Multi-dimensional Density Estimation

David W. Scott [a,*,1], Stephan R. Sain [b,2]

[a] *Department of Statistics, Rice University, Houston, TX 77251-1892, USA*

[b] *Department of Mathematics, University of Colorado at Denver, Denver, CO 80217-3364 USA*

Abstract

Modern data analysis requires a number of tools to undercover hidden structure. For initial exploration of data, animated scatter diagrams and nonparametric density estimation in many forms and varieties are the techniques of choice. This article focuses on the application of histograms and nonparametric kernel methods to explore data. The details of theory, computation, visualization, and presentation are all described.

Key words:
Averaged shifted histograms, Contours, Cross-validation, Curse of dimensionality, Exploratory data analysis, Frequency polygons, Histograms, Kernel estimators, Mixture models, Visualization
PACS: 62-07, 62G07

1 Introduction

Statistical practice requires an array of techniques and a willingness to go beyond simple univariate methodologies. Many experimental scientists today are still unaware of the power of multivariate statistical algorithms, preferring

* Corresponding author.
Email addresses: scottdw@rice.edu, ssain@math.cudenver.edu (Stephan R. Sain).

[1] This research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 02-04723 (non-parametric methodology).
[2] This research was supported in part through the Geophysical Statistics Project at the National Center for Atmospheric Research under National Science Foundation grant DMS 9815344.

the intuition of holding all variables fixed, save one. Likewise, many statisticians prefer the familiarity of parametric statistical techniques and forgo exploratory methodologies. In this chapter, the use of density estimation for data exploration is described in practice with some theoretical justification. Visualization is an important component of data exploration, and examples of density surfaces beyond two dimensions will be described.

We generally consider the analysis of a d-variate random sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from an unknown density function, $f(\mathbf{x})$, where $\mathbf{x} \in \Re^d$. It is worth reminding ourselves that (theoretically) for the analysis of a random sample, perfect knowledge of the density functions $f(\mathbf{x})$ or $f(\mathbf{x}, y)$ means that anything we may need to know can be computed. In practice, the computation may be quite complicated if the dimension of the data is high, but the greatest challenge comes from not knowing a parametric form for the density $f(\mathbf{x})$. Fisher (1932) referred to this step in data analysis as the problem of specification. Nonparametric methodology provides a consistent approach for approximating in a large class of unknown densities, at a cost of less efficient estimation when the correct parametric form is known. Of course, if an incorrect parametric form is specified, then bias will persist.

2 Classical Density Estimators

The statistical analysis of continuous data is a surprisingly recent development. While data collection such as a population census can be traced back thousands of years, the idea of grouping and tabulating data into bins to form a modern frequency curve seems to have only arisen in the seventeenth century. For example, John Graunt (1662) created a crude histogram of the age of death during the English plague years using the bills of mortality, which listed the cause and age of death week by week. In Figure 1, we analyze the closing price, $\{x_t\}$, of the Dow Jones Industrial (DJI) average from 2/3/1930 to 2/27/2004. A plot and histogram of the 18,598 daily change ratios, x_{t+1}/x_t, are shown. While the eye is drawn to the days when the ratio represents more than a 5% change, the histogram emphasizes how rare such events are. The eye is generally a poor judge of the frequency or density of points in a scatter diagram or time series plot.

By the nineteenth century, the histogram was in common use, as was its continuous cousin the frequency polygon. (The frequency polygon interpolates the midpoints of a histogram in a piecewise linear fashion.) The first publication to advocate a systematic rule for the construction of a histogram was due to Sturges (1926). Sturges was motivated by the notion of an ideal histogram in the case of normal data. Observe that the simplest example of a discrete density that is approximately normal is a binomial distribution with $p = 1/2$.

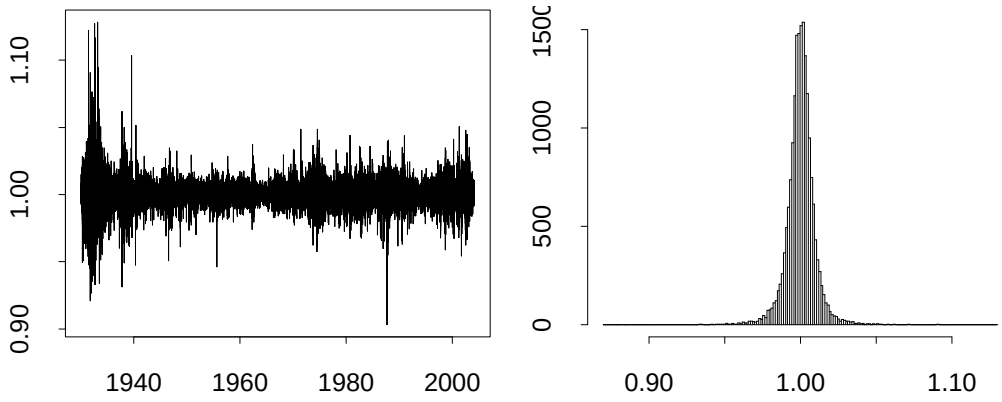

Fig. 1. Plot and histogram of daily change ratio, x_{t+1}/x_t, of the DJI.

Imagine a histogram of some binomial data with k bins, labeled $0, 1, \ldots, k-1$ and $p = 1/2$. Now the binomial density is $f(i) = \binom{k-1}{i} 2^{-(k-1)}$. Sturges argued that the binomial coefficient, $\binom{k-1}{i}$, could be taken as the idealized bin count of normal data as the number of bins, k, grows, since the binomial density looks normal as k increases. Let ν_k denote the bin count of the k-th bin. Then the total bin count is

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

hence, given n data points, the appropriate number of bins, k, may be solved as $k = 1 + \log_2(n)$. Let the bin width of a histogram be denoted by h, then Sturges' rule for the bin width of a histogram of a random sample $\{x_1, x_2, \ldots, x_n\}$ may be expressed as

$$h = \frac{x_{(n)} - x_{(1)}}{1 + \log_2(n)}\,, \tag{1}$$

where $x_{(i)}$ is the i-th order statistic of the sample. Formula (1) is still in common use in most statistical software packages. However, we shall see that this rule-of-thumb is far from optimal in a stochastic setting.

A histogram is a nonparametric estimator because it can successfully approximate almost any density as $n \to \infty$. The only alternative to the histogram before 1900 was a parametric model, $f(x|\theta)$, and θ was usually estimated by the method of moments. Pearson (1902) introduced a hybrid density estimator from the family of solutions to the differential equation

$$\frac{d \log f(x)}{dx} = \frac{x - a}{bx^2 + cx + d}\,. \tag{2}$$

The parameters (a, b, c, d) are estimated by the method of moments. (One parameter serves as a normalization constant, leaving three degrees-of-freedom). The particular "type" or solution of Equation (2) depends upon the roots of the denominator and special values of the parameters. A number of common distributions satisfy this differential equation, including the normal and t densities. In fact, Gosset made use of this fact to derive the t-distribution (Student, 1908).

The Pearson family of densities is not actually nonparametric since many densities are not in the family. However, the Pearson family is still used in many simulations today and is in the modern spirit of letting the data "speak" rather than imposing a fully parametric assumption.

In the following sections, the properties of the modern histogram and frequency polygon are examined.

2.1 Properties of Histograms

In the univariate setting, the frequency histogram plots raw data counts and hence is not actually a density function. We limit our discussion to continuous data and a true density histogram, which is defined over a general mesh on the real line. Let the k-th bin correspond to the interval $B_k = [t_k, t_{k+1})$. Let the number of samples falling into bin B_k be denoted by ν_k. Then the density histogram is defined as

$$\hat{f}(x) = \frac{\nu_k}{n(t_{k+1} - t_k)} \qquad \text{for } x \in B_k \,. \tag{3}$$

In practice, an equally-spaced mesh is often chosen, with $t_{k+1} - t_k = h$ for all k and with bin origin $t_0 = 0$. In this case, the histogram estimator is simply

$$\hat{f}(x) = \frac{\nu_k}{nh} \qquad \text{for } x \in B_k \,. \tag{4}$$

In the latter case, the histogram has only one unknown parameter, namely, the bin width h. Yet as we shall see, the histogram can asymptotically approximate any continuous density and hence earns the label "nonparametric." Some early writers suggested that nonparametric estimators were infinite-dimensional. In fact, the number of parameters has little bearing on whether an estimator is parametric or nonparametric. Rather, the important distinction is the local behavior of the estimator.

2.1.1 Maximum Likelihood and Histograms

How can maximum likelihood be applied to the density histogram? Let us begin with a general binned estimator of the form

$$\hat{f}(x) = f_k \qquad \text{for } x \in B_k = [t_k, t_{k+1}) \,.$$

Then the log-likelihood of the histogram is

$$\sum_{i=1}^{n} \log \hat{f}(x_i) = \sum_k \nu_k \log f_k \qquad \text{subject to } \sum_k (t_{k+1} - t_k) f_k = 1 \,,$$

where we define $f_k = 0$ for bins where $\nu_k = 0$. The Lagrangian is

$$L(\mathbf{f}, \lambda) = \sum_k \nu_k \log f_k + \lambda \left[1 - \sum_k (t_{k+1} - t_k) f_k \right] \,.$$

The stationary point for f_ℓ leads to the equation

$$\frac{\partial L(\mathbf{f}, \lambda)}{\partial f_\ell} = \frac{\nu_\ell}{f_\ell} - \lambda (t_{\ell+1} - t_\ell) = 0 \,.$$

The constraint leads to $\lambda = n$; hence, $\hat{f}_\ell = \nu_\ell / n(t_{\ell+1} - t_\ell)$ as in Equation (3). Thus the histogram is in fact a maximum likelihood estimator (MLE) within the class of simple functions (that is, given a pre-determined mesh).

If we extend the MLE to the estimation of h, or indeed of the entire mesh $\{t_k\}$, then we find that the likelihood is unbounded as $h \to 0$ or as $t_{k+1} - t_k \to 0$. Thus Duin (1976) introduced the leave-one-out likelihood. In the context of estimation of h in Equation (4), the log-likelihood becomes

$$h^* = \arg\max_h \sum_{i=1}^{n} \hat{f}_{-i}(x_i | h) \,, \tag{5}$$

where the leave-one-out density estimate is defined by

$$\hat{f}_{-i}(x_i) = \frac{\nu_k - 1}{(n-1)h} \qquad \text{assuming } x_i \in B_k \,.$$

While good results have been reported in practice, the procedure cannot be consistent for densities with heavy tails. Consider the spacing between the first two order statistics, $x_{(1)}$ and $x_{(2)}$. If $h < x_{(2)} - x_{(1)}$ and the bin count in the first bin is 1, then the likelihood in Equation (5) will be zero, since

$\hat{f}_{-(1)}(x_{(1)}) = 0$. Since a necessary condition for a histogram to be consistent will be shown to be that $h \to 0$ as $n \to \infty$, the spacings between all adjacent order statistics must vanish as well; however, such is not the case for many densities.

Thus most theoretical work on histograms (and other nonparametric estimators) has focused on distance-based criteria rather than the likelihood criterion. There are four common distance criteria between an estimator $\hat{f}(x)$ and the true but unknown density $g(x)$ (switching notation from $f(x)$) including:

$\int \lvert\hat{f}(x) - g(x)\rvert \, dx$	integrated absolute error
$\int \hat{f}(x) \, \log[\hat{f}(x)/g(x)] \, dx$	Kullback-Liebler distance
$\int[\hat{f}(x)^{1/2} - g(x)^{1/2}]^2 dx$	Hellinger distance
$\int[\hat{f}(x) - g(x)]^2 dx$	integrated squared error

The first three are dimensionless, a characteristic which provides many potential benefits in practice. The second is basically the likelihood criterion. The fourth is the most amenable to theoretical investigation and calibration in practice. Integrated squared error (ISE) is also the L_2 distance between the estimator and true densities.

2.1.2 L_2 Theory of Histograms

We spend a little time outlining the derivation of the ISE for a histogram, since every estimator shares this problem. The derivation is quite straightforward for the histogram, and we will not provide these details for other estimators. We limit our discussion to equally-spaced histograms. Rosenblatt (1956) showed in a general fashion that no nonparametric estimator can be unbiased, and that the rate of convergence of any measure of error cannot achieve the parametric rate of $O(n^{-1})$. Since the histogram estimate, $\hat{f}(x)$, for a fixed x cannot be unbiased, then mean square error (MSE) is a natural criterion pointwise. Globally, MSE can be integrated over x to give the integrated mean square error (IMSE). By Fubini's theorem, IMSE is the same as mean integrated square error (MISE)

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

Finally, since MSE = Var + Bias^2, the IMSE is the sum of the mean integrated variance (IV) and the mean integrated squared bias (ISB).

The bin count, ν_k, is a Binomial random variable, $B(n, p_k)$, with probability given by the actual bin probability, $p_k = \int_{t_k}^{t_{k+1}} g(x)\,dx$. Hence, for $x \in B_k$, the pointwise variance of the histogram estimate given in Equation (4) equals

$$\mathrm{Var}\,\hat{f}(x) = \frac{\mathrm{Var}(\nu_k)}{(nh)^2} = \frac{p_k(1-p_k)}{nh^2}\,.$$

Since the variance is identical for any $x \in B_k$, the integral of the variance over B_k multiplies this expression by the bin width, h. Therefore,

$$\mathrm{IV} = \sum_k \int_{B_k} \mathrm{Var}(\hat{f}(x))dx = \sum_k \frac{p_k(1-p_k)}{nh^2} \times h = \frac{1}{nh} - \sum_k \frac{p_k^2}{nh}\,, \tag{6}$$

since $\sum_k p_k = \int g(x)\,dx = 1$. Next, by the mean value theorem, $p_k = \int_{B_k} g(x)dx$ $= h \cdot g(\xi_k)$ for some $\xi_k \in B_k$; thus, the final sum equals $n^{-1}\sum_k g(\xi_k)^2 h$, or approximately $n^{-1}\int g(x)^2 dx$. Thus the variance of a histogram pointwise or globally may be controlled by collecting more data (larger n) or having sufficient data in each bin (wider h).

The bias is only a little more difficult to analyze. Clearly,

$$\mathrm{Bias}\,\hat{f}(x) = \mathrm{E}\,\hat{f}(x) - g(x) = \frac{p_k}{h} - g(x) \qquad \text{for } x \in B_k\,.$$

Again using the fact that $p_k = h \cdot g(\xi_k)$, Bias $\hat{f}(x) = g(\xi_k) - g(x) = O(hg'(x))$, assuming the unknown density has a smooth continuous derivative, since a Taylor's series of $g(\xi_k)$ equals $g(x) + (\xi_k - x)g'(x) + o(h)$ and $|\xi_k - x| < h$ as both ξ_k and x are in the same bin whose width is h.

Thus the squared bias of $\hat{f}(x)$ is of order $h^2 g'(x)^2$, and the integrated squared bias is of order $h^2 \int g'(x)^2 dx$. In contrast to the manner by which the variance is controlled, the bias is controlled by limiting the size of the bin width, h. In fact, we require that $h \to 0$ as $n \to \infty$. From Equation (6), $nh \to \infty$ is also necessary.

Combining, we have that the global error of a fixed-bin-width histogram is

$$\mathrm{IMSE} = \mathrm{IV} + \mathrm{ISB} = \frac{1}{nh} + \frac{1}{12}h^2 \int g'(x)^2 dx + O(\frac{1}{n} + h^4)\,, \tag{7}$$

where the factor $\frac{1}{12}$ results from a more careful analysis of the difference between $g(\xi_k)$ and $g(x)$ for all x in B_k; see Scott (1979).

The IMSE in Equation (7) is minimized asymptotically by the choice

$$h^* = \left[\frac{6}{nR(g')}\right]^{1/3} \qquad \text{and} \quad \text{IMSE}^* = (9/16)^{1/3} R(g')^{1/3} n^{-2/3}, \tag{8}$$

where the "roughness" of $g(x)$ is summarized by $R(g') \equiv \int g'(x)^2 dx$. Indeed, the rate of convergence of the IMSE falls short of the parametric rate, $O(n^{-1})$.

2.1.3 Practical Histogram Rules

We support the idea that histograms constructed with bin widths far from h^* are still of potential value for exploratory purposes. Larger bandwidths allow for a clearer picture of the overall structure. Smaller bandwidths allow for fine structure, which may or may not be real (only a larger sample size will clarify the true structure). Smaller bandwidths may also be useful when the true density is a mixture of components, say, for example, a normal mixture $w\,N(0,1) + (1-w)\,N(\mu, \sigma^2)$. Obviously there is a different bandwidth appropriate for each component, and h^* represents a compromise between those bandwidths. Using a smaller bandwidth may allow excellent estimation of the narrower component, at the price of making the other component undersmoothed. Unless one uses the more general mesh in Equation (3), such compromises are inevitable. Data-based algorithms for the general mesh are considerably more difficult than for a single parameter and may in fact perform poorly in practice.

Expressions such as those for the optimal parameters in Equation (8) may seem of limited utility in practice, since the unknown density $g(x)$ is required. However, a number of useful rules follow almost immediately. For example, for the normal density, $N(\mu, \sigma^2)$, the roughness equals $R(g') = (4\sqrt{\pi}\sigma^3)^{-1}$. Therefore,

$$h^* = \left[\frac{24\sqrt{\pi}\sigma^3}{n}\right]^{1/3} \approx 3.5\,\sigma\,n^{-1/3}\,. \tag{9}$$

Compare this formula to Sturges' rule in Equation (1). Since the logarithm is a very slowly increasing function, Sturges' bin width is too slow to decrease as the sample size increases, at least with respect to IMSE.

Formulae such as Scott's rule (9), with σ replaced by an estimate $\hat{\sigma}$, are variations of so-called normal reference rules. Almost any other density can be shown to be more complex, resulting in an optimal bin width that is narrower. In fact, we can make this idea quite explicit. Specifically, a calculus of variations argument can be formulated to find the smoothest ("easiest")

density that has a given variance, σ^2. Terrell and Scott (1985) showed that this density is given by

$$g_1(x) = \frac{15}{16\sqrt{7}\sigma}\left(1 - \frac{x^2}{7\sigma^2}\right)^2 \qquad -\sqrt{7}\sigma < x < \sqrt{7}\sigma$$

and zero elsewhere. Thus, for all densities with variance σ^2, $R(g') \geq R(g_1')$. Since $R(g_1') = 15\sqrt{7}/343\sigma^3$, the optimal bandwidth for density g_1 is

$$h_1^* = h_{\mathrm{OS}} = \left[\frac{686\sigma^3}{5\sqrt{7}n}\right]^{1/3} \approx 3.73\sigma n^{-1/3}\,,$$

where OS $\equiv$ oversmoothed. The normal reference bandwidth is only 6% narrower, confirming that the normal density is very close to the "oversmoothed" or "smoothest" density, which is in fact Tukey's biweight function. Since any other density is rougher than g_1, the optimal bandwidth satisfies the inequality $h^* \leq h_{\mathrm{OS}}$. Since σ can be estimated quite reliably from the data, we have bounded the search region for h^* to the interval $(0, h_{\mathrm{OS}})$. This is a very useful result in practice.

In fact, we can use cross-validation to obtain an estimate of h^* itself. Rudemo (1982) and Bowman (1984) showed that the integrated squared error can be estimated for each choice of h by replacing the second term in the expanded version of the integrated squared error:

$$\begin{aligned}\mathrm{ISE} &= \int [\hat{f}(x|h) - g(x)]^2 dx \\ &= \int \hat{f}(x|h)^2 dx - 2\int \hat{f}(x|h)g(x)dx + \int g(x)^2 dx\end{aligned} \tag{10}$$

with the unbiased estimator

$$-\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i|h) \sim -2E[\hat{f}(X|h)] = -2\int \hat{f}(x|h)g(x)dx\,.$$

The final integral in (11), $\int g(x)^2 dx$, is unknown but is constant with respect to h and may be ignored. Thus the least-squares or unbiased cross-validation (UCV) criterion which estimates $\mathrm{ISE} - \int g(x)^2 dx$ is

$$\int \hat{f}(x|h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i|h)\,. \tag{11}$$

This approach requires the additional assumption that the true density is square integrable for all choices of the smoothing parameter.

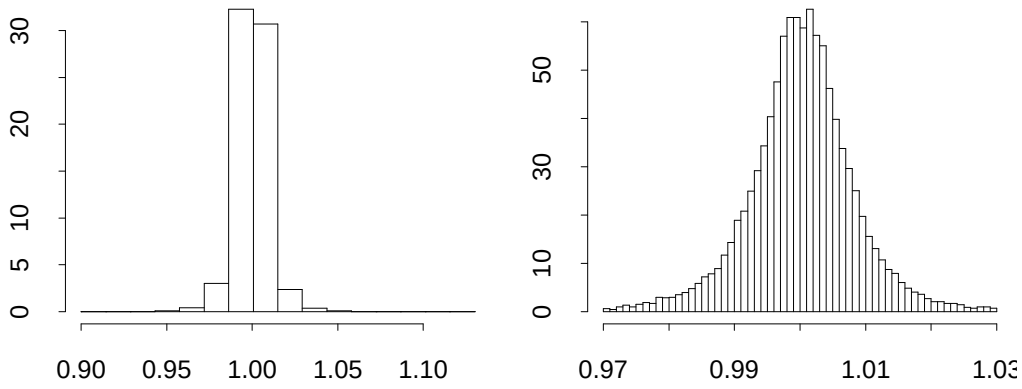

Fig. 2. Histograms of x_{t+1}/x_t of the DJI chosen by Sturges' rule and by eyeball.

There have been a number of further algorithms proposed that focus on improved point estimates of the roughness functional, $R(g')$. However, such approaches are still limited to a single bandwidth. Rudemo (1982) showed how unbiased cross-validation may easily be extended to the variable mesh case. Wegman (1975) demonstrated the strong consistency of a variant of the variable-mesh histogram based upon maintaining a minimum number of samples in each bin. The idea was ultimately the motivation for the random bin width estimators of Hearne and Wegman (1994). Kogure (1987) and Kanazawa (1992) also examine variable mesh histograms. Variable smoothing is discussed further in Section 3.4.

As an example, consider again the ratio of DJI closing prices, $\{x_{t+1}/x_t\}$, since 1930. The oversmoothed bin width is $3.73\hat{\sigma}n^{-1/3} = 0.00137$, which was used to construct the histogram in Figure 1. Sturges' rule suggests $1 + \log_2(18598) = 15.2$ bins. The left frame in Figure 2 uses 16 bins over the sample range. To see if there is structure that might be revealed if $h < h_{\mathrm{OS}}$, we show a detail in the right frame of Figure 2 with $h = 0.001$, which is about 27% narrower than h_{OS}. The minimizer of the UCV criterion (11) occurs at $h = 0.00092$. Note that strictly speaking, this example does not represent a random sample since the data are a correlated time series; however, the rules of thumb still seem useful in this case. Observe that Sturges' rule grossly oversmooths the data in any case.

2.1.4 Frequency Polygons

The use of the piecewise linear frequency polygon (FP) in place of the underlying histogram would seem mainly a graphical advantage. However, Fisher (1932, p. 37) suggested that the advantage was "illusory" and that a frequency polygon might easily be confused with the smooth true density. "The utmost care should always be taken to distinguish" the true curve and the estimated curve, and "in illustrating the latter no attempt should be made to slur over

this distinction."

However, Scott (1985a) showed that the smoothness of the frequency polygon reduced not only the bias but also the variance compared to a histogram. A similar analysis of the estimation errors leads to the expression

$$\text{MISE} = \frac{2}{3nh} + \frac{49}{2880}h^4 R(g'') + O(\frac{1}{n} + h^6)\,,$$

where $R(g'') = \int g''(x)^2 dx$. Thus the best choice of the bin width for the underlying histogram is not that given in Equation (7), but rather

$$h^* = c_0 c_g n^{-1/5} \qquad \text{and} \qquad \text{MISE}^* = c_1 c_g n^{-4/5}\,,$$

where $c_0 = 1.578$, $c_1 = 0.528$, and $c_g = R(g'')^{-1/5}$. For $N(\mu, \sigma^2)$ data, $h^* = 2.15\sigma n^{-1/5}$, which, for large n, will be much wider than the corresponding histogram formula, $h^* = 3.5\sigma n^{-1/3}$. For example, when $n = 10^5$ with normal data, the optimal FP bin width is 185% wider than that of the histogram; the MISE of the FP is reduced by 81%. The wider bins allow the FP to achieve lower variance. The piecewise linear FP more closely approximates the underlying smooth density than the piecewise constant histogram. (In fact, piecewise quadratic estimates can achieve even closer approximation. However, such estimates often take on negative values and do not offer sufficient improvement in practice to recommend their use.).

In Figure 3, we display the common logarithm of the Canadian lynx data together with a histogram using $h = 0.4$, which is slightly less than the normal reference rule bandwidth $h = 0.47$. (The UCV bandwidth is 0.461.) In Figure 4, we display two shifted versions of the histogram in Figure 3. The theoretical analyses of the MISE for both the histogram and FP indicate that the effect of choosing the bin origin t_0 is relegated to the remainder (low-order) terms. However, the graphical impact is not negligible. The bimodal feature varies greatly among these three histograms. For the FP, the wider bins of its underlying histogram suggest that the choice of t_0 matters more with the FP than with the histogram. We revisit this below in Section 3.1.

2.1.5 Multivariate Frequency Curves

The power of nonparametric curve estimation is in the representation of multivariate relationships. In particular, density estimates in dimensions 3, 4, and even 5 offer great potential for discovery. Examples and visualization techniques are described below in Section 5.

Beyond 4 dimensions, the effects of the so-called curse of dimensionality must

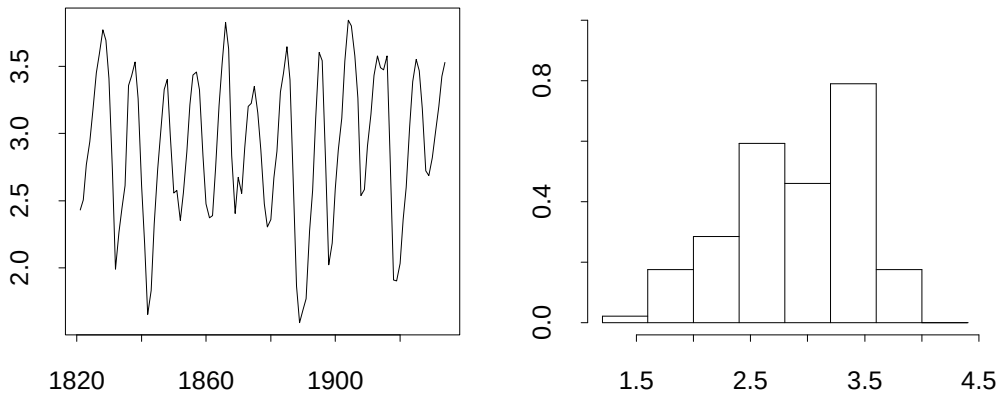

Fig. 3. Canadian lynx data ($\log_{10} x_t$) and its histogram ($h = 0.4$ and $t_0 = 2$).

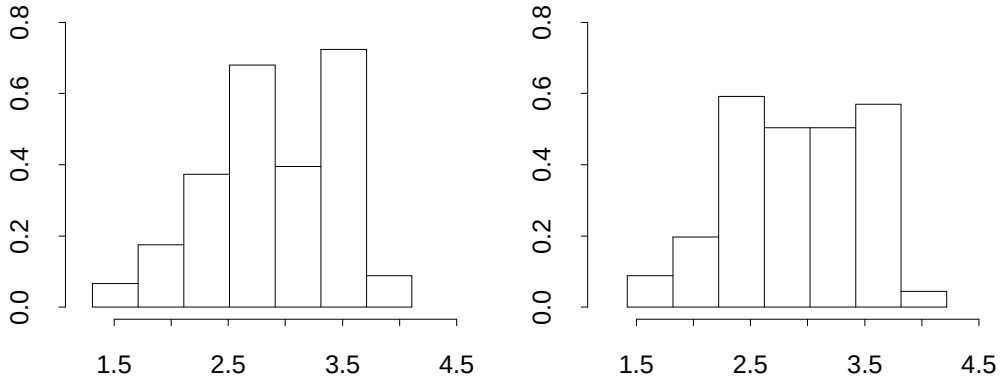

Fig. 4. Two shifted lynx histograms ($t_0 = 2.1$ and $t_0 = 2.2$) with $h = 0.4$.

be considered. The bias-variance tradeoff is subject to failure since the optimal bin widths must be large, and are generally too wide to avoid substantial bias. Imagine a large sample of 10^6 data uniformly distributed on the unit hypercube in $\Re^{10}$. If each axis is divided into 5 equal bins, then the hypercube has 5^{10} or almost ten million bins. Even such a crude binning leaves 90% of the bins empty. If each axis were divided into only 3 bins, then each bin would still have only 17 points on average. Thus these estimators must be quite biased, even with a truly enormous sample size.

However, the extension of the MISE analyses to the multivariate case is straightforward and serves to better quantify the effects of the curse of dimensionality. For a multivariate histogram with cubical bins of volume h^d, the IV is $O(1/nh^d)$ while the ISB remains of $O(h^2)$. Thus

$$\text{Histogram:} \quad h_d^* = O(n^{-1/(d+2)}) \quad \text{and} \quad \text{MISE}_d^* = O(n^{-2/(d+2)}) \,.$$

The IV and ISB for the multivariate frequency polygon are $O(1/nh^d)$ and

$O(h^4)$, respectively. Thus the situation is significantly improved (Scott, 1985a):

$$\text{FP:} \quad h_d^* = O(n^{-1/(d+4)}) \quad \text{and} \quad \text{MISE}_d^* = O(n^{-4/(d+4)}) \,.$$

Perhaps the most encouraging observation is that the MISE convergence rate of order $n^{-2/5}$ is achieved not only by histograms in $d = 3$ dimensions but also by frequency polygons in $d = 6$ dimensions. Since a number of scientists have successfully used histograms in 3 (and even 4) dimensions, we believe that it is reasonable to expect useful nonparametric estimation in at least six dimensions with frequency polygons and other smoother estimators. That is more than sufficient for graphical exploratory purposes in dimensions $d \leq 5$. Complete nonparametric estimation of a density function in more than six dimensions is rarely required.

3 Kernel Estimators

3.1 Averaged Shifted Histograms

The Canadian lynx example in Figures 3 and 4 indicates that both the bin width and the bin origin play critical roles for data presentation. One approach would be to use unbiased cross-validation to search for the best pair (h, t_0). However, Scott (1985b) suggested that t_0 should be viewed as a nuisance parameter which can be eliminated by averaging several shifted histograms. The details of averaging m shifted histograms are easiest if each is shifted by an amount $\delta = h/m$ from the previous mesh. The averaged shifted histogram (ASH) will be constant over intervals of width δ, so we redefine the bin counts, $\{\nu_k\}$, to correspond to the mesh $B'_k = [t_0+k\delta, t_0+(k+1)\delta)$. Then Scott (1985b) shows that

$$\hat{f}_{\text{ASH}}(x) = \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) \nu_{k+i} \qquad \text{for } x \in B'_k \,. \tag{12}$$

In Figure 5, all shifted histograms have $h = 0.4$. The first two frames show individual histograms with $t_0 = 2.0$ and $t_0 = 2.2$. The ASH with $m = 2$ is shown in the third frame, and so on. Eliminating t_0 shows that the data are clearly bimodal, with a hint of a small bump on the left. The limiting ASH is continuous, which provides visual advantages. Connecting the midpoints of the ASH like a frequency polygon (called the FP-ASH) has some theoretical value, although for $m \geq 32$ in Figure 5, the discontinuous nature of the ASH is not visible.

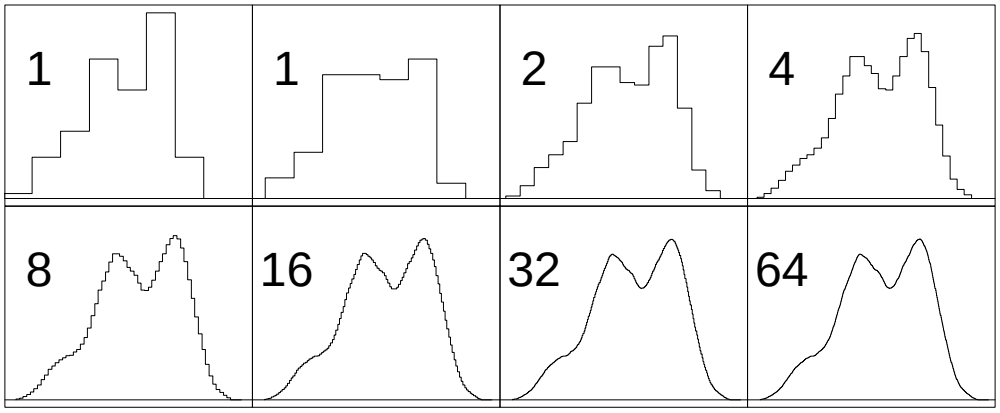

Fig. 5. Original averaged shifted histograms of lynx data.

Uniform weights on the shifted histograms are not the only choice. Choose a smooth symmetric probability density, $K(x)$, defined on $[-1, 1]$ that satisfies $K(\pm 1) = 0$. Define the weight function

$$w_m(i) = \frac{m \cdot K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)} \qquad \text{for } i = 1-m, \ldots, m-1\,. \tag{13}$$

Then the generalized ASH in Equation (12) is

$$\hat{f}_{\mathrm{ASH}}(x) = \frac{1}{nh} \sum_{i=1-m}^{m-1} w_m(i)\, \nu_{k+i} \qquad \text{for } x \in B'_k\,. \tag{14}$$

The weight function for the original ASH in Equation (12) is the triangle kernel, $K(x) = 1 - |x|$, for $|x| < 1$ and zero elsewhere. Kernels in the shifted Beta family, $K_\ell(x) \propto (1-x^2)_+^\ell$ are popular in practice. Tukey's biweight kernel corresponds to $\ell = 2$ while the normal kernel is well-approximated for large ℓ. Some examples of ASH estimates with $\ell = 5$ are shown in Figure 6. Notice that the use of a differentiable kernel makes the ASH visually smoother and the small bump on the left now appears clearer. The ASH of the DJI ratio reveals a small bimodal feature suggesting the DJI tries to avoid closing at the same level two days in a row.

3.2 Kernel Estimators

As the number of shifts $m \to \infty$ in Equation (13), the ASH approximates the so-called kernel estimator

$$\hat{f}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i-1}^{n} K_h(x - x_i)\,, \tag{15}$$

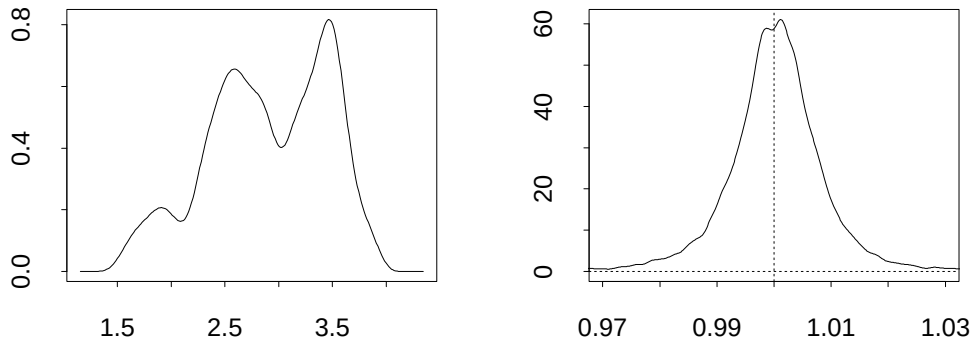

Fig. 6. Averaged shifted histograms of the lynx and DJI data using a smoother weight sequence $w_m(i) \propto (1 - (i/m)^2)_+^5$.

where the kernel, K, corresponds to K in Equation (13) and the scaled kernel function, $K_h(x)$, is defined as $K_h(x) = h^{-1}K(h^{-1}x)$. Thus a kernel estimator is an equal mixture of n kernels, centered at the n data points. For large n, the ASH requires much less work, since determining the bin counts is a linear operation, and the smoothing is a discrete convolution on the bin counts. (The kernel estimator may be viewed as a continuous convolution on all n points.) If one wanted to use the normal kernel, then much of the computational efficiency of the ASH would be lost. However, the Fast Fourier Transform can be used in that case; see Silverman (1982) for details. Using the FFT limits the ability to use boundary kernels or to estimate over a subset of the domain.

Choosing a good value for the bandwidth, h, is the most difficult task. The normal reference rule using a normal kernel is $h = 1.06\,\sigma\,n^{-1/5}$ for univariate data. More sophisticated plug-in rules have been described by Sheather and Jones (1991). However, we continue to recommend least-squares or unbiased cross-validation algorithms, which are well-studied for kernel estimators; see Rudemo (1982), Bowman (1984), and Sain, Baggerly, and Scott (1994). For the lynx data transformed by $\log_{10}$, the unbiased cross-validation function in Equation (11) with the normal kernel suggests using the bandwidth $h = 0.154$; see Figure 7. The corresponding Gaussian kernel estimate is shown in the right frame of this figure. This estimator is slightly less rough than the ASH estimate shown in Figure 6, which was chosen by eyeball to highlight the small bump/mode near $x = 1.9$. However, at that narrow bandwidth, an extra bump seems to be present near $x = 2.8$. Using a single bandwidth for the entire domain implies such compromises. Locally adaptive smoothing is a possibility and is discussed in Section 3.4.

The unbiased cross-validation of the DJI time series also suggests a wider bandwidth than used for the ASH in Figure 6. The slightly bimodal feature at $x = 1$ disappears. However, care should be exercised when using cross-validation on time series data, since serial correlation is present. Specialized

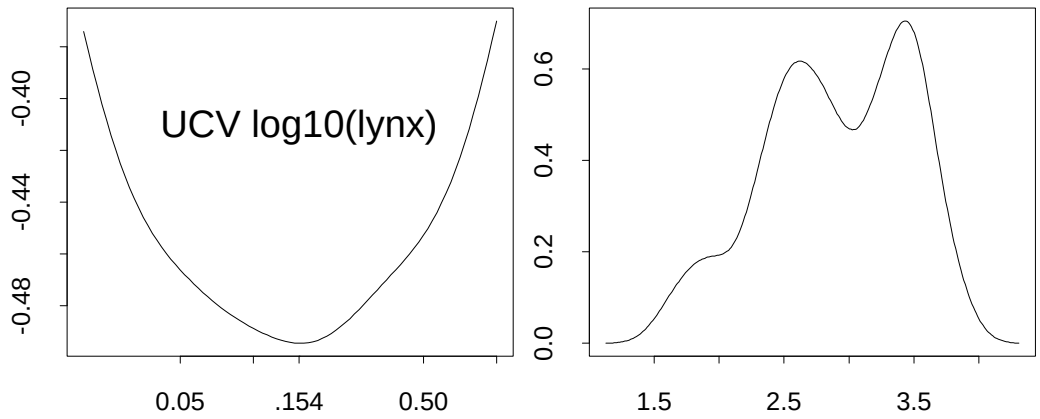

Fig. 7. Unbiased cross-validation and data-based optimal Gaussian kernel estimator.

algorithms exist in this situation; see Hart (1984).

The choice of kernel is largely a matter of convenience. The family of scaled Beta densities provides collection of useful polynomial kernels of the form $K(x) \propto (1-x^2)_+^\ell$ on the interval $(-1,1)$. As $\ell \to \infty$, this kernel converges to the normal kernel. The normal kernel has one advantage in practice; namely, as the smoothing parameter h increases, the number of modes is monotone non-increasing (Silverman, 1981). This property led Minnotte and Scott (1993) to propose the "mode tree," which plots the location of modes of a normal kernel estimator as a function of h. Minnotte (1997) proposed a local bootstrap test for the veracity of individual modes by examining the size of modes at critical points in the mode tree. Chaudhuri and Marron (1999) have introduced a graphical tool called SiZer to test the features in a kernel density.

3.3 Multivariate Kernel Options

The extension of the kernel estimator to vector-valued data, $\mathbf{x} \in \Re^d$, is straightforward for a normal kernel, $K \sim N(0, \Sigma)$:

$$\hat{f}(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}|\Sigma|^{1/2}} \sum_{i=1}^{n} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)'\Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)]\,. \tag{16}$$

It is convenient to separate the "size" of Σ from the "orientation" of Σ. To that end, write $\Sigma = h^2 A$, where $|A| = 1$. Thus, the size of Σ is $|h^2 A| = h^{2d}$. The gaussian kernel estimate becomes

$$\hat{f}(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}h^d} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)' A^{-1}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)\right]. \tag{17}$$

Since A is a symmetric, positive-definite matrix, the symmetric, positive-definite square-root matrix, $A^{-1/2}$ exists. Hence, Equation (17) becomes

$$\hat{f}(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}h^d} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\frac{(A^{-1/2}(\mathbf{x}-\mathbf{x}_i))'}{h}\,\frac{(A^{-1/2}(\mathbf{x}-\mathbf{x}_i))}{h}\right]. \quad (18)$$

This equation proves it is equivalent to rotate the data by the transformation $A^{-1/2}$ and then apply the $N(0, I_d)$ kernel. This transformation is almost into the principal components, except that the final scaling is not applied to make the variances all equal. In this transformed space, the kernel estimate is

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \frac{1}{n(2\pi)^{d/2}h^d} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)'\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{x}_i)\,,\end{aligned}$$

where $K_h(\mathbf{x}) = h^{-d}K(\mathbf{x}/h) = \prod_{k=1}^{d}\phi(x^{(k)}|0,h)$.

We recommend working with transformed data and using either the normal kernel or, more generally, a product kernel, possibly with different smoothing parameter, h_k, in the k-th direction:

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{d} K_{h_k}(x^{(k)} - x_i^{(k)})\right]. \quad (19)$$

The *ashn* software (see Section 5) computes an approximation of the multivariate product kernel estimate with kernels selected from the rescaled Beta family.

The multivariate rule-of-thumb for the bandwidth h is surprisingly simple. Assuming a normal product kernel and a true density that is also normal with $\Sigma = I_d$, then to close approximation

$$h^* = n^{-1/(d+4)} \qquad \text{or} \qquad h_k^* = \hat{\sigma}_k n^{-1/(d+4)}$$

for the general normal product estimator (19). For other choices of kernel, Scott (1992) provides a table of constants by which to multiple h_k^*.

Full cross-validation may be used to estimate h or $(h_1, \ldots, h_d)$ from the data; see Sain, Baggerly, and Scott (1994). Estimating the shape of the kernel parameters in the matrix A is generally not advisable, as there are too many parameters in high dimensions. We demonstrate below in Section 3.4 that useful estimates of A may be obtained in two dimensions. Wand and Jones

(1994) describe multivariate plug-in bandwidth rules, which can be more stable. However, it is important to note that kernel methods cannot handle rank-deficient data. Such degenerate cases can often be detected by computing the principal components and throwing away dimensions where the eigenvalues are essentially zero.

3.4 Locally Adaptive Estimators

As any practitioner will note, more smoothing is needed to counter the excessive variation in the tails of a distribution where data are scarce while less smoothing is needed near the mode of a distribution to prevent important features from being diminished in the resulting estimate. Several situations have been discussed (e.g. multimodal and multivariate distributions) where the bias-variance trade-off that drives most global bandwidth choices can lead to estimates that lack visual appeal and make feature recognition difficult.

These situations have often motivated the notion of a variable bandwidth function that allows different amounts of smoothing depending on the various characteristics of the data and the density being estimated. Two simplified forms of such estimators have been studied extensively. The first, the *balloon* estimator, varies the bandwidth with the estimation point. The second varies the bandwidth with each estimation point and is referred to as the *sample point* estimator. Jones (1990) gives an excellent comparison of such estimators in the univariate case while Terrell and Scott (1992) and Sain (2002) examined each of the two different formulations in the multivariate setting.

3.4.1 Balloon Estimators

The basic form of the balloon estimator is a generalization of Equation (18):

$$\hat{f}_B(\mathbf{x}) = \frac{1}{n|H(\mathbf{x})|^{1/2}} \sum_{i=1}^{n} K(H(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} K_{H(\mathbf{x})}(\mathbf{x} - \mathbf{x}_i)$$

where $H(\mathbf{x})$ is a positive-definite smoothing matrix associated with the estimation point $\mathbf{x}$. Note that H corresponds to hA. At a particular estimation point $\mathbf{x}$, the balloon estimator and the fixed bandwidth are exactly the same. Both place kernels of the same size and orientation at each of the data points and the estimate is constructed by averaging the values of the kernels at $\mathbf{x}$.

Taking K to be a uniform density on the unit sphere with $H(\mathbf{x}) = h_k(\mathbf{x})I_d$ and letting $h_k(\mathbf{x})$ the distance from $\mathbf{x}$ to the k-th nearest data point, one has the k-nearest neighbor estimator of Loftsgaarden and Quesenberry (1965). Much

has been written about this early balloon estimator that tries to incorporate larger bandwidths in the tails. (Where data are scarce, the distances upon which the bandwidth function is based should be larger.) The estimator is not guaranteed to integrate to one (hence, the estimator is not a density) and the discontinuous nature of the bandwidth function manifests directly into discontinuities in the resulting estimate. Furthermore, the estimator has severe bias problems, particularly in the tails (Mack and Rosenblatt, 1979, and Hall, 1983) although it seems to perform well in higher dimensions (Terrell and Scott, 1992).

In general, the identical construction of the balloon estimator and the fixed bandwidth estimator results in identical pointwise error properties. However, there are certain regions of the underlying density, typically in the tails, where the size and orientation of the kernels can be chosen to yield a higher-order bias (Terrell and Scott, 1992) or even eliminate it completely (Sain, 2001; Hazelton, 1998; Sain and Scott, 2002; Devroye and Lugosi, 2000; Sain, 2003).

3.4.2 Sample Point Estimators

The multivariate sample-point estimator is defined to be

$$\begin{aligned}\hat{f}_S(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|H(\mathbf{x}_i)|^{1/2}}K(H(\mathbf{x}_i)^{-1/2}(\mathbf{x}-\mathbf{x}_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n}K_{H(\mathbf{x}_i)}(\mathbf{x}-\mathbf{x}_i)\,,\end{aligned} \tag{20}$$

where $H(\mathbf{x}_i)$ is a positive-definite smoothing matrix associated with the ith data point, $\mathbf{x}_i$. In contrast to the balloon estimator, this estimator still places a kernel at each data point and the estimator is still constructed by averaging the kernel values at $\mathbf{x}$. However, the size and orientation of each kernel is different and is constant over the entire range of the density to be estimated.

Early efforts with such estimators proposed $H(\mathbf{x}_i) \propto f(\mathbf{x}_i)^{-\alpha} I_d$. Breiman et al. (1977) suggested using nearest-neighbor distances which is equivalent to using $\alpha = 1/d$. Abramson (1982) suggested using $\alpha = 1/2$ regardless of the dimension. Pointwise, it can be shown that this parameterization of the bandwidth function can yield a higher-order behavior of the bias (Silverman, 1986; Hall and Marron, 1988; Jones, 1990) and empirical results show promise for smaller sample sizes. However, this higher-order behavior does not hold globally due to bias contributions from the tails (Hall, 1992; McKay, 1993; Terrell and Scott, 1992; Hall et al., 1994; and Sain and Scott, 1996) and any gains can be lost as the sample size increases.

Sain and Scott (1996) and Sain (2002) suggest using a binned version of the

sample-point estimator in (20). Such an estimator has the form

$$\hat{f}_{Sb}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{m} \frac{n_j}{|H(\mathbf{t}_j)|^{1/2}} K(H(\mathbf{t}_j)^{-1/2}(\mathbf{x}-\mathbf{t}_j)) = \frac{1}{n}\sum_{j=1}^{m} n_j K_{H(\mathbf{t}_j)}(\mathbf{x}-\mathbf{t}_j)$$

where n_j is the number of data points in the jth bin centered at $\mathbf{t}_j$ and $H(\mathbf{t}_j)$ is a positive-definite bandwidth matrix associated with the jth bin. Using such an estimator, the MISE can easily be examined by recognizing that only the n_j are random and follow at multinomial distribution with cell probabilities given by $p_j = \int_{B_j} f(\mathbf{x})d\mathbf{x}$ where B_j denotes the jth bin. The MISE for a binned estimator with normal kernels is then given by

$$\begin{aligned}\text{MISE} = &\frac{1}{n(2\sqrt{\pi})^d}\sum_j \frac{p_j(1-p_j)+np_j^2}{|H_j|^{1/2}} + \frac{n-1}{n}\sum_{i\neq j} p_i p_j \phi_{H_i+H_j}(\mathbf{t}_i-\mathbf{t}_j) \\ &-\frac{2}{n}p_j\int \phi_{H_j}(\mathbf{x}-\mathbf{t}_j)f(\mathbf{x})d\mathbf{x} + R(f)\end{aligned}$$

where $H_j = H(\mathbf{t}_j)$. Sain and Scott (1996) used this formulation to examine univariate sample-point estimators and showed that while the MISE did not appear to converge at a faster rate, significant gains over fixed bandwidth estimators could be theoretically obtained for a wide variety of densities. Sain (2002) showed similar results in the multivariate setting.

3.4.3 Parameterization of Sample-Point Estimators

Designing practical algorithms that actually achieve some of the gains predicted in theory has been a difficult task and much of the promise depends on how the bandwidth function is parameterized. It seems to be widely held that the sample-point estimator shows more promise, perhaps since the estimator is a bona fide density by construction. However, n positive-definite smoothing matrices must be estimated for the sample-point estimator and it is clear that some sort of dimension reduction must be utilized.

The binning approach outlined in the previous section is one possible approach to reduce the number of smoothing matrices that must be estimated. In addition, further reduction could be had by restricting the form of the smoothing matrices. For example, one could let the kernels be radially symmetric and just vary the size of the kernels, effectively letting $H(\mathbf{x}_i) = h(\mathbf{x}_i)I_d$. This leaves just one parameter to be estimated for each bin. A step up is to allow different amounts of smoothing for each dimension using the product kernel form. This would reduce the bandwidth function to $H(\mathbf{x}_i) = \text{diag}(h_1(\mathbf{x}_i), \ldots, h_d(\mathbf{x}_i))$ where diag indicates a diagonal matrix. Each kernel would be elliptical with

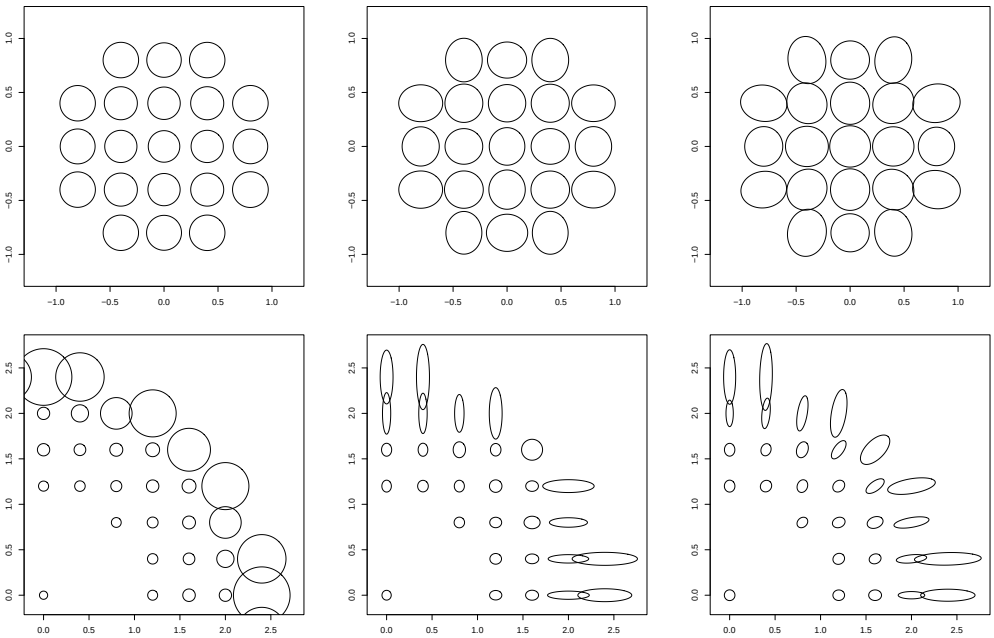

Fig. 8. Ellipses showing relative sizes and shapes of sample point kernels using the binned estimator and a bivariate standard normal density. Left column shows kernels for radially symmetric kernels, middle column shows kernels for diagonal bandwidth matrices, while the right column shows the unrestricted kernels. Top frames show kernels inside the unit circle while the bottom frames shows kernels in the first quadrant and outside the unit circle.

the axis of each ellipse aligned with the coordinate axis and d parameters would be estimated for each bin.

In two dimensions, there are three free parameters in the full smoothing matrix. While the product kernel formulation allows for some dimension reduction, many other formulations are possible. For example, Banfield and Raftery (1993) reparameterize covariance matrix for normal components of a mixture as $\Sigma_k = \lambda_k D_k A_k D_k'$. In this formulation, λ_k controls the volume while the matrix A_k controls the shape and is a diagonal matrix of the form $A_k = \text{diag}(1, \alpha_2, \ldots, \alpha_k)$ for $1 \geq \alpha_2 \geq \cdots \geq \alpha_k$. The matrix D_k is an orthogonal matrix that controls the orientation. The three free parameters for $d = 2$ are then λ_k, α_2, and any one of elements of D_k (the other elements of D_k can be obtained from the constraints imposed by orthogonality, i.e. $D_k D_k' = I_d$). Any combination of these terms could be held constant or allowed to vary yielding a great many different parameterizations that could be effective for densities of different shapes.

A comparison of the size and shape of optimal kernels for the three basic forms is given in Figure 8 for a bivariate standard normal density $n = 1000$ and the binned sample-point estimator. A fixed mesh is laid down over the range of the density, bin probabilities are calculated, and the MISE is minimized.

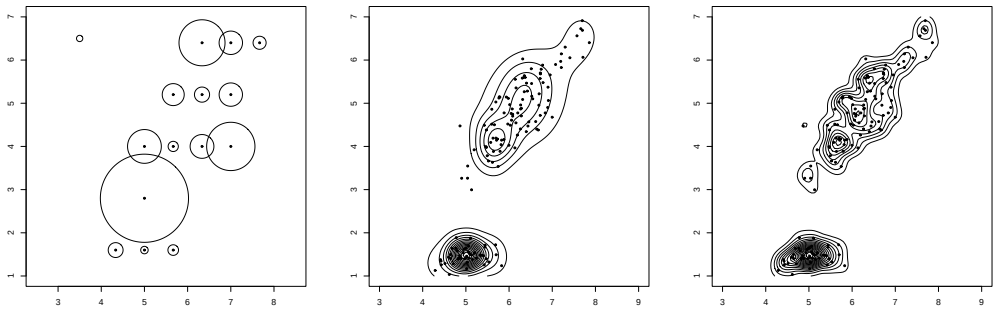

Fig. 9. Example using sepal and petal length for the iris data. Left frame shows ellipses representing the cross-validation estimated kernels with the middle frame the resulting density. The ellipse in the upper left corner of the left frame represents the fixed bandwidth kernel and the resulting estimate is in the right frame.

Bins in the corners with very low bin probabilities were excluded from the optimization.

Kernels near the mode (inside the unit circle) are nearly circular and are very similar, regardless of the parameterization. As the bins move further out into the tails, the size of the kernels get larger and the product kernels and fully parameterized kernels become more and more elliptical. As expected, the kernels for the product kernel estimator are circular on the diagonals and elliptical on the coordinate axis reflecting the nature of that particular restriction. As in Sain (2002), the MISE for the sample-point estimator with fully parameterized smoothing matrices is the smallest, followed by the product kernel formulation.

3.4.4 Estimating Bandwidth Matrices

Estimating variable bandwidth matrices from data continues to be a difficult problem. Sain (2002) outlines a cross-validation algorithm based on the binned estimator that involves finding the collection of bandwidth matrices that minimize

$$\mathrm{UCV} = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{x}_i) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j=1}^{m} n_{ij}^{*} K_{H_j}(\mathbf{x}_i - \mathbf{t}_j)\right]$$

where $n_{ij}^{*} = n_j - 1$ if $\mathbf{x}_i \in B_j$ or n_j, otherwise; here n_j is the number of data points in the jth bin, B_j. In practice, a parameterization for the bandwidth matrices is chosen, a mesh is laid down over the data, the bin counts computed, and the UCV criterion minimized.

An example is shown in Figure 9 for Fishers' iris data (sepal length and petal length). A simple parameterization is chosen using radially symmetric smoothing matrices and the left frame shows ellipses representing the estimated ker-

nels. For reference, the ellipse in the upper left-hand corner of the left frame is the cross-validated fixed bandwidth kernel. As expected, kernels in the tails and valleys between the modes are larger than near the modes. The variable bandwidth estimate is shown in the middle frame while the fixed bandwidth estimator is shown in the right frame. At first glance, the fixed-bandwidth estimate appears undersmoothed, possibly resulting from UCV's well-known tendency to pick bandwidths smaller than necessary (an improved estimate could possibly be found using, for example, the multivariate plug-in approach of Wand and Jones, 1994). However, the estimated bandwidth is clearly focusing on the mode in lower left of the frame (note the similarity between the fixed bandwidth and the variable bandwidths corresponding to this mode). This mode represents one of the species of iris present in the data and has a much smaller scale than the other modes in the data corresponding to the other two species. In contrast, the variable bandwidth estimate, despite being based on just a few bins, is clearly able to adapt to the changes in scale between the modes associated with the three species and does a much better job of simultaneously smoothing the different features in the density.

Sain (2002) further experimented with such methods and demonstrated the potential of even this simple formulation of a multivariate sample-point estimator, in particular for picking out important structure and minimizing the number of false modes. However, Sain (2002) also showed that UCV was not as effective when a fully parameterized bandwidth matrix is used. Hazelton (2003) has explored the product kernel formulation using not a piecewise constant bandwidth structure as in the binning case but a linearly interpolated bandwidth function with some promising results.

3.5 Other Estimators

Kernel estimators and orthogonal series density estimators were developed independently (Rosenblatt, 1956; Watson, 1969). It is well-known that an orthogonal series estimator can be re-expressed as a kernel estimator. However, cross-validation algorithms are somewhat different (Wahba, 1981) and spline estimators are also available. More recently, wavelet bases have become available and fall into this category (Donoho et al., 1996). Wahba pioneered splines for density estimators; however, her representation places knots at each sample point. Kooperberg and Stone (1991) describe an alternative spline formulation on a log-scale. A new spline tool called P-splines has recently emerged that like the ASH model greatly reduces computation; see Eilers and Marx (1996) and Ruppert, Carroll, and Wand (2003).

We remarked earlier that maximum likelihood had a role to play in the definition of histograms, but was limited in any role for defining smoothing parame-

ters. This situation has changed in recent years with the development of local likelihood methods for density estimation as well as regression estimation. This promising family of estimators is surveyed in Loader (1999).

4 Mixture Density Estimation

An alternative to the kernel estimator is the so-called mixture model, where the underlying density is assumed to have the form

$$g(\mathbf{x}) = \sum_{i=1}^{k} p_i g_i(\mathbf{x}; \boldsymbol{\theta}_i). \tag{21}$$

The $\{p_i,\ i = 1, \ldots, k\}$ are referred to as mixing proportions or weights and are constrained so that $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. The components of the mixture, $\{g_i,\ i = 1, \ldots, k\}$, are themselves densities and are parameterized by $\boldsymbol{\theta}_i$ which may be vector valued. Often, the g_i are taken to be multivariate normal, in which case $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \Sigma_i\}$.

Mixture models are often motivated by heterogeneity or the presence of distinct subpopulations in observed data. For example, one of the earliest applications of a mixture model (Pearson, 1894; see also McLachlan and Peel, 2000) used a two-component mixture to model the distribution of the ratio between measurements of forehead and body length on crabs. This simple mixture was effective at modeling the skewness in the distribution and it was hypothesized that the two-component structure was related to the possibility of this particular population of crabs evolving into two new subspecies.

The notion that each component of a mixture is representative of a particular subpopulation in the data has led to the extensive use of mixtures in the context of clustering and discriminant analysis. See, for example, the reviews by Fraley and Raftery (2002) and McLachlan and Peel (2000). It was also the motivation for the development of the multivariate outlier test of Wang et al. (1997) and Sain et al. (1999), who were interested in distinguishing nuclear tests from a background population consisting of different types of earthquakes, mining blasts, and other causes.

Often, a mixture model fit to data will have more components than can be identified with the distinct groups present in the data. This is due to the flexibility of mixture models to represent features in the density that are not well-modeled by a single component. Marron and Wand (1992), for example, give a wide variety of univariate densities (skewed, multimodel, e.g.) that are constructed from normal mixtures. It is precisely this flexibility that makes

mixture models attractive for general density estimation and exploratory analysis.

When the number of components in a mixture is pre-specified based on some a priori knowledge about the nature of the subpopulations in the data, mixtures can be considered a type of parametric model. However, if this restriction on the number of components is removed, mixtures behave in nonparametric fashion. The number of components acts something like a smoothing parameter. Smaller numbers of components will behave more like parametric models and can lead to specification bias. Greater flexibility can be obtained by letting the number of components grow, although too many components can lead to overfitting and excessive variation. A parametric model is at one end of this spectrum, and a kernel estimator is at the other end. For example, a kernel estimator with a normal kernel can be simply considered a mixture model with weights taken to be $1/n$ and component means fixed at the data points.

4.1 Fitting Mixture Models

While mixture models have a long history, fitting mixture models was problematic until the advent of the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). Framing mixture models as a missing data problem has made parameter estimation much easier, and maximum likelihood via the EM algorithm has dominated the literature on fitting mixture models. However, Scott (2001) has also had considerable success using the L_2E method, which performs well even if the assumed number of components, k, is too small.

The missing data framework assumes that each random vector $\mathbf{X}_i$ generating from the density (21) is accompanied by a categorical random variable Z_i where Z_i indicates the component from which $\mathbf{X}_i$ comes. In other words, Z_i is a single-trial multinomial with cell probabilities given by the mixing proportions $\{p_i\}$. Then, the density of $\mathbf{X}_i$ given Z_i is g_i in (21). It is precisely the realized values of the Z_i that are typically considered missing when fitting mixture models, although it is possible to also consider missing components in the $\mathbf{X}_i$ as well.

For a fixed number of components, the EM algorithm is iterative in nature and has two steps in each iteration. The algorithm starts with initial parameter estimates. Often, computing these initial parameter estimates involves some sort of clustering of the data, such as a simple hierarchical approach.

The first step at each iteration is the expectation step which involves prediction and effectively replaces the missing values with their conditional expectation given the data and the current parameter estimates. The next step, the max-

imization step, involves recomputing the estimates using both complete data and the predictions from the expectation step.

For normal mixtures missing only the realized component labels z_i, this involves computing in the expectation step

$$w_{ij} = \frac{\hat{p}_j^0 f_j(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_j^0, \hat{\Sigma}_j^0)}{\sum_{j=1}^k \hat{p}_j^0 f_j(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_j^0, \hat{\Sigma}_j^0)} \tag{22}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ and where $\hat{p}_j^0$, $\hat{\boldsymbol{\mu}}_j^0$, and $\hat{\Sigma}_j^0$ are the current parameter estimates. The maximization step then updates the sufficient statistics

$$T_{j1} = \sum_{i=1}^n w_{ij}\,; \quad \mathbf{T}_{j2} = \sum_{i=1}^n w_{ij}\mathbf{x}_i\,; \quad \mathbf{T}_{j3} = \sum_{i=1}^n w_{ij}\mathbf{x}_i\mathbf{x}_i'$$

for $j = 1, \ldots, k$ to yield the new parameter estimates

$$\hat{p}_j^1 = T_{j1}/n\,; \quad \hat{\boldsymbol{\mu}}_j^1 = \mathbf{T}_{j2}/T_{j1}\,; \quad \hat{\Sigma}_j^1 = (\mathbf{T}_{j3} - \mathbf{T}_{j2}\mathbf{T}_{j2}'/T_{j1})/T_{j1}$$

for $j = 1, \ldots, k$. The process cycles between these two steps until some sort of convergence is obtained. The theory concerning the EM algorithm suggests that the likelihood is increased at each iteration. Hence, at convergence, a local maximum in the likelihood has been found.

A great deal of effort has been put forth to determine a data-based choice of the number of components in mixture models and many of these are summarized in McLachlan and Peel (2000). Traditional likelihood ratio tests have been examined but a breakdown in the regularity conditions have made implementation difficult. Bootstrapping and Bayesian approaches have also been studied. Other criterion such as Akaike's information criterion (AIC) have been put forth and studied in some detail. In many situations, however, it has been found that AIC tends to choose too many components. There is some criticism of the theoretical justification for AIC, since it violates the same regularity conditions as the likelihood ratio test. An alternative information criterion is the Bayesian information criterion (BIC) given by

$$\text{BIC} = -2\ell + d\log n$$

where ℓ is the maximized log-likelihood, d is the number of parameters in the model, and n is the sample size. While there are some regularity conditions for BIC that do not hold for mixture models, there is much empirical evidence that supports its use. For example, Roeder and Wasserman (1997) show that

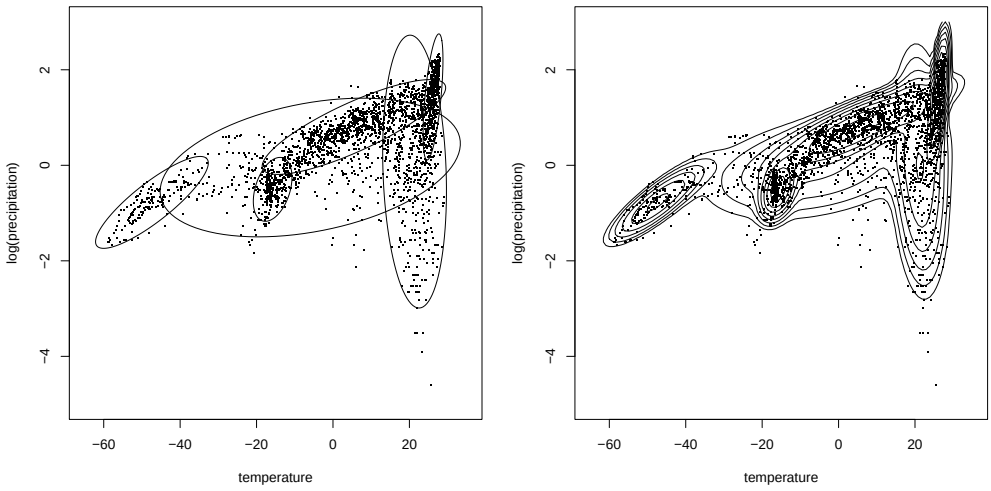

Fig. 10. Example using climate data. The best BIC model uses six components and these components are represented by ellipses in the left frame. The right frame shows a contour plot of the resulting density estimate.

when a normal mixture model is used as a nonparametric density estimate, the density estimate that uses the BIC choice of the number of components is consistent.

Other sophisticated procedures for choosing the number of components in mixture models have also been explored. For example, Priebe and Marchette (1991; 1993) and Priebe (1994) discuss what the authors' refer to as "adaptive mixtures" that incorporate the ideas behind both kernel estimators and mixture models and that use a data-based method for adding new terms to a mixture. The adaptive mixture approach can at times overfit data and Solka et al. (1998) combine adaptive mixtures with a pruning step to yield more parsimonious models. These methods have also been shown, both theoretically and through simulation and example, to be effective at determining the underlying structure in the data.

4.2 An Example

Figures 10 and 11 show an example of an application of a mixture model using bivariate data consisting of twenty-year averages of temperature and precipitation measured globally on a 5° grid (Covey et al., 2003; Wigely, 2003). An initial scatterplot of the measurements shows clearly the presence of multiple groupings in the data. It is hypothesized that this multimodality can be attributed to climatic effects as well as latitude and land masses across the globe.

A sequence of multivariate normal mixture models was fit to the data using

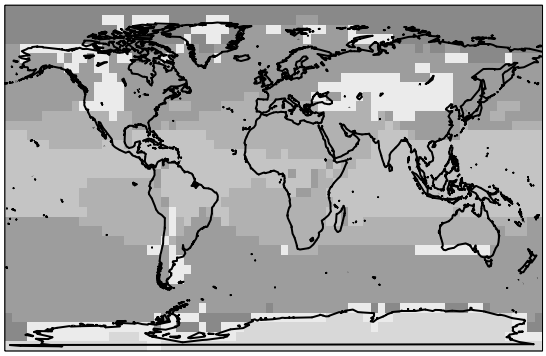

Fig. 11. An image plot displaying the results of the clustering based on the mixture estimate. The effects of land masses and latitude are clearly present in the clusters.

various numbers of components. BIC suggested a six component model. The ellipses in the left frame of Figure 10 indicate location and orientation of the individual components while the right frame shows the contours of the resulting density overlaid on the data.

It seems clear from the contour plot that some components are present to model non-normal behavior in the density. However, Figure 11 shows the result of classifying each observation as coming from one of the six components. This is done by examining the posterior probabilities as given by the w_{ij} in (22) at the end of the EM iterations. The groupings in the data do appear to follow latitude lines as well as the land masses across the globe.

5 Visualization of Densities

The power of nonparametric curve estimation is in the representation of multivariate relationships. While univariate density estimates are certainly useful, the visualization of densities in two, three, and four dimensions offers greater potential in an exploratory context for feature discovery. Visualization techniques are described here.

We examine the zip code data described by Le Cun et al. (1990). Handwritten digits scanned from USPS mail were normalized into 16×16 grayscale images. Training and testing data (available at the U.C. Irvine data repository) were combined into one data set, and the digits 1, 3, 7, and 8 were extracted for analysis here (1269, 824, 792, and 708 cases, respectively). We selected these digits to have examples of straight lines (1 and 7) as well as curved digits (3 and 8). In Figure 12, some examples of the digits together with summary statistics are displayed. Typical error rates observed classifying these data are high, in the 2.5% range.

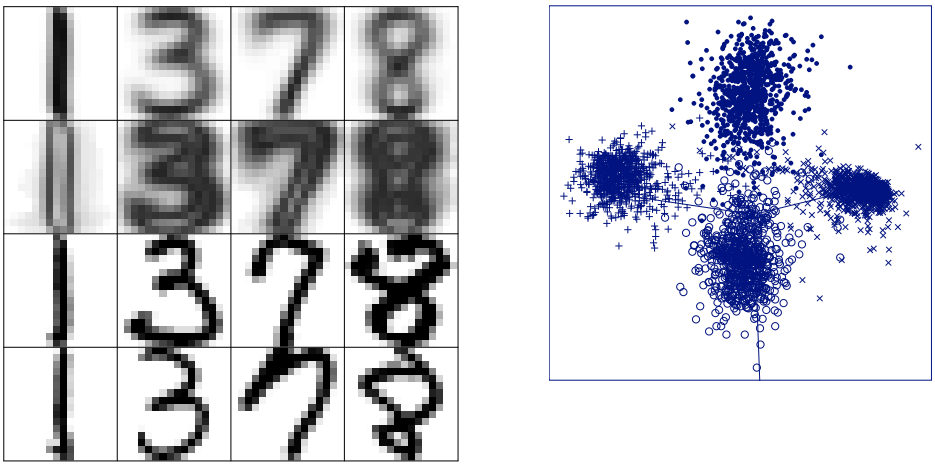

Fig. 12. (Left) Mean, standard deviation, and examples of zip code digits 1, 3, 7, and 8. (Right) LDA subspace of zip code digits 1 (×), 3 (•), 7 (+), and 8 (O).

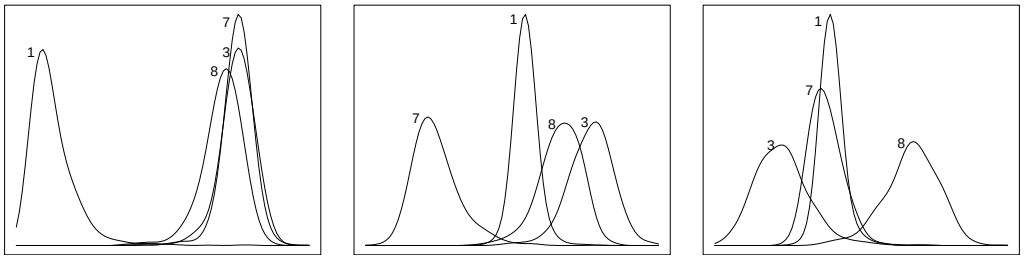

Fig. 13. ASH's for each of the 4 digits for the 1st, 2nd, and 3rd LDA variable (L-R).

To analyze and visualize these data, we computed the Fisher linear discriminant analysis (LDA) subspace. We sphered the data using a pooled covariance estimate, and computed the LDA subspace as the three-dimensional span of the four group means. The right frame of Figure 12 displays a frame from xgobi (Swayne, Cook, and Buja, 1991) and shows that the four groups are reasonably well-defined and separated in the LDA variable space.

If we examine averaged shifted histograms of each digit for each of the three LDA variables separately, we observe that the first LDA variable separates out digit 1 from the others; see the left frame Figure 13. In the middle frame, the second LDA variable separates digit 7 from digits 3 and 8. Finally, in the right frame, the third LDA variable almost completely separates digits 3 and 8 from each other (but not from the others).

We can obtain a less fragmented view of the feature space by looking at pairs of the LDA variables. In Figures 14 and 15, averaged shifted histograms for each digit were computed separately and are plotted. Contours for each ASH were drawn at 10 equally-spaced levels. The left frame in Figure 14 reinforces the notion that the first two LDA variables isolate digits 1 and 7. Digits 3 and 8 are separated by the first and third LDA variables in the right frame of Figure 14; recall that digit 7 can be isolated using the second LDA variables. Interestingly, in Figure 15, all four digits are reasonably separated by the

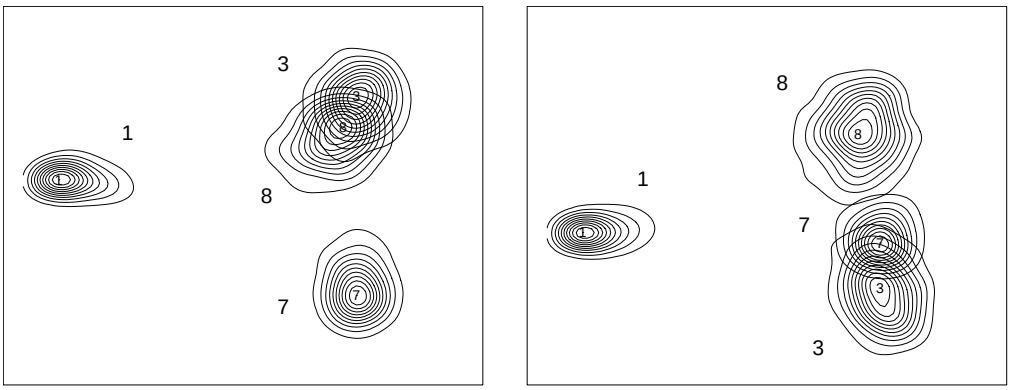

Fig. 14. Bivariate ASH's of the 4 digits using LDA variables (v_1, v_2) (left) and (v_1, v_3) (right) .

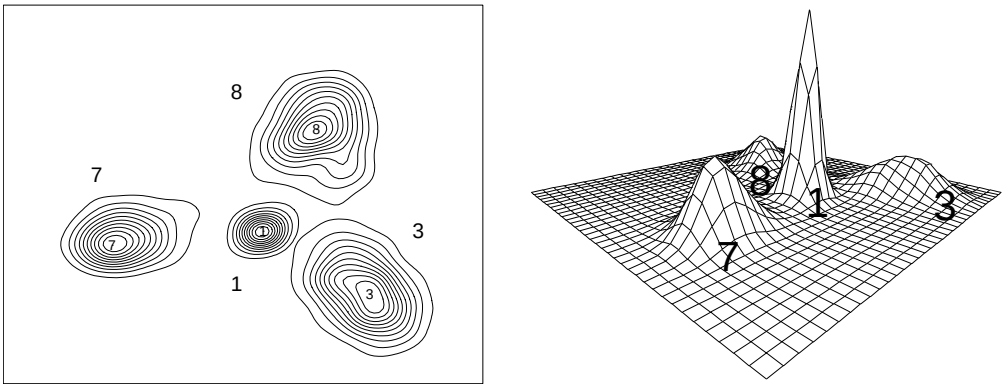

Fig. 15. ASH's for LDA variables (v_2, v_3).

second and third LDA variables alone. We also show a perspective plot of these ASH densities. (The perspective plot in Figure 14 does not display the full 100×100 mesh at this reduced size for clarity and to avoid overplotting the lines.)

Visualization of univariate and bivariate densities has become a fairly routine task in most modern statistical software packages. The figures in this chapter were generated using the Splus package on a Sun under the Solaris operating system. The ASH software is available for download at the ftp software link at author's homepage www.stat.rice.edu/∼scottdw. The ASH software contains separate routines for the univariate and bivariate cases. Visualization of the *ash1* and *ash2* estimates was accomplished using the built-in Splus functions *contour* and *persp*.

A separate function, *ashn*, is also included in the ASH package. The *ashn* function not only computes the ASH for dimensions $3 \leq d \leq 6$, but it also provides the capability to visualize arbitrary three-dimensional contours of a level set of any four-dimensional surface. In particular, if $f_{\max}$ is the maximum value of an ASH estimate, $\hat{f}(x, y, z)$, and α takes values in the interval $(0, 1)$,

then the α-th contour or level set is the surface

$$C_\alpha = \{(x, y, z) : \hat{f}(x, y, z) = \alpha\, f_{\max}\}\,.$$

The mode of the density corresponds to the choice $\alpha = 1$. The *ashn* function can compute the fraction of data within any specified α-contour.

Some simple examples of C_α contours may be given for normal data. If the covariance matrix $\Sigma = I_d$, then contours are spheres centered at μ:

$$C_\alpha = \{(x, y, z) : e^{-0.5((x-\mu_1)^2+(y-\mu_2)^2+(z-\mu_3)^2)} = \alpha\}$$

or $C_\alpha = \{(x, y, z) : (x - \mu_1)^2 + (y - \mu_2)^2 + (z - \mu_3)^2 = -2\log\alpha\}$. For a general covariance matrix, the levels sets are the ellipses $C_\alpha = \{(x, y, z) : (\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu) = -2\log\alpha\}$.

With a nonparametric density, the contours do not follow a simple parametric form and must be estimated from a matrix of values, usually on a regular three-dimensional mesh. This mesh is linearly interpolated, resulting in a large number of triangular mesh elements that are appropriately sorted and plotted in perspective. Since the triangular elements are contiguous, the resulting plot depicts a smooth contour surface. This algorithm is called marching cubes (Lorensen and Cline, 1987).

In Figure 16, a trivariate ASH is depicted for the data corresponding to digits 3, 7, and 8. (The digit 1 is well-separated and those data are omitted here.) The triweight kernel was selected with $m = 7$ shifts for each dimension. The contours shown correspond to the values $\alpha = 0.02$, 0.1, 0.2, 0.35, 0.5, 0.7, and 0.9. The *ashn* function also permits an ASH to be computed for each of the digits separated and plotted in one frame. For these data, the result is very similar to the surfaces shown in Figure 16.

This figure can be improved further by using stereo to provide depth of field, or through animation and rotation. The *ashn* software has an option to output this static figure in the so-called *QUAD* format used by the *geomview* visualization package from the previous NSF Geometry Center in Minneapolis. This software is still available from www.geomview.org and runs on SGI, Sun, and Linux platforms (Geomview, 1998).

5.1 Higher Dimensions

Scott (1992) describes extensions of the three-dimensional visualization idea to four dimensions or more. Here we consider just four-dimensional data,

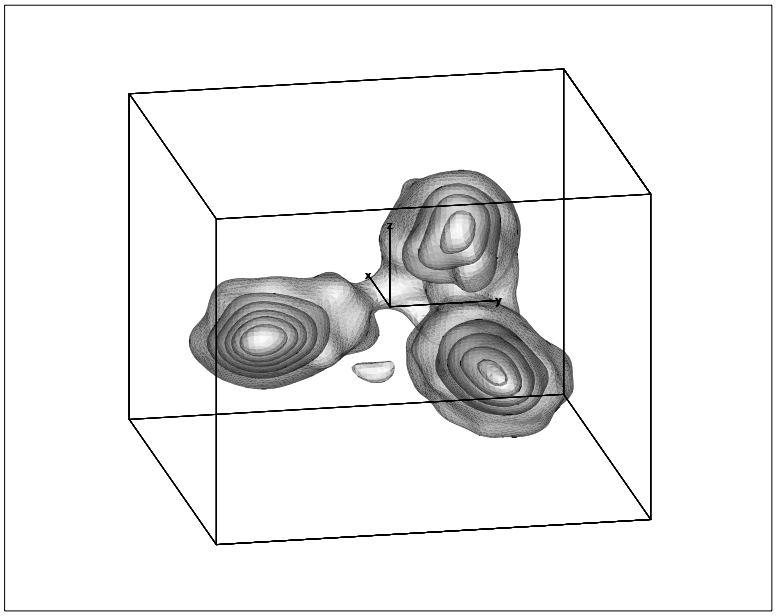

Fig. 16. Trivariate ASH of LDA variables (v_1, v_2, v_3) and digits 3, 7, and 8. The digit labels were not used in this plot. The digit 7 is in the left cluster; the digit 8 in the top cluster; and the digit 3 in the lower right cluster.

(x, y, z, t). The α-th contour is defined as above as

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

Since only a 3-dimensional field may be visualized, we propose to depict *slices* of the four-dimensional density. Choose a sequence of values of the fourth variable, $t_1 < t_2 < \cdots < t_m$, and visualize the sequence of slices

$$C_\alpha(k) = \{(x, y, z) : \hat{f}(x, y, z, t = t_k) = \alpha\, f_{\max}\} \qquad \text{for } k = 1, \ldots, m\,.$$

With practice, observing an animated view of this sequence of contours reveals the four-dimensional structure of the five-dimensional density surface. An important detail is that $f_{\max}$ is not recomputed for each slice, but remains the constant value of maximum of the entire estimate $\hat{f}(x, y, z, t)$. A possible alternative is viewing the conditional density, $\hat{f}(x, y, z|t = t_k)$; however, the renormalization destroys the perception of being in the low-density or tails of the distribution.

To make this idea more concrete, let us revisit the trivariate ASH depicted in Figure 16. This ASH was computed on a $75 \times 75 \times 75$ mesh. We propose as an alternative visualization of this ASH estimate $\hat{f}(x, y, z)$ to examine the sequence of slices

$$C_\alpha(k) = \{(x, y, z) : \hat{f}(x, y, z = z_k)\} \qquad \text{for } k = 1, \ldots, 75\,.$$

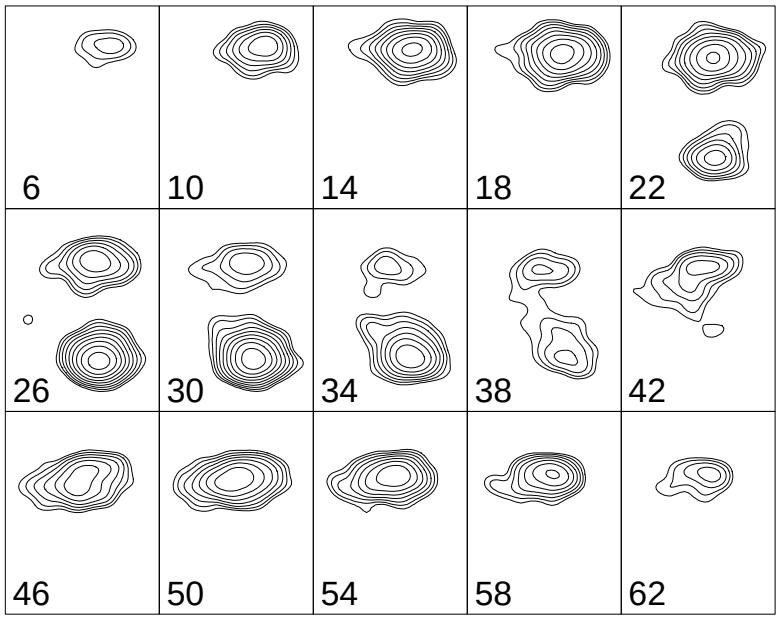

Fig. 17. A sequence of slices of the three-dimensional ASH of the digits 3, 7, and 8 depicted in Figure 16. The z-bin number is shown in each frame from the original 75 bins.

In Figure 17, we display a subset of this sequence of slices of the trivariate ASH estimate. For bins numbered less than 20, the digit 3 is solely represented. For bins between 22 and 38, the digit 7 is represented in the lower half of each frame. Finally, for bins between 42 and 62, the digit 8 is solely represented.

We postpone an actual example of this slicing technique for 4-dimensional data, since space is limited. Examples may be found in the color plates of Scott (1992). The extension to five-dimensional data is straightforward. The *ashn* package can visualize slices such as the contours

$$C_\alpha(k,\ell) = \{(x,y,z) : \hat{f}(x,y,z,t=t_k,s=s_\ell) = \alpha\, \hat{f}_{\max}\}\,.$$

Scott (1986) presented such a visualization of a five-dimensional dataset using an array of ASH slices on the competition data exposition at the Joint Statistical Meetings in 1986.

5.2 Curse of Dimensionality

As noted by many authors, kernel methods suffer from increased bias as the dimension increases. We believe the direct estimation of the full density by kernel methods is feasible in as many as six dimensions.

However, this does not mean that kernel methods are not useful in dimensions beyond six. Indeed, for purposes such as statistical discrimination, kernel

methods are powerful tools in dozens of dimensions. The reasons are somewhat subtle. Scott (1992) argued that if the smoothing parameter is very small, then comparing two kernel estimates at the same point $\mathbf{x}$ is essentially determined by the closest point in the training sample. It is well-known that the nearest-neighbor classification rule asymptotically achieves half of the optimal Bayesian misclassification rate. At the other extreme, if the smoothing parameter is very large, then comparing two kernel estimates at the same point $\mathbf{x}$ is essentially determined by which sample mean is closer for the two training samples. This is exactly what Fisher's LDA rule does in the LDA variable space. Thus, at the extremes, kernel density discriminate analysis mimics two well-known and successful algorithms. Thus there exist a number of choices for the smoothing parameter between the extremes that produce superior discriminate rules.

What is the explanation for the good performance for discrimination and the poor performance for density estimation? Friedman (1997) argued that the optimal smoothing parameter for kernel discrimination was much larger than for optimal density estimation. In retrospect, this result is not surprising. But it emphasizes how suboptimal density estimation can be useful for exploratory purposes and in special applications of nonparametric estimation.

6 Discussion

There are a number of useful references for the reader interested in pursuing these ideas and others not touched upon in this chapter. Early reviews of nonparametric estimators include Wegman (1972a, b) and Tarter and Kronmal (1976). General overviews of kernel methods and other nonparametric estimators include Tapia and Thompson (1978), Silverman (1986), Härdle (1990), Scott (1992), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Bowman and Azzalini (1997), Eubank (1999), Schimek(2000), and Devroye and Lugosi (2001).

Scott (1992) and Wegman and Luo (2002) discuss a number of issues with the visualization of multivariate densities. Classic books of general interest in visualization include Wegman and DePriest (1986), Cleveland (1993), Wolff and Yaeger (1993), and Wainer (1997).

Applications of nonparametric density estimation are nearly as varied as the field of statistics itself. Research challenges that remain include handling massive datasets and flexible modeling of high-dimensional data. Mixture and semiparametric models hold much promise in this direction.

References

[1] Abramson, I.: On Bandwidth Variation in Kernel Estimates-A Square Root Law. The Annals of Statistics, **10**, 1217–1223 (1982)

[2] Banfield, J.D. and Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics, **49**, 803–821 (1993)

[3] Bowman, A.W.: An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. Biometrika, **71**, 353–360 (1984).

[4] Bowman, A.W., and Azzalini, A.: Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford: (1997)

[5] Breiman, L., Meisel, W., and Purcell, E.: Variable kernel estimates of multivariate densities. Technometrics, **19**, 353–360 (1977)

[6] Chaudhuri, P. and Marron, J.S.: SiZer for exploration of structures in curves. Journal of the American Statistical Association, **94**, 807–823 (1999)

[7] Cleveland, W.S.: Visualizing Data. Hobart Press, Summit, NJ (1993)

[8] Covey, C., AchutaRao, K.M., Cubasch, U., Jones, P.D., Lambert, S.J., Mann. M.E., Phillips, T.J. and Taylor, K.E.: An overview of results from the Coupled Model Intercomparison Project (CMIP). Global and Planetary Change, **37**, 103–133 (2003)

[9] Dempster, A.P., Laird, N.M., and Rubin, D.B.: Maximum likelihood for incomplete data vi the EM algorithm (with discussion). Journal of the Royal Statistical Society, Ser. B, **39**, 1–38 (1977)

[10] Devroye, L. and Lugosi, T.: Variable kernel estimates: On the impossibility of tuning the parameters. In: Gine, E. and Mason, D. (ed) High-Dimensional Probability. Springer, New York (2000)

[11] Devroye, L. and Lugosi, T.: Combinatorial methods in density estimation. Springer-Verlag, Berlin (2001)

[12] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D.: Density estimation by wavelet thresholding. The Annals of Statistics, **24**, 508–539 (1996)

[13] Duin, R.P.W.: On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers, **25**, 1175–1178 (1976)

[14] Eilers, P.H.C. and Marx, B.D.: Flexible smoothing with B-splines and penalties. Statistical Science, **11**, 89–102 (1996)

[15] Eubank, R.L.: Nonparametric Regression and Spline Smoothing. Marcel Dekker, New York (1999)

[16] Fan, J. and Gijbels, I.: Local Polynomial Modelling and Its Applications. Chapman and Hall, London (1996)

[17] Fisher, R.A.: Statistical Methods for Research Workers, Fourth Edition. Oliver and Boyd, Edinburgh (1932)

[18] Fraley, C. and Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, **97**, 611–631 (2002)

[19] Friedman, J.H.: On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery, **1**, 55–77 (1997)
[20] Friedman, J.H. and Stuetzle, W.: Projection pursuit regression. Journal of the American Statistical Association, **76**, 817–823 (1981)
[21] Geomview (1998), http://www.geomview.org/docs/html.
[22] Graunt, J.: Natural and Political Observations Made upon the Bills of Mortality. Martyn, London (1662)
[23] Hall, P.: On near neighbor estimates of a multivariate density. Journal of Multivariate Analysis, **12**, 24–39 (1983)
[24] Hall, P.: On global properties of variable bandwidth density estimators. Annals of Statistics, **20**, 762–778 (1992)
[25] Hall, P., Hu, T.C., and Marron, J.S.: Improved variable window kernel estimates of probability densities. Annals of Statistics, **23**, 1–10 (1994)
[26] Hall, P. and Marron, J.S.: Variable window width kernel estimates. Probability Theory and Related Fields, **80**, 37–49 (1988)
[27] Härdle, W.: Smoothing Techniques with Implementations in S. Springer Verlag, Berlin (1990)
[28] Hart, J.D.: Efficiency of a kernel density estimator under an autoregressive dependence model. Journal of the American Statistical Association, **79**, 110–117 (1984)
[29] Hazelton, M.: Bandwidth selection for local density estimation. Scandinavian Journal of Statistics, **23**, 221–232 (1996)
[30] Hazelton, M.L.: Bias annihilating bandwidths for local density estimates. Statistics and Probability Letters, **38**, 305–309 (1998)
[31] Hazelton, M.L.: Adaptive smoothing in bivariate kernel density estimation. Manuscript (2003)
[32] Hearne, L.B. and Wegman, E.J.: Fast multidimensional density estimation based on random-width bins. Computing Science and Statistics, **26**, 150–155 (1994)
[33] Huber, P.J.: Projection pursuit (with discussion). Ann. Statist., **13**, 435–525 (1985)
[34] Jones, M.C.: Variable kernel density estimates and variable kernel density estimates. Australian Journal of Statistics, **32**, 361–372 (1990)
[35] Jones, M.C., Marron, J.S., and Sheather, S.J.: A Brief Survey of Bandwidth Selection for Density Estimation. Journal of the American Statistical Association, **91**, 401–407 (1996)
[36] Kanazawa, Y.: An optimal variable cell histogram based on the sample spacings. The Annals of Statistics, **20**, 291–304 (1992)
[37] Kogure, A.: Asymptotically optimal cells for a histogram. The Annals of Statistics, **15**, 1023-1030 (1987)
[38] Kooperberg, C. and Stone, C.J. (1991), "A Study of Logspline Density Estimation," *Comp. Stat. and Data Anal.*, **12**, 327–347.
[39] Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R. Hubbard, W., and Jackel, L.: Handwritten digit recognition with a back-propagation network. In: D. Touretzky (ed) Advances in Neural Information Processing

Systems, Vol. 2, Morgan Kaugman, Denver, CO (1990)
[40] Loader, C.: Local Regression and Likelihood. Springer, New York (1999)
[41] Loftsgaarden, D.O. and Quesenberry, C.P.: A nonparametric estimate of a multivariate density. Annals of Mathematical Statistics, **36**, 1049–1051 (1965)
[42] Lorensen, W.E. and Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. Computer Graphics, **21**, 163–169 (1987)
[43] Mack, Y. and Rosenblatt, M.: Multivariate k-nearest neighbor density estimates. Journal of Multivariate Analysis, **9**, 1–15 (1979)
[44] Marron, J.S. and Wand, M.P.: Exact mean integrated squared error. The Annals of Statistics, **20**, 712–536 (1992)
[45] McKay, I.J.: A note on the bias reduction in variable kernel density estimates. Canadian Journal of Statistics, **21**, 367–375 (1993)
[46] McLachlan, G. and Peel, D.: Finite Mixture Models. John Wiley, New York (2000)
[47] Minnotte, M.C.: Nonparametric testing of the existence of modes. The Annals of Statistics, **25**, 1646–1667 (1997)
[48] Minnotte, M.C. and Scott, D.W.: The mode tree: A tool for visualization of nonparametric density features. Journal of Computational and Graphical Statistics, **2**, 51–68 (1993)
[49] Pearson, K.: Contributions to the theory of mathematical evolution. Philosophical Transactions of the Royal Society of London, **185**, 72–110 (1894)
[50] Pearson, K.: On the systematic fitting of curves to observations and measurements. Biometrika, **1**, 265–303 (1902)
[51] Priebe, C.E.: Adaptive mixtures. Journal of the American Statistical Association, **89**, 796-806 (1994)
[52] Priebe, C.E. and Marchette, D.J.: Adaptive mixtures: Recursive nonparametric pattern recognition. Pattern Recognition, **24**, 1197–1209 (1991)
[53] Priebe, C.E. and Marchette, D.J.: Adaptive mixture density estimation. Pattern Recognition, **26**, 771–785 (1993)
[54] Roeder, K. and Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, **92**, 894–902 (1997)
[55] Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., **27**, 832–837 (1956)
[56] Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics, **9**, 65–78 (1982)
[57] Ruppert, D., Carroll, R.J., and Wand, M.P.: Semiparametric Regression. Cambridge University Press (2003)
[58] Sain, S.R.: Bias reduction and elimination with kernel estimators. Communications in Statistics: Theory and Methods, **30**, 1869–1888 (2001)
[59] Sain, S.R.: Multivariate locally adaptive density estimation. Computational Statistics & Data Analysis, **39**, 165–186 (2002)
[60] Sain, S.R.: A new characterization and estimation of the zero-bias band-

width. Australian & New Zealand Journal of Statistics, **45**, 29–42 (2003)
[61] Sain, S.R., Baggerly, K.A., and Scott, D.W.: Cross-Validation of Multivariate Densities. Journal of the American Statistical Association, **89**, 807–817 (1994)
[62] Sain, S.R., Gray, H.L., Woodward, W.A., and Fisk, M.D.: Outlier detection from a mixture distribution when training data are unlabeled. Bulletin of the Seismological Society of America, **89**, 294–304 (1999)
[63] Sain, S.R. and Scott, D.W.: On locally adaptive density estimation. Journal of the American Statistical Association, **91**, 1525–1534 (1996)
[64] Sain, S.R. and Scott, D.W.: "Zero-bias bandwidths for locally adaptive kernel density estimation. Scandinavian Journal of Statistics, **29**, 441–460 (2002)
[65] Schimek, M.G. (ed): Smoothing and Regression. Wiley, New York (2000)
[66] Scott, D.W.: On optimal and data-based histograms. Biometrika, **66**, 605–610 (1979)
[67] Scott, D.W.: On optimal and data-based frequency polygons. J. Amer. Statist. Assoc., **80**, 348–354 (1985a)
[68] Scott, D.W.: "Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. Ann. Statist., **13**, 1024–1040 (1985b)
[69] Scott, D.W.: Data exposition poster. 1986 Joint Statistical Meetings (1986)
[70] Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley, New York (1992)
[71] Scott, D.W.: Incorporating density estimation into other exploratory tools." Proceedings of the Statistical Graphics Section, ASA, Alexandria, VA, 28–35 (1995)
[72] Scott, D.W.: Parametric statistical modeling by minimum integrated square error. Technometrics, **43**, 274–285 (2001)
[73] Scott, D.W. and Wand, M.P.: Feasibility of multivariate density estimates. Biometrika, **78**, 197–205 (1991)
[74] Sheather, S.J. and Jones, M.C. (1991), "A Reliable Data-Based Bandwidth Selection Method For Kernel Density Estimation," *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.
[75] Silverman, B.W.: Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, Series B, **43**, 97–99 (1981)
[76] Silverman, B.W.: Algorithm AS176. Kernel density estimation using the fast Fourier transform. Appl. Statist., **31**, 93–99 (1982)
[77] Silverman, B.W. (1986), Density estimation for statistics and data analysis. Chapman & Hall Ltd, London (1986)
[78] Simonoff, J.S.: Smoothing Methods in Statistics. Springer Verlag, Berlin (1996)
[79] Solka, J.L., Wegman, E.J., Priebe, C.E., Poston, W.L., and Rogers, G.W.: Mixture structure analysis using the Akaike information criterion and the bootstrap. Statistics and Computing, **8**, 177–188 (1998)
[80] Student: On the probable error of a mean. Biometrika, **6**, 1–25 (1908)

[81] Sturges, H.A.: The choice of a class interval. J. Amer. Statist. Assoc., **21**, 65–66 (1926)
[82] Swayne, D., Cook, D. and Buja, A.: XGobi: Interactive dynamic graphics in the X window system with a link to S. ASA Proceedings of the Section on Statistical Graphics, ASA, Alexandria, VA, 1–8 (1991)
[83] Tapia, R.A. and Thompson, J.R.: Nonparametric Probability Density Estimation. Johns Hopkins University Press, Baltimore (1978)
[84] Tarter, E.M. and Kronmal, R.A.: An introduction to the implementation and theory of nonparametric density estimation. The American Statistician, **30**, 105-112 (1976)
[85] Terrell, G.R. and Scott, D.W.: Oversmoothed nonparametric density estimates. Journal of the American Statistical Association, **80**, 209–214 (1985)
[86] Terrell, G.R. and Scott, D.W.: Variable kernel density estimation. Annals of Statistics, **20**, 1236–1265 (1992)
[87] Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT (1983)
[88] Wahba, G.: "Data-based optimal smoothing of orthogonal series density estimates. Ann. Statist., **9**, 146–156 (1981)
[89] Wainer, H.: Visual Revelations. Springer-Verlag, New York (1997)
[90] Wand, M.P. and Jones, M.C. (1994), "Multivariate plug-in bandwidth selection," *Computational Statistics*, **9**, 97–116.
[91] Wand, M.P. and Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)
[92] Wang, S.J., Woodward, W.A., Gray, H.L., Wiechecki, S., and Sain, S.R.: A new test for outlier detection from a multivariate mixture distribution. Journal of Computational and Graphical Statistics, **6**, 285–299 (1997)
[93] Watson, G.S.: Density estimation by orthogonal series. Ann. Math. Statist., **40**, 1496–1498 (1969)
[94] Wegman, E.J.: Nonparametric probability density estimation I: A summary of available methods. Technometrics, **14**, 513–546 (1972)
[95] Wegman, E.J.: Nonparametric probability density estimation II: A comparison of density estimation methods. Journal of Statistical Computation and Simulation, **1**, 225–245 (1972)
[96] Wegman, E.J. and DePriest, D. (ed): Statistical Image Processing and Graphics. Marcel Dekker, New York (1986)
[97] Wegman, E.J. and Luo, Q.: On methods of computer graphics for visualizing densities. Journal of Computational and Graphical Statistics, **11**, 137–162 (2002)
[98] Wigley, T.: MAGICC/SCENGEN 4.1: Technical manual. (2003) http://www.cgd.ucar.edu/cas/wigley/magicc/index.html
[99] Wolff, R. and Yaeger, L.: Visualization of Natural Phenomena. Springer-Verlag, New York (1993)