The Annals of Statistics
1985, Vol. 13, No. 3, 1024–1040

AVERAGED SHIFTED HISTOGRAMS: EFFECTIVE NONPARAMETRIC DENSITY ESTIMATORS IN SEVERAL DIMENSIONS[1]

By David W. Scott

Rice University

We introduce two nonparametric multivariate density estimators that are particularly suitable for application in interactive computing environments. These estimators are statistically comparable to kernel methods and computationally comparable to histogram methods. Asymptotic theory of the estimators is presented and examples with univariate and simulated trivariate Gaussian data are illustrated.

1. Introduction.

In this paper we introduce two new nonparametric multivariate density estimators designed for data analysis in three and four dimensions, but useful in one and two dimensions as well. The basic construction in one dimension is easy to describe: form several histograms with equal bin widths but different bin locations and average these shifted histograms, hence the name averaged shifted histograms. In what follows we examine the need for these estimators, demonstrate their close relationship to kernel methods, investigate their asymptotic properties, and present some examples.

1.1 *Motivation for a new estimator.*

These new estimators were developed to analyze large multivariate data sets. One of our first applications involved LANDSAT IV remove sensing data. The data were known to be non-Gaussian, coming from images covering many fields of agricultural crops. Hence a nonparametric technique seemed appropriate. The number of samples was large and organized into units of 22,932 points—117 scan lines with 196 picture elements (pixels) per line, 1.1 acres per pixel. The LANDSAT IV satellite has a four-channel sensor so that the raw data are quadrivariate. Further, a NASA scientist had developed an agronomic growth model that nonlinearly transformed multiple images, usually five overpasses in a single growing season, into a trivariate data set; see Badhwar (1980). I had decided to represent these trivariate data by drawing contours of an estimated trivariate kernel density function on a color graphics terminal, having had success applying bivariate kernel density estimation to medical data (Scott et al., 1980); see Figure 3 for an example of the contour representation. Graphing the bivariate estimates had required only a minute or two of CPU with several hundred points. However, it quickly became apparent that several hours of CPU might be required for the trivariate kernel

Received March 1984; revised March 1985.

[1] Research supported by the Army Research Office, NASA/Johnson Space Center, and Office of Naval Research under DAAG-29-82-K0014, NCC 9-10, and N00014-85-K-0100, respectively.

AMS 1980 *subject classifications.* 62G05, 62E10.

Key words and phrases. Nonparametric density estimation, histograms, frequency polygons, kernel estimators, integrated mean squared error, binned data, multivariate data analysis.

estimates, making real time display and interactive adjustment of the three smoothing parameters almost impossible.

Improved computer graphics technology has encouraged the development of a variety of graphical techniques for the display for multivariate data; see Tukey and Tukey (1981) and Chambers et al. (1983). We believe density estimation has an important role in this field, particularly with data (and projected data) in three and four dimensions. The complexity of structure possible in data increases rapidly from one to four dimensions. High-dimensional structure presents problems well beyond the scope of this paper; by high-dimensional structure we mean density features that cannot be adequately described in any subspace of dimension less than five.

1.2 *Density estimation background.* A variety of nonparametric probability density estimators has been proposed and studied since the pioneering work on kernel methods by Rosenblatt (1956) and Parzen (1962); see Tapia and Thompson (1978). Existing estimators are not well-suited to modern data analysis that requires interactive computing for very large data sets with many variables. Most density estimation research has dealt with the one-dimensional case and applications to relatively small data sets. Multivariate extensions of histogram, kernel, and nearest neighbor density estimators have been studied theoretically and are usually applied to bivariate data. Kernel estimates with five or more variables have been reported in medical applications; see Habbema et al. (1974).

The histogram is extremely efficient computationally compared to kernel methods, but it is quite inefficient statistically. Relative sample sizes required for histograms having errors comparable to kernel estimates increase rapidly for increasing sample size and dimension. Recently the frequency polygon, which is formed by linear interpolation of adjacent mid-bin values of a histogram, was studied (Scott, 1985) as a bridge between these estimators, sharing the computational simplicity of the histogram and the same order of statistical efficiency as the kernel estimator. Thus, the frequency polygon can be useful for interactive graphics; however, there are two difficulties with its direct use on ordinary histograms for our purposes. First, for any sampling density satisfying the set of conditions (C2) given in Section 2.2, the univariate frequency polygon requires 27% more samples to achieve the same statistical efficiency as the optimal Epanechnikov kernel estimator with respect to integrated mean squared error (IMSE). The bivariate frequency polygon requires 51% more independent Gaussian samples than the bivariate Epanechnikov product kernel estimator. In several dimensions, the statistical inefficiency of the frequency polygon may be unacceptable for moderate sample sizes. The second problem deals with arbitrariness in the choice of bin edge locations. The bins of the optimal frequency polygon are wider than those for the optimal histogram. With large bins in several dimensions, bin edge effects are more pronounced and the choice of mesh position more important since estimators with different mesh origins may have different subjective features or number of modes.

For these reasons we propose two new density estimators: first, the averaged shifted histogram (ASH), which we will show has the computational efficiency of a histogram and approaches the statistical efficiency of a kernel estimator. Our second estimator is a frequency polygon of the averaged shifted histogram (FP-ASH), which we will show is functionally identical to a related interpolated kernel estimator of binned data.

2. The averaged shifted histogram.

2.1 *Construction and relationship with kernel methods.* Consider a histogram constructed over an equally spaced mesh with bin width h using a random sample $\{x_1, \cdots, x_n\}$ from the density $f(x)$. Without loss of generality, we shall assume that 0 is a bin edge. Similar histograms could be constructed by choosing a bin edge between 0 and h. Consider m such histograms of the same data set, each with bin width h, but with bin edges given by $\{ih/m, i = 0, \cdots, m-1\}$. Let this smaller (bin) width be denoted by $\delta = h/m$. Define the finer mesh $\{t_k\}$ by $t_k = k\delta$. Let the kth bin be denoted by $I_k = [t_{k-1}, t_k)$. Let n_k be the number of values from the random sample $\{x_1, \cdots, x_n\}$ falling in I_k. Then for $x \in I_k$, the value of the ith shifted histogram, denoted by $\hat{a}_i(x)$, is defined by

$$(2.1) \qquad \hat{a}_i(x) = (1/nh) \textstyle\sum_{j=0}^{m-1} n_{j+i+[(k-i)/m]m} \quad x \in I_k, \quad i = 0, \cdots, m-1,$$

where $[y]$ is the greatest integer less than or equal to y. The sum simply counts the number of points in the larger histogram bin of width $h = m\delta$.

Each of the m histograms defined in (2.1) is a reasonable choice for a histogram of the data with bin width h. Consider their pointwise average:

$$(2.2) \qquad \hat{f}(x) = (1/m) \textstyle\sum_{i=0}^{m-1} \hat{a}_i(x) \quad x \in I_k.$$

An equivalent form is

$$(2.3) \qquad \hat{f}(x) = (1/mnh) \textstyle\sum_{i=1-m}^{m-1} (m - |i|) n_{k+i} \quad x \in I_k.$$

The behavior of this estimator for increasing m is clear if we rewrite

$$(2.4) \qquad \hat{f}(x) = (1/nh) \textstyle\sum_{i=1-m}^{m-1} (1 - (|i|/m)) n_{k+i}$$

and recall that the kernel density estimator is given by:

$$(2.5) \qquad \hat{f}_K(x) = (1/nh) \textstyle\sum_{i=1}^{n} K((x - x_i)/h).$$

As $m \to \infty$, (2.4) converges to (2.5) for the particular kernel

$$(2.6) \qquad K(t) = (1 - |t|) I_{[-1,1]}(t).$$

Thus the ASH with uniform weighting on the shifted histograms approximates a triangle kernel estimator. We may consider a general bin weighting function $w_m(i)$ rather than $(1 - |i|/m))$ in (2.4) that satisfies

$$(2.7) \qquad w_m(-i) = w_m(i), \quad \textstyle\sum_{i=1-m}^{m-1} w_m(i)/m) = 1, \quad w_m(i) \geq 0.$$

The second condition insures that $\hat{f}(x)$ integrates to one and the third that $\hat{f}(x)$ is nonnegative. Each function $w_m(i)$ must correspond to a finite support kernel satisfying the set of conditions (C3) given in Section 2.2.

The extension of the averaged shifted histogram to higher dimensions is straightforward. Let h_j be the histogram bin width in the jth of p dimensions and suppose we shift m_j times in the jth dimension. Then, for example, the ASH estimator for trivariate data $p = 3$ is

$$\hat{f}(\underset{\sim}{x}) = (1/nh_1h_2h_3) \sum_{i_1} \sum_{i_2} \sum_{i_3} w(i_1, i_2, i_3) n_{k_1+i_1, k_2+i_2, k_3+i_3},$$

for $\underset{\sim}{x}$ in bin I_{k_1,k_2,k_3} and where the sum over i_j goes from $1 - m_j$ to $m_j - 1$ with

$$w(i_1, i_2, i_3) = (1 - (|i_1|/m_1))(1 - (|i_2|/m_2))(1 - (|i_3|/m_3)).$$

As $m_i \to \infty$ the multivariate ASH converges to a product triangle kernel estimator. Again other weighting functions may be constructed.

2.2 *Regularity conditions.* We are interested in conditions that are sufficient to bound remainder terms in an IMSE expression of a density estimator. Such conditions are usually stronger than those required just for consistency. For convenience, the "roughness" or squared L_2 norm of a function ϕ will be denoted by

$$R(\phi) = \int_{-\infty}^{\infty} \phi(x)^2 \, dx.$$

For a histogram, let $S(f)$ be the support of f and suppose that $S(f)$ is the union of equal-width bins. Then a set of sufficient conditions is (Freedman and Diaconis, 1981):

(C1): f' absolutely continuous on $S(f)$; $\int_{S(f)} f''(x)^2 < \infty$; $\int_{S(f)} f'(x)^2 \, dx > 0$.

For a frequency polygon, a set of sufficient conditions is (Scott, 1985):

(C2): f'' absolutely continuous on $(-\infty, \infty)$; $f''' \in L_2$ or $R(f''') < \infty$.

For a finite-support nonnegative kernel estimator, we require (see the Appendix):

(C3): conditions (C2), and $K \geq 0$; $K \in L_2$; K continuous on the interior of its support (a, b); $\int K(x) \, dx = 1$; $\mu_K = 0$ and $\sigma_K^2 > 0$,

where μ_K and σ_K^2 are the first two moments of the kernel, which is itself a density. Notice that if $f^{(k)}(x) \in L_2$ then $f, f^{(1)}, \cdots, f^{(k-1)} \in L_2$. This follows from a result in Rosenblatt (1971), in which conditions similar to (C3) are given.

3. Integrated mean squared error of the averaged shifted histogram. We shall evaluate the performance of the one-dimensional ASH by computing its integrated mean squared error (IMSE) as a function of both the bin width h and the shifting parameter m. Walter and Blum (1979) have shown that the ordinary histogram is a kernel estimator and have given the explicit form of the kernel. Similarly, the averaged shifted histogram is a kernel estimator. The exact form of the kernel is not of immediate interest here, but we note that the kernel is piecewise constant and has finite support. Theorem 5, which is stated in the Appendix and generalizes Walter and Blum's pointwise results, shows that conditions (C3) are sufficient to obtain the usual kernel IMSE expression and to insure that the remainder term is bounded. In order to apply Theorem 5 to the averaged shifted histogram, we require two modifications in the proof. First, the ASH kernel is not continuous but only piecewise continuous. Thus when computing the bias, we have many equations like (A.3), one for each interval where the kernel is constant. Adding these equations gives a result very similar to (A.5) after using a convexity argument, but with a constant different from σ_K^2. Second, additional bias terms arise because $\mu_K \neq 0$ for the ASH kernel. Thus Theorem 5 may be used to guarantee that the remainder terms in the IMSE of the ASH are well-behaved but does not provide a constructive method for determining the constants in the leading terms in the IMSE. This we do in a straightforward manner in Sections 3.1 and 3.2.

We obtain the leading terms in the IMSE by finding the mean squared error, the sum of the variance and squared bias, at every point in a typical bin, integrating over each bin, and finally combining all bins over the real line. For convenience in our Taylor's series analysis, we examine a bin centered on 0 rather than bordering on 0; hence, in this section we define bin $I_i = [(i - \tfrac{1}{2})\delta, (i + \tfrac{1}{2})\delta)$. The number of points n_i falling in bin I_i is Binomial, $B(n, p_i)$, where $p_i = \int_{I_i} f(t)\,dt$. We shall prove the following theorem:

THEOREM 1. *Assume that the sampling density f satisfies the set of regularity conditions* (C2). *Then the integrated mean squared error of the averaged shifted histogram estimator* (2.4) *with bin width h and shift parameter m is given by*

$$\begin{aligned} IMSE = {} & \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) - \frac{1}{n}R(f) + \frac{h^2}{12m^2}R(f') \\ & + \frac{1}{144}h^4\left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right)R(f'') + O\left(\frac{h}{n} + h^5\right), \end{aligned} \tag{3.1}$$

which, for sufficiently large n and ignoring terms of order n^{-1}, is minimized by choosing $m = \infty$ and smoothing parameter h as for a triangle kernel estimator.

3.1 *Bias of the averaged shifted histogram.* For convenience, we analyze the bias in bin I_0, which is centered about 0. Since $En_i = np_i$,

$$E\hat{f}(x) = (1/mh)\textstyle\sum_{i=1-m}^{m-1}(m - |i|)p_i \quad x \in I_0.$$

Taking a Taylor's series of p_i about x, we find

$$\begin{aligned} E\hat{f}(x) = {} & f(x) - xf'(x) + \left[\frac{x^2}{2} + \frac{\delta^2}{12}\left(m^2 - \frac{1}{2}\right)\right]f''(x) \\ & - \left[\frac{x^3}{6} + \frac{x\delta^2}{12}\left(m^2 - \frac{1}{2}\right)\right]f'''(x) + o(\delta^3). \end{aligned}$$

The squared bias at x has four leading terms: using approximations indicated by

$$\int_{-\delta/2}^{\delta/2} x^2 f'(x)^2\,dx \approx f'(0)^2 \int_{-\delta/2}^{\delta/2} x^2\,dx = \frac{\delta^3}{12}f'(0)^2 \approx \frac{\delta^2}{12}\int_{-\delta/2}^{\delta/2} f'(x)^2,$$

the total integrated and squared bias for bin I_0 is

$$\int_{-\delta/2}^{\delta/2} [E\hat{f}(x) - f(x)]^2\, dx$$

$$(3.2)\qquad = \frac{\delta^2}{12}\int_{I_0} f'(x)^2\, dx + \delta^4\left[\frac{1}{720} + \frac{m^4}{144}\right]\int_{I_0} f''(x)^2\, dx + \delta^4\left[\frac{m^2}{72} - \frac{1}{360}\right]\int_{I_0} f'(x)f'''(x)\, dx + O(\delta^6).$$

Expression (3.2) is valid for every bin I_k. Summing over all bins and noting $\int f'(x)f'''(x)\, dx = -\int f''(x)^2\, dx$ we obtain

$$(3.3)\qquad \int_{-\infty}^{\infty} \mathrm{Bias}(x)^2\, dx = \frac{h^2}{12m^2} R(f') + \frac{h^4}{144}\left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right) R(f'') + O(h^5).$$

We may check (3.3) for two extreme cases. When $m = 1$, the ASH is the ordinary histogram and the bias in (3.3) is dominated by $h^2R(f')/12$, the same expression obtained by Scott (1979). When $m = \infty$, the ASH becomes a kernel estimator with kernel (2.6). Parzen (1962) has shown that the leading bias term for the kernel estimator is

$$(3.4)\qquad \tfrac{1}{4}h^4\sigma_K^4 R(f'').$$

For kernel (2.6), $\sigma_K^2 = \tfrac{1}{6}$, and it follows that (3.3) and (3.4) are equal.

For relatively small values of m, we can essentially eliminate the portion of the bias in (3.3) due to binning. For moderate sample sizes, m greater than 5 or 10 seems sufficient.

3.2 *Integrated variance.* The variance in bin I_0 of the ASH, which is given by (2.4) with $k = 0$, may be computed by noting $\mathrm{Var}(n_i) = np_i(1 - p_i)$ and $\mathrm{Cov}(n_i, n_j) = -np_ip_j$:

$$\mathrm{Var}\,\hat{f}(x) = (1/m^3n\delta h)\textstyle\sum_{i=1-m}^{m-1}(m - |i|)^2p_i - (1/m^4n\delta^2)\{\sum_{i=1-m}^{m-1}(m - |i|)p_i\}^2.$$

Substituting $\delta f(0)$ for p_i and summing we obtain

$$(3.5)\qquad \mathrm{Var}\,\hat{f}(x) = (2f(0)/3nh)(1 + (1/2m^2)) - (f(0)^2/n) + O(h/n).$$

Equation (3.5) is valid in bin I_k if we replace 0 by t_k. Integrating over I_0 multiplies (3.5) by δ. Summing over all bins and using midpoint numerical quadrature, we obtain

$$(3.6)\qquad \int_{-\infty}^{\infty} \mathrm{Var}\,\hat{f}(x)\, dx = \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) - \frac{1}{n}R(f) + O\left(\frac{h}{n}\right).$$

For the ordinary histogram, Scott (1979) proved that the integrated variance is $1/(nh)$, with which (3.6) agrees when $m = 1$. Parzen (1962) showed the integrated variance for the kernel estimator is dominated by $R(K)/nh$. Now

$R(K) = 2/3$ for the triangle kernel; hence (3.6) agrees with Parzen's result when $m = \infty$. Finally, adding (3.3) and (3.6), we have proven the theorem.

4. Frequency polygon ASH estimator. The ASH requires the specification of not only the smoothing parameter h but also the shift parameter m. The interplay of these two parameters is highly nonlinear. Although our theorem suggests choosing $m = \infty$ to eliminate the $O(h^2)$ term, we would lose computational efficiency. The ASH estimator, when graphed, looks like an ordinary histogram with bin width δ rather than $h = m\delta$. For graphical reasons, the discontinuity of the ASH is not appealing in more than one dimension. These drawbacks can be eliminated by constructing a frequency polygon of the ASH, an estimator we will denote by $\hat{g}(x)$. The FP-ASH is well-suited for graphics hardware because all continuous functions are drawn as piecewise linear functions. Multivariate frequency polygons of p variables connect $p + 1$ adjacent histogram mid-bin values (above a p-simplex basis element) with a hyperplane. In the bivariate case, take two adjacent triangular (2-simplex) basis elements, which form a rectangle (with one diagonal) defined by four adjacent histogram bin centers, and reflect (not translate) this rectangle about its sides to cover the plane with basis elements; see Figure 1. With p variables, triangulate the rectangle formed by $2p$ adjacent bin centers to obtain $p!$ p-simplex basis elements. Again, extend by reflecting the rectangle.

4.1 *Integrated mean squared error.* The FP-ASH is also a kernel estimator. The kernel has finite support, zero mean, and is a continuous piecewise linear

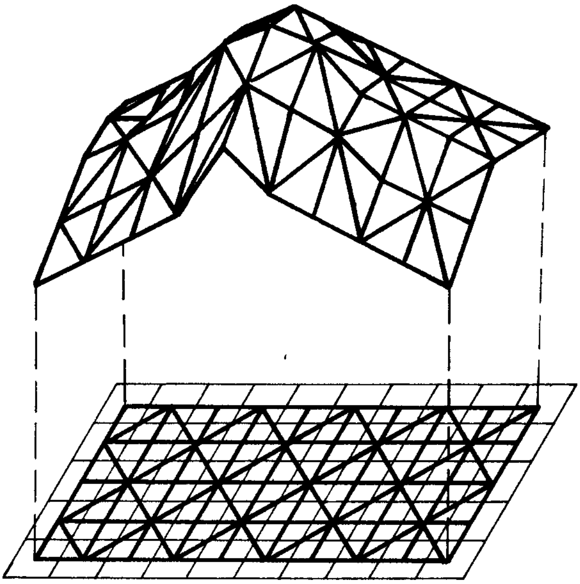

FIG. 1. *Illustration of the construction of a bivariate frequency polygon ASH surface. In the plane are shown the rectangular histogram bins (thin lines) and triangular basis elements (thick lines) formed by projecting the frequency polygon edges onto the plane (see text).*

function (Terrell, 1984). Thus Theorem 5 is directly applicable. We find it easier to proceed as in Section 3, computing the pointwise bias and variance in a typical bin and combining to obtain the IMSE. In this section, we return to our original bin mesh definition given in Section 2.1 with $I_k = [(k-1)\delta, k\delta)$. We first compute the IMSE over the interval $[-\delta/2, \delta/2]$. The FP-ASH is given by

$$\hat{g}(x) = (1/2 - x/\delta)\hat{f}_0 + (1/2 + x/\delta)\hat{f}_1, \quad x \in [-\delta/2, \delta/2] \tag{4.1}$$

where $\hat{f}_0$ and $\hat{f}_1$ are the values of the ASH in bins I_0 and I_1, respectively. Taking a Taylor's expansion of p_i about x, we find

$$E\hat{g}(x) = f(x) + [(\delta^2/12)(1 + m^2) - (x^2/2)]f''(x) + o(\delta^2).$$

Integrating the squared bias over $[-\delta/2, \delta/2]$ as in equation (3.2), we obtain

$$\int_{-\delta/2}^{\delta/2} \operatorname{Bias}(x)^2\, dx = \frac{1}{2880}(20m^4 + 20m^2 + 9)\delta^4 \int_{-\delta/2}^{\delta/2} f''(x)^2\, dx + O(\delta^6). \tag{4.2}$$

Now an expression similar to (4.2) is valid between the midpoints of bins I_k and I_{k+1}. Summing over all such intervals, we obtain

$$\int_{-\infty}^{\infty} \operatorname{Bias}(x)^2\, dx = \frac{1}{144}\left(1 + \frac{1}{m^2} + \frac{9}{20m^4}\right)h^4 R(f'') + O(h^5). \tag{4.3}$$

For the ordinary frequency polygon $m = 1$, Scott (1985) obtained 49/2880 as the constant in the leading bias term, which checks. As $m \to \infty$ we again approach the triangle kernel estimate and the bias term (4.3) agrees with (3.4).

To compute the variance in $[-\delta/2, \delta/2]$, we rewrite (4.1) as

$$g(x) = (1/mnh)\{\textstyle\sum_{i=1-m}^{m}[(m - |i|)(1/2 - x/\delta) + (m - |i-1|)(1/2 + x/\delta)]n_i\}.$$

A computation similar to that in Section 3.2 reveals

$$\int_{-\infty}^{\infty} \operatorname{Var} \hat{g}(x)\, dx = \frac{2}{3nh} - \frac{1}{n}R(f) + O\left(\frac{h}{n}\right). \tag{4.4}$$

Thus the variance term is essentially independent of m and is the same as for the ordinary frequency polygon. Combining (4.3) and (4.4) we have proven:

THEOREM 2. *Under the assumptions of Theorem 1, the integrated mean squared error of the frequency polygon of the averaged shifted histogram is given by*

$$\begin{aligned} IMSE &= (2/3nh) - (1/n)R(f) \\ &\quad + (1/144)h^4(1 + (1/m^2) + (9/20m^4))R(f'') + O(h/n + h^5). \end{aligned} \tag{4.5}$$

4.2 *FP-ASH vs. ASH.* Comparing (4.5) and (3.1), we see that we have eliminated the $O(h^2/m^2) = O(\delta^2)$ term from the bias and the $O(1/nm^2h)$ binning term from the variance. The remaining bias terms involving m are easily controlled with modest values of m, values smaller than required by the ASH. Next we evaluate statistical efficiency. It is easy to see from equation (A.1) that if we ignore terms of order n^{-1} and minimize the IMSE with respect to h, the relative contribution to the optimal IMSE of the variance term is four times the bias term, for all densities f. Therefore, if in (4.5) we use this same h, which is a slightly suboptimal choice, the IMSE of the FP-ASH is greater by the factor $1 + (1/5\mathrm{m}^2) + (9/100\mathrm{m}^4)$. For $m = 1$ the increase is 29% (compared to the optimal 27% given in Section 1.2), but for $m \geq 5$ the increase is less than 1%. Similar calculations for the ASH may be done for specific densities. For example, with 100 Gaussian data points, the increase is 164% for $m = 1$, but not until $m \geq 13$ does the increase fall below 1%.

5. Bivariate results. Bivariate IMSE results are given below. Explicit multivariate IMSE results for the ASH and FP-ASH are not generally available, although, for large m_i, we may use the fact that these estimators converge to the product kernel estimator. The bivariate formulae parallel the univariate results (3.1) and (4.5) in interesting ways and indicate that the choices of the parameters m_x and m_y can be about the same as for m in the univariate case. Before stating the theorems, let us introduce the following notation:

$$I_{ij,kl} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial^{i+j} f(x, y)}{\partial x^i \partial y^j} \times \frac{\partial^{k+l} f(x, y)}{\partial x^k \partial y^l}\, dx\, dy.$$

THEOREM 3. *Asymptotically, for the bivariate ASH,*

$$\begin{aligned} IMSE = {} & \frac{4}{9nh_xh_y}\left(1 + \frac{1}{2m_x^2}\right)\left(1 + \frac{1}{2m_y^2}\right) - \frac{1}{n} I_{00,00} + \frac{1}{12}\frac{h_x^2}{m_x^2} I_{10,10} \\ & + \frac{1}{12}\frac{h_y^2}{m_y^2} I_{01,01} + \frac{1}{144} h_x^4\left(1 - \frac{2}{m_x^2} + \frac{3}{5m_x^4}\right) I_{20,20} \\ & + \frac{1}{144} h_y^4\left(1 - \frac{2}{m_y^2} + \frac{3}{5m_y^4}\right) I_{02,02} \\ & + \frac{1}{72} h_x^2 h_y^2\left(1 - \frac{1}{m_x^2} - \frac{1}{m_y^2} + \frac{1}{2m_x^2m_y^2}\right) I_{20,02} + o\left(\frac{1}{n}\right). \end{aligned}$$

THEOREM 4. *Asymptotically, for the bivariate FP-ASH,*

$$\begin{aligned} IMSE = {} & \frac{4}{9nh_xh_y}\left(1 + \frac{1}{8m_x^2m_y^2}\right) - \frac{1}{n} I_{00,00} \\ & + \frac{1}{144} h_x^4\left(1 + \frac{1}{m_x^2} + \frac{9}{20m_x^4}\right) I_{20,20} + \frac{1}{144} h_y^4\left(1 + \frac{1}{m_y^2} + \frac{9}{20m_y^4}\right) I_{02,02} \\ & + \frac{1}{72} h_x^2h_y^2\left(1 + \frac{1}{2m_x^2}\right)\left(1 + \frac{1}{2m_y^2}\right) I_{20,02} + \frac{1}{90}\frac{h_x^2h_y^2}{m_x^2m_y^2} I_{11,11} + o\left(\frac{1}{n}\right). \end{aligned}$$

These expressions match previously obtained results for the special cases $m_x = m_y = 1$ (bivariate histogram and frequency polygon; see Scott, 1985) and $m_x = m_y = \infty$ (bivariate product triangle kernel estimator). One may choose values of m_x and m_y between 2–5, about the same as for m in the one-dimensional case. We have found similar values of m_i in the FP-ASH to work with data in three and four dimensions.

6. Application considerations.

6.1 *Choice of smoothing parameters and kernel.* The FP-ASH is in a form convenient for plotting on a graphics device, namely, piecewise linear. Our approach in previous sections has suggested that h is fixed and known and that m is increased (δ decreased) in order to get as close as desired to the triangle kernel estimate. In practice, h is the most difficult parameter to choose. We advocate choosing δ based on graphical considerations (visual smoothness of the FP-ASH) or based on a certain moderate number of bins M (say 30, 50 or 100) covering the sample range. This determines the bin width δ. Then look at an "oversmoothed" FP-ASH, that is, an FP-ASH with smoothing parameter set to a theoretical upper bound; see Terrell and Scott (1985). Now an oversmoothed frequency polygon (FP-ASH with $m = 1$) has $(147n/2)^{1/5}$ bins of width h over the sample range. An oversmoothed triangle kernel estimator (ASH or FP-ASH with $m = \infty$) has $(30n)^{1/5}$ bins of width h over the sample range. Hence we can find an effective upper bound m_u, on m given by $m_u = M/(30n)^{1/5}$. We shall not discuss data-based procedures for choosing $m \in [1, m_u]$; see Scott and Factor (1981), Rudemo (1982) and Bowman (1984). Similar procedures for our estimates are under development. If m is much greater than 5, then adjacent bins of width δ can be aggregated to further reduce computations required for smoothing. Although the multivariate oversmoothed density theory does not yet exist, such estimates can be approximated by assuming that the data are multivariate Gaussian. Epanechnikov's (1969) theory for multivariate product kernels may then be applied. Surprisingly, the choice of h_i or m_i by graphical means does not seem too difficult in the multivariate case, but the representation of the FP-ASH or any density is quite challenging beyond two dimensions. We have been able to construct FP-ASHs to examine data sets in one to four dimensions with sample sizes as large as 400,000 points in an interactive graphics environment and to focus on interesting representational problems; see Scott (1983) and Scott and Thompson (1983).

The choice of weighting function is not critical (Epanechnikov, 1969). A popular choice for constructing a kernel estimate is the quartic (biweight) kernel $K(t) = {}^{15}\!/_{16}(1 - t^2)^2 I_{[-1,1]}(t)$. The ASH weighting function corresponding to the quartic kernel is

$$(6.1) \quad w_m(i) = (15m^4/(16m^4 - 1))((1 - (i^2/m^2))^2 \quad 1 - m \leq i \leq m - 1.$$

By computing finite differences over the bins of the ASH with this weighting function, we obtain essentially the same information about the first and second derivatives of the true density as with derivatives of the quartic kernel estimate. Therefore, we may not always wish to use the triangle weighting function. Estimators with higher order convergence may be obtained if we relax the restriction that the weighting function be nonnegative.

Another weighting function, which is piecewise quadratic, results if we consider the estimator constructed by averaging shifted frequency polygons. This may be useful for dealing with densities with discontinuities at boundary points as illustrated for negative exponential data in Scott (1985).

We can extend the results of our theorems to general weighting functions satisfying (2.7). The formulae are similar to equation (A.1) for kernel estimators. Let $R_w = \sum w_m(i)^2/m$, $\sigma_w^2 = \sum (i/m)^2 w_m(i)/m$, and $\gamma_w = \sum w_m(i)w_m(i-1)/m$, where the sums are for $i = 1 - m$ to $m - 1$.

PROPOSITION 1. *Equations* (3.1) *and* (4.5) *generalize, respectively, to*

$$R_w/nh + (h^2/12m^2)R(f') + \tfrac{1}{4}h^4(\sigma_w^4 - (1/90m^4))R(f'') \tag{6.2}$$

and

$$(1/nh)(\tfrac{2}{3}R_w + \tfrac{1}{3}\gamma_w) + \tfrac{1}{4}h^4(\sigma_w^4 + (\sigma_w^2/2m^2) + (49/720m^4))R(f''). \tag{6.3}$$

6.2 *Computational and statistical efficiency of the FP-ASH.* We have not proven that the FP-ASH is more computationally tractable than a kernel estimate. In this section we show that the FP-ASH is identical to a kernel estimator that (i) uses binned data, (ii) is evaluated over a mesh and then linearly interpolated, and finally (iii) renormalized to integrate to one. Operation counts do not tell the whole story since disk I/O may be the most costly item for large data sets. One-pass algorithms, which are important for large multivariate data sets, exist for computing ordinary kernel estimates over a grid. However, if smoothing parameters are unknown, many passes through the data may be required for kernel estimates. Silverman (1982) has developed an algorithm that greatly reduces the work required for changing smoothing parameters for univariate Gaussian kernel estimates based on the Fast Fourier transform. The procedure provides estimates on a mesh by binning of data into between 32 and 2048 bins; see Jones and Lotwick (1984) for some reduced error results.

Pre-binning or rounding of data on a mesh of width δ can mean tremendous memory and computational saving with very large samples. The IMSE of the binned kernel estimator (assuming the kernel estimator is evaluated *everywhere* and not just over a mesh) is the same as equation (A.1) except that the bias term is multiplied by the factor $(1 + \delta^2/12h^2\sigma_K^2)^2$; see Scott and Sheather (1985), which follows a result of Hall (1982). For a triangle kernel and assuming $h/\delta = m$, this factor is $(1 + 1/m^2 + 1/4m^4)$, which should be compared to the factor $(1 + 1/m^2 + 9/20m^4)$ in (4.5) for the FP-ASH.

In order to realize computational savings with the binned kernel estimator, we should evaluate the estimator only over a mesh and interpolate these values. If we use the same mesh as used for data binning, choose linear interpolation, and restrict attention to smoothing parameters $h = m\delta$, we have the following interesting results. Let us further restrict ourselves to kernels K supported on $[-1, 1]$. The linearly interpolated binned kernel estimator will integrate to one if and only if

$$(1/m)\textstyle\sum_{i=-m}^{m} K(i/m) = \displaystyle\int_{-1}^{1} K(x)\,dx = 1.$$

This is the case for the triangle kernel (2.6) but not for most nonlinear kernels such as the biweight (6.1). When the interpolated binned kernel estimator fails to integrate to one, an obvious fix is to renormalize the estimator. However, further analysis of perturbations in the IMSE is not required because of the following easily proven result:

PROPOSITION 2. *A renormalized linearly interpolated binned kernel estimator over a mesh of width δ with $h = m\delta$ is identical to the FP-ASH with the same parameters.*

Thus we see there are no "computational or statistical gaps" between these two approaches and the FP-ASH answers questions about the use of binned data in a kernel estimator. The FP-ASH deals directly with binned data and avoids introducing notions of numerical interpolation error of a sampled kernel estimator. Higher-order kernel interpolation schemes will introduce negativity problems in the tails and still involve renormalization. The FP-ASH could be modified to study this case. Negative estimates seem to be less desirable in the multivariate case.

Computationally the FP-ASH is relatively efficient because smoothing operations are required only at the end of the data scanning phase and are determined by m and the number of nonempty bins rather than the sample size. Thus recomputing an FP-ASH with different smoothing parameters requires only a modest amount of work for small values of m.

7. Examples. Our first example deals with 63 annual snowfall accumulations in Buffalo; see Table 1 (Parzen, 1979). The sample range is 101.4, so we chose $\delta = 1$. The oversmoothed ASH corresponds to $m_u = 102/(30 \times 63)^{1/5} \sim 22$. In Figures 2A and 2B, we display ASHs with $m = 22$ and 14, respectively. Three bumps are suggested in both graphs; see Good and Gaskins (1980) for a lengthy discussion of a similar data set. The ASH becomes visibly rough for $m < 10$. A FP-ASH for these data might use $\delta = 2$ or 3 since fewer bins are required.

Our second example is an FP-ASH estimate of 1000 pseudo-random points from an independent trivariate Gaussian density. Thirty bins were constructed along each axis with $\delta_i = 0.2$ and $m_i = 5$ in each dimension. Contours of the FP-ASH at levels 10, 35, 60, and 85% of the sample mode (0.050) are shown in Figure 3, with darker shades of gray for higher density contours. Contours for the true density are concentric spheres. The slightly nonspherical shape of each contour is the result of sampling variation. A similar picture of the FP-ASH with $m_i = 3$,

TABLE 1
Yearly snowfall in Buffalo, New York (1910–1973) *in inches*

126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5	25.0
69.3	53.5	39.8	63.6	46.7	72.9	79.6	83.6	80.7	60.3
79.0	74.4	49.6	54.7	71.8	49.1	103.9	51.6	82.4	83.6
77.8	79.3	89.6	85.5	58.0	120.7	110.5	65.4	39.9	40.1
88.7	71.4	83.0	55.9	89.9	84.8	105.2	113.7	124.7	114.5
115.6	102.4	101.4	89.8	71.5	70.9	98.3	55.5	66.1	78.4
120.5	97.0	110.0							

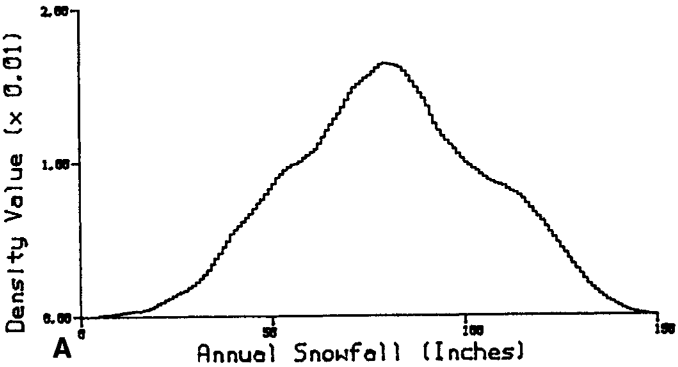

FIG. 2A. *ASH estimate of snowfall data with $m = 22$ and $\delta = 1$.*

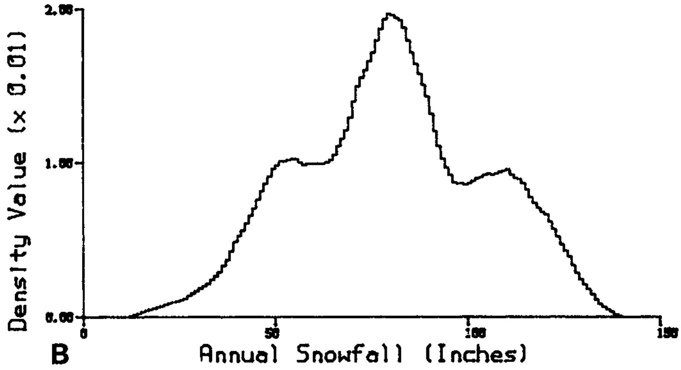

FIG. 2B. *ASH estimate of snowfall data with $m = 14$ and $\delta = 1$.*

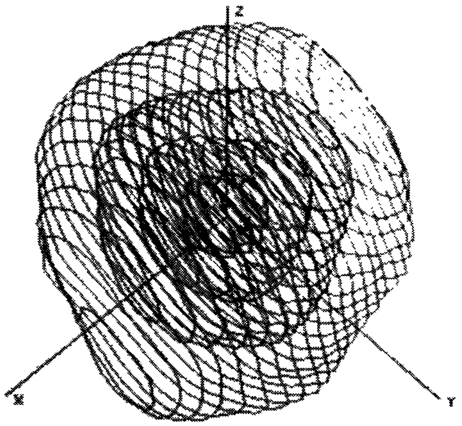

FIG. 3. *Contours of trivariate FP-ASH for Normal data with $m_i = 5$.*

for which the sample mode is 0.069, is very rough. The FP-ASH with $m_i = 4$ is slightly rough.

8. Discussion and conclusions.

We have introduced a conceptually simple density estimator based on averaging several different histograms of the same data. A direct relationship with product kernel methods has been demonstrated. The averaged shifted histogram has several appealing features: it is computationally simple, performs smoothing operations on the bin counts rather than the raw data, and may be quickly and exactly evaluated at any point in the sample space. This last feature has particular importance in pattern recognition where ordinary histograms continue to be heavily used. The computation required for any density estimator can be reduced by binning the data, but then approximation errors must be considered. The modest statistical inefficiency of the FP-ASH due to binning is given exactly in our theorems.

The use of the estimated density function for examining three- and four-dimensional data is a data smoothing operation. For higher-dimensional data, we envision two distinct kinds of smoothing: dimension reduction and density estimation. Dimension reduction options range from classical principal components to modern projection pursuit algorithms (Friedman and Tukey, 1974; Huber, 1985); the latter are particularly powerful for non-Gaussian data. Dimension reduction can result in a significant improvement in the signal-to-noise ratio, although we must be careful not to lose too much signal. Dimension reduction does not reduce the number of data points. Hence additional smoothing may be required. We believe density estimation is well-suited to this purpose. Our density contour representation is the same whether we have a few hundred or a million points.

An interesting question is how much of each kind of smoothing to apply. Projection all the way to one or two dimensions is more likely to provide too much smoothing (loss of signal) than is projection to three or four dimensions. On the other hand, as the dimension increases density estimation becomes less efficient statistically and smoothing parameter selection involves many combinations of m_i choices. Clearly the variety of possible structure with trivariate and quadrivariate data is much more complex and interesting than with univariate or bivariate data. We can only speculate whether with real data it will prove worthwhile to estimate the density function in five or more dimensions where the curse of dimensionality will be felt, or whether an appropriate projection into four or fewer dimensions will be sufficient.

Finally, we note that there are interesting parallels between spectral density windows and ASH weighting functions and between certain digital image processing formulae and the bivariate ASH weighting scheme. Such comparisons with image processing algorithms could be used to suggest how resistant ASH density estimates could be constructed and parallel computer architecture employed.

Acknowledgements. I would like to thank George Terrell, Frank Jones, Dick Heydorn, and Jim Thompson for their help, and also the referees for their careful reading and comments.

APPENDIX

THEOREM 5. *For a finite-support kernel estimator* (2.5) *satisfying conditions* (C3),

$$IMSE = (1/nh)R(K) - (1/n)R(f) + \tfrac{1}{4}h^4\sigma_K^4R(f'') + O[(h/n)R(f)R(f') + h^5R(f'')R(f''')] \tag{A.1}$$

PROOF. We assume without loss of generality that $[-1, 1]$ is the support of K.

$$\begin{aligned} E\hat{f}_K(x) &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right)f(t)\,dt = \int_{-1}^{1} K(w)f(x-hw)\,dw \\ &= \int_{-1}^{1} K(w)\left[f(x) - hwf'(x) + \frac{1}{2}h^2w^2f''(x-\gamma_{xw}hw)\right]dw, \end{aligned} \tag{A.2}$$

where $0 \le \gamma_{xw} \le 1$ using an exact Taylor's expansion for $f(x - hw)$. Notice $f''(x - \gamma_{xw}hw)$ is a continuous function of w. Since K is continuous over the interior of its (finite) support, we may use the generalized mean value theorem (GMVT) to obtain:

$$\int_{-1}^{1} w^2K(w)f''(x-\gamma_{xw}hw)\,dw = f''(x-\gamma_{xc}hc)\int_{-1}^{1} w^2K(w)\,dw, \tag{A.3}$$

where $c \in [-1, 1]$ is a particular value of w. Let $\gamma_x = c\gamma_{xc}$; then $|\gamma_x| < 1$. Hence (A.2) becomes

$$E\hat{f}_K(x) = f(x) - h\mu_K f'(x) + \tfrac{1}{2}h^2\sigma_K^2 f''(x-\gamma_x h). \tag{A.4}$$

By assumption $\mu_K = 0$. Now compute the bias contribution of the IMSE:

$$\begin{aligned} \int_{-\infty}^{\infty} \text{Bias}(x)^2\,dx &= \frac{1}{4}h^4\sigma_K^4\int_{-\infty}^{\infty} f''(x-\gamma_x h)^2\,dx \\ &= \frac{1}{4}h^4\sigma_K^4\int_{-\infty}^{\infty} f''(x)^2\,dx + O(h^5\,\|f''\|_2\,\|f'''\|_2) \end{aligned} \tag{A.5}$$

using the lemma below with $\phi = (f'')^2$ which is absolutely continuous and noting that

$$\|\phi'\|_1 = \|2f''f'''\|_1 \le 2\,\|f''\|_2\,\|f'''\|_2 < \infty$$

by the Cauchy-Schwarz inequality and since $f'', f''' \in L_2$.

To compute the integrated variance, we note that

$$\begin{aligned} \text{Var}\,\hat{f}_K(x) &= \frac{1}{nh^2}EK\left(\frac{x-x_i}{h}\right)^2 - \frac{1}{nh^2}\left[EK\left(\frac{x-x_i}{h}\right)\right]^2 \\ &= \frac{1}{nh}\int_{-1}^{1} K(w)^2f(x-hw)\,dw - \frac{1}{n}\left[\int_{-1}^{1} K(w)f(x-hw)\,dw\right]^2. \end{aligned} \tag{A.6}$$

Integrating the first term in (A.6) over x gives exactly

$$\frac{1}{nh}\int_{-1}^{1} K(w)^2\,dw. \tag{A.7}$$

The second bracketed term in (A.6) equals

$$f(x-hw_x)\int_{-1}^{1} K(w)\,dw = f(x-hw_x) \tag{A.8}$$

using the GMVT, where $-1 \le w_x \le 1$. Therefore the second term when squared and integrated over x becomes

$$-\frac{1}{n}\int_{-\infty}^{\infty} f(x-hw_x)^2\,dx = -\frac{1}{n}\int_{-\infty}^{\infty} f(x)^2\,dx + O\left(\frac{h}{n}\,\|f\|_2\,\|f'\|_2\right), \tag{A.9}$$

using the lemma below with $\phi = f^2$, which is absolutely continuous and noting that $\|\phi'\|_1 = \|2ff'\|_1 \le 2\,\|f\|_2\,\|f'\|_2 < \infty$, since $f, f' \in L_2$. Adding (A.5), (A.7), and (A.9) proves the theorem and ensures that the remainder terms are bounded.

LEMMA 1. *Suppose ϕ is absolutely continuous, $\phi' \in L_1$, and $w_t = w(t)$ is a measurable function onto* $[-1, 1]$. *Then*

$$\int_{-\infty}^{\infty} \phi(t-hw_t)\,dt = \int_{-\infty}^{\infty} \phi(t)\,dt + O(h\,\|\phi'\|_1).$$

PROOF.

$$\begin{aligned}\left|\int_{-\infty}^{\infty} [\phi(t-hw_t)-\phi(t)]\,dt\right| &= \left|\int_{-\infty}^{\infty}\int_{s=t}^{t-hw_t} \phi'(s)\,ds\,dt\right| \\ &\le \int_{t=-\infty}^{\infty}\int_{s=t-h}^{t+h} |\phi'(s)|\,ds\,dt \\ &= \int_{s=-\infty}^{\infty}\int_{t=s-h}^{s+h} |\phi'(s)|\,ds\,dt = 2h\,\|\phi'\|_1.\end{aligned}$$

REFERENCES

BADHWAR, G. G. (1980). Crop emergence data determination from spectral data. *Photogram. Eng. Remote Sens.* **46** 369–377.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. and TUKEY, P. A. (1983). *Graphical Methods for Data Analysis.* Duxbury, Boston.

EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multi-dimensional probability density. *Theory Probab. Appl.* **14** 153–158.

FREEDMAN, D. and DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahrsch. verw. Gebiete* **57** 453–476.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.

GOOD, I. J. and GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–73.

HABBEMA, J. D. F., HERMANS, J. and VAN DEN BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 74.* (G. Bruckmann, ed.) 101–110. Physica-Verlag, Vienna.

HALL, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM J. Appl. Math.* **42** 390–399.

HUBER, P. J. (1983). Projection pursuit. Technical Report MSRI 009-83, Berkeley.

JONES, M. C. and LOTWICK, H. W. (1984). A remark on Algorithm AS 176: Kernel density estimation using the Fast Fourier transform. *Appl. Statist.,* **33** 120–122.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66** 605–610.

SCOTT, D. W. (1983). Nonparametric probability density estimation for data analysis in several dimensions. In *Proc. Twenty-Eighth Conference on the Design of Experiments in Army Research, Development and Testing.* Report No. 83-2, 387–397. U.S. Army Research Office, Research Triangle Park.

SCOTT, D. W. (1985). Frequency polygons: Theory and application. *J. Amer. Statist. Assoc.* to appear.

SCOTT, D. W. and FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric density estimators. *J. Amer. Statist. Assoc.* **76** 9–15.

SCOTT, D. W., GORRY, G. A., HOFFMAN, R. G., BARBORIAK, J. J. and GOTTO, A. M. (1980). Plasma lipid concentrations and the severity of coronary artery disease: A study of 1847 males with angiographically-demonstrated disease. *Circulation* **62** 477–484.

SCOTT, D. W. and SHEATHER, S. J. (1985). Kernel density estimation with binned data. *Comm. Statist.* to appear.

SCOTT, D. W. and THOMPSON, J. R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proc. Fifteenth Symposium on the Interface.* (J. E. Gentle, ed.) 173–179. North-Holland, Amsterdam.

SILVERMAN, B. W. (1982). Algorithm AS 176: Kernel density estimation using the Fast Fourier transform. *Appl. Statist.* **31** 93–99.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation.* The Johns Hopkins University Press, Baltimore.

TERRELL, G. R. (1984). Efficiency of nonparametric density estimators. Unpublished manuscript.

TERRELL, G. R. and SCOTT, D. W. (1985). Oversmoothed nonparametric density estimates. *J. Amer. Statist. Assoc.* **80** 209–214.

TUKEY, P. A. and TUKEY, J. W. (1981). Graphical display of data sets in 3 or more dimensions. In *Interpreting Multivariate Data* (V. Barnett, ed.) 187–275. Wiley, New York.

WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.

DEPARTMENT OF MATHEMATICAL SCIENCES
RICE UNIVERSITY
HOUSTON, TEXAS 77251