LONDHE, ANIL RAMCHANDRA

NONPARAMETRIC DENSITY ESTIMATION USING KERNELS WITH VARIABLE SIZE WINDOWS

Iowa State University PH.D. 1980

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

Nonparametric density estimation using kernels with variable size windows

by

Anil Ramchandra Londhe

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1980

TABLE OF CONTENTS

1. A REVIEW OF DENSITY ESTIMATION METHODS

1.1. Introduction

Density estimation is possibly the most important topic in applied statistics. When we do not know the density, f(x), we must infer its characteristics from a sample X_1, $X_2,\ldots,X_n$ before we can make any inferences or predictions. Classical density estimation involves initial screening of the data, which leads to hypothesizing that the data came from a particular parametric family of density curves. The process of estimating the parameters of that family of densities and hypothesis testing to see if this hypothesis is tenable follows. In the absence of a priori information, the initial screening of the data is usually done with the time honored histogram which leaves a lot to be desired as a density estimate. It was in 1951 that Fix and Hodges (1951) suggested some improvement in the method of producing a histogram, reducing the subjectivity to some extent. This soon led to nonparametric estimates of f(x) which are continuous and, so to some extent, could bypass the usual inference chain of parametric density estimation.

The estimators suggested fall roughly into the following categories:

(a) kernel (or window) estimators,

(b) spline estimators,

(c) series estimators,

(d) maximum likelihood, and histogram type estimators.

We shall examine these nonparametric methods of univariate density estimation in this chapter. Chapter 2 takes a closer look at kernel density estimators, nearest neighbor estimators in particular, and develops two new estimators. Chapter 3 presents simulation results for these estimators and compares their performance to some kernel estimators.

1.2. Kernel Density Estimators

Fix and Hodges (1951), in a paper on nonparametric discrimination, used a "running histogram" as a density estimate rather than assume an underlying normal distribution or choose the usual histogram. They subjectively chose an interval width, h, and then estimated the density at any given point as being proportional to the number of observations falling within an interval of width h centered at the point under consideration. This running histogram, or naive estimator led Rosenblatt (1956) to define a class of univariate estimators, known as kernel, or window, estimators, which can be written as

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \qquad (1.2.1)$$

where $X_1, X_2, \ldots, X_n$ are assumed to be independently and identically distributed with $f(\cdot)$, the unknown density and

K(·) is the kernel. For the naive estimator,

$$K(x) = \frac{1}{2} \quad \text{for} \quad |x| \leq 1$$
$$= 0 \quad \text{otherwise.} \qquad (1.2.2)$$

The larger the value of h, the coarser the grouping, so that as n becomes large, h should become smaller. If, the kernel K satisfies the following conditions:

$$K \text{ is symmetric}$$
$$\int K(u)du = 1$$
$$\int K^2(u)du < \infty \qquad (1.2.3)$$
$$\int K(u)|u|^3du < \infty,$$

then Rosenblatt, showed that this class of estimators was pointwise and integratedly consistent in quadratic mean provided $h = h_n$ is chosen suitably. The optimal choice of h_n, which depends on the unknown density f, leads to the convergence of the MSE, the mean squared error, at the rate of $O(n^{-4/5})$. Parzen (1962) imposing further constraints on K, showed asymptotic unbiasedness, and then listed several forms of the kernel which satisfy these constraints. The kernels considered include the rectangular, triangular, normal and cauchy density functions. If

$$h_n \to 0$$

and

$$nh_n \to \infty, \tag{1.2.4}$$

he also showed that the MSE converges to zero, and under further conditions, showed the asymptotic normality of $\{\hat{f}_n(x)\}$ for fixed x. Many authors followed Parzen's approach, changing the assumptions about f(x), the conditions imposed on K(·), and $\{h_n\}$, and proved consistency properties for $\hat{f}_n$. Amongst them are Nadaraya (1963, 1965), Murthy (1965), Woodroofe (1967), Bhattacharya (1967), Schuster (1969, 1970) and Silverman (1978a), Craswell (1965) generalized Parzen's results to estimation on a topological group, and Borwanker (1971) considered strictly stationary processes.

Other authors found asymptotically optimal forms for K(·), the kernel. Bartlett (1963), for example, proposed

$$\begin{aligned} K(u) &= \frac{9}{8h}(1 - \frac{5u^2}{3h^2}) \quad \text{for} \quad |u| \leq h \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{1.2.5}$$

which optimizes a larger group of terms in the asymptotic expansion of the mean squared error. Watson and Leadbetter (1963) used the integrated mean squared error as a criterion and arrived at

$$\phi_K(t) = \frac{|\phi_f(t)|^2}{(\frac{1}{n} + \frac{[(n-1)/n]}{|\phi_f(t)|^2})} , \tag{1.2.6}$$

where $\phi_f(t)$ is the Fourier transform of $f(\cdot)$, assuming ϕ_f to be square integrable. They demonstrated the optimal form for K corresponding to various densities f and showed that the integrated mean squared error cannot be better than $O(1/n)$. Woodroofe (1968) presented a two-stage procedure to estimate $f(\cdot)$ when the kernel K has been specified. After two initial guesses for h_n, which are used to obtain rough estimates for f and the first nonvanishing moment of f, a new value h_n for h is computed. This h_n is used to estimate $f(\cdot)$ in the usual way. Woodroofe showed that this method converged asymptotically in mean squared error. Nadaraya (1974) provided a similar two-stage procedure, but was based on a different optimality criterion. Whittle (1958) suggested a linear estimator of f(x) of the form

$$\hat{f}(x) = \frac{\Sigma W_x(X_j)}{N} \tag{1.2.7}$$

where W is a weight function to be optimized and then considered a Poissonization of the problem. Consider N to be distributed as Poisson (M) and estimate $\phi(x) = Mf(x)$ by

$$\hat{\phi}(x) = \frac{\Sigma W_x(X_j) \cdot M}{N} , \tag{1.2.8}$$

using as an optimization criterion

$$\min E_p E_s |\hat{\phi}(x) - \phi(x)|^2. \tag{1.2.9}$$

The suffices in (1.2.9) refer to the prior distribution of the parameters and the sampling fluctuations, respectively. Anderson (1969a,b), after a fairly extensive study, concluded that the actual kernel K(·) used makes little difference to the optimum value of the integrated mean squared error, but that the optimal value of h_n differs for different kernels. The normal kernel performs satisfactorily when estimating normal and relatively symmetric densities, but not when estimating the negative exponential. Fryer (1977) recommended that skew data should be transformed nearer to symmetry before using the estimation procedure, and then the resulting estimate should be transformed back. Nadaraya (1964) and Watson (1964) considered estimating the regression curve of Y on X,

$$E(Y|X) = \frac{\int yf(x,y)dy}{\int f(x,y)dy} \equiv m(x). \tag{1.2.10}$$

As an estimator, they used

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}, \tag{1.2.11}$$

with a symmetric $K(\cdot)$ and proved some asymptotic results. In a later paper, Nadaraya (1965) considered the regression problem of Y on X where

$$X = Y + Z, \quad Z \sim N(0,\sigma^2), \tag{1.2.12}$$

and the density of Y being unknown. Since

$$E(Y|x) = \sigma^2 \frac{f'(x)}{f(x)} + x, \tag{1.2.13}$$

he proposed

$$\hat{m}(x) = \sigma^2 \frac{\psi_n(x)}{\hat{f}_n(x)} + x, \tag{1.2.14}$$

$$\psi_n(x) = \frac{\hat{f}_n(x+h) - \hat{f}_n(x-h)}{2h}$$

and proved some consistency properties with

$$h_n = n^{-\theta}, \quad 0<\theta<\tfrac{1}{2}. \tag{1.2.15}$$

Another group of papers is concerned with estimating the hazard function

$$z(x) = \frac{f(x)}{1-F(x)} . \tag{1.2.16}$$

Watson and Leadbetter (1964a,b) used as estimators,

$$\frac{\hat{f}_n(x)}{(1-\int_0^x \hat{f}_n(t)dt)} \tag{1.2.17}$$

and

$$\frac{f_n(x)}{1-F_n(x)},$$

$F_n(\cdot)$ being the sample distribution function. Both the estimators are asymptotically unbiased under suitable conditions.

Nearest neighbor estimators, which are kernel estimators also, are considered in Chapter 2.

1.3. Spline Estimators

Boneva, Kendall and Stefanov (1971) used the histogram as a starting point for a smoothed estimator of an unknown probability density function. Starting with the data in the form of a histogram, with the cell width ε specified, they found a one to one, linear invariant and bi-continuous mapping onto a Hilbert space of smooth functions and called the resulting form a histospline. The histospline is a quadratic spline, i.e., a continuous and continuously differentiable function which is a quadratic in each fixed interval and square integrable. It is also constrained to integrate to the same value as the original histogram over every cell, but is not necessarily nonnegative. To apply it to raw data, it can be written in the form of a kernel estimator and hence enjoys the attributes of that class of estimators. Schoenberg (1972) and Lii and Rosenblatt (1975) presented similar modifications of the histospline estimator, producing a less wiggly and nonnegative estimator, Wahba (1975) considered the statistical properties of a slight generalization of the histospline for densities with finite support. Instead of the second derivatives being assumed zero at the end points of the interval, their values can be estimated from the data.

Wahba (1971) considered another approach. A local estimate of f at x can be based on the derivative of an mth degree polynomial estimate of F in the neighborhood of x obtained by the Lagrangian interpolation formulae. Wahba showed that this estimator is pointwise consistent in mean squared error at a slightly faster rate than the kernel estimators under stated conditions. Van Ryzin (1969) earlier had derived a special case of this estimator.

1.4. Series Estimators

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed with the unknown probability density function, $f(x)$, $x \varepsilon R$. Let $r(\cdot)$ be a known weight function so that the inner product

$$(\phi,\psi) = \int_R \phi(x)\psi(x)r(x)dx \qquad (1.4.1)$$

defines a Hilbert space $L_2(r)$ and let an orthonormal basis $\{\phi_{R,N}\}$, $k=1,2,\ldots,N$ exist for the N-dimensional subspace E_n, then

$$f_n(x) = \sum_{k=1}^{N} a_{kN}\phi_{k,N}$$

$$\equiv \sum_{k=1}^{N} (\phi_{k,N},f)\phi_{k,N} \qquad (1.4.2)$$

is the mean squared approximation to f(x). Cencov (1962a,b) considered

$$a^*_{kN} = \frac{1}{n}\sum_{k=1}^{n}\phi_{k,N}(X_i)r(X_i) \qquad (1.4.3)$$

as a strongly consistent estimator for a_{kN} and proposed

$$\hat{f}_n(x) = \sum_{k=1}^{N} a^*_{kN}\phi_{k,N} \qquad (1.4.4)$$

as an estimator for f(x). The choice of E_N and n contribute to the closeness of $\hat{f}_n(x)$ to f(x). He proved several theorems relating to the degree of approximation and proposed a stopping rule for N, the number of terms in the series. The resulting estimator is not necessarily always nonnegative. Several authors considered the same formulation as Cencov and studied the properties of the resulting estimators. Schwartz (1967) considered the case when $r(x) \equiv 1$, $N = N(n)$ and $\phi_k(\cdot)$ is the kth Hermite function over the real line and proved several consistency properties requiring conditions such as $\frac{N}{n} \to 0$ as $n \to \infty$. Blaydon (1967) considered a generalization, estimating both $F(\cdot)$ and $f(\cdot)$ by a linear combination of functions using the criterion of minimum least squares. Kashyap and Blaydon (1968) evaluated a^*_{kN} by a gradient type technique and gave an example using the first three Laguerre polynomials over [0,4] to estimate the distribution function corresponding to an exponential density. Watson (1969) introduced a general weight function λ in the estimator,

$$\hat{f}_n(x) = \sum_{k=1}^{\infty} \lambda_{k(N)} a_{kN} \phi_{k,N}(x); \qquad (1.4.5)$$

however,

$$\begin{aligned} \lambda_{k(N)} &= 1, \quad \text{for } k = 1,2,\ldots,n \\ &= 0 \quad \text{elsewhere} \end{aligned}$$

is a better form as far as application is concerned.

Tarter and Kronmal (1967) extend the Cencov model to cover both $F(\cdot)$ and $f(\cdot)$, but only with finite support. They chose the trigonometric functions $\{\cos k\pi x\}$, $\{\sin k\pi x\}$, and $\{\cos k\pi x, \sin k\pi x\}$ for their ϕ's, as they require the orthogonal series used to estimate $F(\cdot)$, still to be orthogonal when differentiated to yield the estimate for $f(\cdot)$. They give a stopping rule for the number of terms to be included in the series and suggest that ten should be sufficient for all practical purposes. Fellner (1974), in a synthesis of the papers by Whittle (1958), Tarter and Kronmal (1967) produced a multistep estimating procedure which bypassed the usual stopping rule problem by using a hypothesis testing technique. Crain (1974) proposed a maximum likelihood approach to estimating the coefficients of the orthogonal series. Several authors, e.g. Watson (1969), Fellner and Tarter (1971), and Tarter and Raman (1971), noted the theoretical equivalence under stated conditions of the Fourier estimators and kernel estimators, but in practice, the resulting estimators are locally very different, especially when n is relatively small.

1.5. Maximum Likelihood Estimators

Wegman (1969; 1970a,b) employed a maximum likelihood approach to obtain a modified histogram estimator for an unknown density f with domain (a,b]. His estimators are of the form

$$f(x|\underline{c}) = \sum_{j=0}^{k-1} c_j I_{(a_j,a_{j+1})}(x), \qquad (1.5.1)$$

where

$$I_{(c,d]}(x) = 1 \quad \text{if} \quad x \varepsilon (c,d]$$
$$= 0 \quad \text{otherwise,}$$

$$c_j > 0, \; j = 0,1,\ldots,k,$$

$$\sum_{j=0}^{k-1} c_j(a_{j+1}-a_j) = 1,$$

$$a_0 = a$$
$$a_k = b$$
$$\underline{c} = (c_1,c_2,\ldots,c_k).$$

The criterion function to be maximized based on a sample of size n, $\{X_1,X_2,\ldots,X_n\}$ is

$$L(\underline{c}) = \prod_{i=1}^{n} \hat{f}(X_i/\underline{c}), \tag{1.5.2}$$

subject to the constraint that at least m(n) observations must fall into each of the k intervals where

$$k \leq [\frac{n}{m(n)}] \quad \text{and}$$

$$m(n) \to \infty \quad \text{faster than } O(\sqrt{\log(\log n)}). \tag{1.5.3}$$

For k = nm(n), the solution is

$$\hat{f}(x) = \frac{1}{k}\sum_{j=0}^{k-1} \frac{1}{a_{j+1}-a_j} \; I_{(a_j,a_{j+1}]}(x),$$

$$\begin{aligned} a_0 &= a \\ a_1 &= X_{(1m)} \\ a_2 &= X_{(2m)} \\ a_{k-1} &= X_{\{(k-1)m\}} \end{aligned} \tag{1.5.4}$$

and

$$a_k = b.$$

Thus, Wegman used interval widths of the histogram, which varied across the data base in a manner inversely proportional to the density of the data points in the interval. This approach to density estimation, maximizing the likelihood function over a certain space, was initially proposed by Grenander (1956). He derived the maximum likelihood estimate for a nonincreasing density $f(\cdot)$ corresponding to an absolutely continuous distribution function, $F(\cdot)$. It arose from the studies of the force of mortality determined from mortality tables. He showed the estimate $\hat{f}_n(\cdot)$ to be a step function, the derivative of the greatest convex minorant of the empirical distribution function. Other authors who have used similar techniques are Marshall and Proschan (1965), Robertson (1967), Weiss and Wolfowitz (1967), Rao (1969), McGilchrist (1975).

Good (1971) and Good and Gaskins (1971, 1972) considered maximizing a score function

$$\omega = L - \phi(f), \tag{1.5.5}$$

where L is the sample log likelihood function and ϕ a non-negative roughness penalty function. They proved pointwise consistency in probability under stated conditions. They let

$$\phi(f) = 4\alpha \int \gamma'^2 dx + \beta \int \gamma''^2 dx. \tag{1.5.6}$$

where

$f = \gamma^2$,

$\alpha \geq 0$,

$\beta \geq 0$,

$\alpha+\beta > 0$, and assumed

$$\gamma(x) = \sum_{m=0}^{\infty} \gamma_m \phi_m(x), \tag{1.5.7}$$

γ_m are real coefficients and $\phi_m(x)$ are the Hermite polynomials. This leads to a set of simultaneous nonlinear equations to be solved iteratively for $\gamma_1, \gamma_2, \ldots, \gamma_R$, where $\gamma_{R+1}, \ldots,$ are assumed zero.

1.6. Remarks

So far, we have discussed various methods of density estimation. One method not yet mentioned is a subclass of kernel estimators. In kernel density estimators, regardless of the point at which the density is to be estimated, the window used is the same for a given sample size. Various authors have used the distance to the $k(n)^{th}$ nearest neighbor as the window instead, thus making the choice of the window data dependent. Such estimators are referred to in the literature as the nearest neighbor estimators. These are the estimators of main interest in this thesis and are discussed fully in the next chapter. The remainder of the thesis is concerned with developing two new nearest neighbor estimators, proving their theoretical properties, and comparing their performance by a Monte Carlo simulation study.

2. THEORETICAL RESULTS

2.1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample, each identically and independently distributed as the random variable X, with unknown density function $f(\cdot)$. In Section 2.2 we review some of the available kernel estimators of the function f. For the case when the function f is to be estimated at a particular point, say x, we propose the nearest neighbors estimator $\hat{f}_n(x)$ of $f(x)$, in Section 2.3 and develop consistency results for this estimator. When f is to be estimated over its entire range, we propose the spheres of influence estimator, $\tilde{f}_n(\cdot)$, in Section 2.4 and develop its theoretical properties; Section 2.5 is devoted to some concluding remarks.

2.2. Kernel Density Estimators

A method which is often used to estimate probability densities of unknown functional form is the histogram. Let $X_1, X_2, \ldots, X_n$ be a random sample from an unknown absolutely continuous probability density with domain of positivity $[a,b]$. If the unknown density has an infinite range, we estimate the truncated density on $[a,b]$ only. We assume that the sample points all lie in the interval $[a,b]$.

Partition $[a,b]$ by $a = a_0 < a_1 < \ldots < a_m = b$. Consider an estimator f_H of the form

$$f_H(x) = c_i \quad \text{for} \quad a_i \leq x < a_{i+1},$$
$$i = 0,1,\ldots,m-1$$
$$f_H(b) = c_{m-1} \qquad (2.2.1)$$
$$f_H(x) = 0 \quad \text{for} \quad x \notin [a,b],$$

where

$$f_H(x) \geq 0 \quad \text{and} \quad \int_a^b f_H(x)dx = 1.$$

Defining q_i to be the number of observations falling in the i^{th} interval, then the histogram estimator, $\hat{f}_H(x)$, is obtained by letting

$$c_i = \frac{q_i}{n(a_{i+1}-a_i)}, \quad i = 0,1,\ldots,m-1 .$$

The intuitive appeal of $\hat{f}_H(x)$ is clear. As f is assumed to be absolutely continuous, if $a_{i+1}-a_i$ is small,

$$f(x) \sim f(a_i) \quad \text{for} \quad a_i \leq x < a_{i+1}, \qquad (2.2.2)$$

and hence $\frac{q_i}{n}$ estimates $(a_{i+1}-a_i)f(x)$. It can be shown that among estimators of the form (2.2.1), $\hat{f}_H$ uniquely maximizes the likelihood (Tapia and Thompson, 1978, p. 45),

$$L(a_0,a_1,\ldots,a_m) = \prod_{j=1}^{n} f_H(X_j) .$$

Most of the time, the partition $a_0,a_1,\ldots,a_m$ is taken to be equally spaced, say of length $2h_n$. If

(i) f has continuous derivatives up to order three except at the endpoints of $[a,b]$,

(ii) f is bounded on $[a,b]$ and

(iii) $h_n \to 0$, $nh_n \to \infty$ and $n \to \infty$, then for $x \varepsilon [a,b]$,

$$\mathrm{MSE}(\hat{f}_H(x)) = E[(\hat{f}_H(x)-f(x))^2] \to 0 \quad \text{as} \quad n \to \infty,$$

i.e., $\hat{f}_H(x)$ is a consistent estimator for $f(x)$, (Tapia and Thompson, 1978, pp. 46-48). By a proper choice of h_n, which depends upon the unknown density $f(x)$, the global measure, integrated mean square error,

$$\mathrm{IMSE} = \int \mathrm{MSE}(\hat{f}_H(x))dx$$

can be made to decrease at the rate of $n^{-2/3}$.

The histogram suffers a number of drawbacks; namely, the arbitrariness in choosing the size, number and location of the intervals. However, it is a good tool in preliminary data analysis.

Rosenblatt (1956), in a very insightful paper, extended the histogram estimator of a probability density. Assuming the unknown density f to be continuous, the Rosenblatt estimator is given by

$$\hat{f}_n(x) = \frac{\#\text{ of sample points in } (x-h_n,\ x+h_n)}{2nh_n}$$

where h_n is a real valued number constant for each n. $\hat{f}_n(x)$ can also be written as

$$\hat{f}_n(x) = \frac{F_n(x+h_n)-F_n(x-h_n)}{2h_n}$$

where

$$F_n(x) = \frac{\text{\# of sample points} \leq x}{n}$$

If $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then it can be shown that the $\text{MSE}(\hat{f}_n(x)) \to 0$ as $n \to \infty$. If h_n can be chosen to minimize the $\text{MSE}(\hat{f}_n(x))$, which then depends upon the unknown density f, the rate of convergence of $\text{MSE}(\hat{f}_n(x))$ is seen to be of the order of $n^{-4/5}$. Studies of various measures of consistency of density estimators are given in Bickel and Rosenblatt (1973), Kim and Van Ryzin (1974), Nadaraya (1965), Schuster (1970), Van Ryzin (1969), and Woodroofe (1970).

Rosenblatt estimator is essentially a histogram which, for estimating the density at x, say, has been shifted so that x lies at the center of one of the partition intervals. For estimating the density at another point, the mesh is shifted again so as to make that point, the midpoint of one of the mesh intervals. This shifted histogram estimator can also be written as

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n} K\left(\frac{x-X_j}{h_n}\right) \qquad (2.2.3)$$

where

$$K(u) = \begin{cases} \frac{1}{2} & \text{if } |u| < 1 \\ 0 & |u| \geq 1 \end{cases}$$

and $\{X_j\}$ are the data points. Rosenblatt suggested generalizing the above representation to use other functions $K(\cdot)$, the detailed derivation of the kernel estimators, however, is due to Parzen (1962). He considered kernels of the form

$$\int_{-\infty}^{\infty} |K(y)|dy < \infty,$$

$$\sup_{-\infty<y<\infty} |K(y)| < \infty,$$

$$\lim_{y\to\infty} |yK(y)| = 0,$$

$$K(y) \geq 0$$

and

$$\int_{-\infty}^{\infty} K(y)dy = 1.$$

If f is assumed to be continuous, then $\hat{f}_n(x)$ can be shown to be limiting unbiased and consistent under the usual conditions that

$$h_n \to 0 \quad \text{as} \quad n\to\infty$$

and

$$nh_n \to \infty \text{ as } \quad n\to\infty \quad .$$

The rate of convergence of the mean squared error at some point x or the integrated mean squared error is of the order of $n^{-4/5}$, the same as the one obtained for the shifted histogram, if h_n is chosen to minimize MSE(x) or IMSE, respectively. Note that h_n then depends upon the unknown density $f(\cdot)$.

The problem of determining the sequence $\{h_n\}$ to achieve best results is not answered in the absence of the knowledge about the unknown density $f(\cdot)$. Various authors have tried to answer this question in one manner or another, e.g., Silverman (1978b), Schuster and Gregory (1978). However, the estimate is still not satisfactory. Moreover, if the density f is very low in some region and if only one sample point falls in that region, then the kernel estimator will have a peak at that sample point and be too low over the remainder of that region. Similarly, in the region where f is large, the sample points are more densely packed together and the kernel estimator will tend to spread out the high density region. To overcome these drawbacks, kernel density estimators with variable windows were developed.

Loftsgaarden and Quesenberry (1965) studied an estimator $g_n(x)$ which is the ratio of the empiric measure and the Lebesgue measure of the sphere $S_k(x)$ centered at x and having radius $R(k,x)$ equal to the distance from x to the $k(n)^{th}$ nearest of $X_1,X_2,\ldots,X_n$. $\{k(n)\}$ is a sequence of positive integers such that

$$\lim_{n\to\infty} k(n) \to \infty$$

and

$$\lim_{n\to\infty} \frac{k(n)}{n} = 0 \quad .$$

In the univariate case,

$$g_n = \frac{\{\text{\# of points in } (x-R(k,x),\ x+R(k,x))\}/n}{2R(k,x)}$$

Loftsgaarden and Quesenberry showed that g_n is pointwise consistent in probability at continuity points of f. Estimators utilizing the distance to the $k(n)^{th}$ nearest neighbor are denoted in the literature as the nearest neighbor estimators. Wagner (1973) established pointwise consistency with probability one under an additional assumption equivalent to

$$\lim_{n\to\infty} \frac{k(n)}{\log n} \to \infty \ . \tag{2.2.4}$$

In fact, Moore and Yackel (1977a,b) point out that g_n is pointwise consistent with probability one under the weaker condition that

$$\lim_{n\to\infty} \frac{k(n)}{\log \log(n)} \to \infty$$

and show that this condition is the weakest possible. Moore and Henrichon (1969) prove uniform consistency with probability one of g_n in the univariate case under the condition that (2.2.4) holds. A general nearest neighbor density estimator, Moore and Yackel (1977a), is defined as

$$f_n(\underline{x}) = \frac{1}{nR(k,\underline{x})^p} \sum_{j=1}^{n} K\left(\frac{\underline{x}-\underline{X}_i}{R(k,\underline{x})}\right)$$

where $(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n)$ is a random sample from a p-variate distribution with unknown probability density function f. They show that any consistency theorem true for the bandwidth estimator using kernel K and also true for the uniform kernel, remains true for f_n also. Wagner (1975) studied a similar estimator which replaces $R(k,\underline{x})$ with a random radius Γ_n independent of $\underline{x}$ and showed pointwise consistency in probability at all continuity points of f. Moore and Yackel's results allow almost all the results available for the bandwidth estimator, i.e., fixed window estimators, to be transferred to the nearest neighbor estimators. However, they leave the question of the choice of the sequence $\{k(n)\}$ unanswered. Schuster and Gregory (1978) and Breiman, Meisel and Purcell (1977) suggest ways of obtaining k(n) for practical situations. However, it involves a lot of computations.

Let us examine how Parzen's (bandwidth) kernel estimators and the nearest neighbor estimators tend to estimate the unknown density at some point, say x. When the kernel used is the uniform kernel, the bandwidth estimator takes an interval of fixed length around x, and counts the number of sample points falling in it; i.e.,

$$\hat{f}_n(x) = \frac{\#\text{ of sample points in } (x-h_n,\ x+h_n)}{2nh_n}$$

So, the interval is fixed, whereas the random number of points falling in it leads to an estimate of f(x). Using the uniform kernel, the nearest neighbor estimator is given by

$$\hat{g}_n(x) = \frac{k(n)}{n\mu\{R(k,x)\}}$$

where $\mu\{R(k,x)\}$ is the Lebesgue measure of the sphere of radius R(k,x) around x, R(k,x) being the distance from x to its $k(n)^{th}$ nearest neighbor in $X_1,X_2,\ldots,X_n$. In this case, the randomness is provided by the distance to the $k(n)^{th}$ nearest neighbor of x. Combining these two ideas, so as to have a random interval around x and a random number of points falling in this interval, both being determined by the sample, leads us to our estimators:

(i) the nearest neighbors estimator, which is studied in Section 2.3, and

(ii) the spheres of influence estimator which is studied in Section 2.4.

2.3. Nearest Neighbors Estimator

Let $X_1,X_2,\ldots,X_n$ be a random sample of size n from a distribution with unknown, continuous and bounded density f. It is desired to estimate the density at a point x. Let K be a Borel function satisfying:

$$K(y) \geq 0 \quad \text{for all } y$$

$$\int_{-\infty}^{\infty} K(y)dy = 1$$

$$\sup_{-\infty<y<\infty} |K(y)| < \infty \qquad (2.3.1)$$

$$\lim_{y\to\infty} |yK(y)| = 0.$$

K is called a kernel. Let $\{k(n)\}$ be a sequence of positive integers such that

$$k(n) = [n^{\alpha'}], \quad 0<\alpha'<1$$

Define

$$h_n(x) = \gamma \sum |x-X_j| \qquad (2.3.2)$$

j ranging over
k(n) nearest neighbors
of x

where $\gamma>0$ is a constant. The nearest neighbors estimator of f(x) is defined to be

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n(x)} K(\frac{x-X_i}{h_n(x)}).$$

This estimator is similar to the nearest neighbor estimator discussed in the previous section except the window defined by (2.3.2) considers all neighbors up to and including the $k(n)^{th}$ nearest neighbor whereas only the $k(n)^{th}$ neighbor was considered before. The motivation behind varying window estimators is that the size of the window at a particular point should be proportional to the denseness of the sample around that point. Moreover, the sample points are clustered around the point in question, a smaller window would result in higher weights being given to those points in the estimate reflecting the higher density in the region. The nearest neighbor estimators achieve this by considering the distance to the $k(n)^{th}$ nearest neighbor as the window at a particular point. However, a better measure of the denseness of the sample would be obtained if all the neighbors up to $k(n)$ are utilized. The nearest neighbors estimator achieves this by considering the sum of the distances to each of the neighbors up to $k(n)$ and taking the window proportional to this quantity.

To establish consistency results for $f_n(x)$, we need the following lemmas.

Lemma 2.3.1: Let K be a Borel function satisfying:

$$\int_{-\infty}^{\infty} K(y)\,dy < \infty$$

$$\sup_{-\infty<y<\infty} |K(y)| < \infty$$

and

$$\lim_{y\to\infty} |yK(y)| = 0.$$

Let $g\varepsilon L^1$, i.e., $\int |g(y)|dy < \infty$ and let

$$g_n(x) = \frac{1}{h_n(x)} \int K\left(\frac{y}{h_n(x)}\right) g(x-y)\,dy$$

where $\{h_n(x)\}$ is a sequence of positive constants having

$$\lim_{n\to\infty} h_n(x) = 0$$

for each x.

If x is a point of continuity of g, then

$$\lim_{n\to\infty} g_n(x) = g(x) \int_{-\infty}^{\infty} K(y)\,dy.$$

Proof: Consider

$$|g_n(x) - g(x) \int_{-\infty}^{\infty} K(y)\,dy|$$

$$= |\int_{-\infty}^{\infty} K(\frac{y}{h_n(x)})\, g(x-y) \frac{1}{h_n(x)}\, dy - \int_{-\infty}^{\infty} K(\frac{y}{h_n(x)})\, g(x) \frac{1}{h_n(x)} dy|$$

$$= |\int_{-\infty}^{\infty} K(\frac{y}{h_n(x)}) \frac{1}{h_n(x)}\, \{g(x-y) - g(x)\} dy\,|$$

$$\leq |\int_{|y|\leq\delta} \{g(x-y) - g(x)\} \frac{1}{h_n(x)}\, K(\frac{y}{h_n(x)})\, dy|$$

$$+ |\int_{|y|>\delta} \{g(x-y) - g(x)\} \frac{1}{h_n(x)}\, K(\frac{y}{h_n(x)})\, dy|$$

$$\leq \sup_{|y|\leq\delta} |g(x-y) - g(x)| \cdot \int_{-\infty}^{\infty} |K(z)|\,dz$$

$$+ \int_{|y|>\delta} |\frac{g(x-y)}{y} \cdot \frac{y}{h_n(x)}\, K(\frac{y}{h_n(x)})|\,dy$$

$$+ |g(x)| \int_{|y|>\delta} |\frac{1}{h_n(x)}\, K(\frac{y}{h_n(x)})|\,dy$$

$$\leq \sup_{|y|\leq\delta} |g(x-y) - g(x)| \int_{-\infty}^{\infty} |K(z)|\,dz$$

$$+ \frac{1}{\delta} \sup_{|z|>\frac{\delta}{h_n(x)}} |zK(z)| \cdot \int_{-\infty}^{\infty} |g(y)|\,dy$$

$$+ |g(x)| \int_{|z|>\frac{\delta}{h_n(z)}} |K(z)|\,dz$$

As g is continuous at x, one can take δ small enough so that

$$\sup_{|y|\leq\delta} |g(x-y)-g(x)| < \frac{\varepsilon}{3\int_{-\infty}^{\infty}|K(z)|dz},$$

$\int_{-\infty}^{\infty}|K(z)|dz$ being finite. Therefore, the first term on the right hand side can be made $< \frac{\varepsilon}{3}$. Also, because $\lim_{y\to\infty} |yK(y)| = 0$, and $\sup_{|z| > \frac{\delta}{h_n(x)}} |zK(z)|$ gets smaller and smaller as $n\to\infty$, and $h_n(x)\to 0$. Since $g\varepsilon L^1$, the second term on the right hand side can be made $< \frac{\varepsilon}{3}$ for large enough n. Lastly, since $\int_{-\infty}^{\infty}|K(y)|dy < \infty$, the third term on the right hand side can be made smaller than $< \frac{\varepsilon}{3}$ for large enough n. Hence, for all $\varepsilon>0$, there exists M such that for all $n\geq M$,

$$|g_n(x)-g(x)\int_{-\infty}^{\infty}K(y)dy| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Lemma 2.3.2: Let $X_1,X_2,\ldots,X_n$ be independent and identically distributed as

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases}$$

then

$$e^{-nX} \geq \begin{cases} P[\frac{1}{n} \sum_{i=1}^{n} X_i \geq d]; & 1 \geq d > p \\ P[\frac{1}{n} \sum_{i=1}^{n} X_i \leq d]; & 0 \leq d < p \end{cases} ,$$

where

$$X = -d \ln p - (1-d) \ln(1-p)$$
$$+d \ln d + (1-d) \ln(1-d) \qquad (2.3.3)$$

Proof: Using the methods of Chernoff (1952), Wozencraft and Jacobs (1965) have derived the Chernoff bound as follows:

If $X_1, X_2, \ldots, X_n$ are independently and identically distributed with unknown mean μ, then

$$[E(e^{\lambda_0 (X-d)})]^n \geq \begin{cases} P[\frac{1}{n} \sum_{i=1}^{n} X_i \geq d]; & d > \mu \\ P[\frac{1}{n} \sum_{i=1}^{n} X_i \leq d]; & d < \mu \end{cases} \qquad (2.3.4)$$

with λ_0 defined implicitly by

$$\frac{E[Xe^{\lambda_0 X}]}{E[e^{\lambda_0 X}]} = d. \qquad (2.3.5)$$

Therefore, in our case,

$$E[e^{\lambda_0 X}] = (1-p) + pe^{\lambda_0}$$
and
$$E[Xe^{\lambda_0 X}] = pe^{\lambda_0} \quad . \qquad (2.3.6)$$

So, for $0 \leq d \leq 1$, by (2.3.5) and (2.3.6)

$$d = \frac{pe^{\lambda_0}}{(1-p) + pe^{\lambda_0}}$$

giving

$$\lambda_0 = \ln\left[\frac{d(1-p)}{p(1-d)}\right]$$

Now,

$$E(e^{\lambda_0(X-d)}) = e^{-\lambda_0 d} E(e^{\lambda_0 X})$$

$$= e^{-\lambda_0 d}[(1-p)+pe^{\lambda_0}]$$

$$= \left[\frac{p(1-d)}{d(1-p)}\right]^d [(1-p) + p\,\frac{d(1-p)}{p(1-d)}]$$

$$= \left(\frac{p}{d}\right)^d \left(\frac{1-p}{1-d}\right)^{1-d} .$$

Substituting in (2.3.4) we get the desired result.

Note that, from (2.3.3),

$$X = T_p(d) - H(d)$$

where

$$T_p(\alpha) = -\alpha \ln p - (1-\alpha)\ln(1-p)$$

and

$$H(\alpha) = -\alpha \ln \alpha - (1-\alpha)\ln(1-\alpha)$$

$T_p(\alpha)$ is a linear function of α and $H(\alpha)$ increases from 0 to ln 2 at $\alpha = .5$, and decreases to 0 again at $\alpha = 1$ as shown in Figure 2.1. $T_p(\alpha)$ is tangent to the surface at $\alpha = p$ and

X is as shown in Figure 2.1.

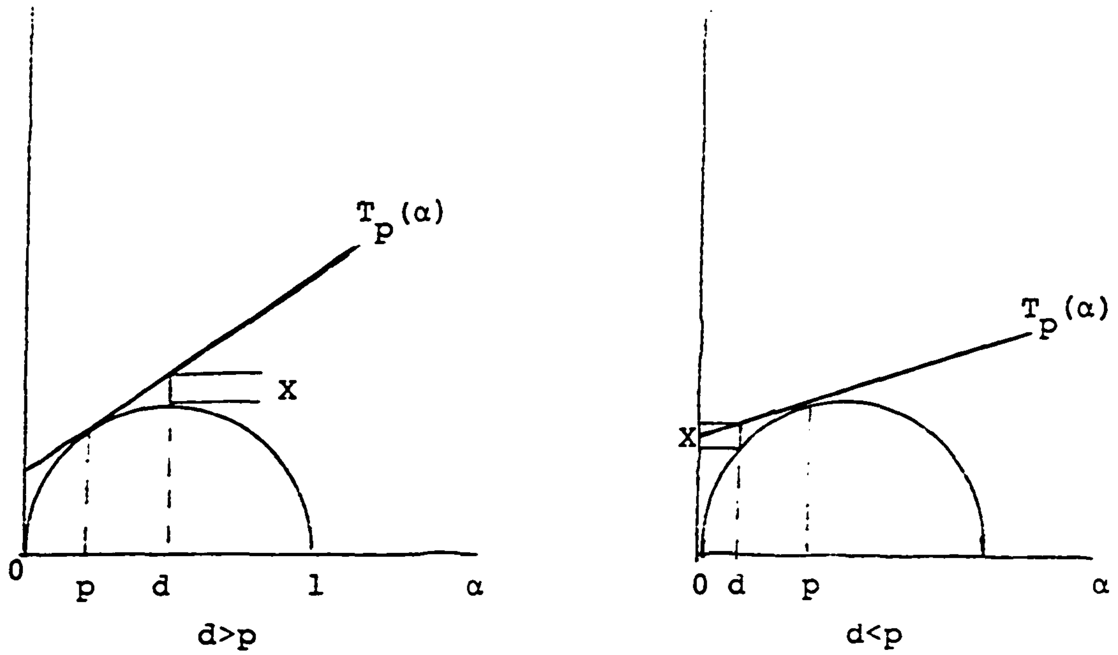

Figure 2.1. Function $H(\alpha)$ and $T_p(\alpha)$

Notice that X increases as $|d-p|$ increases.

Lemma 2.3.3: Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with unknown bounded density function f.

Let

$$k(n) = [n^{\alpha'}], \quad 0<\alpha'<1,$$

$$h_n(x) = \Sigma|x-X_\ell|, \qquad (2.3.7)$$

ℓ ranging over k(n) nearest neighbors of x in $(X_1, \ldots, X_n)$

$f(x) > 0,$

$h_n^j = h_n(X_j), \quad j = 1,2,\ldots,n,$

and h_n is selected at random from $\{h_n^j\}$, $j = 1,2,\ldots,n$, then

(a) for $0<\alpha'<\frac{1}{2}$,

$h_n \to 0$ in probability

and $h_n(x) \to 0$ with probability 1;

and

(b) $n^\alpha h_n \to \infty$ in probability for $1-\alpha'<\alpha$;

and $n^\alpha h_n(x) \to \infty$ with probability 1

for $1-\alpha'<\alpha$.

Proof: Consider $h_n^*(x)$ = distance from x to the $k(n)^{th}$ nearest neighbor in $(X_1,\ldots,X_n)$.

Then for any positive δ,

$$P(n^{\beta} h_n^*(x) \geq \delta) = P(h_n^*(x) \geq \delta_n)$$

where $\delta_n = \frac{\delta}{n^{\beta}}$

= P(sphere S of radius δ_n around x contains at most k(n)-1 number of points in a sample of size n drawn at random)

$$= \sum_{j=0}^{k(n)-1} \binom{n}{j} (P(S_{\delta_n}(x)))^j (1-P(S_{\delta_n}(x)))^{n-j}$$

where

$$P(S_{\delta_n}(x)) = \int_{S_{\delta_n}(x)} f(x')dx'$$

$$= \int_{(x-\frac{\delta_n}{2}\,,\; x+\frac{\delta_n}{2})} f(x')dx'$$

$$= 0(\delta_n) = 0(\frac{1}{n^{\beta}}) \qquad (2.3.8)$$

Therefore,

$$P(n^{\beta} h_n^*(x) \geq \delta) = P(\sum_{i=1}^{n} Y_i \leq k(n)-1)$$

$$= P(\frac{1}{n} \sum_{i=1}^{n} Y_i \leq d_n)$$

where $Y_1,\ldots,Y_n$ are independent identically distributed random variables with probability $p_n = P(S_{\delta_n}(x))$ and $d_n = \frac{k(n)-1}{n}$. Now,

$$d_n = \frac{n^{\alpha'}-1}{n} = 0(\frac{1}{n^{1-\alpha'}}) \qquad (2.3.9)$$

From (2.3.8) and (2.3.9), for $\beta<1-\alpha'$, d_n is less than p_n for large enough n and hence appealing to Lemma 2.3.2, it can be shown that for x such that $P(S_{\delta_n}(x)) > 0$,

$$P(n^{\beta}h_n^*(x) \geq \delta) \leq e^{-nc} \quad \text{for all} \quad n\geq M,$$

for some positive constant M. Hence, by Borel-Cantelli lemma,

$n^{\beta}h_n^*(x)\to 0$ with probability 1 as $n\to\infty$ for $\beta<1-\alpha'$. Now, α' being less than 1/2, the condition $\beta<1-\alpha'$ can be satisfied by $\beta>\alpha'$ also.

Therefore, from (2.3.7), we have that

$$h_n(x) \leq k(n)h_n^*(x) = n^{\alpha'}h_n^*(x).$$

And, as $n^{\alpha'}h_n^*(x) \to 0$ with probability 1, so does $h_n(x)$.

Now,

$$P(h_n\geq\delta) = \int P(h_n(x)\geq\delta)\ f(x)dx.$$

$P(h_n(x)\geq\delta)$ is bounded by 1; for each x for which $P(S_\delta(x))>0$, it tends to zero with probability 1; and $\{x: P(S_\delta(x))>0\}$ has measure one with respect to f. Therefore, the Lebesque dominated convergence theorem,

$$\lim_{n\to\infty} P(h_n \geq \delta) = 0 \quad \text{for each } \delta>0.$$

This implies that $h_n \to 0$ in probability, thus proving (a).

To prove (b), it suffices to consider only the $k(n)^{th}$ nearest neighbor in the definition of $h_n(x)$. Now, consider $M>0$, then

$$P(n^\alpha h_n(x) \leq M) = P(h_n(x) \leq \frac{M}{n^\alpha})$$

$$= \sum_{k(n)}^{n} \binom{n}{j} (P(S_{\delta_n}(x)))^j (1-P(S_{\delta_n}(x)))^{n-j},$$

as the event $\{x: h_n(x) \leq \frac{M}{n^\alpha}\}$ is equivalent to saying that a sphere of radius $\delta_n = \frac{M}{n^\alpha}$ around x should contain at least k(n) points. Therefore,

$$P(n^\alpha h_n(x) \leq M) = P(\frac{1}{n} \sum_{i=1}^{n} Y_i \geq \frac{k(n)}{n}),$$

$\{Y_i\}$ being independent identically distributed Bernouli variables with probability $P(S_{\delta_n}(x))$. Now,

$$P(S_{\delta_n}(x)) = \int_{(x - \frac{M}{2n^\alpha},\ x + \frac{M}{2n^\alpha})} f(x')dx'$$

$$= 0(\frac{1}{n^\alpha}) \quad \text{and}$$

$$\frac{k(n)}{n} = \frac{1}{n^{1-\alpha'}}, \quad \text{hence for} \quad 1-\alpha'<\alpha\leq 1,$$

the conditions of Lemma 2.3.2 are satisfied and it can be shown that

$$P(n^{\alpha}h_n(x) \leq M) \leq e^{-nc} \quad \text{for all} \quad n \geq N,$$

a positive integer. This implies, by the Borel-Cantelli lemma, that $n^{\alpha}h_n(x) \to \infty$ with probability one for $1-\alpha' < \alpha \leq 1$. Also, by the Lebesgue dominated convergence theorem, $n^{\alpha}h_n \to \infty$ in probability. Note that the restriction $\alpha > 1-\alpha'$ in (b) will yield a value of $\alpha < \frac{1}{2}$ only if α' is $> \frac{1}{2}$. However, for (a) to hold, α' has to be $< \frac{1}{2}$. A slight modification will do the trick.

Let $k(n) = [n^{1/2}]+1$, then $k(n) = n^{\alpha'}$, for a value of $\alpha' > \frac{1}{2}$. Also, as (a) holds true when $[n^{1/2}]$ neighbors are considered and when $k(n)^{\text{th}}$ nearest neighbor is considered by itself, it holds true when $k(n)$ nearest neighbors are considered too. Hence, we have a value of $\alpha' > \frac{1}{2}$ for which (a) holds. Now, we shall state and prove consistency results for the nearest neighbors estimator, $f_n(x)$.

<u>Theorem 2.3.1</u>: Let K be the uniform density on $[-1,1]$ and let $h_n(x)$ satisfy

(a) $h_n(x) \to 0$ as n tends to ∞ with probability 1, and

(b) $n^{\alpha}h_n(x) \to \infty$ as n tends to ∞ for some $0<\alpha<\frac{1}{2}$,

then

(i) at every continuity point of f, $f_n(x) \to f(x)$ with probability 1, and

(ii) if f is uniformly continuous on R, and $n^{\alpha} \inf_x h_n(x) \to \infty$ with probability 1 for some $0<\alpha<\frac{1}{2}$, then

$$\sup_x |f_n(x)-f(x)| \to 0 \quad \text{with probability 1.}$$

Proof: Consider $f_n(x) = \frac{1}{n}\sum_1^n \frac{1}{h_n(x)} K(\frac{x-X_i}{h_n(x)})$,

where

$$K(x) = \begin{cases} \frac{1}{2}, & |x| \le 1 \\ 0, & |x| > 1 \end{cases}$$

Let

$$S = \{j \,|\, |x-X_j| < h_n(x)\}.$$

Then

$$f_n(x) = \frac{1}{2nh_n(x)} \sum_{i\varepsilon S} 1$$

and

$$\begin{aligned} |f_n(x)-f(x)| &= |\frac{1}{n}\sum_{i\varepsilon S} 1-2h_n(x)f(x)|/2h_n(x) \\ &= |\frac{1}{n}\sum_{i\varepsilon S} 1 - \int_{x-h_n(x)}^{x+h_n(x)} f(x')dx' \\ &\qquad + \int_{x-h_n(x)}^{x+h_n(x)} f(x')dx' \\ &\qquad - 2h_n(x)f(x)|/2h_n(x) \\ &\le \frac{1}{2h_n(x)}|\frac{1}{n}\sum_{i\varepsilon S} 1 - \int_{x-h_n(x)}^{x+h_n(x)} f(x')dx'| \\ &\qquad + \frac{1}{2h_n(x)}|\int_{x-h_n(x)}^{x+h_n(x)} f(x')dx'-2h_n(x)f(x)| \end{aligned} \tag{2.3.10}$$

If f is continuous at x, then the second term on the right hand side of (2.3.10) converges to zero with probability 1 if $h_n(x) \to 0$ with probability 1. Let

$$F_n(x) = \frac{1}{n} \{\# \text{ of points} \leq x\}.$$

Hence, the first term in (2.3.10) is

$$= \frac{1}{2h_n(x)} |F_n(x+h_n(x)) - F_n(x-h_n(x))$$

$$- F(x+h_n(x)) + F(x-h_n(x))|,$$

where F is the distribution function associated with f. Therefore, it is

$$\leq \frac{\sup_x |F_n(x)-F(x)|}{h_n(x)} = \frac{n^\alpha \sup_x |F_n(x)-F(x)|}{n^\alpha h_n(x)} \qquad (2.3.11)$$

Now, $n^\alpha \sup_x |F_n(x)-F(x)| \to 0$ with probability 1 for $0 < \alpha < \frac{1}{2}$, Keifer and Wolfowitz (1958). Also, by the conditions of the theorem, $n^\alpha h_n(x) \to \infty$ with probability 1 for some $0<\alpha<\frac{1}{2}$. Therefore, both the terms in (2.3.10) converge to zero with probability 1 and hence $|f_n(x)-f(x)| \to 0$ with probability 1 for x a continuity point of f. This proves (i). To prove (ii), note that from (2.3.10) and (2.3.11),

$$\sup_x |f_n(x)-f(x)| \leq \sup_x\{\frac{\sup_x |F_n(x)-F(x)|}{h_n(x)}\}$$

$$+ \sup_x\left|\frac{1}{2h_n(x)}\int_{x-h_n(x)}^{x+h_n(x)} f(x')dx'-f(x)\right|$$

$$\leq \frac{\sup_x |F_n(x)-F(x)|}{\inf_x h_n(x)} + \sup_x\left|\frac{F(x+h_n(x)-F(x-h_n(x)}{2h_n(x)} - f(x)\right|$$

(2.3.12)

If f is uniformly continuous, then the second term in (2.3.12) tends to zero with probability 1 as $h_n(x)\to 0$ with probability 1 for all x. The first term in (2.3.12) tends to zero with probability 1 if $n^\alpha \inf_x h_n(x)\to\infty$ with probability 1 for some $0<\alpha<\frac{1}{2}$, which then proves (ii).

Note that by Lemma 2.3.3, α' and α can be chosen so as to satisfy the conditions of the theorem.

<u>Theorem 2.3.2</u>: If the kernel K has bounded variation, and $h_n(x)$ satisfies:

(a) $h_n(x) \to 0$ w.p. 1 for some $0<\alpha<\frac{1}{2}$, and

(b) $n^\alpha h_n(x) \to \infty$ for some $0<\alpha<\frac{1}{2}$, then

(i) at every continuity point of f, $f_n(x)\to f(x)$ w.p.1

(ii) if $n^\alpha \inf_x h_n(x)\to\infty$ w.p. 1 for some $0<\alpha<\frac{1}{2}$, and f is uniformly continuous, then

$$\sup_x|f_n(x)-f(x)| \to 0 \text{ w.p. } 1$$

Proof: $|f_n(x)-f(x)| = |\frac{1}{h_n(x)} \sum_{i=1}^{n} K[\frac{x-X_i}{h_n(x)}] - f(x)|$

$$= |\frac{1}{h_n(k)} \int K[\frac{x-y}{h_n(x)}] dF_n(y) - f(x)|$$

$$= |\frac{1}{h_n(x)} \int K[\frac{x-y}{h_n(x)}] dF_n(y) - \frac{1}{h_n(x)} \int K[\frac{x-y}{h_n(x)}] dF(y)$$

$$+ \frac{1}{h_n(x)} \int K[\frac{x-y}{h_n(x)}] dF(y) - f(x)|$$

$$\leq \frac{1}{h_n(x)} |\int K[\frac{x-y}{h_n(x)}] dF_n(y) - \int K[\frac{x-y}{h_n(x)}] dF(y)|$$

$$+ |\frac{1}{h_n(x)} \int K[\frac{x-y}{h_n(x)}] dF(y) - f(x)|. \qquad (2.3.13)$$

Integrating by parts we see that the first term is

$$\leq \frac{1}{h_n(x)} \int |F_n(y)-F(y)| dK[\frac{x-y}{h_n(x)}]$$

$\leq \frac{V}{h_n(x)} \sup_y |F_n(y)-F(y)|$, where

V is the total variation of K.

Hence it is $\leq \frac{V n^\alpha}{n^\alpha h_n(x)} \sup_y |F_n(y)-F(y)|, \quad 0<\alpha<\frac{1}{2}$

and $n^\alpha \sup_y |F_n(y)-F(y)| \to 0$ with probability one for $0<\alpha<\frac{1}{2}$.

Also, if $n^\alpha h_n(x) \to \infty$ with probability 1, the first term $\to 0$ with probability 1. Now, the second term in (2.3.13) is equal to

$$|\frac{1}{h_n(x)} \int K[\frac{x-y}{h_n(x)}] f(y) dy - f(x)|$$

which →0 by Lemma 2.3.1.

Hence, $f_n(x) \to f(x)$ with probability 1. This proves (i).

To prove (ii), note that

$$\sup_x |f_n(x)-f(x)| \le \frac{V \sup_y |F_n(y)-F(y)|}{\inf_y h_n(y)}$$

$$+ \sup_x \left|\frac{1}{h_n(x)} \int K\left[\frac{x-y}{h_n(x)}\right] f(y)dy - f(x)\right|$$

If for some $0<\alpha<\frac{1}{2}$, $n^\alpha \inf_y h_n(y) \to \infty$ w.p. 1, then the first term →0 w.p. 1. Also, if f is uniformly continuous, then Nadaraya (1965) has shown that the second term →0 w.p. 1, if $h_n(x) \to 0$ w.p. 1. Thus $\sup_x |f_n(x)-f(x)| \to 0$ w.p. 1.

Note that α' and α can be chosen according to Lemma 2.3.3 to satisfy the conditions of the theorem.

Following the method of Schuster and Gregory (1978), an estimator very similar to the nearest neighbors estimator can be defined as follows.

Divide the sample of size n into two parts randomly of size n_1 and n_2, $n_1+n_2 = n$. Let $X_1,\ldots,X_{n_1}$ and $Y_1,\ldots,Y_{n_2}$ be the two parts of the original sample.

Define $h_1(x) = \gamma_1 \Sigma |x-Y_j|$ (2.3.14)

j ranging over

$k_1(n)$ nearest neighbors

of x in $(Y_1,\ldots,Y_{n_2})$

and

$h_2(x) = \gamma_2 \ \Sigma|x-X_j|$

j ranging over

$k_2(n)$ nearest neighbors

of x in $(X_1,\ldots,X_{n_1})$,

where

$k_1(n) = [n_1^{\alpha'}]$, $0<\alpha'<1$,

$k_2(n) = [n_2^{\alpha'}]$, for some $0<\alpha'<1$, and γ_1, γ_2 are some positive constants.

Then

$$\hat{f}_n(x) = \frac{n_1 \ f_{1,n_1}(x) + n_2 \ f_{2,n_2}(x)}{n} \qquad (2.3.15)$$

where $f_{i,n_i}(x)$ is the nearest neighbors estimator, with $h_i(x)$ as the window and the ith subsample used with it, i = 1,2. Then the results of Theorems 2.3.1 and 2.3.2 hold true for $\hat{f}_n(x)$ also, since it is a linear combination of two nearest neighbors estimators. Moreover, it can be shown now that $\hat{f}_n(x)$ is limiting unbiased and consistent in mean squared error. This is established in Theorems 2.3.3 and 2.3.4.

<u>Theorem 2.3.3</u>: If

(i) $h_1(x)$ and $h_2(x)\to 0$ with probability 1 and

(ii) f is continuous at x, then

$\lim_{n\to\infty} E(\hat{f}_n(x)) = f(x)$ with probability 1, where $\hat{f}_n(x)$ is defined by (2.3.15).

Proof:

$$E(f_{1,n_1}(x)) = E\{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{h_1(x)}K(\frac{x-X_i}{h_1(x)})\}$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{h_1(x)}E\{K(\frac{x-X_i}{h_1(x)})\}$$

as $h_1(x)$ does not depend upon $X_1,X_2,\ldots,X_{n_1}$ by definition, (2.3.14).

Therefore,

$$E(f_{1,n_1}(x)) = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{h_1(x)}\int K(\frac{x-X_i}{h_1(x)})f(X_i)dX_i$$

$$= \int K(y)f(x-h_1(x)y)dy. \tag{2.3.16}$$

By Lemma (2.3.1), (2.3.16) converges to $f(x)\int_{-\infty}^{\infty}K(y)dy = f(x)$, which implies that

$$E(f_{1,n_1}(x)) = f(x) + 0(h_1(x))\cdot$$

and

$$E(f_{2,n_2}(x)) = f(x) + 0(h_2(x)).$$

Therefore,

$$E(\hat{f}_n(x)) = E(\frac{n_1f_{1,n_1}(x) + n_2f_{2,n_2}(x)}{n})$$

$$\to f(x) \text{ with probability 1}$$

Theorem 2.3.4: The estimator $\hat{f}_n(x)$ defined in (2.3.15) is consistent if

(i) $h_1(x)$, $h_2(x) \to 0$ with probability 1,

(ii) $n_1 h_1(x)$, $n_2 h_2(x) \to \infty$ with probability 1, and

(iii) x is a continuity point of f.

Proof: Consider

$$\begin{aligned}
\text{Var}(f_{1,n_1}(x)) &= \text{Var}\Big(\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{h_1(x)} K\Big(\frac{x-X_i}{h_1(x)}\Big)\Big) \\
&= \frac{1}{n_1} \text{Var}\Big[\frac{1}{h_1(x)} K\Big(\frac{x-X_1}{h_1(x)}\Big)\Big] \\
&\leq \frac{1}{n_1} E\Big[\frac{1}{h_1(x)} K\Big(\frac{x-X_i}{h_1(x)}\Big)\Big]^2 \\
&= \frac{1}{n_1 h_1(x)}\Big[\frac{1}{h_1(x)} \int K^2\Big(\frac{x-y}{h_1(x)}\Big) f(y)\,dy\Big] \\
&\to 0 \text{ by Lemma (2.3.1).}
\end{aligned}$$

Now,

$$\begin{aligned}
\text{MSE}(f_{1,n_1}(x)) &= E[f_{1,n_1}(x) - f(x)]^2 \\
&= \text{Var}(f_{1,n_1}(x)) + \text{Bias}^2(f_{1,n_1}(x))
\end{aligned}$$

and by Theorem (2.3.3),

Bias $(f_{1,n_1}(x)) \to 0$ with probability 1 if $h_1(x) \to 0$ with probability 1 and thus,

$\text{MSE}(f_{i,n_i}(x)) \to 0$ with probability 1 for $i = 1,2$. And since $\hat{f}_n(x)$ is a linear combination of $f_{i,n_i}(x)$,

$\text{MSE}(\hat{f}_n(x)) \to 0$ with probability 1.

Note that conditions of Theorems 2.3.3 and 2.3.4 can be satisfied by choosing α' and α according to Lemma 2.3.3.

2.4. Spheres of Influence Estimators

When the density is to be estimated at several points, the computations of windows around each one of them is necessary before the nearest neighbors estimator studied in Section 2.3 can be used. Instead, if each sample point is considered to have a "sphere of influence" in which it contributes to the estimation of density, then only these spheres of influence of each sample point need be computed and used over and over again to estimate the density at any point in the range of the unknown density f. The estimator discussed in this section uses this concept and takes this sphere of influence around each sample point to be proportional to the window at that point.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with unknown density, f. Let K be a Borel function satisfying (2.3.1). Let $\{k(n)\}$ be a sequence of positive integers, such that

$$k(n) = [n^{\alpha'}], \quad 0<\alpha'<1$$

Define

$$h_{jn} = \gamma \sum_{\substack{\ell \text{ ranges over} \\ k(n) \text{ neighbors of } X_j}} |X_j - X_\ell| \;, \qquad j = 1,\ldots,n \tag{2.4.1}$$

where $\gamma>0$ is a constant.

The spheres of influence estimator is defined as

$$g_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_{in}}K(\frac{\cdot-X_i}{h_{in}})$$

Note that

$$\int g_n(x)dx = \frac{1}{n}\sum_{i=1}^{n}\int\frac{1}{h_{in}}K(\frac{x-X_i}{h_{in}})dx$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int K(y_i)dy_i$$

$$= 1$$

and $g_n(\cdot) \geq 0$, so that g_n is a density. Consider the estimates

$$f_{jn}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_{jn}}K(\frac{x-X_i}{h_{jn}}), \quad j = 1,\ldots,n.$$

By Wagner's results (1975), $f_{jn}(x)$, $j = 1,2,\ldots,n$, is a consistent estimator of $f(x)$ in the sense that $f_{jn}(x)\rightarrow f(x)$ in probability as $n\rightarrow\infty$ if $1-\alpha'<\alpha<1$ and K is of bounded variation. Wagner proved the results for the case when only the $k(n)^{th}$ nearest neighbor is used in (2.4.1) but to extend it to the present situation is similar to the methods employed in Section 2.3.

Therefore, for each $\varepsilon>0$,

$$P(|f_{jn}(x)-f(x)| \leq \varepsilon) \rightarrow 1 \quad \text{as} \quad n\rightarrow\infty, \quad j = 1,2,\ldots,n$$

i.e., $P(-\varepsilon \leq f_{jn}(x)-f(x) \leq \varepsilon) \rightarrow 1$ as $n\rightarrow\infty$, which implies that

$$P(-n\varepsilon \le \sum_{j=1}^{n} f_{jn}(x)-nf(x) \le n\varepsilon)$$

$$\to 1 \quad \text{as} \quad n\to\infty$$

and

$$-n\varepsilon \le \sum_{j=1}^{n} f_{jn}(x)-nf(x) \le n\varepsilon \tag{2.4.2}$$

for all $n\ge N_1$, a positive constant, with probability arbitrarily close to 1.

Now

$$\begin{aligned}
&\sum_{j=1}^{n} f_{jn}(x)-nf(x) \\
&= \sum_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_{jn}} K(\frac{x-X_i}{h_{jn}}) - nf(x) \\
&= \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_{jn}} K(\frac{x-X_j}{h_{jn}}) \\
&\quad + \frac{1}{n} \sum_{j\ne i}\sum \frac{1}{h_{jn}} K(\frac{x-X_i}{h_{jn}}) - nf(x) \\
&= g_n(x)-f(x) + \frac{1}{n} \sum_{j=1}^{n} \sum_{i\ne j} \frac{1}{h_{jn}} K(\frac{x-X_i}{h_{jn}}) \\
&\quad - (n-1)f(x) \\
&= \{g_n(x)-f(x)\} + (n-1)\{\frac{1}{n} \sum_{j=1}^{n} \frac{1}{(n-1)} \sum_{k\ne j} \frac{1}{h_{jn}} K(\frac{x-X_i}{h_{jn}}) \\
&\quad - f(x)\}
\end{aligned} \tag{2.4.3}$$

Now, for each j,

$$\frac{1}{n-1} \sum_{i\ne j} \frac{1}{h_{jn}} K(\frac{x-X_i}{h_{jn}})$$

is a consistent estimator of f(x) and hence their average is also consistent. Therefore, the second term in (2.4.3) converges to zero in probability and hence for the $\varepsilon>0$ chosen before,

$$(n-1)\left|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n-1}\sum_{i\neq j}\frac{1}{h_{jn}}K\left(\frac{x-X_i}{h_{jn}}\right)\right\}-f(x)\right| \le (n-1)\varepsilon \tag{2.4.4}$$

for all $n\geq N_2$, for some $N_2>0$, with probability arbitrarily close to 1. Substituting (2.4.4) in (2.4.2), we have

$$-n\varepsilon \le \sum_{j=1}^{n} f_{jn}(x)-nf(x) \le n\varepsilon$$

if and only if

$$-n\varepsilon \le \{g_n(x)-f(x)\} + (n-1)\{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{(n-1)}\sum_{i\neq j}\frac{1}{h_{jn}}K(\frac{x-X_i}{h_{jn}}) -f(x)\} \le n\varepsilon \quad .$$

Hence, for all $n \geq \max(N_1,N_2)$,

$$-n\varepsilon + (n-1)\varepsilon \le g_n(x)-f(x) \le n\varepsilon - (n-1)\varepsilon$$

i.e., $|g_n(x)-f(x)| \le \varepsilon$ with probability arbitrarily close to 1. Therefore, Theorems 2.3.1 and 2.3.2 can be carried over to the spheres of influence estimator.

Theorem 2.4.1: Let K be the uniform density on [-1,1] and let $\{h_{jn}\}$, j = 1,2,...,n satisfy the following conditions:

(a) $h_{jn} \to 0$ in probability, j = 1,2,...,n and

(b) $n^{\alpha} h_{jn} \to \infty$ in probability for some $0<\alpha<\frac{1}{2}$, j = 1,2,...,n,

then

(i) at every continuity point of f, $g_n(x) \to f(x)$ in probability

(ii) if f is uniformly continuous on R, then

$$\sup_x |g_n(x) - f(x)| \to 0 \text{ in probability}$$

Theorem 2.4.2: If the kernel K has bounded variation, and $\{h_{jn}\}$, j = 1,2,...,n, satisfy,

(a) $h_{jn} \to 0$ in probability, j = 1,2,...,n

and

(b) $n^{\alpha} h_{jn} \to \infty$ in probability for some $0<\alpha<\frac{1}{2}$, j = 1,2,...,n.

Then,

(i) at every continuity point of f, $g_n(x) \to f(x)$ in probability, and

(ii) if f is uniformly continuous, then

$$\sup_x |g_n(x) - f(x)| \to 0 \text{ in probability.}$$

2.5. Remarks

The nearest neighbors and spheres of influence estimators use the data in determining windows for kernel density estimators in such a way that the windows and the number of points falling in them are both random. In various empirical studies, we have found that the optimal choice of k(n) lies around $n^{1/2}$ and that $k(n) = n^{1/2}$ gives the optimal estimate for all practical purposes. The choice of the constant multiple γ can be determined by a search over a grid of values maximizing the likelihood type function

$\prod_{i=1}^{n} \hat{f}_n(x_i)$, where $\hat{f}_n(x_i)$ is the estimate based on the sample with X_i removed. The value of γ is usually close to unity. In the next chapter, we examine various possibilities regarding the choice of γ and suggest a way of fixing γ. Results of simulation studies using this technique and comparison with the fixed window kernel estimator using optimal window size are presented in Chapter 3.

3. SIMULATION RESULTS

3.1. Introduction

In the preceding chapter, we have shown that the nearest neighbors and the spheres of influence estimators are consistent estimators of the unknown continuous univariate density f. However, the parameters α' and γ in the definition of

$$k(n) = [n^{\alpha'}], \quad 0<\alpha'<1$$

and

$$h(x) = \gamma \, \Sigma |x-X_j|$$

j ranging over k(n)
nearest neighbors of x in
$(X_1, X_2, \ldots, X_n)$

are yet to be determined before these estimators can be used in practice. It is desirable to have estimators which do not involve constants depending upon the unknown density f. A one shot approach which provides an estimate on obtaining a sample of size n is what we hope to achieve.

One method of choosing (α', γ) is to use the method suggested by Shuster and Gregory (1978) for the nearest neighbor estimator which utilizes the distance to the $k(n)^{th}$ nearest neighbor in determining the window at a point. They used a grid of values of (α', γ) to maximize

$$\prod_{i=1}^{n} \hat{f}(X_i) \tag{3.1.1}$$

$\hat{f}(X_i)$ is the estimate at X_i based on (n-1) observations with X_i deleted. In test runs with various distributions, the optimal values of α' and γ for the nearest neighbors and spheres of influence estimators were found to lie around (.5, 1). The use of the optimality criterion (3.1.1) did not give any great improvement in mean squared error and required a large number of computations to implement. Moreover, fixing the values of

$$\left.\begin{array}{l} \alpha' = .5 \\ \gamma \ = 1 \end{array}\right\}$$

provides a quick and easy method of obtaining density estimates. It is interesting to note that for nearest neighbor estimator which uses only the $k(n)^{th}$ nearest neighbor to define the window at a point, the optimal value of α' can generally be as high as .8.

In evaluating nearest neighbors and spheres of influence estimators, the sample has to be sorted in ascending order so that the neighbors of a point can be found. The computations required to sort and then to find neighbors of a point increase rapidly as the sample size increases. This may not be much of a drawback since other methods require specification of some parameters before an estimate of the density can be obtained. Knowledge about these parameters may not be available and in its absence good estimates will not be possible. However, for a large enough sample size, the naive shifted histogram estimator will yield good estimates and as the computing time required is small, a few test runs with different window sizes will be sufficient. For small sample sizes, say up to 200, the nearest neighbors estimators are superior to the other kernel estimators and provide estimates without any information besides the sample.

In this chapter, we study the mean squared error rates and efficiency comparisons of the nearest neighbors and spheres of influence estimators by generating samples from different distributions. Section 3.2 discusses the distribution considered and the method of comparison, Section 3.3 presents the results of the simulation study and Section 3.4 is devoted to some concluding remarks.

3.2. Method of Comparison

To determine the efficiency of the estimators, the kernel fixed window estimator (2.2.3) is used as a standard. It is defined as

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)$$

where $\{h_n\}$ is a sequence of constants, and K is the kernel used. The integrated mean squared error for this estimator is of the order of $0(n^{-4/5})$ and this convergence is fastest when

$$h_n = n^{-1/5}\alpha(K)\beta(f), \qquad (3.2.1)$$

and

$$\alpha(K) = [\frac{\int K^2(y)dy}{2r(\int y^r K(y)dy/r!)^2}]^{1/2r + 1}$$

and

$$\beta(f) = [\int |f^{(r)}(y)|^2 dy]^{-1/(2r+1)}.$$

Tapia and Thompson (1978). The value r is called the characteristic exponent of the kernel K and for the kernels used in this simulation, the value of r used was 2. Although $\beta(f)$ is a function of the unknown density, for the purpose of comparing efficiencies we can use the fixed window kernel estimator with the optimal window given by (3.2.1) as a standard technique against which the new estimators are tested.

Three distributions are considered. The normal distribution with mean zero and unit variance; equal mixture of normal distributions with means -1.5 and 1.5 and unit variances; and the t-distribution with 5 degrees of freedom. Three kernels with characteristic exponent r=2 are used. The uniform kernel K_1, the quartic kernel K_2 and the normal kernel K_3. The values of $\alpha(K)$ used in the definition of the optimal window, (3.2.1) are presented in Table 3.1. For each of the distributions listed, each of the kernels is used for sample

Table 3.1. Values of $\alpha(K)$

K	$\alpha(K)$
$K_1(y) = \frac{1}{2} \qquad \lvert y\rvert \leq 1$	1.351
$K_2(y) = \frac{15}{16}(1-y^2)^2 \lvert y\rvert \leq 1$	2.0362
$K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \quad \lvert y\rvert < \infty$	.7764

sizes of 50 and 100. All samples were generated by the random number generators of the IMSL programs. Fifty equidistant points in the range of (-4, 4) for N(0,1) and t_5 distributions, and in the range of (-5, 5) for the mixture of two normal distributions, are selected at which the unknown density is estimated. The estimates of mean squared error at these points are obtained by repeating the estimation procedure a hundred times with a new random sample at each repetition.

3.3. Results

For each distribution, kernel and sample size, we present

(i) the integrated mean squared error,

(ii) efficiency of nearest neighbors, and spheres of influence estimators compared to the optimal fixed window kernel estimator,

(iii) the mean squared errors of the three estimators at the fifty points at which the density is estimated for a sample size and kernel, and

(iv) a sample plot of the estimates obtained by the three methods for a sample size and kernel.

These are given in Tables 3.2, 3.3, 3.4 and Figure 3.1 for the Normal (0,1) distribution, Tables 3.5, 3.6, 3.7, and Figure 3.2 for the equal mixture of normal distributions, and in Tables 3.8, 3.9, 3.10 and Figure 3.3 for the t-distribution with five degrees of freedom.

From the tables, it is seen that the spheres of influence and nearest neighbors estimators are better than the fixed optimal window kernel estimator in most instances. The efficiencies of the estimators for the Normal (0,1) distribution are between 45 and 105% for the uniform and quartic kernels but they are between 113 and 137% for the normal kernel. For estimating the bimodal distribution, equal mixture of normal distributions, the spheres of influence and nearest neighbor estimators are superior for the uniform and normal kernels and the fixed optimum window kernel estimators is only slightly better for the quartic kernel. For estimating the t-distribution, the spheres of influence and nearest neighbor estimators are at least as good or better than the fixed optimal window kernel estimator. Considering that the fixed window estimator uses the optimal window size which is a function of the unknown density f, our estimators are preferable and perform better also.

It is natural to compare these estimators to the nearest neighbor estimator which utilizes only the $k(n)^{th}$ nearest neighbor of a point in determining the window. However, no efficient algorithm is available to determine the values of α' and γ. Using the values of $\alpha' = .5$ and $\gamma = 1$ leads to absurd estimates and efficiencies of the order of 400%-800% for the nearest neighbors estimators. The nearest neighbor estimator is sensitive to the choice of α' and the optimal value is around .8. Also, for the small sample sizes considered, the nearest neighbor estimator is not very superior to the fixed optimal window kernel estimator and hence comparison of our estimators to the fixed window estimator is sufficient for establishing their usefulness and preferability.

The nearest neighbors estimator is more efficient in the region where the density is large compared to the other two estimators. The spheres of influence estimator, though more efficient than the fixed window estimator, is less efficient than the nearest neighbors estimator in the region where the density is away from zero. In sparse areas, the fixed window estimator is more efficient than the other two and the spheres of influence estimator is more efficient than the nearest neighbors estimator. Among the kernels considered, the quartic and normal kernel gave better results for the fixed window estimator. For estimating the Normal (0,1) distribution, the normal kernel gave better results for our estimators. However, the uniform kernel performed at least as good as the other two considered for estimating the mixture of normals and the t-distribution.

3.4. Conclusions

The nearest neighbors and spheres of influence estimators provide good estimates for small samples from the unknown continuous distribution without the need to supply any parameters before the estimation procedure can be started. These estimators are especially good for densities with long drawn out tail areas. As the nearest neighbors estimator requires computation of neighbors of the points at which the estimates are desired, it can be used when it is desired to estimate the density at a few points only. The spheres of influence estimator lends itself to the situation when the density is to be estimated over the entire range. For large samples, the nearest neighbors estimator is preferable as the number of neighbors to consider increases at the rate of $n^{.5}$ only.

In this thesis, we have presented a one-shot approach to univariate density estimation which requires only the sample, and its sample size n. Further research to extend the results of this thesis to multivariate distributions and finding efficient algorithms to implement them is a further area of research.

Table 3.2. Integrated mean squared errors, Normal (0,1) distribution

Sample size	Spheres of influence estimator	Nearest neighbors estimator	Fixed optimal window estimator
Uniform kernel, $K_1(y) = \frac{1}{2}, \lvert y\rvert \leq 1$			
50	.031112	.023278	.024549
100	.021649	.018975	.014468
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2, \lvert y\rvert \leq 1$			
50	.041825	.042415	.022865
100	.024769	.029851	.013598
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, -\infty<y<\infty$			
50	.018669	.018333	.023916
100	.010370	.012530	.014234

Table 3.3. Efficiency of the estimators, Normal (0,1) distribution

Sample size	Spheres of influence vs. fixed window	Nearest neighbors vs. fixed window	Spheres of influence vs. nearest neighbors
Uniform kernel, $K_1(y) = \frac{1}{2}, \lvert y\rvert \leq 1$			
50	78.88	105.46	74.8
100	66.83	76.25	87.65
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2, \lvert y\rvert \leq 1$			
50	54.67	53.9	101.41
100	54.89	45.55	120.51
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, -\infty<y<\infty$			
50	128.11	130.45	98.2
100	137.26	113.599	120.83

Table 3.4. Mean squared errors, N(0,1) distribution, sample size = 100 (normal kernel)

X-values	Spheres of influence	Nearest neighbors	Fixed window
-3.84	0.000103	0.000383	0.000000
-3.68	0.000125	0.000440	0.000006
-3.52	0.000152	0.000508	0.000017
-3.36	0.000184	0.000590	0.000019
-3.2	0.000220	0.000686	0.000016
-3.04	0.000260	0.000800	0.000032
-2.88	0.000300	0.000933	0.000067
-2.72	0.000333	0.001088	0.000108
-2.56	0.000353	0.001266	0.000196
-2.4	0.000350	0.001469	0.000294
-2.24	0.000320	0.001689	0.000489
-2.08	0.000268	0.001872	0.000645
-1.92	0.000218	0.001964	0.000846
-1.76	0.000222	0.001889	0.000992
-1.6	0.000350	0.001614	0.001285
-1.44	0.000678	0.001043	0.001750
-1.28	0.001246	0.000614	0.002156
-1.12	0.002023	0.000451	0.003052
-.96	0.002884	0.000716	0.003590
-.8	0.003610	0.001211	0.003641
-.64	0.003906	0.001829	0.003969
-.48	0.003651	0.002721	0.003592
-.32	0.003246	0.003733	0.004197
-.16	0.003406	0.003851	0.004501
0.00	0.003852	0.005107	0.005765
.16	0.003629	0.004581	0.005542
.32	0.003668	0.004736	0.005324

Table 3.4 (Continued)

X-values	Spheres of influence	Nearest neighbors	Fixed window
.48	0.003947	0.004049	0.005552
.64	0.003640	0.003824	0.005741
.8	0.003153	0.002416	0.003735
.96	0.003032	0.001761	0.003175
1.12	0.002771	0.001098	0.003560
1.28	0.002252	0.000618	0.003157
1.44	0.001623	0.000548	0.002933
1.6	0.001037	0.000590	0.002330
1.76	0.000592	0.000880	0.002009
1.92	0.000322	0.001297	0.001553
2.08	0.000210	0.001616	0.000970
2.24	0.000205	0.001718	0.000676
2.4	0.000247	0.001639	0.000472
2.56	0.000293	0.001488	0.000293
2.72	0.000320	0.001338	0.000228
2.88	0.000321	0.001188	0.000209
3.04	0.000302	0.001038	0.000140
3.2	0.000270	0.000900	0.000071
3.36	0.000232	0.000776	0.000037
3.52	0.000195	0.000669	0.000023
3.68	0.000161	0.000576	0.000013
3.84	0.000132	0.000498	0.000007
4.00	0.000108	0.000432	0.000008
Integrated MSE	0.010370	0.012530	0.014234

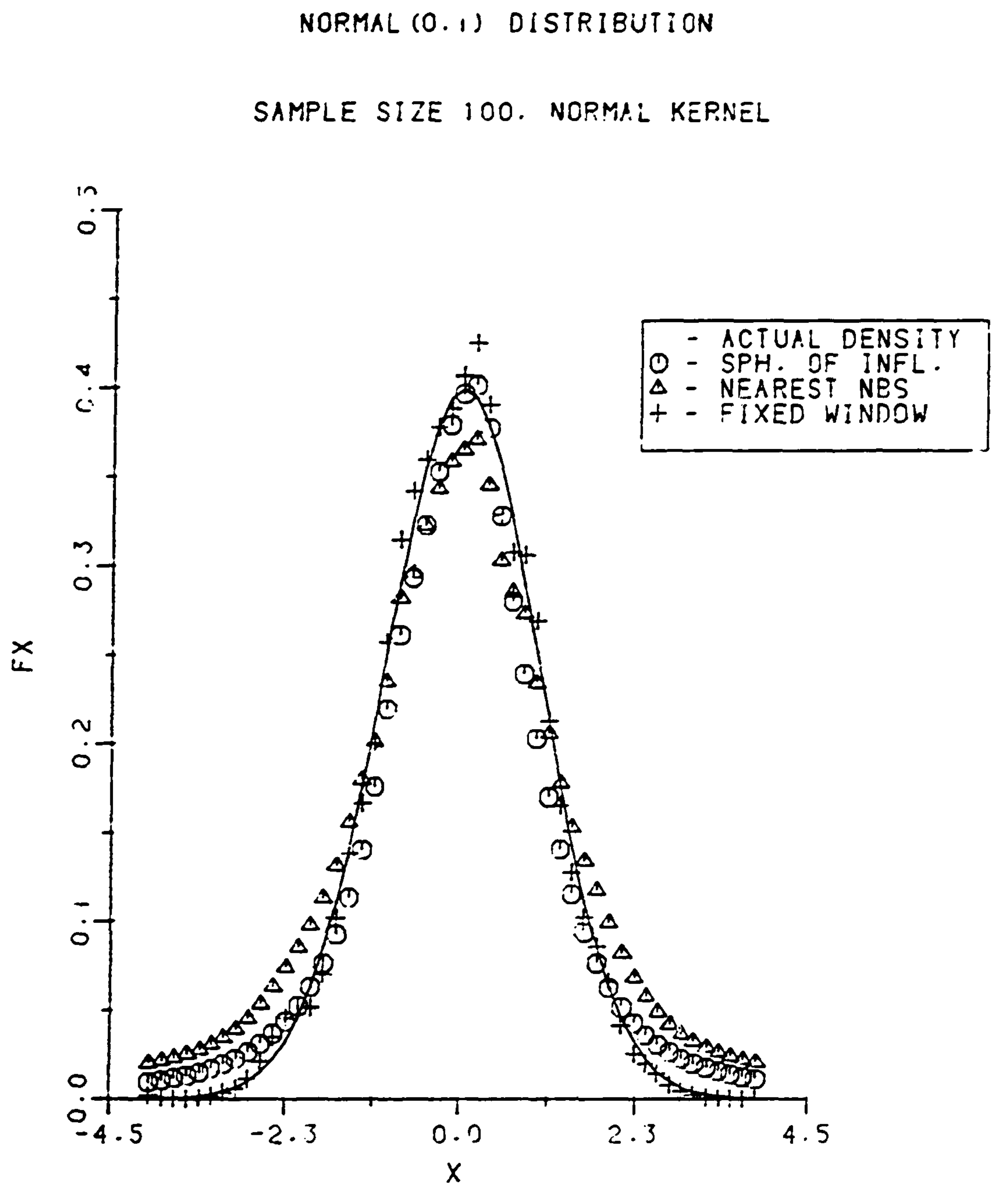

Figure 3.1. Sample plot

Table 3.5. Integrated mean squared errors, equal mixture of normal (-1.5, 1) and normal (1.5, 1)

Sample size	Spheres of influence estimator	Nearest neighbors estimator	Fixed optimal window estimator
Uniform kernel, $K_1(y) = \frac{1}{2}$, $\lvert y \rvert \leq 1$			
50	.014220	.013641	.020892
100	.009538	.010888	.012315
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2$, $\lvert y \rvert \leq 1$			
50	.017512	.017859	.019506
100	.012142	.013419	.011409
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$, $-\infty<y<\infty$			
50	.014586	.012375	.020371
100	.009566	.009471	.011948

Table 3.6. Efficiency of the estimators, equal mixture of normal (-1.5, 1) and normal (1.5, 1)

Sample size	Spheres of influence of fixed window	Nearest neighbors vs. fixed window	Spheres of influence vs. nearest neighbors
Uniform kernel, $K_1(y) = \frac{1}{2}$, $\lvert y \rvert \leq 1$			
50	146.92	153.16	95.93
100	129.12	113.11	114.15
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2$, $\lvert y \rvert \leq 1$			
50	111.39	109.22	101.98
100	93.96	85.02	110.52
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$, $-\infty<y<\infty$			
50	139.66	164.61	84.84
100	124.9	126.15	99.01

Table 3.7. Mean squared errors, equal mixture of N (-1.5, 1) and N (1.5, 1), sample size = 100, uniform kernel

X-values	Spheres of influence	Nearest neighbors	Fixed window
-4.8	0.000075	0.000736	0.000009
-4.6	0.000099	0.000887	0.000019
-4.4	0.000128	0.001081	0.000030
-4.2	0.000149	0.001332	0.000061
-4.0	0.000175	0.001641	0.000091
-3.8	0.000200	0.002019	0.000178
-3.6	0.000199	0.002294	0.000227
-3.4	0.000204	0.002230	0.000363
-3.2	0.000208	0.001714	0.000518
-3.0	0.000325	0.000985	0.000802
-2.8	0.000518	0.000438	0.001077
-2.6	0.000842	0.000159	0.001210
-2.4	0.001143	0.000190	0.001553
-2.2	0.001594	0.000481	0.001543
-2.0	0.001809	0.000894	0.001768
-1.8	0.002035	0.001289	0.001650
-1.6	0.001866	0.001455	0.001959
-1.4	0.001305	0.001645	0.002021
-1.2	0.002042	0.001504	0.002430
-1.0	0.002013	0.001259	0.002235
-.8	0.001286	0.000798	0.002334
-.6	0.001274	0.000397	0.001765
-.4	0.001202	0.000424	0.001509
-.2	0.001377	0.000851	0.001493
0.00	0.001397	0.001315	0.001927
.2	0.001632	0.001771	0.001587
.4	0.001759	0.001434	0.001503
.6	0.001656	0.000792	0.002080

Table 3.7 (Continued)

X-values	Spheres of influence	Nearest neighbors	Fixed window
.8	0.001604	0.000402	0.002185
1.0	0.001527	0.000461	0.001930
1.2	0.001405	0.000633	0.002408
1.4	0.001280	0.001114	0.002517
1.6	0.001375	0.001365	0.002398
1.8	0.001546	0.001392	0.002511
2.0	0.001665	0.001235	0.002561
2.2	0.001954	0.001343	0.002257
2.4	0.001689	0.000746	0.002060
2.6	0.001577	0.000437	0.001665
2.8	0.000998	0.000182	0.001550
3.0	0.000695	0.000142	0.001053
3.2	0.000452	0.000399	0.000889
3.4	0.000325	0.001011	0.000578
3.6	0.000245	0.001795	0.000368
3.8	0.000219	0.002163	0.000256
4.0	0.000160	0.002092	0.000164
4.2	0.000152	0.001805	0.000120
4.4	0.000123	0.001502	0.000103
4.6	0.000104	0.001227	0.000053
4.8	0.000084	0.001005	0.000020
5.0	0.000071	0.000831	0.000005
Integrated MSE	0.009538	0.010888	0.012315

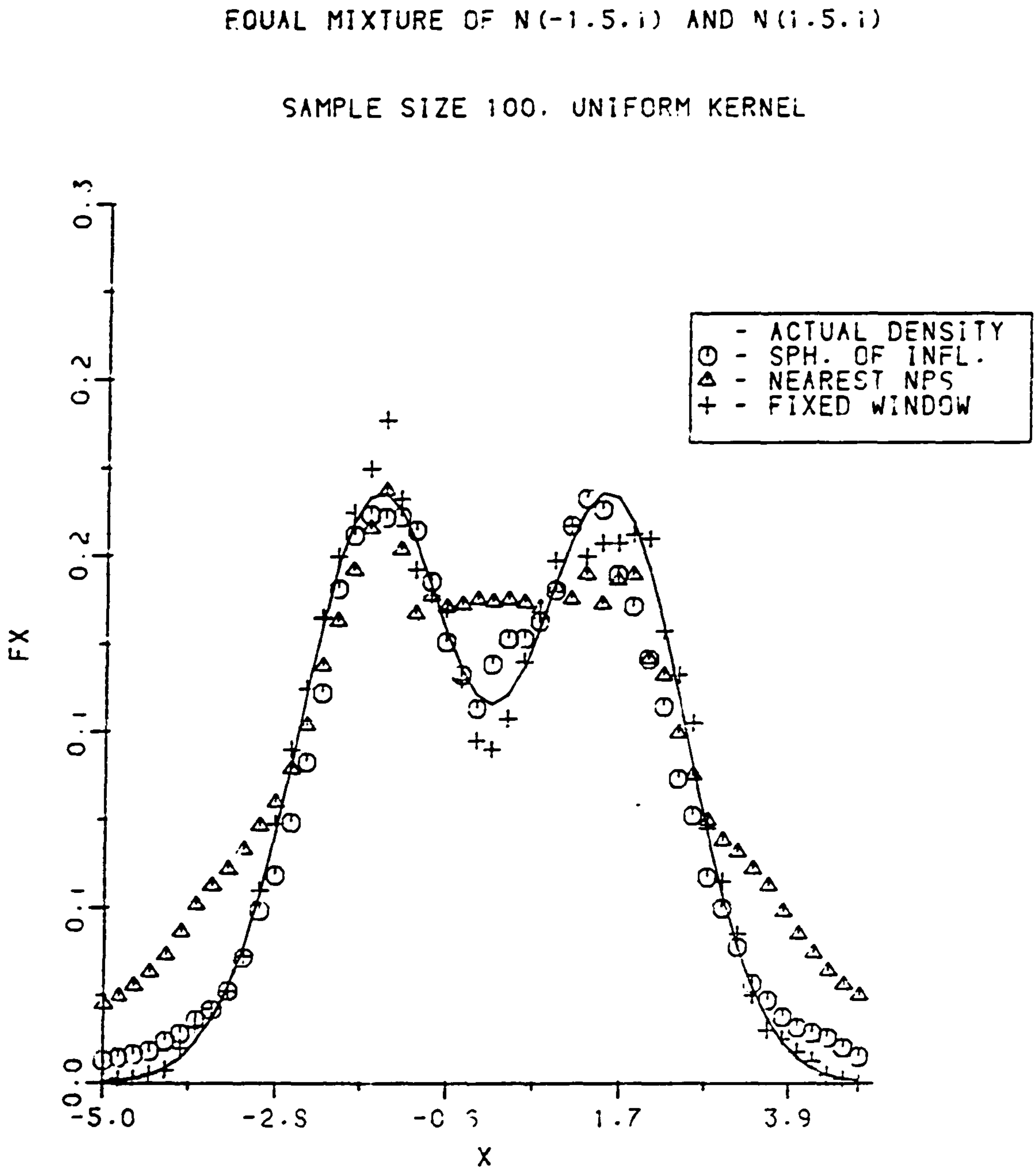

Figure 3.2. Sample plot

Table 3.8. Integrated mean squared errors, t-distribution with 5 degrees of freedom

Sample size	Spheres of influence vs. fixed window	Nearest neighbors vs. fixed window	Spheres of influence vs. nearest neighbor
Uniform kernel, $K_1(y) = \frac{1}{2}, \ |y| \leq 1$			
50	.034185	.049596	.063060
100	.025284	.047735	.048488
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2, \ |y| \leq 1$			
50	.034796	.045330	.060647
100	.028953	.041328	.047370
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \ -\infty<y<\infty$			
50	.051387	.055205	.062000
100	.040059	.051348	.048118

Table 3.9. Efficiency of the estimators, t-distribution with 5 degrees of freedom

Sample size	Spheres of influence vs. fixed window	Nearest neighbors vs. fixed window	Spheres of influence vs. nearest neighbor
Uniform kernel, $K_1(y) = \frac{1}{2}, \ |y| \leq 1$			
50	184.47	127.15	145.08
100	191.77	101.58	188.79
Quartic kernel, $K_2(y) = \frac{15}{16}(1-y^2)^2, \ |y| \leq 1$			
50	174.29	133.79	130.27
100	163.61	114.62	142.74
Normal kernel, $K_3(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \ -\infty<y<\infty$			
50	120.65	112.31	107.43
100	120.12	93.71	128.18

Table 3.10. t-distribution with 5 degrees of freedom, sample size = 100, quartic kernel

X-values	Spheres of influence	Nearest neighbors	Fixed window
-3.84	0.000253	0.002954	0.000898
-3.68	0.000302	0.003048	0.000821
-3.52	0.000358	0.003124	0.000865
-3.36	0.000421	0.003165	0.001249
-3.20	0.000487	0.003175	0.001537
-3.04	0.000554	0.003184	0.002053
-2.88	0.000615	0.003208	0.002488
-2.72	0.000668	0.003253	0.002749
-2.56	0.000709	0.003251	0.002548
-2.40	0.000759	0.003123	0.002798
-2.24	0.000845	0.002961	0.002803
-2.08	0.000958	0.002746	0.002958
-1.92	0.001065	0.002587	0.003380
-1.76	0.001187	0.002230	0.003224
-1.60	0.001350	0.001566	0.002635
-1.44	0.001451	0.001266	0.003699
-1.28	0.001452	0.000694	0.003797
-1.12	0.001606	0.000530	0.002751
-.96	0.002357	0.001217	0.003397
-.8	0.003974	0.002849	0.005753
-.64	0.006049	0.005764	0.008712
-.48	0.008526	0.009764	0.013865
-.32	0.011849	0.014002	0.016024
-.16	0.015612	0.018748	0.019375
0.00	0.018033	0.022303	0.024908
.16	0.018971	0.023218	0.027414
.32	0.018112	0.021805	0.025650
.48	0.015432	0.018204	0.021371
.64	0.012319	0.013687	0.018162

Table 3.10 (Continued)

X-values	Spheres of influence	Nearest neighbors	Fixed window
.8	0.008411	0.008918	0.012060
.96	0.005408	0.004907	0.007650
1.12	0.003955	0.002870	0.005678
1.28	0.002524	0.001187	0.003686
1.44	0.001648	0.000630	0.003553
1.6	0.001436	0.000834	0.003214
1.76	0.001431	0.001384	0.003362
1.92	0.001412	0.001840	0.003095
2.08	0.001342	0.002363	0.002863
2.24	0.001222	0.002790	0.003033
2.4	0.001071	0.003195	0.003504
2.56	0.000910	0.003229	0.002954
2.72	0.000750	0.003285	0.002235
2.88	0.000628	0.003352	0.001944
3.04	0.000557	0.003390	0.002117
3.2	0.000503	0.003391	0.002015
3.36	0.000449	0.003337	0.001538
3.52	0.000394	0.003298	0.001442
3.68	0.000341	0.003256	0.001100
3.84	0.000294	0.003191	0.001002
4.00	0.000253	0.003099	0.001135
Integrated MSE	0.028953	0.041328	0.047370

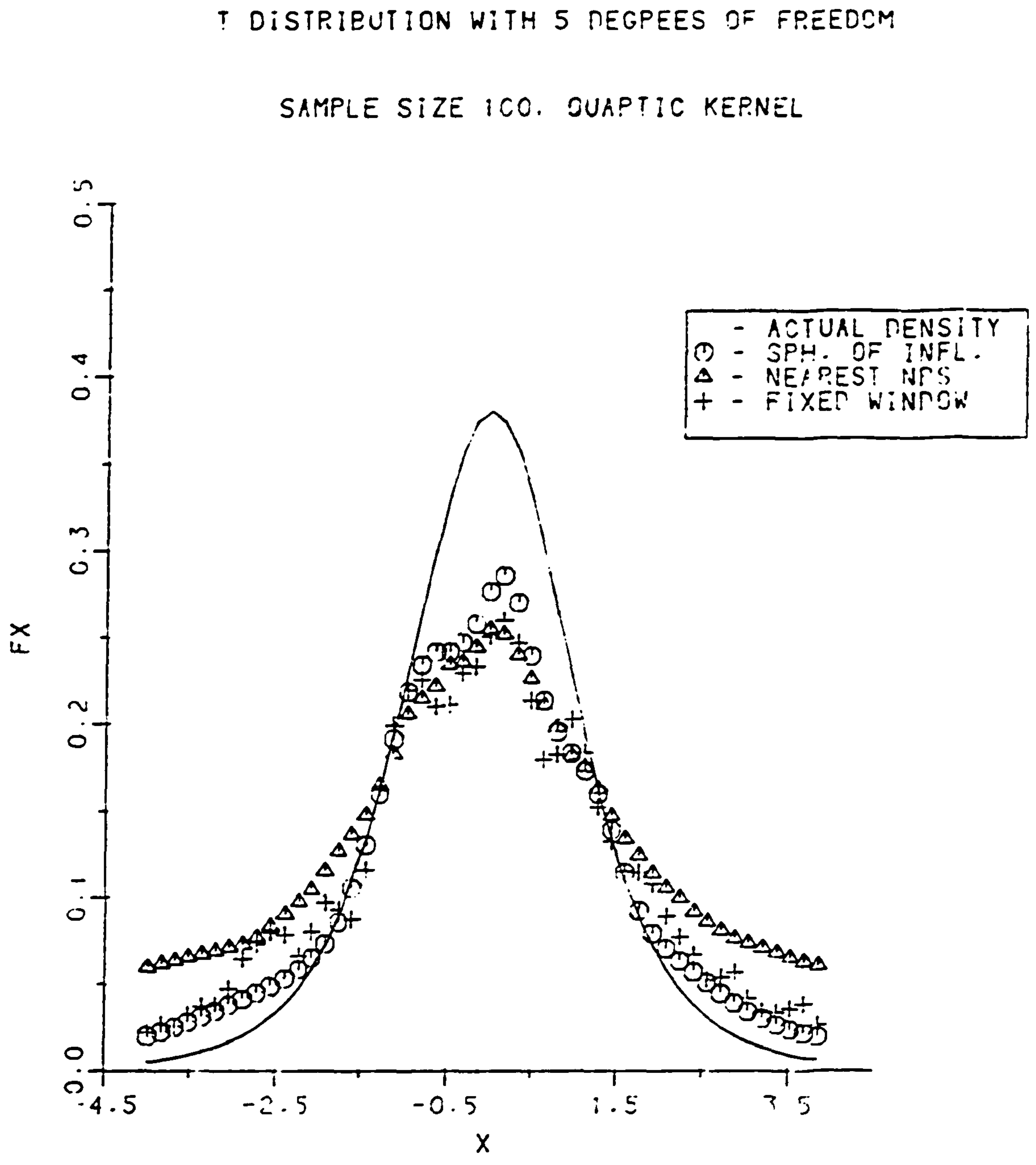

Figure 3.3. Sample plot

4. BIBLIOGRAPHY

Anderson, G. D. 1969a. A comparison of probability density estimates. Presented at IMS annual meetings, New York, August 1969.

Anderson, G. D. 1969b. Nonparametric density estimation. Ph.D. dissertation. University of Washington.

Bartlett, M. S. 1963. Statistical Estimation of Density Functions. Sankhya, Series A, 25:245-254.

Bennett, J. O. 1974. Estimation of a Multivariate Probability Density Function Using B-Splines. Doctoral Dissertation. Rice University, Houston, Texas.

Bhattacharya, P. K. 1967. Estimation of a probability density function and its derivatives. Sankhya, Series A, 29:373-382.

Bickel, P. J. and Rosenblatt, M. 1973. On Some Global Measures of the Deviations of Density Function Estimates. Annals of Statistics 1:1071-1095.

Blaydon, C. C. 1967. Approximation of Distribution and Density Functions. Proceedings of IEEE 55:231-232.

Boneva, L. L., Kendall, D. G. and Stefanov, I. 1971. Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst. Journal of the Royal Statistical Society, Series B, 33:1-70.

Borwanker, J. D. 1971. Asymptotic Theory of Density Estimation. Z. Wahrscheinlichkeitsch 20:182-188.

Breiman, Leo, Meisel, William and Purcell, Edward. 1977. Variable Kernel Estimates of Multivariate Densities. Technometrics 19 (2):135-144.

Brunk, H. D. 1976. Univariate Density Estimation by Orthogonal Series. Technical Report #51. Statistics Department, Oregon State University.

Cacoullos, T. 1966. Estimation of a Multivariate Density, Ann. Inst. Statist. Math. 18:179-189.

Carrol, R. J. 1976. On Sequential Density Estimation. Z. Wahrscheinlichkeit Theorie Verw. Gebiete 36:137-151.

Cencov, N. N. 1962. Evaluation of an Unknown Distribution Density from Observations. Soviet Mathematics 3: 1559-1562.

Chernoff, H. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on a Sum of Observations. Annals of Math. Stat. 23:493-507.

Chi, P. Y. and Van Ryzin, John. 1977. A Histogram Method for Nonparametric Classification. Classification and Clustering. Academic Press Inc., New York.

Cover, T. M. and Hart, P. E. 1967. Nearest Neighbor Pattern Classification. IEEE Trans. IT-13:21-27.

Crain, B. R. 1974. Estimation of distributions using orthogonal expansions. Annals of Stat. 2:454-463.

Crain, B. R. 1976. Matrix Density Estimation. Commun. Statist. A5(1):89-96.

Craswell, K. J. 1965. Density Estimation in a Topological Group. Annals of Math. Stat. 56:1047-1048.

Davies, H. I. 1973. Strong Consistency of a Sequential Estimator of a p.d.f. Bull. Math. Statist. 15:49-54.

Davies, H. I. and Wegman, E. J. 1975. Sequential Nonparametric Density Estimation. IEEE Trans. IT-21: 619-628.

Davis, K. B. 1975. Mean Square Error Properties of Density Estimates. Annals of Statistics 3:1025-30.

de Montricher, G. M. 1973. Nonparametric Bayesian Estimation of Probability Densities by Function Space Techniques. Doctoral Dissertation. Rice University, Houston, Texas.

de Montricher, G. M., 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.

Duin, R. P. 1976. On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions. IEEE Transactions on Computers 25:1175-1179.

Epanechnikov, V. A. 1969. Nonparametric Estimates and a Multivariate Probability Density. Theory of Probability and Its Applications 14:153-158.

Farrell, R. H. 1972. On Best Obtainable Asymptotic Rates and Convergence in Estimates of a Density Function at a Point. Annals of Math. Stat. 43:170-80.

Feinberg, S. and Holland, P. 1972. On the Choices of Flattening Constants for Estimating Multinomial Probabilities. J. Multivariate Anal. 2(1):127-134.

Fellner, W. H. 1974. Heuristic Estimation of Probability Densities. Biometrika 61:485-492.

Fellner, W. H. and Tarter, M. E. 1971. Some New Results Concerning Density Estimates based upon Fourier Series. Proceedings of the Fifth Interface Symposium between Statistics and Computer Science 5:54-64.

Fix, E., and Hodges, J. L., Jr. 1951. Nonparametric Discrimination: Report #4, Project #21-49-004. USAF School of Aviation Medicine, Randolph AFB, Texas.

Fix, E., and Hodges, J. L., Jr. 1952. Nonparametric Discrimination II: Report #11, Project #21-49-004. USAF School of Aviation Medicine, Randolph AFB, Texas.

Fryer, M. J. 1977. A Review of some Nonparametric Methods of Density Estimation. Journal of the Institute of Mathematics and Its Applications 20:335-354.

Good, I. J. 1971. A Nonparametric Roughness Penalty for Probability Densities. Nature 229:29-30.

Good, I. J. and Gaskins, R. A. 1971. Nonparametric Roughness Penalties for Probability Densities. Biometrika 58:255-277.

Good, I. J. and Gaskins, R. A. 1972. Global Nonparametric Estimation of Probability Densities. Virginia Journal of Science 23:171-93.

Gregory, G. C. and Schuster, E. F. 1979. Contributions to Nonparametric Maximum Likelihood Methods of Density Estimation. Proceedings of the Twelfth Annual Symposium on the Interface of Computer Science and Statistics, University of Waterloo 12:427-431.

Grenander, U. 1956. On the theory of mortality measurement. Skand. Aktuarietidskr. 39:125-153.

Habbema, J. D. F., Hermans, J., and Van Den Broek, K. 1974. A Stepwise Disciminant Analysis Program Using Density Estimation. COMPSTAT-74, Physica, Vienna, 101-110.

Habbema, J. D. F., Hermans, J., and Remme, J. 1978. Variable Kernel Density Estimation in Discriminant Analysis. COMPSTAT-78:178-185.

Kashyap, R. L. and Blaydon, C. C. 1968. Estimation of Probability Density and Distribution Functions. IEEE Transactions. IT-14:549-56.

Kiefer, J. and Wolfowitz, J. 1958. On the Deviations of the Empirical Distribution Function of Vector Chance Variables. Transactions of the American Math. Soc. 87:173-86.

Kim, Bock Ki, and Van Ryzin, J. 1974. Uniform Consistency of a Histogram Density Estimator and Model Estimation. MRC Report 1494.

Koontz, W. L. G. and Fukunaga, K. 1972. Asymptotic Analysis of a Nonparametric Clustering Technique. IEEE Transactions on Computers 21:967-974.

Kronmal, R. A., and Tarter, M. E. 1968. The Estimation of Probability Densities and Cumulatives by Fourier Series Methods. Journal of the American Statistical Association 63:925-952.

Lii, Keh-Shin and Rosenblatt, M. 1975. Asymptotic Behavior of a Spline Estimate of a Density Function. Computation and Mathematics with Applications 1:223-235.

Londhe, Anil R., and Gentle, James E. 1979. Density Estimation Using Kernels Over Variable Size Windows. Proceedings of A.S.A. Meetings, American Statistical Association, Washington, D.C.

Loftsgaarden, P. O. and Quesenberry, C. P. 1965. A Nonparametric Estimate of a Multivariate Probability Density Function. Ann. Math. Statist. 28:1049-1051.

Marshall, A. W. and Proschan, F. 1965. Maximum Likelihood Estimation for Distributions with Monotone Failure Rate. Annals of Math. Stat. 36:69-77.

McGilchrist, C. A. 1975. Estimation of Unimodal Probability. Sankhya, Series A, 37:139-149.

Moore, D. S. and Henrichon, E. G. 1969. Uniform Consistency of some Estimates of a Density Function. Annals of Math. Stat. 40: 1499-1502.

Moore, D. S. and Yackel, J. W. 1977a. Consistency Properties of Nearest Neighbor Density Function Estimators. Annals of Statistics 5:143-154.

Moore, D. S. and Yackel, J. W. 1977b. Large Sample Properties of Nearest Neighbor Density Function Estimators. Statistical Decision Theory and Related Topics, II. Academic Press, New York.

Murthy, V. K. 1965. Estimation of probability density. Annals of Math. Stat. 36:1027-1031.

Nadaraya, E. A. 1963. On estimation of density functions of random variables. Soobshch. Akad. Nauk Gruzin SSR XXXII(2):277-280.

Nadaraya, E. A. 1964. Estimation of a convolution component. SIAM Theory of Prob. and Applications 9: 141-142.

Nadaraya, E. A. 1965. On Nonparametric Estimates of Density Functions and Regression Curves. SIAM Journal of Probability and Applications 10:186-190.

Nadaraya, E. A. 1974. On the Integral Mean Square of Some Nonparametric Estimates for the Density Function. SIAM Journal of Probability Theory and Applications 19: 133-141.

Parzen, E. 1962. On the Estimation of a Probability Density Function and the Mode. Annals of Math. Stat. 33:1065-1076.

Rao, B. L. S. P. 1969. Estimation of a unimodal density. Sankhya, Series A,31:23-36.

Robertson, T. 1967. On estimating a density which is measurable with respect to a σ-lattice. Annals of Math. Stat. 38:482-493.

Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. Annals of Mathematical Statistics 27:832-835.

Rosenblatt, M. 1971. Curve Estimates. Annals of Mathematical Statistics 42:1815-42.

Scheult, A. H. and Quesenberry, C. P. 1971. On Unbiased Estimation of Density Functions. Annals of Mathematical Statistics 42:1434-1438.

Schoenberg, I. J. 1946. Contributions to the Problem of Approximation of Equidistant Data by Analytical Functions. Quarterly of Applied Mathematics 4:45-99, 112-14.

Schoenberg, I. J. 1972. Notes on Spline Functions II: On the Smoothing of Histograms. MRC Technical Report #1222, Madison, Wisconsin.

Schucany, W. R. and Sommers, John P. 1977. Improvement of Kernel Type Density Estimators. JASA, Theory & Methods Section, 72 (358):420-423.

Schucany, W. R., Gray, H. L. and Owen, D. B. 1971. On Bias Reduction in Estimation. JASA 66:524-533.

Schuster, E. F. 1969. Estimation of a Probability Density Function and its Derivatives. Annals of Mathematical Statistics 40:1187-1195.

Schuster, Eugene F. 1970. Note on the Uniform Convergence of Density Estimates. Annals of Mathematical Statistics 41:1347-1348.

Schuster, E. F. and Gregory, G. G. 1978. Choosing the Shape Factor(s) When Estimating a Density. Bulletin of the Institute of Mathematical Statistics 7(5):292.

Schwartz, S. C. 1967. Estimation of Probability Density by an Orthogonal Series. Annals of Math. Stat. 38: 1261:1265.

Scott, David W. 1976. Nonparametric Probability Density Estimation by Optimization Theoretic Techniques. Doctoral Dissertation. Rice University, Houston, Texas.

Scott, D. W., Tapia, R. A., and Thompson, J. R. 1976. An Algorithm for Density Estimation. Computer Science and Statistics. Ninth Annual Symposium on the Interface, Harvard University, Cambridge, Massachusetts.

Scott, D. W., Tapia, R. A. and Thompson, J. R. 1977. Kernel Density Estimation Revisited. Nonlinear Analysis 1:339-372.

Silverman, B. W. 1978a. Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives. Annals of Statistics 6:177-184.

Silverman, B. W. 1978b. Choosing the Window Width when Estimating a Density. Biometrika 65:1-12.

Sommers, John P. 1972. Improved Density Estimation. Technical Report #114. Department of Statistics Series, Southern Methodist University.

Specht, D. F. 1971. Series Estimation of a Probability Density Function. Technometrics 13:409-424.

Srivastava, R. C. 1973. Estimation of p.d.f. based on Random Number of Observations with Applications. International Statistician, Rev. 41:77-86.

Tapia, R. A. and Thompson, J. R. 1978. Nonparametric Probability Density Estimation. The Johns Hopkins University Press, Baltimore.

Tarter, M. E. and Kronmal, R. A. 1967. After the histogram what? A Description of New Computer Methods for Estimating the Population Density. Proceedings of the A.C.M. 22:511-519.

Tarter, M. E. and Kronmal, R. A. 1970. On Multivariate Density Estimates based on Orthogonal Expansions. Annals of Math. Stat. 41:718-722.

Tarter, M. E. and Raman, S. 1971. A systematic approach to graphical methods in biometry. Proceedings of the 6th Berkeley Symposium on Math. Stat. and Probability 4: 199-222.

Tarter, M. E. and Kronmal, R. A. 1976. An Introduction to the Implementation and Theory of Nonparametric Density Estimation. American Statistician 30:105-112.

Van Ryzin, J. 1969. On Strong Consistency and Density Estimates. Annals of Math. Stat. 40:1765-1772.

Van Ryzin, J. 1973. A histogram method of density estimation. Comm. in Stat. 12:493-506.

Wagner, T. J. 1973. Strong Consistency of a Nonparametric Estimate of a Density Function. IEEE Trans. Systems, Man and Cybermetrics 3:289-290.

Wagner, T. J. 1975. Nonparametric Estimates of Probability Densities. IEEE Trans. Information Theory IT-21(4): 438-440.

Wahba, Grace. 1971. A Polynomial Algorithm for Density Estimation. Annals of Math. Stat. 42:1870-1886.

Wahba, Grace. 1975. Optimal Convergence Properties of Variable Knot, Kernel and Orthogonal Series Methods for Density Estimation. Annals of Statistics 3:15-29.

Wahba, Grace. 1977. Optimal Smoothing of Density Estimates. Classification and Clustering. Academic Press, Inc., New York.

Wahba, Grace. 1978. Data-based Optimal Smoothing of Orthogonal Series Density Estimates. Technical Report #509. Department of Statistics, University of Wisconsin, Madison, Wisconsin.

Watson, G. S. 1964. Smooth regression analysis. Sankhya, Series A, 26:359-372.

Watson, Geoffrey S. 1969. Density Estimation by Orthogonal Series. Annals of Math. Stat. 40:1496-1498.

Watson, G. S. and Leadbetter, M. R. 1963. On the Estimation of the Probability Density I. Annals of Math. Stat. 34:480-491.

Watson, G. S. and Leadbetter, M. R. 1964a. Hazard analysis II. Sankhya, Series A, 26:101-116.

Watson, G. S. and Leadbetter, M. R. 1964b. Hazard analysis I. Biometrika, 51:175-184.

Wegman, Edward J. 1969. A Note on Estimating a Unimodal Density. Annals of Math. Stat. 40:1661-1667.

Wegman, Edward J. 1970a. Maximum Likelihood Estimation of a Unimodal Density Function. Annals of Math. Stat. 41: 457-471.

Wegman, Edward J. 1970b. Maximum Likelihood Estimation of a Unimodal Density II. Annals of Math. Stat. 41:2169-2174.

Wegman, E. J. 1972a. Nonparametric Probability Density Estimation I. A Summary of Available Methods. Technometrics 11:533-546.

Wegman, E. J. 1972b. Nonparametric Probability Density Estimation II. A Comparison of Density Estimation Methods. Journal of Stat. Computations and Simulation 1:225-245.

Wegman, Edward J. 1977. Maximum Likelihood Estimation of a Probability Density Function. To appear in Sankhya, Series A.

Wegman, E. J. and Davies, H. I. 1979. Remarks on some Recursive Estimators of a Probability Density. Annals of Statistics 7(2):316-327.

Weiss, L. and Wolfowitz, J. 1967. Estimation of a Density Function at a Point. Z. Wahrscheinlickeitsthieor. Verw. Geb., 7:327-335.

Whittle, P. 1958. On the Smoothing of Probability Density Functions. Journal of the Royal Statistical Society (B)20:334-343.

Wolverton, C. T. and Wagern, T. J. 1969. Asymptotically Optimal Discriminant Functions for Pattern Classification. IEEE Trans. Information Theory IT-15(2):258-265.

Woodroofe, M. 1967. On the maximum deviation of the sample density. Annals of Math. Stat. 38:475-481.

Woodroofe, M. 1970. On Choosing a Delta-Sequence. Annals of Math. Stat. 41:1665-1671.

Wozencraft, John M., and Jacobs, Irwin Mark. 1965. Principles of Communication Engineering. John Wiley & Sons, Inc., New York, N.Y.

5. ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. J. E. Gentle for suggesting this dissertation topic and supervising the research in the initial stages while he was at Iowa State University. I would also like to thank Dr. V. A. Sposito who graciously accepted to be my major professor after Dr. Gentle's departure and painstakingly supervised the final preparation of the dissertation, even after my leaving Ames. I also wish to thank the other committee members, Dr. W. J. Kennedy, Dr. D. H. Harville, Dr. K. J. Koehler and Dr. C. G. Maple for their willingness to serve on the committee.

Thanks are due to the Statistical Numerical Analysis section for the financial support provided while I was at Iowa State University and Schering Corporation for providing computing facilities to complete the simulation study.

Thanks go to my colleagues and friends, Mr. Vaithianathan, Mr. Skarpness and Mr. Escobar for the help they provided in getting the dissertation typed, and completing other formalities.

Finally, thanks are due to my wife and parents for their patience and understanding during my years of graduate study.