. Notice that likelihood CV produces a smoother density estimate having lower peaks than least squares CV. With permission of the Biometrika trustees.

density f of measurements, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, observed on the surface of a d-dimensional unit sphere S_d, $d \geq 2$. Kernel density estimators for such "directional data" have the forms

$$\hat{f}_{\kappa,K_1}(\mathbf{x}) = n^{-1}c(\kappa) \sum_{i=1}^{n} K_1(\kappa \mathbf{x}^T \mathbf{X}_i), \quad (10.3)$$

$$\hat{f}_{\kappa,K_2}(\mathbf{x}) = n^{-1}d(\kappa) \sum_{i=1}^{n} K_2(\kappa(1 - \mathbf{x}^T \mathbf{X}_i)), \quad (10.4)$$

where K_1 and K_2 are known kernel functions typically defined on $[0, \infty)$, $\kappa > 0$ is an unknown smoothing parameter, $c(\kappa)$ and $d(\kappa)$ are positive numbers, and $\mathbf{x} \in S_d$. Asymptotic properties of (10.3) and (10.4) were studied by Hall, Watson, and Cabrera (1987) and Bai, Rao, and Zhao (1988). For a discussion of the related problem of nonparametric density estimation on Riemannian manifolds using Fourier transform methods, see Hendriks (1990). As an example, three-dimensional perspective plots of kernel density estimators of different cometary orbits regarded as directional data are given in Figure 7 using likelihood and least squares cross-validation for determining the smoothing parameter.

Censored Data. Often, in biomedical and industrial studies, censored survival or lifetime data are recorded, and it is of interest to estimate density and hazard functions for such data. Padgett and McNichols (1984) provided an excellent survey paper on this topic. Since then, the kernel (Marron and Padgett 1987), nearest-neighbor (Mielniczuk 1986), and penalized likelihood (Lubecke and Padgett 1985) methods have been used to obtain nonparametric estimates of the density f in the presence of censored data. The hazard function (intensity function, failure rate) was estimated for censored data by the kernel method (Blum and Susarla 1980; Liu and Van Ryzin 1985; Schafer 1985; Tanner 1983; Tanner and Wong 1983; Yandell 1983) and by the MPL method (Anderson and Senthilselvan 1980; Bartoszynski, Brown, McBride, and Thompson 1981).

Incomplete Data. Kernel density estimation from incomplete data was considered by Titterington and Mill (1983).

Time Series Data. For dependent observations generated by a strictly stationary process, kernel density estimators were studied by Roussas (1969), Rosenblatt (1970, 1971), Nguyen (1979), and Hart (1984), recursive density estimators were studied by Masry (1986, 1989) and Masry and Gyorfi (1987), and survival function and hazard rate estimators were studied by Roussas (1989, 1990) and Izenman and Tran (1990).

REFERENCES

Abramson, I. S. (1982a), "On Bandwidth Variation in Kernel Estimates—A Square Root Law," *The Annals of Statistics,* 10, 1217–1223.

——— (1982b), "Arbitrariness of the Pilot Estimate in Adaptive Kernel Methods," *Journal of Multivariate Analysis,* 12, 562–567.

——— (1984), "Adaptive Density Flattening—A Metric Distortion Principle for Combating Bias in Nearest Neighbor Methods," *The Annals of Statistics,* 12, 880–886.

Aitchison, J., and Lauder, I. J. (1985), "Kernel Density Estimation for Compositional Data," *Applied Statistics,* 34, 129–137.

Anderson, G. L., and de Figueiredo, R. J. P. (1980), "An Adaptive Orthogonal-Series Estimator for Probability Density Functions," *The Annals of Statistics,* 8, 347–376.

Anderson, J. A., and Senthilselvan, A. (1980), "Smooth Estimates for the Hazard Function," *Journal of the Royal Statistical Society,* Ser. B, 42, 322–327.

Anderson, T. W. (1966), "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," in *Multivariate Analysis: Proceedings of an International Symposium,* ed. P. R. Krishnaiah, New York: Academic Press, pp. 5–27.

Bai, Z. D., Rao, C. R., and Zhao, L. C. (1988), "Kernel Estimators of Density Function of Directional Data," *Journal of Multivariate Analysis,* 27, 24–39.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference Under Order Restrictions,* New York: John Wiley.

Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. (1981), "Some Nonparametric Techniques for Estimating the Intensity Function of a Cancer Related Nonstationary Poisson Process," *The Annals of Statistics,* 9, 1050–1060.

Bean, S. J., and Tsokos, C. P. (1980), "Developments in Nonparametric Density Estimation," *International Statistical Review,* 48, 267–287.

Beran, R. (1977a), "Robust Location Estimates," *The Annals of Statistics,* 5, 431–444.

——— (1977b), "Minimum Hellinger Distance Estimates for Parametric Models," *The Annals of Statistics,* 5, 445–463.

Birge, L. (1983), "Approximation Dans Les Espaces Metriques et Theorie de l'Estimation," (in French), *Zeitschrift fur Wahrkeinlichkeitstheorie und verwandte Gebeite,* 65, 181–237.

——— (1986), "On Estimating a Density Using Hellinger Distance and

Some Other Strange Facts," *Probability Theory and Related Fields*, 71, 271–291.
——— (1987a), "Estimating a Density Under Order Restrictions: Nonasymptotic Minimax Risk," *The Annals of Statistics*, 15, 995–1012.
——— (1987b), "On the Risk of Histograms for Estimating Decreasing Densities," *The Annals of Statistics*, 15, 1013–1022.
——— (1989), "The Grenander Estimator: A Nonasymptotic Approach," *The Annals of Statistics*, 17, 1532–1549.
Blum, J. R., and Susarla, V. (1980), "Maximal Deviation Theory of Density and Failure Rate Function Estimates Based on Censored Data," in *Multivariate Analysis V*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 213–222.
Boneva, L. I., Kendall, D. G., and Stefanov, I. (1971), "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 33, 1–70.
Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.
Boyd, D. W., and Steele, J. M. (1978), "Lower Bounds for Nonparametric Density Estimation Rates," *The Annals of Statistics*, 6, 932–934.
Breiman, L., Meisel, W., and Purcell, E. (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, 19, 135–144.
Broniatowski, M., Deheuvels, P., and Devroye, L. (1989), "On the Relationship Between Stability of Extreme Order Statistics and Convergence of the Maximum Likelihood Kernel Density Estimate," *The Annals of Statistics*, 17, 1070–1086.
Brunk, H. B. (1978), "Univariate Density Estimation by Orthogonal Series," *Biometrika*, 65, 521–528.
Butler, W. J., and Kronmal, R. A. (1985), "Discrimination with Polychotomous Predictor Variables Using Orthogonal Functions," *Journal of the American Statistical Association*, 80, 443–448.
Cacoullos, T. (1966), "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics*, 18, 178–189.
Carroll, R. J. (1976), "On Sequential Density Estimation," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebeite*, 36, 136–151.
Cencov, N. N. (1962), "Evaluation of an Unknown Distribution Density From Observations," *Soviet Mathematics*, 3, 1559–1562.
Chow, Y. S., Geman, S., and Wu, L. D. (1983), "Consistent Cross-Validated Density Estimation," *The Annals of Statistics*, 11, 25–38.
Cline, D. (1988), "Admissible Kernel Estimators of a Multivariate Density," *The Annals of Statistics*, 16, 1421–1427.
Crain, B. R. (1973), "A Note on Density Estimation Using Orthogonal Expansions," *The Annals of Statistics*, 2, 454–463.
Davies, H. I., and Wegman, E. J. (1975), "Sequential Nonparametric Density Estimation," *IEEE Transactions on Information Theory*, 21, 619–628.
Deheuvels, P. (1973), "Sur l'Estimation Sequentielle de la Densite," *Comptes Rendus de l'Academie des Sciences de Paris*, 276, 1119–1121.
——— (1977), "Estimation nonparametrique de la densite par histogrammes generalises," *Revue de Statistique Appliquée*, 25/3, 5–42.
De Jager, O. C., Swanepoel, J. W. H., and Raubenheimer, B. C. (1986), "Kernel Density Estimators Applied to Gamma Ray Light Curves," *Astronomy and Astrophysics*, 170, 187–196.
De Montricher, G. M., Tapia, R. A., and Thompson, J. R. (1975), "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *The Annals of Statistics*, 3, 1329–1348.
Denby, L., and Vardi, Y. (1986), "The Survival Curve With Decreasing Density," *Technometrics*, 28, 359–367.
Devijver, P. A., and Kittler, J. (1982), *Pattern Recognition: A Statistical Approach*, London: Prentice-Hall.
Devroye, L. (1979), "On the Pointwise and Integral Convergence of Recursive Kernel Estimates of Probability Densities," *Utilitas Mathematica*, 15, 113–128.
——— (1983), "The Equivalence of Weak, Strong, and Complete Convergence in L_1 For Kernel Density Estimates," *The Annals of Statistics*, 11, 896–904.
——— (1985), "A Note on the L_1 Consistency of Variable Kernel Estimates," *The Annals of Statistics*, 13, 1041–1049.
——— (1987), *A Course in Density Estimation*, Boston: Birkhauser.
Devroye, L., and Gyorfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*. New York: John Wiley.
Devroye, L., and Penrod, C. S. (1984), "The Consistency of Automatic Kernel Density Estimates," *The Annals of Statistics*, 12, 1231–1249.
——— (1986), "The Strong Uniform Convergence of Multivariate Variable Kernel Estimates," *The Canadian Journal of Statistics*, 14, 211–219.
Diggle, P. J., and Hall, P. (1986), "The Selection of Terms in an Orthogonal Series Density Estimator," *Journal of the American Statistical Association*, 81, 230–233.
Donoho, D. L. (1988), "One-Sided Inference About Functionals of a Density," *The Annals of Statistics*, 16, 1390–1420.
Donoho, D. L., and Johnstone, I. M. (1989), "Projection-Based Approximation and a Duality with Kernel Methods," *The Annals of Statistics*, 17, 58–106.
Dubuisson, B., and Lavison, P. (1980), "Surveillance of a Nuclear Reactor by Use of a Pattern Recognition Methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, 10, 603–609.
Emerson, J. D., and Hoaglin, D. C. (1983), "Stem and Leaf Displays," in *Understanding Robust and Exploratory Data Analysis*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: John Wiley.
Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.
Falk, M. (1984), "Relative Deficiency of Kernel Type Estimators of Quantiles," *The Annals of Statistics*, 12, 261–268.
Farrell, R. H. (1972), "On the Best Obtainable Asymptotic Rates of Convergence in Estimation of a Density Function at a Point," *The Annals of Mathematical Statistics*, 43, 170–180.
Fix, E., and Hodges, J. L. (1951), "Discriminatory Analysis, Nonparametric Estimation: Consistency Properties," *Report No. 4, Project No. 21-49-004*, Randolph Field, Texas: USAF School of Aviation Medicine.
Foldes, A., and Revesz, P. (1974), "A General Method for Density Estimation," *Studia Scientiarum Mathematicarum Hungarica*, 9, 81–92.
Fraser, D. A. S. (1951), "Sequentially Determined Statistically Equivalent Blocks," *The Annals of Mathematical Statistics*, 22, 372–381.
——— (1953), "Nonparametric Tolerance Regions," *The Annals of Mathematical Statistics*, 24, 44–55.
——— (1957), *Nonparametric Methods in Statistics*, New York: John Wiley.
Fraser, D. A. S., and Guttman, I. (1956), "Tolerance Regions," *The Annals of Mathematical Statistics*, 27, 162–179.
Freedman, D., and Diaconis, P. (1981a), "On the Maximum Deviation Between the Histogram and the Underlying Density," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58, 139–167.
——— (1981b), "On the Histogram as a Density Estimator: L_2 Theory," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebeite*, 57, 453–476.
Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.
Friedman, J. H., and Stuetzle, W. (1982), "Projection Pursuit Methods for Data Analysis," in *Modern Data Analysis*, eds. R. L. Launer and A. F. Siegel, New York: Academic Press, pp. 123–147.
Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984), "Projection Pursuit Density Estimation," *Journal of the American Statistical Association*, 79, 599–608.
Friedman, J. H., and Tukey, J. W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computing*, 23, 881–890.
Fryer, M. J. (1976), "Some Errors Associated With the Nonparametric Estimation of Density Functions," *Journal of the Institute of Mathematics and its Applications*, 18, 371–380.
——— (1977), "A Review of Some Nonparametric Methods of Density Estimation," *Journal of the Institute of Mathematics and Its Applications*, 20, 335–354.
Fukunaga, K. (1972), *Introduction to Statistical Pattern Recognition*. London: Academic Press.
Gajek, L. (1986), "On Improving Density Estimators Which Are Not Bona Fide Functions," *The Annals of Statistics*, 14, 1612–1618.
Gasser, T., Muller, H.-G., and Mammitzsch, V. (1985), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society*, Ser. B, 47, 238–252.
Gawronski, W., and Stadtmuller, U. (1980), "On Density Estimation by Means of Poisson's Distribution," *Scandinavian Journal of Statistics*, 7, 90–94.
Geman, S., and Hwang, C.-R. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 10, 401–414.
Gessaman, M. P. (1970), "A Consistent Nonparametric Multivariate Density Estimator Based on Statistically Equivalent Blocks," *The Annals of Mathematical Statistics*, 41, 1344–1346.
Gessaman, M. P., and Gessaman, P. H. (1972), "A Comparison of Some Multivariate Discrimination Procedures," *Journal of the American Statistical Association*, 67, 468–472.
Ghurye, S. G. and Olkin, I. (1969), "Unbiased Estimation of Some Mul-

tivariate Probability Densities and Related Functions," *The Annals of Mathematical Statistics,* 40, 1261–1271.

Good, I. J., and Deaton, M. L. (1981), "Recent Advances in Bump Hunting," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface,* ed. W. F. Eddy, New York: Springer-Verlag, pp. 92–104.

Good, I. J., and Gaskins, R. A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika,* 58, 255–277.

——— (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data" (with discussion), *Journal of the American Statistical Association,* 75, 42–73.

Greblicki, W., and Pawlak, M. (1981), "Classification Using the Fourier Series Estimate of Multivariate Density Functions," *IEEE Transactions on Systems, Man, and Cybernetics,* 11, 726–730.

Grenander, U. (1956), "On the Theory of Mortality Measurement. Part II," *Skandinavisk Aktuarietidskrift,* 39, 125–153.

——— (1981), *Abstract Inference,* New York: John Wiley.

Groeneboom, P. (1983), "Estimating a Monotone Density," *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer,* eds. L. M. LeCam and R. A. Olshen, 2, 539–555. Belmont, CA: Wadsworth.

Hall, P. (1981), "On Trigonometric Series Estimates of Densities," *The Annals of Statistics,* 9, 683–685.

——— (1982), "Cross-Validation in Density Estimation," *Biometrika,* 69, 383–390.

——— (1983a), "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics,* 11, 1156–1174.

——— (1983b), "Orthogonal Series Methods for Both Qualitative and Quantitative Data," *The Annals of Statistics,* 11, 1004–1007.

——— (1986), "On the Rate of Convergence of Orthogonal Series Density Estimators," *Journal of the Royal Statistical Society,* Ser. B, 48, 115–122.

——— (1987a), "On Kullback–Leibler Loss and Density Estimation," *The Annals of Statistics,* 15, 1491–1519.

——— (1987b), "Cross-Validation and the Smoothing of Orthogonal Series Density Estimators," *Journal of Multivariate Analysis,* 21, 189–206.

——— (1989a), "On Polynomial-Based Projection Indices for Exploratory Projection Pursuit," *The Annals of Statistics,* 17, 589–605.

——— (1989b), "On Convergence Rates in Nonparametric Problems," *International Statistical Review,* 57, 45–58.

——— (1990), "Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems," *Journal of Multivariate Analysis,* 32, 177–203.

Hall, P., Diciccio, T. J., and Romano, J. P. (1989), "On Smoothing and the Bootstrap," *The Annals of Statistics,* 17, 692–704.

Hall, P., and Hannan, E. J. (1988), "On Stochastic Complexity and Nonparametric Density Estimation," *Biometrika,* 75, 705–714.

Hall, P., and Marron, J. S. (1987a), "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Probability Theory and Related Fields,* 74, 567–581.

——— (1987b), "On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator," *The Annals of Statistics,* 15, 163–181.

——— (1988), "Choice of Kernel Order in Density Estimation," *The Annals of Statistics,* 16, 161–173.

Hall, P., and Wand, M. P. (1988), "Minimizing L_1 Distance in Nonparametric Density Estimation," *Journal of Multivariate Analysis,* 26, 59–88.

Hall, P., Watson, G. S., and Cabrera, J. (1987), "Kernel Density Estimation With Spherical Data," *Biometrika,* 74, 751–762.

Hand, D. J. (1982), *Kernel Discriminant Analysis.* Chichester, U.K.: Research Studies Press.

Hart, J. D. (1984), "Efficiency of a Kernel Density Estimator Under an Autoregressive Dependence Model," *Journal of the American Statistical Association,* 79, 110–117.

——— (1985), "On the Choice of Truncation Point in Fourier Series Density Estimation," *Journal of Statistical Computation and Simulation,* 21, 95–116.

Hassani, S., Sarda, P., and Vieu, P. (1986), "Nonparametric Approaches to Hazard Functions: Bibliographical Review (in French)," *Revue de Statistique Appliquée,* 34/4, 27–42.

Hendriks, H. (1990), "Nonparametric Estimation of a Probability Density on a Riemannian Manifold Using Fourier Expansions," *The Annals of Statistics,* 18, 832–849.

Huber, P. J. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics,* 13, 435–525.

International Mathematical and Statistical Libraries, Inc. (1987), *STAT/LIBRARY* (Version 1.0), Houston, TX: Author.

Izenman, A. J., and Sommer, C. J. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association,* 83, 941–953.

Izenman, A. J., and Tran, L. T. (1990), "Kernel Estimation of the Survival Function and Hazard Rate Under Weak Dependence," *Journal of Statistical Planning and Inference,* 24, 233–247.

Jee, J. R. (1987), "Exploratory Projection Pursuit Using Nonparametric Density Estimation," *Proceedings of the Statistical Computing Section of the American Statistical Association,* 335–339.

Joe, H (1987), "Estimation of Entropy and Other Functionals of a Multivariate Density," Technical Report, University of British Columbia.

Johnstone, I. M., and Silverman, B. W. (1990), "Speed of Estimation in Positron Emission Tomography and Related Inverse Problems," *The Annals of Statistics,* 18, 251–280.

Jones, M. C. (1989), "Discretized and Interpolated Kernel Density Estimates," *Journal of the American Statistical Association,* 84, 733–741.

Jones, M. C., and Lotwick, H. W. (1984), "A Remark on Algorithm AS 176. Kernel Density Estimation Using the Fast Fourier Transform," *Applied Statistics,* 33, 120–122.

Jones, M. C. and Sibson, R. (1987), "What is Projection Pursuit?" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 150, 1–36.

Kanazawa, Y. (1988), "An Optimal Variable Cell Histogram," *Communications in Statistics,* 17, 1401–1422.

Kasser, I. S., and Bruce, R. A. (1969), "Comparative Effects of Aging and Coronary Heart Disease on Submaximal and Maximal Exercise," *Circulation,* 39, 759–774.

Klonius, V. K. (1982), "Consistency of Two Nonparametric Maximum Penalized Estimators of the Probability Density Function," *The Annals of Statistics,* 10, 811–824.

——— (1984), "On a Class of Nonparametric Density and Regression Estimators," *The Annals of Statistics,* 12, 1263–1284.

Klonius, V. K., and Nash, S. G. (1983), "On the Computation of a Class of Maximum Penalized Likelihood Estimators of the Probability Density Function," in *Computer Science and Statistics: The Interface,* ed. J. E. Gentle, Amsterdam: North-Holland, pp. 310–314.

Kogure, A. (1987), "Asymptotically Optimal Cells for a Histogram," *The Annals of Statistics,* 15, 1023–1030.

Kronmal, R. and Tarter, M. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," *Journal of the American Statistical Association,* 63, 925–952.

——— (1973), "The Use of Density Estimates Based on Orthogonal Expansions," in *Exploring Data Analysis: The Computer Revolution in Statistics,* eds. W. J. Dixon and W. L. Nicholson, Los Angeles: University of California Press, pp. 365–395.

Lecoutre, J.-P. (1986), "The Histogram with Random Partition," in *New Perspectives in Theoretical and Applied Statistics,* eds. M. L. Puri, J. P. Vilaplana, and W. Wertz, New York: John Wiley, pp. 265–276.

Leonard, T. (1978), "Density Estimation, Stochastic Processes, and Prior Information" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 40, 113–146.

Liu, R. Y. C., and Van Ryzin, J. (1985), "A Histogram Estimator of the Hazard Rate With Censored Data," *The Annals of Statistics,* 13, 592–605.

Lock, M. D. (1990), "Optimizing Density Estimates Based On Unweighted and Weighted Mean Integrated Squared Error," unpublished Ph.D. dissertation, University of California, Berkeley, Group in Biostatistics.

Loftsgaarden, D. O., and Quesenberry, C. P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *The Annals of Mathematical Statistics,* 36, 1049–1051.

Lubecke, A. M., and Padgett, W. J. (1985), "Nonparametric Maximum Penalized Likelihood Estimation of a Density from Arbitrarily Right-Censored Observations," *Communications in Statistics, Part A—Theory and Methods,* 14, 257–271 (corrigendum, p. 2007).

Mack, Y. P. (1980), "Asymptotic Normality of Multivariate K-NN Density Estimates," *Sankhyā,* Ser. A, 42, 53–63.

Mack, Y. P., and Rosenblatt, M. (1979), "Multivariate K-Nearest Neighbor Density Estimates," *Journal of Multivariate Analysis,* 9, 1–15.

Marron, J. S. (1985), "An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation," *The Annals of Statistics,* 13, 1011–1023.

——— (1987a), "A Comparison of Cross-Validation Techniques in Den-

sity Estimation," *The Annals of Statistics*, 15, 152–162.
——— (1987b), "Automatic Smoothing Parameter Selection: A Survey," *Empirical Economics*, 13, 187–208.
Marron, J. S., and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91–113.
Marron, J. S., and Nolan, D. (1987), "Canonical Kernels for Density Estimation," Technical Report, University of North Carolina, Chapel Hill.
Marron, J. S., and Padgett, W. J. (1987), "Asymptotically Optimal Bandwidth Selection for Kernel Density Estimators from Randomly Right-Censored Samples," *The Annals of Statistics*, 15, 1520–1535.
Masry, E. (1986), "Recursive Probability Density Estimation for Weakly Dependent Stationary Processes," *IEEE Transactions on Information Theory*, 32, 254–267.
——— (1989), "Nonparametric Estimation of Conditional Probability Densities and Expectations of Stationary Processes: Strong Consistency and Rates," *Stochastic Processes and Their Applications*, 32, 109–128.
Masry, E., and Gyorfi, L. (1987), "Strong Consistency and Rates for Recursive Probability Density Estimators of Stationary Processes," *Journal of Multivariate Analysis*, 22, 79–93.
Mielniczuk, J. (1986), "Some Asymptotic Properties of Kernel Estimators in Case of Censored Data," *The Annals of Statistics*, 14, 766–773.
Moore, D. S., and Yackel, J. W. (1977), "Consistency Properties of Nearest Neighbor Density Function Estimators," *The Annals of Statistics*, 5, 143–154.
Muller, H.-G. (1988), *Nonparametric Regression Analysis of Longitudinal Data* (Springer Lecture Notes in Statistics), New York: Springer-Verlag pp. 295–298.
Nadarya, E. A. (1989), *Nonparametric Estimation of Probability Densities and Regression Curves*. Dordrecht, Neth.: Kluwer Academic Publishers.
Nguyen, H. T. (1979), "Density Estimation in a Continuous-Time Stationary Markov Process," *The Annals of Statistics*, 7, 341–348.
Olkin, I., and Spiegelman, C. H. (1987), "A Semiparametric Approach to Density Estimation," *Journal of the American Statistical Association*, 82, 858–865.
O'Sullivan, F. (1986), "A Statistical Perspective on Ill-Posed Inverse Problems," *Statistical Science*, 1, 502–527.
Ott, J., and Kronmal, R. A. (1976), "Some Classification Procedures for Multivariate Binary Data Using Orthogonal Functions," *Journal of the American Statistical Association*, 71, 391–399.
Padgett, W. J., and McNichols, D. T. (1984), "Nonparametric Density Estimation From Censored Data," *Communications in Statistics—Theory and Methods*, 13, 1581–1611.
Park, B. U., and Marron, J. S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistial Association*, 85, 66–72.
Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065–1076.
——— (1979), "Nonparametric Statistical Data Modeling" (with discussion), *Journal of the American Statistical Association*, 74, 105–131.
Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*. New York: Academic Press.
Quesenberry, C. P., and Gessaman, M. P. (1968), "Nonparametric Discrimination Using Tolerance Regions," *The Annals of Mathematical Statistics*, 39, 664–673.
Reiss, R.-D. (1976), "On Minimum Distance Estimators for Unimodal Densities," *Metrika*, 23, 7–14.
Robertson, T. (1967), "On Estimating a Density Which is Measurable With Respect to a σ Lattice," *The Annals of Mathematical Statistics*, 33, 482–493.
Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: John Wiley.
Rodriguez, C. C., and van Ryzin, J. (1985), "Maximum Entropy Histograms," *Statistics and Probability Letters*, 3, 117–120.
Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, 27, 832–837.
——— (1970), "Density Estimates and Markov Sequences," in *Nonparametric Techniques in Statistical Inference*, ed. M. L. Puri, Cambridge, U.K.: Cambridge University Press, pp. 199–210.
——— (1971), "Curve Estimates," *The Annals of Mathematical Statistics*, 42, 1815–1842.
——— (1979), "Global Measures of Deviation for Kernel and Nearest Neighbor Density Estimates," in *Smoothing Techniques for Curve Estimation* (Lecture Notes in Mathematics No. 757), eds. T. Gasser and M. Rosenblatt, Berlin: Springer-Verlag, pp. 181–190.
Roussas, G. (1969), "Nonparametric Estimation of the Transition Distribution Function of a Markov Process," *The Annals of Mathematical Statistics*, 40, 1386–1400.
——— (1989), "Hazard Rate Estimation Under Dependence Conditions," *Journal of Statistical Planning and Inference*, 22, 81–93.
——— (1990), "Asymptotic Normality of the Kernel Estimate Under Dependence Conditions: Application to Hazard Rate," *Journal of Statistical Planning and Inference*, 25, 81–104.
Sager, T. W. (1982), "Nonparametric Maximum Likelihood Estimation of Spatial Patterns," *The Annals of Statistics*, 10, 1125–1136.
——— (1986), "An Application of Isotonic Regression to Multivariate Density Estimation," in *Advances in Order Restricted Statistical Inference* (Springer Lecture Notes in Statistics, Vol. 37), eds. R. Dykstra, T. Robertson, and F. T. Wright, New York: Springer-Verlag, pp. 69–90.
Schafer, H. (1985), "A Note on Data-Adaptive Kernel Estimation of the Hazard and Density Function in the Random Censorship Situation," *The Annals of Statistics*, 13, 818–820.
Schuster, E. F., and Gregory, C. G. (1981), "On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimators," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, New York: Springer-Verlag pp. 295–298.
Schwartz, S. C. (1967), "Estimation of Probability Density by an Orthogonal Series," *The Annals of Mathematical Statistics*, 38, 1261–1265.
Scott, D. W. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610.
——— (1985), "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *The Annals of Statistics*, 13, 1024–1040.
——— (1985b), "Frequency Polygons," *Journal of the American Statistical Association*, 80, 348–354.
——— (1988), "A Note on Choice of Bivariate Histogram Bin Shape," *Journal of Official Statistics*, 4, 47–51.
Scott, D. W., and Factor, L. E. (1981), "Monte Carlo Study of Three Data-Based Nonparametric Density Estimators," *Journal of the American Statistical Association*, 76, 9–15.
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., Tapia, R. A., and Thompson, J. R. (1980), "Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria," *The Annals of Statistics*, 8, 820–832.
Scott, D. W., and Terrell, G. R. (1987), "Biased and Unbiased Cross-Validation in Density Estimation," *Journal of the American Statistical Association*, 82, 1131–1146.
Scott, D. W., and Thompson, J. R. (1983), "Probability Density Estimation in Higher Dimensions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, ed. J. E. Gentle, Amsterdam: North-Holland, pp. 173–179.
Sheather, S. J., and Marron, J. S. (1988), "Kernel Quantile Estimators," Working Paper 88-012, Australian Graduate School of Management, The University of New South Wales, Australia.
Silverman, B. W. (1978c), "Density Ratios, Empirical Likelihood, and Cot Death," *Applied Statistics*, 27, 26–33.
——— (1981a), "Density Estimation for Univariate and Bivariate Data," in *Interpreting Multivariate Data*, ed. V. Barnett, New York: John Wiley, Ch. 3, pp. 37–53.
——— (1981b), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Ser. B, 43, 97–99.
——— (1982a), "Algorithm AS 176. Kernel Density Estimation Using the Fast Fourier Transform," *Applied Statistics*, 31, 93–97.
——— (1982b), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, 10, 795–810.
——— (1983), "Some Properties of a Test for Multimodality Based on Kernel Density Estimates," in *Probability, Statistics, and Analysis*, eds. J. F. C. Kingman and G. E. H. Reuter, 248–259, Cambridge: Cambridge University Press.
——— (1984), "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916.
——— (1985), "Two Books on Density Estimation," *The Annals of Statistics*, 13, 1630–1638.

——— (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Silverman, B. W., and Jones, M. C. (1988), "E. Fix and J. L. Hodges (1951): An Important Unpublished Contribution to Nonparametric Discriminant Analysis and Density Estimation," Technical Report, University of Bath.

Silverman, B. W., and Young, G. A. (1987), "The Bootstrap: To Smooth or Not To Smooth?" *Biometrika*, 74, 469–479.

Simpson, D. G. (1987), "Minimum Hellinger Distance Estimation for the Analysis of Count Data," *Journal of the American Statistical Association*, 82, 802–807.

——— (1989), "Hellinger Deviance Tests: Efficiency, Breakdown Points, and Examples," *Journal of the American Statistical Association*, 84, 107–113.

Singpurwalla, N. D., and Wong, M.-Y. (1983), "Estimation of the Failure Rate—A Survey of Nonparametric Methods, Part I: Non-Bayesian Methods," *Communications in Statistics, Part A—Theory and Methods*, 12, 559–588.

Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, 12, 1285–1297.

Tamura, R. N., and Boos, D. D. (1986), "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance," *Journal of the American Statistical Association*, 81, 223–229.

Tanner, M. A. (1983), "A Note on the Variable Kernel Estimator of the Hazard Function from Randomly Censored Data," *The Annals of Statistics*, 11, 994–998.

Tanner, M. A., and Wong, W. H. (1983), "The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method," *The Annals of Statistics*, 11, 989–993.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, Baltimore, MD: Johns Hopkins University Press.

Tarter, M. E., and Kronmal, R. A. (1976), "An Introduction to the Implementation and Theory of Nonparametric Density Estimation," *The American Statistician*, 30, 105–112.

Taylor, C. C. (1987), "Akaike's Information Criterion and the Histogram," *Biometrika*, 74, 636–639.

——— (1989), "Bootstrap Choice of the Smoothing Parameter in Kernel Density Estimation," *Biometrika*, 76, 705–712.

Terrell, G. R. (1990), "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, 85, 470–477.

Terrell, G. R., and Scott, D. W. (1980), "On Improving Convergence Rates for Nonnegative Kernel Density Estimators," *The Annals of Statistics*, 8, 1160–1163.

——— (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 80, 209–214.

Titterington, D. M., and Mill, G. M. (1983), "Kernel-Based Density Estimates from Incomplete Data," *Journal of the Royal Statistical Society*, Ser. B, 45, 258–266.

Tukey, J. W. (1947), "Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions—The Continuous Case," *The Annals of Mathematical Statistics*, 18, 529–539.

——— (1948), "Nonparametric Estimation, III. Statistically Equivalent Blocks and Multivariate Tolerance Regions—The Discontinuous Case," *The Annals of Mathematical Statistics*, 19, 30–39.

Tukey, P. A., and Tukey, J. W. (1981), "Data-Driven View Selection; Agglomeration and Sharpening," in *Interpreting Multivariate Data*, ed. V. Barnett, New York: John Wiley, Ch. 11, pp. 215–243.

Van Es, A. J. (1990), *Aspects of Nonparametric Density Estimation*, CWI Tract, Amsterdam: Centre for Mathematics and Computer Science.

Van Ryzin, J. (1973), "On a Histogram Method of Density Estimation," *Communications in Statistics*, 2, 493–506.

Vitale, R. A. (1975), "A Bernstein Polynomial Approach to Density Estimation," in *Statistical Inference and Related Topics* (Vol. 2), ed. M. L. Puri, San Francisco: Academic Press, pp. 87–99.

Wagner, T. J. (1975), "Nonparametric Estimates of Probability Densities," *IEEE Transactions on Information Theory*, 21, 438–440.

Wahba, G. (1971), "A Polynomial Algorithm for Density Estimation," *The Annals of Mathematical Statistics*, 42, 1870–1886.

——— (1975a), "Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation," *The Annals of Statistics*, 3, 15–29.

——— (1975b), "Interpolating Spline Methods for Density Estimation I. Equi-Spaced Knots," *The Annals of Statistics*, 3, 30–48.

——— (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," *The Annals of Statistics*, 9, 146–156.

Wald, A. (1943), "An Extension of Wilk's Method for Setting Tolerance Limits," *The Annals of Mathematical Statistics*, 14, 45–55.

Walter, G. (1977), "Properties of Hermite Series Estimation of Probability Density," *The Annals of Statistics*, 5, 1258–1264. [Addendum: *Annals of Statistics*, 8, 454–455 (1980)].

Walter, G., and Blum, J. R. (1979), "Probability Density Estimation Using Delta Sequences," *The Annals of Statistics*, 7, 328–340.

——— (1984), "A Simple Solution to a Nonparametric Maximum Likelihood Estimation Problem," *The Annals of Statistics*, 12, 372–379.

Watson, G. S. (1969), "Density Estimation by Orthogonal Series," *The Annals of Mathematical Statistics*, 40, 1496–1498.

Watson, G. S., and Leadbetter, M. R. (1964), "Hazard Analysis I," *Biometrika*, 51, 175–184.

Wegman, E. J. (1969), "Maximum Likelihood Histograms," Technical Report, University of North Carolina, Chapel Hill.

——— (1972), "Nonparametric Probability Density Estimation: I. A Summary of Available Methods," *Technometrics*, 14, 533–546.

——— (1975), "Maximum Likelihood Estimation of a Probability Density Function," *Sankhyā*, Ser. A, 37, 211–224.

——— (1982), "Density Estimation," in *Encyclopedia of Statistical Sciences*, (Vol. 2), eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp. 309–315.

Wegman, E. J., and Davies, H. I. (1979), "Remarks on Some Recursive Estimators of a Probability Density," *The Annals of Statistics*, 7, 316–327.

Wertz, W. (1978), *Statistical Density Estimation: A Survey*, Göttingen, F.R.G.: Vanderhoeck and Ruprecht.

Wertz, W., and Schneider, B. (1979), "Statistical Density Estimation: A Bibliography," *International Statistical Review*, 47, 155–175.

Whittle, P. (1958), "On the Smoothing of Probability Density Functions," *Journal of the Royal Statistical Society*, Ser. B, 20, 334–343.

Wilks, S. S. (1962), *Mathematical Statistics*, New York: John Wiley.

Wolverton, C. T., and Wagner, T. J. (1969), "Recursive Estimates of Probability Densities," *IEEE Transactions on Systems, Science, and Cybernetics*, 5, 307.

Yamato, H. (1971), "Sequential Estimation of a Continuous Probability Density Function and the Mode," *Bulletin of Mathematical Statistics*, 14, 1–12.

Yandell, B. S. (1983), "Nonparametric Inference for Rates With Censored Survival Data," *The Annals of Statistics*, 11, 1119–1135.

Yang, S.-S. (1985), "A Smooth Nonparametric Estimator of a Quantile Function," *Journal of the American Statistical Association*, 80, 1004–1011.