VARIABLE KERNEL DENSITY ESTIMATION

By George R. Terrell and David W. Scott [1]

Virginia Polytechnic Institute and State University and Rice University

May 21, 1992

0. Summary. We investigate some of the possibilities for improvement of univariate and multivariate kernel density estimates by varying the window over the domain of estimation, pointwise and globally. Two general approaches are to vary the window width by the point of estimation, and by point of the sample observation. The first possibility is shown to be of little efficacy in one variable. In particular, nearest neighbor estimators in all versions perform poorly in one and two dimensions, but begin to be useful in three or more variables. The second possibility is more promising. We give some general properties and then focus on the popular Abramson estimator. We show that in many practical situations such as normal data, a nonlocality phenomenon limits the commonly applied version of the Abramson estimator to bias of $O([h/\log h]^2)$ instead of the hoped for $O(h^4)$.

1. Introduction. Among the plethora of multivariate nonparametric density estimators is the fixed kernel estimator

$$\hat{f}(\mathbf{y}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h}\right)$$

where $\{\mathbf{x}_i\}$ is an i.i.d. random sample of size n, $\mathbf{x}_i \in \mathbb{R}^d$, $K: \mathbb{R}^d \to \mathbb{R}^1$ is a function centered at zero that integrates to one, and h is a smoothing parameter that would usually tend to zero as the sample size n goes to infinity. The first univariate estimator of this type with $K \equiv U(-1,1)$ was proposed by Fix and Hodges (1951); the general class was investigated by Rosenblatt (1956) and Parzen (1962) with multivariate extension by Cacoullos (1966) and Epanechnikov (1969).

Though in most applied work h has been held constant, two important proposals have been put forth to vary h in hopes of improving the resulting density estimates. The first

[1] Research supported in part by ONR and ARO under grants N00014-90-J-1176 and DAAL03-88-K-0131, respectively.

AMS 1980 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. Kernel estimators, adaptive estimation, nearest neighbor estimators, balloongrams, nonparametric smoothing.

proposal was the k–th nearest neighbor (k–NN) of Loftsgaarden and Quesenberry (1965) given by

$$\hat{f}(\mathbf{y}) = \frac{k}{n\, V_d\, h_k(\mathbf{y})^d} \quad , \tag{1.1}$$

where $h_k(\mathbf{y})$ is the Euclidean distance from $\mathbf{y}$ to the k–th nearest sample point, and V_d is the volume of the unit sphere S_d in $\mathbb{R}^d$. The k–NN estimator can be written as a kernel estimator if K is chosen to be a uniform density on the unit d–sphere S_d; then

$$\hat{f}(\mathbf{y}) = \frac{1}{nh_k(\mathbf{y})^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k(\mathbf{y})}\right) . \tag{1.2}$$

A second proposal was the adaptive kernel estimate of Breiman, Meisel, and Purcell (1977) given by

$$\hat{f}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^d} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_i}\right) \tag{1.3}$$

where h_i is the Euclidean distance from x_i to the k-th nearest other sample point. This is asymptotically equivalent to choosing $h_i \propto f(\mathbf{x}_i)^{-1/d}$; while Abramson (1982) proposed using

$$h_i \propto f(\mathbf{x}_i)^{-1/2} \tag{1.4}$$

for all dimensions. In practice, a pilot estimate is obtained for the unknown density f at the sample points. Also, Breiman et al. were interested in estimating the entire density while Abramson limited his proposal to pointwise estimation; however, given the attractiveness of Abramson's proposal, most workers have applied Abramson's choice to estimation of the entire density. Abramson also "clipped" the pilot estimate away from zero, a condition ignored in most subsequent theoretical and practical work. We shall refer to the application of formula (1.4) without the clipping as the "non–clipped Abramson estimator," and again wish to emphasize that this was not his proposal. Hall and Marron (1988) considered the theoretical properties of a practical global implementation of Abramson's estimator when the unknown density is bounded away from zero.

This paper will explore some of the implications of several classes of schemes for letting the smoothing parameter vary over the real line. A primary conclusion will be that it is surprisingly difficult to do significantly better than the original fixed kernel scheme. For the simplest case of the univariate and bivariate histogram, Scott (1982), Kogure (1987) and Hüesemann and Terrell (1991) have investigated the related problem of varying bin widths over the domain of the data.

Our criteria for good estimates will be asymptotic mean–squared error (AMSE) at a single point of estimation $\mathbf{y}$; and that function integrated over the entire real line, the asymptotic mean integrated squared error (AMISE). Other measures of quality have been studied; for example, L_1 error (see Devroye and Györfi, 1985). We chose ours for convenience; conclusions usually seem to be similar under different choices (see Hall and Wand, 1988, and Scott and Wand, 1991). Since our criteria are asymptotic, it will be assumed that we mean by a density estimator a family of techniques, one for each sample size. Certain finite sample issues will be addressed by simulation. There is also a large literature dealing with the issue of calibration or cross–validation; that is, how to match the degree of smoothing to the unknown underlying density, using sample information. We will not address that important issue here.

The paper will first extend the informal observations of Walter and Blum (1979) by showing that every multivariate density estimator that is in any reasonable sense nonparametric may be written in the form

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} K_n(\mathbf{x}_i, \mathbf{y})\,, \tag{1.5}$$

where K_n is asymptotically a Dirac evaluation functional at $\mathbf{y}$. This almost, but not quite, says that all nonparametric density estimators are kernel estimators, since K_n may be dependent to second order on the other sample points. An estimator of this form which is independent of the observations will be called a *generalized kernel estimator*. The form displays the possibility that kernel shape may vary over the line in a large variety of ways.

For pointwise estimation of $\hat{f}(\mathbf{y})$, examination of equations (1.2) and (1.3) suggests the study of two simple rules for variability of the kernel (see also Jones (1990). Yet another is discussed in Wand, Marron, and Ruppert (1991)). First, the scale of the kernel may depend only on $\mathbf{y}$, the point at which the estimate is taken; or, secondly, only on $\mathbf{x}_i$, the sample point. The first says that

$$\hat{f}_1(\mathbf{y}) = \frac{1}{nh(\mathbf{y})^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h(\mathbf{y})}\right)\,. \tag{1.6}$$

We shall call this a *balloon estimator*, generalizing a suggestion due to Tukey and Tukey (1981). We will show that the Loftsgaarden–Quesenberry–style nearest neighbor estimators are asymptotically of this type. This estimator, $\hat{f}_1(\cdot)$, has the advantage of having a straightforward asymptotic analysis, since it uses standard pointwise results (Mack and Rosenblatt, 1979); on the other hand, when applied globally, the estimate typically does not integrate to one and thus is usually not itself a density, even when K is. We investigate the degree of improvement that this approach allows over fixed kernel estimates; in common cases, the improvement is seen to be very modest. The asymptotic behavior of the univariate k–th nearest neighbor estimator turns out to be particularly poor; simulations performed by the authors suggest that this holds also for reasonable finite sample sizes. However, the multivariate extensions of the balloon estimator exhibit some interesting new phenomena. For example, above dimension three, the nearest neighbor method becomes competitive with fixed kernels for normal data.

The second approach gives

$$\hat{f}_2(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(\mathbf{x}_i)^d} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h(\mathbf{x}_i)}\right) . \tag{1.7}$$

We shall call this a *sample smoothing estimator*; it is a mixture of identical but individually scaled kernels centered at each observation. The Abramson estimator is of this form; as is the Breiman–Meisel–Purcell nearest neighbor estimator, in an asymptotic sense. The main advantage of the sample smoothing estimator is that if K is a density, then so is $\hat{f}_2(\cdot)$. The primary disadvantage is that the estimates generally exhibit *nonlocality*; that is, the estimate at a point may be significantly influenced by observations very far away and not just by points nearby. We give an asymptotic analysis of these estimators. The Breiman et al. estimator turns out to be even more unsatisfactory in important cases than the Loftsgaarden–Quesenberry estimator; simulations show that this is reflected even in moderate sample sizes. The nonlocality phenomenon is shown to prevent the non–clipped Abramson estimator from achieving the $O(n^{-8/9})$ convergence rate that has been claimed for it (Silverman, 1986). However, simulations still show good behavior for small–to–moderate sample sizes, but deterioration in performance compared to fixed estimates as the sample size grows.

Our often negative results should not be too discouraging; for one thing, multivariate variable kernels show some promise. For another, we are far from exhausting the possible variety of such estimates.

2. Generalized Kernels. We may bring some order into the world of nonparametric density estimation:

Theorem 1. *Any multivariate density estimator that is a continuous and Gâteaux differentiable functional on the empirical distribution function may be written as*

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{x}_i, \mathbf{y}, \hat{F}_n) ,$$

where K is the Gâteaux derivative of $\hat{f}$ under variation of $\mathbf{x}_i$.

This theorem is proved in the Appendix where a general constructive formula for finding K is given. The conditions seem to be essentially vacuous: all proposals for nonparametric density estimators with which we are familiar meet them. K is essentially the influence function of $\hat{f}(\mathbf{y})$ at $\mathbf{x}_i$ (Hampel, 1974). It is interesting to note that deMontricher, Tapia, and Thompson (1975) show that a certain maximum penalized likelihood estimator is a linear combination of Laplace kernels, with weights that depend on all the other observations. Walter and Blum (1979) have cataloged many "equivalent kernels," which generally coincide with the K's of Theorem 1, and called the general class delta methods. Since $\hat{F}_n$ converges to F, we see that K is asymptotically independent of those other observations. Thus, any continuous density estimator may be written asymptotically as

$$\hat{f}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} K_n(\mathbf{x}_i, \mathbf{y}) \,.$$

We restrict attention to those that may be written exactly in this form for finite n; these are the generalized kernel estimators of the last section.

Generalized kernels cover an enormous variety of methods; including, for example, histograms and frequency polygons. Figure 1 displays the kernels $K(x, y)$ for a fixed–bin–width univariate frequency polygon. This paper will concentrate on estimators that generalize fixed kernels in straightforward ways. Notice that even a fixed kernel method may be thought of in two different ways: (1) as a system of weights centered at y that specifies the degree of influence an observation at each x will have on the estimate at y; and (2) as an equal mixture of densities centered at each observation. Generalizing these two points of view leads to balloon estimators and to sample smoothing estimators.

FIGURE 1. *The equivalent kernel function for a frequency polygon.*

3. Univariate Balloon Estimators. The asymptotic pointwise behavior of the univariate balloon estimator

$$\hat{f}_1(y) = \frac{1}{nh(y)} \sum_{i=1}^{n} K\left(\frac{x_i - y}{h(y)}\right)$$

at y is well–known, because it is equivalent pointwise to a fixed kernel. We shall not restrict ourselves to positive kernels. Thus we need

Definition 1. *An univariate* order–p kernel *is a function K such that*

$$\int K(x)\,dx = 1 \qquad \lim_{|z|\to\infty} |z\,K(z)| = 0$$

$$\int |K(z)| dz < \infty \qquad \int K(z)^2\,dz < \infty \tag{3.1}$$

$$\int z^r K(z)\,dz = 0 \quad \text{for} \quad 1 \le r \le p-1 \qquad \int z^p K(z) = \pm 1\,.$$

In the most common case, in which K is a density of mean zero and finite variance, it is an order–2 kernel.

We use the usual variance/squared–bias decomposition of the mean squared error (MSE). We focus attention on the the asymptotic MSE (AMSE), which is the leading nonzero terms in the MSE. The AMSE is the asymptotic limit as h goes to 0 as $n\,h$ goes to infinity of the sum of the asymptotic pointwise variance and squared bias. Integrating the pointwise AMSE(y) over the line, we get a global criterion, the asymptotic mean integrated squared error (AMISE).

Then for a density f that meets the conditions: $f^{(p)}$, its p-th derivative, is continuous and nonzero at y, and $\int |f^{(p)}(y)|\,dy < \infty$, we have

$$\text{asymptotic variance}(y) = \frac{f(y)\int K^2}{nh(y)}$$

$$[\text{asymptotic bias}(y)]^2 = \left[\frac{h(y)^p}{p!} f^{(p)}(y)\right]^2 \tag{3.2}$$

where $nh(y) \to \infty$ as $h(y) \to 0$; see for example Parzen (1962).

Proposition 1. *The asymptotically best balloon estimator optimizes the AMSE pointwise; it achieves a minimum where*

$$h^*(y) = \left[\frac{(p!)^2 f(y)\int K^2}{2p\left(f^{(p)}(y)\right)^2}\right]^{1/(2p+1)} n^{-1/(2p+1)} \tag{3.3}$$

$$\text{AMSE}^*(y) = (2p+1)\left[\frac{f(y)^p f^{(p)}(y)\left(\int K^2\right)^p}{(2p)^p p!}\right]^{2/(2p+1)} n^{-2p/(2p+1)}. \tag{3.4}$$

This generalizes the result of Rosenblatt (1979) for nonnegative kernels. Figure 2 shows the shape of this optimal h function for the standard Gaussian density (see Dodge and Lejeune, 1986). Integrating over the line, we get

$$\text{AMISE}^* = (2p+1)\left[\frac{\left(\int K^2\right)^p}{(2p)^p p!}\right]^{\frac{2}{2p+1}} \left[\int f(y)^{\frac{2p}{2p+1}} \left(f^{(p)}(y)\right)^{\frac{2}{2p+1}} dy\right] n^{\frac{-2p}{2p+1}}. \tag{3.5}$$

This is asymptotically the best performance we can get from a balloon estimator. A kernel of order 1 behaves somewhat like a histogram. The case where $p = 2$ was investigated by Terrell and Scott (1983).

FIGURE 2. *Optimal local bandwidths for a standard normal density.*

To put the possible improvement in perspective, let us compare it to the asymptotic error in a fixed kernel. It may be readily checked that only the integral changes in the expression above; the other terms, in particular, the rate of convergence in n, are unchanged. For the asymptotic relative efficiency of a fixed order–p kernel we get

$$\frac{\text{AMISE}^*_{\text{adapt}}}{\text{AMISE}^*_{\text{fixed}}} = \frac{\int f(y)^{\frac{2p}{2p+1}} \left(f^{(p)}(y)\right)^{\frac{2}{2p+1}} dy}{\left[\int \left(f^{(p)}(y)\right)^2 dy\right]^{\frac{1}{2p+1}}} . \tag{3.6}$$

An application of Jensen's inequality shows that this is at most 1, as we would expect.

Examples. Table 1 shows the value of this efficiency when f is univariate Gaussian or Cauchy, for various values of p. The numerator was computed by numerical integration using *Mathematica* (Wolfram, 1988). The integral in the denominator may be shown to equal $\frac{1}{2\pi}\Gamma\left(p+\frac{1}{2}\right)$ for normal densities by an application of Parseval's theorem.

Table 1.

Kernel Order	**Gaussian**	**Cauchy**
1	89.3%	84.0%
2	91.5%	76.7%
4	94.2%	72.0%
6	95.6%	70.0%
8	96.5%	68.9%

Table 1. Asymptotic relative efficiencies of optimal fixed to optimal adaptive kernel estimators for standard Gaussian and Cauchy densities.

We conclude that balloon estimators allow very little improvement for normal data. However, the efficiency may be made arbitrarily close to zero by considering an infinitely separated equal mixture of two normal components with differing scale parameters. Nevertheless, for densities of the usual shapes, balloon estimators are mostly hot air.

The most popular univariate balloon estimator is the Loftsgaarden–Quesenberry k–th nearest neighbor kernel of the form

$$\hat{f}(y) = \frac{1}{nh_k(y)} \sum_{i=1}^{n} K\left(\frac{x_i - y}{h_k(y)}\right) ;$$

cf. equation (1.2). Notice that $k/\left(2nh_k(y)\right)$ is a nonparametric estimate of $f(y)$ consistent whenever $k \to \infty$ while $k/n \to 0$. We will therefore analyze the asymptotically equivalent balloon estimator for which $h_k(y) \equiv k/\left(2\,nf(y)\right)$, thereby avoiding the issue of data dependence. Pointwise, this method is equivalent to the usual kernel estimator, because we have simply reparameterized h as k. However, when we treat it as a global estimator, something new happens. Make the same substitution for h above, so that we find a constant number of neighbors k as we scale the kernel. When we write down the AMISE, the integrated squared bias contains the factor

$$\int \frac{\left(f^{(p)}(y)\right)^2}{f(y)^{2p}}\, dy\,. \tag{3.7}$$

This is infinite unless the tail exponents of the density are less than $1 + 1/2(p+1)$; for example, for nonnegative kernels when $p = 2$ the only Student's t–densities that meet this requirement have less than 0.5 degrees of freedom! We conclude that the asymptotic efficiency of the k–th nearest neighbor density estimator is zero in all cases usually encountered. This suggests that it is of very little practical value. This problem isn't just a phenomenon associated with huge samples or the far tails; Silverman (1986) has illustrated the roughness of this estimator for small univariate samples.

We will return to balloon estimators in the multivariate setting in Section 5. We will see that the situation is significantly different in several variables than in one dimension.

4. Sample Smoothing Estimators.

4.1. *Univariate Asymptotic Errors and Monte Carlo Example.* If we instead let the the scale of a kernel vary with the location of the data points, we have

$$\hat{f}_2(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_i)} K\left(\frac{x_i - y}{h(x_i)}\right)\,. \tag{4.1}$$

Theorem 2. *As the sample size grows, let $h(x) \to 0$ and $nh(x) \to \infty$ for all x; and let $(x-y)/h(x)$ be a monotone increasing function of x onto the entire real line. Then for f and h p–times differentiable we have*

$$\text{asymptotic variance}(y) = \frac{f(y)\int K^2}{nh(y)} \tag{4.2}$$

$$[\text{asymptotic bias}(y)]^2 = \left[\frac{1}{p!}\left[h(y)^p\, f(y)\right]^{(p)}\right]^2 \tag{4.3}$$

for an order–p kernel sample smoothing density estimate.

This is proved in the Appendix. Note that, compared to the fixed or balloon estimates, the asymptotic variance is the same, but the power of h has been moved under the differential operator. A close examination of the bias expression in equation (4.3) reveals the rationale for Abramson's choice (1.4) for the bandwidth:

$$h(x_i) = hf(x_i)^{-1/2} \, . \tag{4.4}$$

If we use a second–order positive kernel ($p = 2$), the asymptotic bias (4.3) becomes

$$\frac{1}{2}\left[h^2 f(y)^{-1} f(y)\right]'' = 0 \, .$$

Silverman (1986) has given an expression for the next higher order term in the asymptotic bias and shown that it is of order $O(h^4)$, which leads to AMISE $= O(n^{-8/9})$, a rate usually reserved for kernels that are not nonnegative. Several authors, including Abramson (1982), Silverman (1986), and Worton (1989), have shown that the algorithm works well for small samples even using a pilot estimator. In order to investigate its asymptotic behavior, we performed a numerical integration to obtain the exact MISE of the non–clipped Abramson estimator for Gaussian data and a biweight kernel, $\frac{15}{16}(1-x^2)_+^2$. In Figure 3, we plot the numerical estimates of exact optimal MISE (using the best bandwidth, which was also determined numerically) together with the exact optimal MISE for a Gaussian kernel estimate (Wertz, 1978), the latter being asymptotically $O(n^{-4/5})$.

FIGURE 3. *Computed optimal exact MISE for the non–clipped Abramson estimator (solid line) for standard normal data and exact MISE for optimal fixed bandwidth estimator (dashed line) with normal kernel. The two small parallel vertical lines represent factors of* 1.5 *and* 1.2.

An examination of these curves reveals that the adaptive estimator enjoys a significant advantage for small n, for example, 35% smaller MISE when $n = 100$. However, no practical difference exist when $n > 1000$; and in fact the fixed kernel estimator has smaller MISE when $n > 30,000$. Any $O(n^{-8/9})$ behavior seems lost. What is our explanation? The special case considered in the next section clarifies matters.

4.2. **Example:** Gaussian Data and Uniform Kernel Example. The theoretical analysis of the non-clipped Abramson estimator (4.4) is quite interesting, as we shall see. It is also surprisingly different from the usual Taylor series arguments presented for estimating MSE of kernel estimators.

Closed form pointwise AMSE expressions have not been generally available, but can be obtained in the special case where $K \sim U(-1,1)$ and $f \sim \phi(0,1)$. The relevant results can be obtained by examining the error of the pointwise estimator $\hat{f}(0)$, which from equations (4.1) and (1.4) is given by

$$\hat{f}(0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sqrt{\phi(x_i)}}{h} K \left(\frac{(0 - x_i)\sqrt{\phi(x_i)}}{h} \right), \tag{4.5}$$

assuming the exact adaptive bandwidth is known (cf. Abramson, 1982). Then

$$E\, \hat{f}(0) = \int \frac{\sqrt{\phi(x)}}{h} K \left(\frac{x\sqrt{\phi(x)}}{h} \right) \phi(x)\, dx\,, \tag{4.6}$$

since $K(-x) = K(x)$. In the usual analysis, only a small neighborhood around the point $x = 0$ need be considered in the integral (4.6); in fact, that neighborhood is $(-h, h)$ if the support of the kernel is $(-1, 1)$. However, it is easy to check that with the non–clipped Abramson estimator, this is not necessarily so. The integrand in (4.6) vanishes whenever the argument of the kernel exceeds 1; hence, the integral is over those points x satisfying

$$\left| \frac{x\sqrt{\phi(x)}}{h} \right| \leq 1. \tag{4.7}$$

It is easy to check that the left–hand side is maximized when $x = \sqrt{2}$. If $h > .5418$, then all values of x satisfy the inequality and the kernel *never vanishes* on $(-\infty, \infty)$. When $h < .5418$ (which roughly corresponds to $n > 445$ in the fixed bandwidth case), then equation (4.7) has 4 (not 2) solutions; call them $-b < -a < a < b$ (see Figure 4). In this situation the kernel is nonzero when x falls in the intervals $(-\infty, -b)$, $(-a, a)$, and (b, ∞). Previous analyses have only considered the "local" interval $(-a, a)$ and have neglected the influence of the other 2 intervals, which represent the influence from very distant points in the tails. We consider the contribution of these intervals to the bias separately.

FIGURE 4. *Regions where data contribute to the non–clipped Abramson estimate at origin.*

Consider the interval $(-a, a)$ containing the point $x = 0$. From (4.7), it follows that a is a solution to

$$x\, e^{-x^2/4} = (2\pi)^{1/4} h \equiv c. \tag{4.8}$$

Taking a Taylor's series on the left–hand side of (4.8), it can be shown that

$$a = c + c^3/4 + 5c^5/32 + O(c^7)\,.$$

Hence, from equation (4.6) and by symmetry of $\phi(x)^{3/2}$,

$$E\,\hat{f}(0) = \int_{-a}^{a} \frac{\phi(x)^{3/2}}{h}\frac{1}{2}dx = \frac{(2\pi)^{-3/4}}{h}\int_{0}^{a}(1 - 3x^2/4 + 9x^4/32 - 9x^6/128 + \cdots)\,dx$$

$$= \frac{(2\pi)^{-3/4}}{h}(c + c^5/40 + \cdots) = (2\pi)^{-1/2} + \frac{(2\pi)^{1/2}}{40}h^4 + O(h^6). \tag{4.9}$$

The first term is $\phi(0)$, hence

$$\text{Bias}\hat{f}(0) = \frac{(2\pi)^{1/2}}{40}h^4 + O(h^6). \tag{4.10}$$

REMARK. The bias contribution from the interval $(-a, a)$ is $O(h^4)$; in fact, equation (4.10) exactly matches Silverman's (1986, pp. 104–105) equations (5.11) and (5.12).

For the intervals $(-\infty, -b)$ and (b, ∞), let us assume that n is large, or equivalently, h is small. It is easy to see that $b \to \infty$ as $h \to 0$. To find an approximate solution for the root $x = b$ in equation (4.8), we take logarithms, which implies

$$\log x - x^2/4 = \log c \qquad \text{or equivalently} \qquad x = (-4\log c + 4\log x)^{1/2}\,.$$

Since $c \to 0$, a first order solution for the root is $b \approx \sqrt{-4\log c}$; however, the next order terms are important. A sufficient approximation is afforded by

$$b \approx (-4\log c + 4\log\sqrt{-4\log c})^{1/2}\,.$$

Again, by symmetry of the integral, the remaining contribution to the bias is

$$2\int_{-\infty}^{-b} \frac{\phi(x)^{3/2}}{h}\frac{1}{2}dx = 2^{1/4}\,3^{-1/2}\,\pi^{-1/4}\,\frac{1}{h}\Phi(-b\sqrt{3/2})\,,$$

where Φ is the cumulative distribution function of ϕ. Using the approximation

$$\Phi(x) \approx -\frac{1}{x}\phi(x) \qquad \text{for} \quad x \ll 0\,,$$

it follows that the remainder of the bias contribution from the tails is

$$\frac{h^2}{24[\log\{(2\pi)^{1/4}h\}]^2}\,, \tag{4.11}$$

which is $o(h^2)$, but just barely.

Thus we have shown that the contribution from the tail dominates the bias asymptotically in the non–clipped Abramson estimate and that the squared bias is just a bit faster than $O(h^4)$. In fact, the squared bias is asymptotically

$$O([h/\log(h)]^4)\,.$$

For small samples, the contribution from the tails appears to be negligible in many cases, so that the squared bias initially looks like $O(h^8)$ as in equation (4.10). This partially explains the good small sample behavior observed by many authors for this estimator (Abramson, 1982; Silverman, 1986; and Worton, 1989) as well as our simulation result in Figure 3.

One fix that is obviously suggested by this argument is to eliminate the influence of points outside the interval $(-a, a)$. This is a workable solution; however, closer examination reveals that the estimate no longer integrates to one. Such an estimator would add to the existing list of higher order nonnegative estimators that do not integrate to 1 (Terrell and Scott, 1980). The original clipping idea of Abramson deserves closer examination.

The pointwise variance may be computed in a similar fashion over the three intervals $(-\infty, -b)$, $(-a, a)$, and (b, ∞):

$$\operatorname{Var} \hat{f}(0) \approx \frac{1}{n} \int \frac{\phi(x)}{h^2} K\left(\frac{x\sqrt{\phi(x)}}{h}\right)^2 \phi(x)\, dx \approx \frac{0.1260}{nh} + o(\frac{h^2}{n})\,. \tag{4.12}$$

The tails do not contribute much to the variance (the $o(h^2/n)$ term). The MSE$[\hat{f}(0)]$ is the sum of the variance in equation (4.12) and the square of the bias, where the bias is the sum of equations (4.10) and (4.11). This can be minimized numerically over h for each n and compared to the MSE(0) of a fixed uniform kernel estimator of $f(0)$, which can be shown to asymptotically equal

$$\text{AMSE}(0) = 0.1536 n^{-4/5}\,.$$

When $n = 10^3$, the optimal bandwidth is $h^*(0) \approx 0.228$ (determined numerically) and the corresponding optimal AMSE(0) of the non–clipped Abramson estimator is about 1% less than that for the optimal fixed kernel estimator, but by the time $n = 10^6$, it is 49% less. The theoretical cross–over occurs when $n \approx 887$.

For $n < 500$, the asymptotic theoretical expressions are not useful, of course. Several authors have commented on the very good performance of the non–clipped Abramson estimator for small samples. When $h > 0.5418$ (i.e. for small samples), the argument of the kernel in equations (4.6) and (4.12) never exceeds 1 in magnitude and the limits of integration extend over the entire real line. It is easy to check that

$$\operatorname{E} \hat{f}(0) = \frac{0.2579}{h} \qquad \text{and} \qquad \operatorname{Var} \hat{f}(0) = \frac{1}{8\sqrt{\pi} n h^2}\,. \tag{4.13}$$

This is a very unusual bias/variance arrangement. The "optimal bandwidth" $h^*(0) \approx 0.646$ since the bias is zero for this choice. In Figure 5, we have plotted non–clipped Abramson estimates for several bandwidths with simulated normal data ($n = 1000$) but using the true density in Equation (4.5). Compare the two estimates for our 2 different notions of an optimal bandwidth, $h = .228$ based on asymptotic ideas and $h = .646$ based on small sample ideas: the latter seems superior. For sufficiently large bandwidths, the estimator is *constant* over the "middle" of the estimate. (This is easy to predict since all of the adaptive bandwidths h_i are sufficiently large to cover a neighborhood surrounding 0.) More work is required to understand the theoretical behavior of the MISE of the Abramson estimator for small sample sizes.

FIGURE 5. *Four bandwidths for 1000 standard normal data in the non–clipped Abramson estimator with a $U(-1,1)$ kernel. The middle two are optimal in different senses; see text.*

There is nothing about our results specific to the uniform kernel. (Abramson excluded it for theoretical convenience.) The only significant difference is the graphical appearance of the estimate with small samples and large values of the smoothing parameter h.

We could extend our pointwise analysis for points other than the origin, but there is little additional knowledge to be gained. Clearly, a global AMISE analysis of the Abramson estimator would be interesting. We chose to compute the MISE error numerically rather than in asymptotic form. We made this choice in part because of our interest in the small–sample behavior of the estimator. Other numerical experiments have been reported recently; see Bowman and Foster (1991).

The reason that the MISE of the non–clipped Abramson estimator is asymptotically greater than the fixed kernel estimator can be explained as follows. While the bias $= O(h^2/\log(h))$ is smaller than the $O(h^2)$ bias of fixed kernels, the use of the same fixed h in $hf(x_i)^{-1/2}$ throughout the estimation line is quite suboptimal; and when aggregated, the adaptive estimator's MISE performance is worse asymptotically. That it performs well for small samples is still of practical importance. In fact, Abramson's (1982, p. 1217) discussion on this observation led him to question the practical application of his pointwise procedure globally. Abramson did show that his clipped pointwise will have squared bias of $O(h^8)$. It is left as an exercise for the reader to show that if the Cauchy density is chosen in equation (4.6) that the pointwise bias is exactly zero for all $y \in \mathbb{R}^1$. Thus the Abramson rule adapts perfectly in this setting. Such results also follow from a version of Theorem 2 identifying higher order bias terms.

Theorem 2 suggests we may generalize Abramson's procedure. A *generalized Abram-*

son estimator is a sample smoothing kernel estimator of order p when

$$h(x) = f(x)^{-1/p} q(x)^{1/p}$$

where $q(x)$ is a polynomial of degree less than p. The asymptotic bias expression from the theorem is zero for this class of estimators; we therefore expect the AMSE to be of lower than usual order. Abramson's proposal is for order $p = 2$ kernels, where he lets q be a constant. For order $p = 4$ kernels, we have a richer choice of polynomials; use for example $q(x) = a(1 + x^2)$.

4.3. *Other Sample-Smoothing Estimates.* The monotonicity condition in Theorem 2 is very strong, and is by no means met by all sample smoothing estimators in which we will be interested (as we saw in the previous subsection). If it does not hold, then two different x's will generate the same argument for K. This may lead, as we have seen, to observations that are far away having a substantial influence on the density estimate at a point; this is a disturbing property for a nonparametric method. We will call an estimate that meets the monotonicity condition *local.* Unfortunately,

Proposition 2. *There exist no nonconstant adaptivity functions $h(x)$ that are local for all y.*

Proof: The condition says that, for $x > y$, $(\log h)' \leq 1/(x - y)$ and the reverse for $x < y$. This says that h is necessarily constant. ∎

The proof also shows that the condition is not very limiting if we are concerned only with estimation at a point; it disallows only rapid growth of h in each direction from the point of estimation.

The proposition raises doubts as to the usefulness of the error estimates of Theorem 2. We have shown by example, however, that our error expressions, which deal with the influence of nearby data points, seem often to be the dominant terms in the asymptotic error even for nonlocal sample smoothing estimators.

The nearest neighbor estimators of Breiman et al. (BMP) may be thought of as data driven sample smoothing estimators in which h is an estimate of $k/2nf(x)$. The bias term from our theorem is then

$$\frac{k^p}{2^p p!\, n^p} \left(f^{1-p}\right)^{(p)}$$

for order–p estimators with an h of this form. Then the integrated squared bias is necessarily infinite for any density with a tail exponent less than $1 + 1/2(p - 1)$, as with the Loftsgaarden–Quesenberry case. This is catastrophic for a global density estimate; and the only claimed advantage of the BMP estimator, integrating to one, is relevant only in the global case. In fact

Proposition 3. *For any strongly unimodal density (convex logarithm) the Breiman–Meisel–Purcell estimator of order 2 has uniformly greater AMSE than the Loftsgaarden–Quesenberry estimator of the same order.*

This may be checked by comparing the two AMSE expressions in light of the convexity condition $(\log f)'' \leq 0$. The condition is met by a number of important, though rather structureless, densities. However, it also applies in a neighborhood of any mode. It is also straightforward to consider the univariate Breiman et al. estimator as in Section 4.3. The bias turns out to be

$$O(h^2 + h/\log h)\,,$$

which is substantially worse than the rate of a fixed kernel estimator. We conclude that the BMP estimator is of doubtful utility for univariate densities.

4.4. *Other Results.* We have not attempted a full multivariate analysis of the Abramson estimator. We have however checked the performance when estimating $\phi(0,0)$, the bivariate independent Gaussian density, with a bivariate uniform kernel supported on the unit circle. It is easy to see that the kernel contribution is nonzero for samples (x, y) with $x^2 + y^2 \leq r_1^2$ or with $x^2 + y^2 \geq r_2^2$ for large n. The contribution to the bias from the neighborhood of $(0,0)$ is

$$\frac{1}{12}\pi h^4 + O(h^6)\,,$$

but the contribution from the tails can be shown to be

$$O\left(\frac{h^2}{(\log\sqrt{2\pi}h)^2}\right)\,. \tag{4.14}$$

Thus asymptotically, the tails contribute the most to the bias and the story is much the same as in the univariate example. The general multivariate pattern seems clear from (4.11) and (4.14).

5. Multivariate Balloon Estimators. Generalized kernel estimators were described in Equation (1.5). An appropriate generalization of a fixed kernel would be

$$\hat{f}(\mathbf{y}) = \frac{1}{n|\mathbf{H}|}\sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{y})\right)\,, \tag{5.1}$$

where $\mathbf{H}$ is a nonsingular matrix that generalizes the smoothing parameter, and K is a multivariate function centered at the origin which integrates to one. It will be convenient to write $\mathbf{H} = h\mathbf{A}$ where h is a positive scalar and $|\mathbf{A}| = 1$. With this parameterization, the shape of the multivariate kernel is determined by $\mathbf{A}$ and the volume by h.

Definition 2. *A multivariate order–p kernel meets the conditions for the univariate case, except that the moment conditions are*

$$\int z_1^{n_1} \cdots z_d^{n_d} K(\mathbf{z})\, d\mathbf{z} = 0 \quad \text{for} \quad 0 < \sum_{i=1}^{d} n_i < p$$

$$\int z_i^p K(\mathbf{z})\, d\mathbf{z} = \pm 1 \quad \text{for} \quad i = 1, \cdots, d\,.$$

There are two obvious approaches to constructing multivariate order–p kernels: the first just constructs a coordinatewise product. Let K_i be a univariate order–p kernel. Then

$$K(\mathbf{z}) = \prod_{i=1}^{d} K_i(z_i)$$

is a multivariate order–p product kernel. The second allows for the full range of linear scaling by making the kernel spherically symmetric; that is, the value of the kernel depends only on the Euclidean distance from the origin. Of course, Gaussian kernels satisfy both notions. Consider a univariate kernel K_1 symmetric about zero. Define its absolute moments about zero

$$a_k = \int |x|^k K_1(x)\, dx\,.$$

Then if $a_{d-1} > 0$ we may let

$$K(\mathbf{z}) = \frac{\Gamma(d/2)}{\pi^{d/2} a_{d-1}} K_1(||\mathbf{z}||)\,,$$

which is a multivariate kernel. To make it standard and order–p we need

Proposition 4. *The central moments of a spherically symmetric kernel K are determined by the spherical moments*

$$s_k = \int ||\mathbf{z}||^k K(\mathbf{z})\, d\mathbf{z}$$

by the relation

$$\int \prod_{i=1}^{d} z_i^{2n_i} K(\mathbf{z})\, d\mathbf{z} = s_{2n} \frac{\Gamma(d/2)}{\Gamma(1/2)^d} \frac{\prod_{i=1}^{d} \Gamma(n_i + 1/2)}{\Gamma(n + d/2)}$$

where $n = \sum_{i=1}^{d} n_i$.

Proof: A general position argument shows that the moments are all in fixed proportion to the corresponding spherical moment. The constants of proportionality are readily calculated from the Gaussian case. ■

Moments with any factor with exponent odd are zero. Therefore, we may construct a spherically symmetric order–p kernel from a univariate kernel K_1 with $a_{2k+d-1} = 0$ for $0 \le k \le p/2$ by rescaling so that

$$a_{p+d-1} = \frac{a_{d-1}\pi^{1/2}\Gamma\left((p+d)/2\right)}{\Gamma(d/2)\Gamma\left((p+1)/2\right)}$$

by the construction method described above.

Now a balloon estimate simply lets the scaling matrix in (5.1) depend on the point of estimation:

$$\hat{f}(\mathbf{y}) = \frac{1}{n|\mathbf{H}(\mathbf{y})|}\sum_{i=1}^{n} K\left(\mathbf{H}(\mathbf{y})^{-1}(\mathbf{x}_i - \mathbf{y})\right) . \tag{5.2}$$

For a balloon estimate at a point $\mathbf{y}$ where the density has order p derivatives continuous at $\mathbf{y}$ and absolutely integrable, we have

$$\text{asymptotic variance}(\mathbf{y}) = \frac{f(\mathbf{y})\int K^2}{nh(\mathbf{y})^d} \tag{5.3}$$

$$[\text{asymptotic bias}(\mathbf{y})]^2 = [h(\mathbf{y})^p g(\mathbf{y})]^2 ,$$

where g is a sum of terms depending on the p–th partial derivatives of f, the p–th central moments of K, and the scaling matrix $\mathbf{A}$. This makes possible an AMSE of $O(n^{-2p/(2p+d)})$. (The general expression for g, obtained by a Taylor's series argument, is complicated and will be omitted since we will not need it in the sequel.)

This looks like a simple generalization of the univariate case, but an important new phenomenon appears. The bias has several terms, involving the various moments of K and the partial derivatives of f in ways that depend on our choice of scaling matrix. This raises the possibility that the matrix might be chosen so that the terms cancel; and the asymptotic bias would then be of lower order. Consider the case $p = 2$, where K may be itself a density. In the univariate case, the asymptotic bias is zero precisely when the second derivative of f is zero; this occurs in general only at isolated points. In the multivariate case, let K have mean zero and identity covariance matrix. Let the scaling matrix for the fixed point $\mathbf{y}$ be $\mathbf{H} = h\mathbf{A}$, and let the matrix of second partial derivatives of f at that point be $\mathbf{S_y}$.

Proposition 5. *The bias of (5.2) may be approximated by*

$$|\text{asymptotic bias}| = \frac{h^2}{2} tr(\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A}) , \tag{5.4}$$

where tr denotes trace of a matrix. A proof is given in the Appendix. Our power to adjust the shape of the kernel through $\mathbf{A}$ has different implications at points $\mathbf{y}$ depending on the second derivative matrix $\mathbf{S_y}$. There are three cases:

Case I: $\mathbf{S_y}$ is positive definite or negative definite. In Figure 6 we display such a portion of the bivariate normal density. Recall that $|\mathbf{A}| = 1$ by construction. Then no choice of $\mathbf{A}$ makes the asymptotic bias zero, and the best possible AMSE($\mathbf{y}$) is of order $O(n^{-4/(d+4)})$. We need to minimize the absolute trace of a symmetric definite matrix, $\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A}$, whose determinant equals $|\mathbf{S_y}|$. The trace is the sum of the eigenvalues and the determinant is their product. Then by the standard fact that the arithmetic mean exceeds the geometric mean except when all values are equal, the minimum is obtained when all the eigenvalues are equal, at any of the solutions of

$$\mathbf{A}\mathbf{A}^{\mathrm{T}} = |\mathbf{S_y}|^{1/d}\mathbf{S_y}^{-1} .$$

It is easy to check that with this choice, $\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A} \propto I_d$, $|\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A}| = |\mathbf{S_y}|$, and $tr\,(\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A}) = d|\mathbf{S_y}|^{1/d}$. For spherically symmetric kernels, all these solutions are identical; notice that such a kernel is scaled completely by its covariance matrix $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. Following (5.3) and (5.4), we may now compute and then minimize the AMSE($\mathbf{y}$) with respect to h to get

Proposition 6.

$$h^*(\mathbf{y}) = \left[\frac{f(\mathbf{y})\int K^2}{nd|\mathbf{S_y}|^{2/d}}\right]^{\frac{1}{d+4}} \qquad \textit{and}$$

$$\mathrm{AMSE}^*(\mathbf{y}) = n^{\frac{-4}{d+4}} \left(\frac{1}{4}+\frac{1}{d}\right)^{\frac{2(d+2)}{d+4}} \left(\int K^2\right)^{\frac{4}{d+4}} f(\mathbf{y})^{\frac{4}{d+4}} |\mathbf{S_y}|^{\frac{2}{d+4}} .$$

FIGURE 6. *Negative definite (Case I) region of a bivariate normal density.*

Case II: $\mathbf{S_y}$ has both positive and negative eigenvalues. We will call the density *saddle–shaped* at any such point; it is curved upwards in some directions and downwards in others. A saddle–shaped portion of a bimodal bivariate normal density is shown in Figure 7.

Proposition 7. *Where the density is saddle-shaped, A may be chosen so that the bias is* $o(h^2)$.

The idea of the proof is that the sign of the bias is determined by the direction of curvature of the density. Where the curvature is upward in some directions and downward in others, a clever choice of kernel scale allows the two kinds of bias to cancel. The proof is given in the Appendix. The contribution of such a point to the AMISE may therefore be made negligible compared to that of the points in Case I, by choosing h of order $n^{-1/(d+8)}$. Notice that this cannot happen in 1 dimension; we need at least 2 different kinds of curvature to achieve cancellation.

FIGURE 7. *Saddle–shaped (Case II) region of a bimodal bivariate normal density.*

Case III: $\mathbf{S_y}$ is positive or negative semidefinite, with some eigenvalues zero. Thus, the Hessian for the density is of lower rank than the dimension of the space. The asymptotic bias, and therefore the AMSE, corresponds to that for a problem with fewer variables. As a result, it is of lower order in n than in case I. The contribution of these points to the AMISE may be made negligible by choosing appropriate scales along null and nonnull eigenvectors.

Example. Let f be a multinormal density with mean at the origin and identity covariance matrix. Then $\mathbf{S_y} = f(\mathbf{y})(\mathbf{y}\mathbf{y}^{\mathrm{T}} - \mathbf{I})$. Then for an arbitrary vector $\mathbf{z}$ compute $\mathbf{z}^{\mathrm{T}}\mathbf{S_y}\mathbf{z} = (\mathbf{z}^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{z} - \mathbf{z}^{\mathrm{T}}\mathbf{z})f$. We see by Cauchy's inequality that the expression is negative definite for $||\mathbf{y}||$ less than 1 (Case I); It is saddle for $||\mathbf{y}|| > 1$ (compare the direction toward the origin and perpendiculars to it) (Case II); and it is flat in the direction of the origin for $||\mathbf{y}|| = 1$ (Case III). Thus the usual rate of convergence of kernel density estimates will only apply to a central sphere corresponding to chi–squared values less than one (whose probability decreases with dimension). The rest of the normal density makes only a negligible contribution to AMISE.

Let us compare these asymptotically best–possible balloon estimators to a corresponding multivariate fixed kernel estimator. That is, we select scale h and proportions $\mathbf{A}$ for the kernel freely, but use the same one over all values of $\mathbf{y}$. The optimal choices for h and $\mathbf{A}$ were characterized by Deheuvels (1977). Integrating the variance and the square of the bias over space gets

$$\mathrm{AMISE} = \frac{\int K^2}{nh^d} + \frac{h^4}{4}\int tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A})\,.$$

Minimize over h and characterize $\mathbf{A}$ implicitly to get

$$\mathrm{AMISE_{opt}} = n^{\frac{-4}{d+4}}\left(\frac{1}{4}+\frac{1}{d}\right)^{\frac{4}{d+4}}\left(\int K^2\right)^{\frac{4}{d+4}}\left[\min_{\mathbf{A}}\int tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S_y}\mathbf{A})\right]^{\frac{d}{d+4}}\,. \qquad (5.5)$$

Example. Table 2 shows the ratio of the best for the balloon estimator to this fixed best for multivariate normal data, for some values of d. The integral (5.5) was obtained as follows: we may always transform to an identity covariance matrix. Then a standard calculation with normal moments expresses the integral as a term involving the sum of the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and a term involving the sum of the squared eigenvalues. Since the product of these eigenvalues is one, the integral is minimized when all eigenvalues are equal, so that $\mathbf{A} = \mathbf{I}$. Then the integral is simply $3\,d \cdot 2^{-(d+2)}\pi^{-d/2}$. The integral in the numerator was evaluated numerically using Mathematica. Clearly the balloon estimator is potentially enormously better than the fixed kernel for multivariate density estimates. The reason is equally clear: the two compete only over the unit ball; elsewhere the bias of the balloon estimator is of lower order. This advantage is offset to some extent by our interest in modes; these occur in areas of negative definite curvature, where a fixed kernel has error of the same order in n as an optimal balloon estimate. In Case II and III regions, the fixed kernel has the same order error as in the Case I region, while the optimal estimate has higher order (better) error there. Nevertheless, multivariate optimal balloon estimators are worth exploring. Perhaps they could be calibrated by a pilot density estimate. We will not address the issue here.

Table 2

Dimension	**Efficiency of Fixed to Balloon**
2	45.46%
3	30.21%
4	18.59%
5	10.64%
6	5.70%

Table 2. Relative efficiencies of optimal fixed to optimal balloon estimator for the multinormal density with identity covariance matrix as function of dimension.

Loftsgaarden and Quesenberry (1965) suggested a multivariate k–th nearest neighbor estimate of the form

$$\hat{f}(\mathbf{y}) = \frac{1}{nh_k^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) ,$$

where h_k is the radius of the smallest sphere containing k observations centered at $\mathbf{y}$. Our more general kernel density estimates suggest instead

$$\hat{f}(\mathbf{y}) = \frac{1}{n|\mathbf{H}_k|} \sum_{i=1}^{n} K\left(\mathbf{H}_k^{-1}(\mathbf{x}_i - \mathbf{y})\right) , \tag{5.6}$$

where $(\mathbf{x}_i - \mathbf{y})^{\mathrm{T}}(\mathbf{H}_k\mathbf{H}_k^{\mathrm{T}})^{-1}(\mathbf{x}_i - \mathbf{y}) \leq 1$ is the smallest such ellipsoid that contains k sample points. As before we write $\mathbf{H}_k = h_k\mathbf{A}$ where $|\mathbf{A}| = 1$; the eccentricity of the ellipsoid thus remains fixed over space while we let the volume adapt to the local density of sample points. We call this a Mahalanobis k–th nearest neighbor estimator. An even greater degree of local adaptivity would be achieved by letting $\mathbf{A}$ vary freely as well. This raises computational difficulties, and will not be pursued here.

As $k \to \infty$ and $k/n \to 0$ in the Mahalanobis estimator,

$$h_k \to \left[\frac{k}{nf(\mathbf{y})V_d}\right]^{1/d},$$

where V_d is the volume of the unit d–ball, at any point $\mathbf{y}$ where the density is positive. As in the univariate case, we will study the balloon estimator in which h is replaced by its asymptotic value. For a kernel of order 2

$$\text{AMSE}(\mathbf{y}) = \frac{f(\mathbf{y})^2 V_d \int K^2}{k} + \frac{k^{4/d}}{4n^{4/d}V_d^{4/d}} \frac{tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A})}{f(\mathbf{y})^{4/d}}$$

and when we integrate over all values of $\mathbf{y}$

$$\text{AMISE} = \frac{\int f(\mathbf{y})^2\, d\mathbf{y} V_d \int K^2}{k} + \frac{k^{4/d}}{4n^{4/d}V_d^{4/d}} \int \frac{tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A})}{f(\mathbf{y})^{4/d}}\, d\mathbf{y}\,.$$

As before we may now minimize this quantity with respect to k and $\mathbf{A}$ to get

Proposition 8.

$$k_{\text{opt}} = n^{\frac{4}{d+4}} V_d \left[\frac{d\int f^2 \int K^2}{\min_{\mathbf{A}} \int \frac{tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A})}{f^{4/d}}}\right]^{d/(d+4)}$$

$$\text{AMISE}_{\text{opt}} = n^{\frac{-4}{d+4}} \left(\frac{1}{4}+\frac{1}{d}\right)^{\frac{4}{d+4}} \left(\int K^2\right)^{\frac{4}{d+4}} \left(\int f^2\right)^{\frac{4}{d+4}} \left[\min_{\mathbf{A}} \int \frac{tr^2(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A})}{f^{4/d}}\right]^{\frac{d}{d+4}}.$$

Example. To see how well nearest neighbor estimates can perform, it is natural to compare them to fixed kernel estimates; since optimizing k and $\mathbf{A}$ is comparable to optimizing h and $\mathbf{A}$. We will look at the ratio of the AMISE's for the two estimators for multivariate normal data. Because of our freedom of scaling, the result is invariant under changes in the covariance matrix; we will therefore do all our calculations for identity covariance matrix. By symmetry, in both numerator and denominator $\mathbf{A} = \mathbf{I}$. We calculate

$$\int tr^2\mathbf{S}_{\mathbf{y}} = \frac{d(d+2)}{2^{d+2}\pi^{d/2}}$$

$$\int f(\mathbf{y})^2 = \frac{1}{s^d\pi^{d/2}}$$

$$\int \frac{tr^2\mathbf{S}_{\mathbf{y}}}{f(\mathbf{y})^{4/d}} = \left(\frac{d}{d-2}\right)^{(d+4)/2} \frac{d^2-6d+16}{2^d\pi^{(d-4)/2}}$$

for $d > 2$. Therefore the asymptotic relative efficiency of nearest neighbor estimates compared to fixed kernel estimates for Gaussian variables is

$$\frac{\text{AMISE}_{\text{opt}(h)}}{\text{AMISE}_{\text{opt}(k)}} = 2^{2d/(d+4)} \left(\frac{d-2}{d}\right)^{\frac{d(d+2)}{2(d+4)}} \left[\frac{d^2-4}{d^2-6d+16}\right]^{\frac{d}{d+4}} .$$

Table 3 shows this efficiency for various numbers of variables. The maximum is achieved with fifteen variables; the limit as the number of variables grows is $4/e$, or approximately 1.472. Our experience is qualitatively similar with other smooth densities. We conclude that the nearest neighbor balloon estimator is of little interest in the usual, low–dimensional cases; but may well be worth considering for many variables. This bolsters the conclusions of Boswell (1983).

Table 3.

Number of Variables	**Efficiency of k–NN to Fixed**
1	0
2	0
3	0.483
4	0.866
5	1.146
15	1.545
100	1.491

Table 3. Relative efficiencies of optimal multivariate k–NN estimates to fixed kernel for multinormal densities. Both estimators take as the kernel the uniform density on the unit sphere in $\mathbb{R}^d$.

6. Discussion. We have examined three classes of multivariate adaptive estimators. We have seen some dramatic differences between the univariate and multivariate cases with adaptive estimation. We have shown that k–NN estimators behave well only in higher dimensions. We have seen that the non–clipped sample smoothing estimator of Abramson is not necessarily a local procedure for densities supported on the real line, in which case the pointwise bias of the procedure is only slightly faster than $O(h^2)$ but not $O(h^4)$. This finding does not seem inconsistent with the results of Hall and Marron (1988), who considered the extension of Abramson's procedure to the entire density. They also introduced the condition that the density is bounded locally from below; this requires that the density have finite support, which would asymptotically eliminate the tail problem.

Existing techniques all tend to adapt based on some function of the *level* of the unknown density function. Optimal adaptive estimation requires attention be paid not only to the level of the unknown density but also to its local or global curvature. Optimal adaptive estimation leads to local linear transformations of the data. For certain dimensions, we have seen that very significant gains are possible using balloon–type estimators. Roughly speaking, this is because the MISE is dominated by errors only from regions surrounding peaks while contributions elsewhere are of lower order. This phenomena does not exist in the univariate setting. It is possible to use nearest neighbor distances to estimate not only the level of the unknown density function but also its curvature. Hall (1983) has given a theoretical demonstration of such a procedure.

For regression, more dramatic improvements are surely possible since no integral or positivity constraints exist for the curve as in density estimation. However, much of the machinery is the same and there are some important similarities in our conclusions.

We conclude with the following (perhaps not so obvious) observation about adaptive estimators. Adaptive estimation in a nonoptimal fashion can be inferior to nonadaptive methods for sufficiently large samples. Adapting on level alone seems to work very well for small samples, but asymptotically curvature cannot be ignored in general. Given the complexity of correctly applying optimal multivariate adaptive algorithms, we suspect the use of fixed but higher order kernel algorithms will gain favor. The fact that nearest neighbor estimators are superior to fixed kernel estimators beyond 4 dimensions should be reassuring to workers in classification using these estimators. Further work understanding the new adaptive estimators proposed will be required. Designing practical algorithms based upon these results will be quite challenging. However, some univariate work suggests it is not an impossible task even with modest sample sizes.

APPENDIX

Proof of Theorem 1. Given a real sample $\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n$, then its empirical cumulative distribution function is

$$\hat{F}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X} \left(\prod_{j=1}^{d} \left[x_i^j, \infty \right) \right),$$

where $\mathbf{X}$ is the indicator function on the given orthant. Write the density estimator as an operator $\hat{f}(\mathbf{y}) = T_{\mathbf{y}}(\hat{F}_n)$. Then define

$$K(\mathbf{x}, \mathbf{y}, \hat{F}_n) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[T_{\mathbf{y}} \left(\{1 - \epsilon\} \hat{F}_n + \epsilon \mathbf{X} \left\{ \prod_{j=1}^{d} [x^j, \infty) \right\} \right) - (1 - \epsilon) T_{\mathbf{y}}(\hat{F}_n) \right]$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[T_{\mathbf{y}} \left(\hat{F}_n + \epsilon \left\{ \mathbf{X} \left(\prod_{j=1}^{d} [x^j, \infty) \right) - \hat{F}_n \right\} \right) - T_{\mathbf{y}}(\hat{F}_n) \right] + T_{\mathbf{y}}(\hat{F}_n)$$

$$= DT_{\mathbf{y}}(\hat{F}_n) \left[\mathbf{X} \left(\prod_{j=1}^{d} [x^j, \infty) \right) - \hat{F}_n \right] + \hat{f}(\mathbf{y}),$$

where $DT(\mathbf{z})[\mathbf{w}]$ is the Gateaux derivative of T at $\mathbf{z}$ in the direction $\mathbf{w}$. Proposition 2.7 of Tapia (1971) gives us that $DT_{\mathbf{y}}$ is linear in the second argument, so

$$\frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x}_i, \mathbf{y}, \hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} DT_{\mathbf{y}}(\hat{F}_n) \left[\mathbf{X} \left(\prod_{j=1}^{d} [x_i^j, \infty) \right) - \hat{F}_n \right] + \hat{f}(\mathbf{y})$$

$$= DT_{\mathbf{y}}(\hat{F}_n) \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X} \left(\prod_{j=1}^{d} [x_i^j, \infty) \right) - \hat{F}_n \right] + \hat{f}(\mathbf{y}) = \mathbf{0} + \hat{f}(\mathbf{y}).$$

PROOF OF THEOREM 2. Without loss of generality, let $y = 0$. Then

$$E\left(\hat{f}(0)\right) = \int \frac{1}{h(x)} K\left(\frac{x}{h(x)}\right) f(x)\, dx\,.$$

Make the one–to–one change of variables $z = x/h(x)$ to get

$$E\left(\hat{f}(0)\right) = \int K(z) \frac{f(x[z])}{1 - zh'(x[z])} dz\,.$$

We will expand the second term in the integrand in a variant of a Lagrange expansion. For a fixed choice of z, let $x^* = x(z)$ be the solution of $x = zh(x)$. Consider the contour integral in the complex plane

$$\frac{1}{2\pi i} \oint_C \frac{f(x)}{x - zh(x)} dx\,,$$

where the contour makes a single right–hand loop about x^*. By the residue theorem this integral becomes

$$\frac{f(x^*)}{1 - zh'(x^*)} = \frac{f\,(x(z))}{1 - zh'\,(x(z))}.$$

Now evaluate the contour integral again by expanding the integrand as a power series in z; apply the generalized residue theorem term by term with the pole at zero:

$$f(0) + z(fh)'(0) + \cdots + \frac{z^p}{p!}(fh^p)^{(p)}(0) + o(z^p)\,.$$

Now substitute this series into the expectation and integrate term–by–term to get

$$E\left(\hat{f}(0)\right) = f(0) + \frac{1}{p!}(fh^p)^{(p)}(0) + o(h^p)\,.$$

PROOF OF PROPOSITION 5.

$$\text{Bias} = E[\hat{f}(\mathbf{y})] - f(\mathbf{y}) = \int \frac{1}{h^d} K(h^{-1}\mathbf{A}^{-1}[\mathbf{x}-\mathbf{y}])\, f(\mathbf{x})\, d\mathbf{x} - f(\mathbf{y})\,.$$

Now make the change of variables $\mathbf{z} = h^{-1}\mathbf{A}^{-1}[\mathbf{x}-\mathbf{y}]$; then continuing

$$= \int K(\mathbf{z})\, f(\mathbf{y} + h\,\mathbf{A}\,\mathbf{z})\, d\mathbf{z} - f(\mathbf{y})\,.$$

Expand f is a 3-term Taylor's series; then

$$= \int K(\mathbf{z})[f(\mathbf{y}) + h\nabla f(\mathbf{y})\,\mathbf{A}\,\mathbf{z} + \frac{h^2}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A}\,\mathbf{z} + o(h^2)]\, d\mathbf{z} - f(\mathbf{y})$$

where ∇f is the gradient, a row vector

$$= \int K(\mathbf{z})[f(\mathbf{y})]\, d\mathbf{z} - f(\mathbf{y}) + h\nabla f(\mathbf{y})\,\mathbf{A}\int \mathbf{z}\, K(\mathbf{z})\, d\mathbf{z} + \frac{h^2}{2}\int \mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A}\,\mathbf{z}\, K(\mathbf{z})\, d\mathbf{z} + o(h^2)\,.$$

The first two terms sum to zero and the third is zero, by assumption on the kernel, so

$$= \frac{h^2}{2}\int tr\,[\mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A}\,\mathbf{z}]\, K(\mathbf{z})\, d\mathbf{z} + o(h^2)$$

since the trace of a scalar is just the scalar,

$$= \frac{h^2}{2}\int tr\,[\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A}\int \mathbf{z}\,\mathbf{z}^{\mathrm{T}} K(\mathbf{z})\, d\mathbf{z}] + o(h^2)$$

since $tr\,(\mathbf{BC}) = tr\,(\mathbf{CB})$. But the integral is the identity matrix, by assumption on the kernel.

PROOF OF PROPOSITION 7. Write the spectral decomposition

$$\mathbf{S}_{\mathbf{y}} = \mathbf{O}^{\mathrm{T}}\, diag\,(\lambda_1, \ldots, \lambda_k, -\lambda_{k+1}, \ldots, -\lambda_{k+\ell})\,\mathbf{O}$$

where $\lambda_i > 0$ and $\mathbf{O}$ is the orthogonal matrix of normalized eigenvectors. Then let

$$\mathbf{A} = \mathbf{O}^{\mathrm{T}}\, diag\left(\sqrt{a\,\ell\,\lambda_1^{-1}}, \ldots, \sqrt{a\,\ell\,\lambda_k^{-1}}, \sqrt{a\,k\,\lambda_{k+1}^{-1}}, \ldots, \sqrt{a\,k\,\lambda_{k+\ell}^{-1}} \right) \mathbf{O}$$

where a is chosen so that $|\mathbf{A}| = 1$. Then

$$\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\mathbf{A} = diag\,(\, a\,\ell, \ldots, a\,\ell, -a\,k, \ldots, -a\,k\,)$$

and

$$tr\,(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathbf{y}}\,\mathbf{A}\,) = a\,\ell\,k - a\,k\,\ell = 0\,.$$

Compare this with (5.4). Q.E.D.

Acknowledgment. We would like to thank the referees for their helpful remarks and, in particular, for pointing out the importance of clipping in the Abramson estimator.

REFERENCES

Abramson, I.S. (1982). On bandwidth variation in kernel estimates – a square root law. *Annals of Statistics* **10** 1217–1223.

Boswell, S.B. (1983). Nonparametric mode estimation for higher dimensional densities. Doctoral Dissertation, Rice University, Houston.

Bowman, A.W. and Foster, P.J. (1991). Adaptive smoothing and density based tests of multivariate normality. Manuscript.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135–144.

Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 179–189.

Deheuvels, P. (1977). Estimation nonparametrique de la densite par histogrammes generalises II. *Publications de l'Institute Statistique de l'Universite Paris* **22** 1–23.

deMontricher, G.F., Tapia, R.A., and Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics* **3** 1329–1348.

Devroye, L., and Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View.* New York: John Wiley & Sons.

Dodge, Y. and Lejeune, M. (1986). Some difficulties involving the nonparametric estimation of a density function. *Journal of Official Statistics* **2** 193–202.

Fix, E. and Hodges, J.L. (1951) Nonparametric discrimination: Consistency properties. Report No. 11, USAF School of Aviation Medicine, Randolph Field, Texas.

Hall, P. (1983). On near neighbour estimates of a multivariate density. *Journal of Multivariate Analysis* **13** 24–39.

Hall, P. and Marron, J.S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields* **80** 37–49.

Hall, P. and Wand, M.P. (1988). Minimizing L_1 distance in nonparametric density estimation. *Journal of Multivariate Analysis* **26** 59–88.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69** 383–393.

Hüsemann, J.A., and Terrell, G.R. (1991). Optimal parameter choice for error minimization in bivariate histograms. *Journal of Multivariate Analysis.* **37** 85–103.

Jones, M.C. (1990). Variable Kernel Density Estimates. *Australian Journal of Statistics* **32** 361–371.

Kogure, A. (1987). Asymptotically optimal cells for a histogram. *Annals of Statistics* **15** 1023–1030.

Loftsgaarden, D.O., and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics* **36** 1049–1051.

Mack, Y.P. and Rosenblatt, M. (1979). Multivariate K-nearest neighbor density estimates. *Journal of Multivariate Analysis* **9** 1–15.

Parzen, E. (1962). On the estimation of probability density function and mode. *Annals of Mathematical Statistics* **33** 1065–1076.

Rosenblatt, M. (1956). On some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27** 832–837.

Scott, D.W. (1982). Optimal meshes for histograms using variable–width bins. Technical Report, Dept. of Math Sciences, Rice University, Presented at ASA Annual Meetings, Cincinnati, Ohio.

Scott, D.W. and Wand, M.P. (1991). Feasibility of multivariate density estimates. *Biometrika.* **78** 197–206.

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

Tapia, R.A. (1971). The differentiation and integration of nonlinear operators. In *Nonlinear Functional Analysis and Applications* (ed. L.B. Rall) 45–101.

Terrell, G.R. and Scott, D.W. (1980). On improving convergence rates for nonnegative kernel density estimators. *Annals of Statistics* **8** 1160–1163.

Terrell, G.R. and Scott, D.W. (1983). Variable window density estimates. Paper presented at ASA meetings in Toronto.

Tukey, P.A. and Tukey, J.W. (1981). Data–driven view selection: Agglomeration and sharpening. In *Interpreting Multivariate Data*, V. Barnett, ed. Chichester: John Wiley & Sons, pp. 215–243.

Walter, G., and Blum, J. (1979). Probability density estimation using delta sequences. *Annals of Statistics* **7** 328–340.

Wand, M.P., Marron, J.S., and Ruppert, D. (1991). Transformations in Density Estimation. *J. Amer. Statist. Assoc.* **86** 343–353.

Wertz, W. (1978). *Statistical Density Estimation: A Survey.* Göttingen: Vandenhoeck and Ruprecht.

Wolfram, S. (1988). *Mathematica: A System for Doing Mathematics by Computer.* Redwood City, California: Addison–Wesley.

Worton, B.J. (1989). Optimal smoothing parameters for multivariate fixed and adaptive kernel methods. *J. Statist. Comput. Simul.* **32** 45-57.

TITLE FOR FIGURES

1. The equivalent kernel function for a frequency polygon.

2. Optimal local bandwidths for a standard normal density.

3. Computed optimal exact MISE for the non–clipped Abramson estimator (solid line) for standard normal data and exact MISE for optimal fixed bandwidth estimator (dashed line) with normal kernel. The two small parallel vertical lines represent factors of 1.5 and 1.2.

4. Regions where data contribute to the non–clipped Abramson estimate at origin.

5. Four bandwidths for 1000 standard normal data in the non–clipped Abramson estimator with a $U(-1,1)$ kernel. The middle two are optimal in different senses; see text.

6. Negative definite (Case I) region of a bivariate normal density.

7. Saddle–shaped (Case II) region of a bimodal bivariate normal density.