KERNEL–BASED NONPARAMETRIC DENSITY ESTIMATION AND REGRESSION WITH STATISTICAL APPLICATIONS

by

PETER JOHN FOSTER

A thesis submitted to the

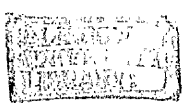
UNIVERSITY OF GLASGOW

for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics

July 1990



Acknowledgements

I would like to thank Dr. A.W. Bowman for all his encouragement and support given while supervising this work.

Thanks also to Lynn Pittendrigh for her careful typing of the final manuscript.

I am grateful to the Science and Engineering Research Council for their provision of a three-year studentship.

Contents

Summary

This thesis is concerned with nonparametric kernel density estimation and regression. In particular, techniques for obtaining better estimates than those produced by the standard fixed kernel approach are examined as well as the use of nonparametric estimates in certain other statistical procedures.

Allowing the degree of smoothing to adapt to the "local" density of the data has been suggested as a means of reducing bias and mean integrated squared error (MISE) in comparison with the levels for fixed kernel density estimators. In chapter two the finite sample properties of two particular adaptive estimators are investigated and compared with those of the fixed kernel method. This is carried out for both univariate and multivariate data from a number of different underlying distributions which are assumed to be of a known form. Numerical integration techniques are used to calculate exact values for the bias, variance and MISE. A simple smoothing strategy based on Normality is also derived.

In the first part of chapter three techniques for obtaining fixed kernel density estimators with smaller bias than that of the standard fixed approach are described and their asymptotic properties studied. These fall into three classes which are using "higher order" kernels, subtracting a bias reducing correction factor and using a multiplicative correction factor. Those with the best asymptotic properties in each class are compared with the standard fixed and the adaptive approaches via a simulation study. In the second part of this chapter methods for reducing the bias inherent in the Priestley-Chao fixed kernel regression estimator are similarly explored. These techniques

are generally analogous to those studied for density estimators except for a two-stage procedure called "twicing" which is also considered.

In chapter four the problem of obtaining pointwise confidence intervals for the unknown density function is examined. The sampling distributions of the estimators are unknown but can be approximated in two ways. These are firstly by assuming Normality and secondly by the use of the bootstrap method. Competing approaches are again compared via a simulation study.

In chapter five two density based tests of multivariate Normality are described. The first is based on a measure of integrated squared error and the second utilises the entropy property of the multivariate Normal (MVN) distribution. Critical values for the test statistics are obtained and a power study carried out. These powers are also compared with those for an omnibus procedure due to Koziol based on the "radii and angle" properties of the MVN distribution.

In chapter six a procedure for graphically exploring a multivariate set based on finding directions of high multivariate density is proposed. The three main aims are to explore the main features of a p-dimensional ($p > 2$) density function, seek non-linear features in the data and use pairs of directions in the construction of two-dimensional representations. This approach is illustrated by application to real data sets.

In chapter seven the goodness-of-fit of a logistic regression model based on multiple covariates is assessed by comparing the parametric probabilities with estimates obtained by nonparametric regression. The global discrepancy is assessed using a pseudo-

likelihood ratio test statistic and significance determined through a simulation procedure. The degree of smoothing plays an important role so methods for choosing the value of the smoothing parameter are disucssed. Also, the use of partial residual plots to determine if the functional form of a covariate effect has been specified correctly are explored and a test of linearity to aid in this is proposed and investigated.

Chapter 1. Introduction

Density estimation is an important topic in applied statistics because in general the underlying density $f(x)$ is unknown so that its characteristics need to be inferred from a random sample $X_1,\ldots,X_n$ before any analysis is carried out. A long standing approach to this, particularly for univariate data, is to construct a histogram. While this provides a useful description of the actual sample it is not appropriate for describing features of the population such as skewness, truncation or bimodality. This is not only due to the stepwise nature of the final figure but also to the essentially arbitrary decisions which have to be made prior to the actual counting and drawing, i.e. the number and size of intervals (or cells) together with their location must be decided. Silverman (1986, Section 3.2) illustrates how different decisions can reshape the final histogram.

Using histograms to present bivariate or trivariate data introduces a number of further problems. It is difficult, because of the block nature, to assess the structure of the data and a contour diagram representation cannot easily be obtained. Also, the final form of a multivariate histogram will be dependent on the co-ordinate direction of the cells.

Two modifications to this process greatly improve it. Firstly, instead of placing the 'boxes' for each observation over the centre of the appropriate histogram cell they can be centred on the actual observed value. Secondly, the 'boxes' can be replaced by a general 'kernel function' K which is usually chosen to be a symmetric probability density function such as a normal density. The kernel

estimator is then defined by

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

where h is the smoothing parameter or bandwidth. The resulting density estimate will integrate to one and inherit the smoothness properties of the particular kernel function used. It is widely regarded that the precise choice of K is not crucial to the performance of such estimators. On the other hand, choice of the value of h plays a critical role both in the performance of the estimator and the form of the final estimate. As h tends to zero the estimate takes on a spikey appearance while as h becomes large all detail is obscured.

A problem associated with such kernel estimates arises from using the same smoothing parameter across the whole sample. This is that spurious noise tends to appear in the tails of long-tailed distributions but increasing h to overcome this masks detail in the main body of the distribution. In order to deal with this problem various authors such as Breiman et al (1977) and Abramson (1982) have proposed adaptive methods which try to adapt the degree of smoothing to the 'local' density of the data. These approaches are discussed in chapter 2 and in particular those which scale h by an estimate of $f(x_i)^{\alpha}$, $0 < \alpha < 1$. There has been much emphasis in the literature on the asymptotic properties of estimators which, in the fixed case, have also proved useful for understanding its small sample behaviour. This is investigated for two particular adaptive estimators by assuming f to be of a known form and then using numerical integration to obtain exact values for bias, variance and mean integrated squared error. A smoothing strategy based on

normality is derived and the results are extended to multivariate data.

Results in Rosenblatt (1956) imply that any unbiased estimate of a continuous density function is either not continuous or not a density function. Hence, in practice a certain degree of bias has to be accepted. In chapter 3 a number of methods for finding estimators with smaller bias, and hopefully MISE as well, are described. These fall into three general classes of approach which are subtracting a bias reducing correction factor, using a 'higher order' kernel which can take negative values (Bartlett (1963)) and using a multiplicative correction factor. The properties of each technique are firstly evaluated asymptotically and then the small sample performances of those with the best asymptotic properties in each class are compared with the simple fixed and the adaptive approaches via a simulation study.

In the second part of chapter 3 the univariate regression problem is discussed. It is assumed that we have observations $(Y_1,x_1)....(Y_n,x_n)$ which satisfy the relationship

$$Y_i = g(x_i) + e_i, \quad i = 1,...,n$$

where the errors e_i are uncorrelated with zero mean and constant variance σ^2, the x_i are equally spaced at intervals of length δ and $g(x)$ is an unknown function for which an estimate is required. Rather than assuming it is of a particular form, e.g. linear, it will be estimated nonparametrically using the Priestley and Chao (1972) fixed kernel estimator

$$\hat{g}(x) = \delta \sum_{i=1}^{n} \frac{1}{b} K\left[\frac{x-x_i}{b}\right] Y_i$$

where K is the kernel function and b the smoothing parameter. This is a weighted average of the Y_i's with the value of b determining the amount of local averaging carried out. Such estimators are, however, inherently biased so in addition to bias reduction for density estimators the problem of reducing the bias of $\hat{g}(x)$ is also considered. The techniques which are studied are in fact analogous to those studied for density estimation with the exception of an additional two stage procedure called 'twicing', first suggested by Tukey (1977), which is also investigated.

In chapters 2 and 3 the main emphasis is on obtaining better estimates than those provided by the fixed kernel approach. The rest of the thesis though is concerned with using estimates, including some of those studied earlier, as part of certain other statistical procedures. This kind of approach dates back to Fix and Hodges (1951) who proposed a form of density estimation to be used as part of a nonparametric discrimination procedure. Since then the number of applications has grown extensively with density or regression estimates being used in a wide variety of areas such as survival analysis, cluster analysis, projection pursuit and parametric model checking.

In chapter 4 the problem of obtaining pointwise confidence intervals for the unknown density f is considered. To try and obtain accurate coverage probabilities it is important the estimate is centered correctly and hence certain of the more effective bias reducing estimators studied in chapter 3 are employed. The exact sampling distributions of the estimators are unknown but can be approximated in two different ways which are either by assuming

normality or by using the bootstrap resampling procedure. A simulation study is again carried out to compare competing methods.

Many 'classical' multivariate statistical procedures assume the data arise from a multivariate normal (MVN) distribution so it is important to check this assumption before an anaylsis is carried out. Many existing techniques tend to concentrate on the 'radii and angles' properties of the MVN distribution which are examined graphically, as by Healy (1968), or alternatively, for example, by comparing the empirical distribution function of the squared radii with the appropriate χ^2 distribution function. An omnibus approach is due to Koziol (1983) who combines a test of the uniformity of the angles with a test based on the radii. In chapter 5 two new density based tests are described which are extensions of univariate tests proposed by Bowman (1988). The first is based on an integrated squared error measure of the discrepancy between an estimate of the density and the expected density under the null hypothesis. The second utilises the property of the MVN distribution that its entropy exceeds that for any other distribution with the same variance structure. Critical values for the test statistics are obtained and a power study carried out. The powers are also compared with those for Koziol's (1983) approach. The possible benefits of using an adaptive estimate in the entropy test statistic is also investigated.

A commonly used technique for investigating the structure in a multivariate dataset is to project the data points onto a lower dimensional subspace, usually of dimension two, and examine plots of the projected data. Principal components analysis (PCA) obtains a subspace which hopefully explains a large percentage of the

variation in the data. Projection pursuit (PP) methods find lower dimensional subspaces such that the projected data maximise an index of 'interestingness' with non-normality being a common choice. Projections may be difficult to interpret however and also may obscure either partly or totally actual structure in the full dimensional data. In chapter 6 a different exploratory approach based on finding directions of high multivariate density is proposed. The principal aims in doing this are to explore the main features of the shape of a p-dimensional ($p > 2$) density function, to find non-linear features in the data and to use pairs of directions for the construction of 2-dimensional representations. These techniques are illustrated by analyses of real data sets.

For data consisting of a binary response, together with the values of a number of covariates, which may arise in many areas of social science and medicine, a commonly used model is the logistic regression model. In order to avoid incorrect conclusions it is important to check goodness of fit and the assumptions underlying the model. In chapter 7 goodness-of-fit is assessed by comparing an estimate of P('success'/$\underline{x}$) based on nonparametric regression with a parametric estimate of this probability function. A pseudo-likelihood ratio test which provides a global measure of the discrepancy is described together with a simulation procedure for assessing significance. The degree of smoothing has an important influence on the results so methods for choosing the value of the smoothing parameter are discussed. Also, the use of partial residual plots to determine if a regressor variable has been specified correctly in the model are explored and a test of linearity is proposed and investigated.

Chapter 2. Adaptive Kernel Density Estimators.

2.1. Introduction.

It is assumed that we have a set of n independent random variables $(X_1, X_2, \ldots, X_n)$ each identically distributed and from a continuous univariate distribution with unknown density f. The problem is to construct an estimate of f based on the sample of observed values $(x_1, x_2, \ldots, x_n)$.

In this chapter the properties of a fixed kernel density estimator will be compared with those of adaptive kernel estimators. This will be done on the basis of both asymptotic and exact small sample results. The aim is to find an estimator with both low bias and variance. This is important when an estimate is to be used for the exploration and presentation of data and especially so if it is to be used as part of another statistical procedure such as a density based test of multivariate normality to be described in chapter 5.

The fixed kernel estimator, introduced by Rosenblatt (1956) is defined by:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K((x-X_i)/h) \qquad (2.1.1)$$

K is the kernel function which satisfies:

(i) $\int_{-\infty}^{\infty} K(t)dt = 1, \quad K(t) \geq 0 \quad$ for every $\quad t \in (-\infty, \infty)$

(ii) $K(t) = K(-t). \quad$ i.e. $\int_{-\infty}^{\infty} tK(t)dt = 0.$

$$\int_{-\infty}^{\infty} t^2 K(t)dt = \beta_2 < \infty$$

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

$$\int_{-\infty}^{\infty} |K(t)|dt < \infty$$

and $\lim_{t\to\infty} |tK(t)| = 0.$ (2.1.2)

These conditions are satisfied by most symmetric probability density functions and in practice a function such as the standard normal density is used.

h is the smoothing parameter or window width and controls the amount of smoothing applied to the data. If h is chosen to be very small then the estimate will take on a spikey appearance as spurious fine structure is highlighted whereas if h is large all detail is obscured. The term "fixed" refers to the fact that the kernel is scaled by the same amount in all parts of the sample. One problem with this for example, is that spurious bumps tend to appear in the tails of long tailed densities. If, however, the value of h is increased in an attempt to remove this effect, structure in the main part of the distribution is masked.

Loftsgaarden and Quesenberry (1965) proposed the nearest neighbour estimator which is given by:

$$\hat{f}_N(x) = \frac{k/(2n)}{R(x;k)} \qquad (2.1.3)$$

where $R(x;k)$ is the distance of x from its kth nearest neighbour among the data. Hence, this estimator does vary the degree of smoothing according to the position of x in the distribution. For example if x is in the tails then $R(x;k)$ will generally be larger than in the main part of the distribution and so a larger amount of smoothing will be carried out. This should remove the bumps characteristic of the estimate based on fixed kernels.

Moore and Yackel (1977) suggested a generalised nearest neighbour estimator by replacing h in (2.1.1) by $R(x;k)$. The main problem with nearest neighbour estimators is that they are complicated functions of x, have discontinuous derivatives at points $(X_{(j)}+X_{(j+k)})/2$, where $X_{(.)}$ are the sample order statistics, and do not integrate to one because the tails approach zero at too slow a rate. Hence, the resulting estimates will have sharp peaks, heavy tails and will therefore not be appropriate if an estimate of the whole density is required.

If equation (2.1.3) is rearranged to give

$$R(x;k) = \frac{k/(2n)}{\hat{f}_N(x)}$$

or

$$R(x;k) \simeq \frac{k/(2n)}{f(x)} \qquad (2.1.4)$$

then the smoothing parameter in the generalised nearest neighbour method can be seen to be proportional to the inverse of another estimator of the underlying density at x, generally referred to as a pilot estimator. If instead, we now consider the distance of X_i from its kth nearest neighbour and use this as a smoothing parameter as suggested by Breiman et al (1977) we have removed the dependence of the window width of the kernel on x. The resulting "variable kernel estimate" will now be a proper p.d.f. (i.e. integrate to one) and inherit the smoothness properties of the kernel function used. The kernel placed over data point x_i will be scaled by an amount proportional to $1/\hat{f}_N(x_i)$. In general though it is not necessary to use a nearest neighbout estimator in the smoothing parameter - any convenient one such as that based on a fixed kernel (2.1.1) may be used instead.

The term "adaptive" has been given to such estimators because the degree of smoothing will adapt to the sparseness of the data, as measured by $\tilde{f}$, about either x or X_i depending on the choice of method. Here, $\tilde{f}$ denotes the particular pilot estimator used.

Abramson (1982) and Silverman (1986) consider more general forms of adaptive estimators where the local smoothing parameter is proportional to a power of the inverse of $\tilde{f}(x_i)$. Abramson (1982) shows that there are theoretical reasons for choosing this power to be $1/2$.

Several simulation studies to examine the performance of adaptive methods are described in the literature. These include those of Breiman et al (1977), Habbema, Hermans and Remme (1978), Bean and Tsokos (1982) and Bowman (1985). Their results show that adaptive methods are better than the fixed kernel approach when the underlying density is heavily skewed or long tailed.

When the data are multivariate the fixed kernel estimator is defined by:

$$\hat{f}(\underline{x}) = n^{-1} h^{-p} \sum_{i=1}^{n} K_p \left[\frac{\underline{x}-\underline{X}_i}{h}\right] \qquad (2.1.5)$$

where K_p is a symmetric p-dimensional density function.

There are additional problems in estimating a density function in a multidimensional setting. Firstly, many observations in a sample will tend to fall at points where the underlying density is low. This means that regions of low density are very important parts of the distribution and thus need to be estimated as accurately as possible. Secondly, many regions of the sample space may be devoid of observations, even those where the underlying density is high. This is

referred to as the "empty space phenomenon" by Scott and Thompson (1983). Silverman (1986) examines the sample sizes required to ensure that the relative mean squared error, $E(\hat{f}(\underline{x})-f(\underline{x}))^2/f(\underline{x})^2$, is less than 0.1 when estimating a standard multivariate normal density at $\underline{x} = \underline{0}$ using a fixed kernel estimator with window width chosen to minimise the mean squared error (MSE). The table of results he includes shows that it increases very rapidly as the dimension increases. For less well behaved distributions and for points in the tails the sizes would probably be much greater still.

In view of this discussion it might be hoped that adaptive methods may help in trying to overcome these extra difficulties. The constant h in (2.1.5) will then be scaled by $1/\tilde{f}(\underline{x})^{1/p}$ and $1/\tilde{f}(\underline{x}_i)^{1/p}$ in the adaptive approaches.

There are many theoretical papers on the fixed kernel method but relatively few on adaptive methods. Moore and Yackel (1977) obtain results on weak and strong consistency, pointwise and uniform, for nearest neighbour estimators. Mack and Rosenblatt (1979) calculate asymptotic expressions for the mean and variance of the nearest neighbour estimators using Taylor series, but it is clear their expression for the mean will be inappropriate in the tails because it tends to infinity as $f(x) \longrightarrow 0$. Hall (1983) considers nearest neighbour estimators with smoothing parameters chosen as a function of nearest neighbour distances which he shows overcome some of the asymptotic problems encountered by Mack and Rosenblatt (1979). Abramson (1982) also used Taylor series to obtain asymptotic expressions for mean and variance of the estimator with bandwidth proportional to $f(x_i)^{-1/2}$ but again that for the mean is inappropriate in the tails. Devroye (1985) proves the weak convergence to 0 of

$\int |\hat{f}-f|$ for all f where $\hat{f}$ is the estimator of Breiman et al (1977).

In Section 2 the means and variances of adaptive estimators are calculated in simple cases by numerical integration and the adequacy of Taylor series approximations are explored.

The asymptotic expressions for the means and variances of adaptive estimators derived in the literature do not permit the construction of a useful strategy for choosing the degree of overall smoothing to be applied to the data. In Section 3 numerical integration is again used for the purpose of comparing the performance of the adaptive methods to that of the fixed when estimating a variety of known shapes of underlying density and the results for the normal distribution are used to construct a specific smoothing strategy.

2.2. Means and Variances of Adaptive Estimators.

The exact mean and variance of the fixed kernel estimator (2.1.1) are:

$$E(\hat{f}(x)) = E\left[\frac{1}{n}\sum K((x-X_i)/h)\right] = \int \frac{1}{h} K((x-y)/h)f(y)dy \qquad (2.2.1)$$

since the (X_i) are independently and identically distributed.

$$V(\hat{f}(x)) = \frac{1}{n} V\left[\frac{1}{h} K(x-X_i)/h\right]$$

$$= \frac{1}{n}\left[\int \frac{1}{h^2} K((x-y)/h)^2 f(y)dy - \left\{\int \frac{1}{h} K((x-y)/h)f(y)dy\right\}^2\right] \qquad (2.2.2)$$

Rosenblatt and Parzen (1962) made the change of variable $t = (x-y)/h$ and used Taylor series to obtain the following asymptotic expressions:

$$E(\hat{f}(x)) = f(x) + 1/2.h^2.f^{(2)}(x).\int t^2K(t)dt + o(h^2), \qquad (2.2.3)$$

where $f^{(2)}(x)$ denotes the second derivative of f evaluated at x and

$$V(\hat{f}(x)) = (nh)^{-1}.f(x).\int K(t)^2dt + o((nh)^{-1}). \qquad (2.2.4)$$

These asymptotic expressions have been found to generally give good guidance as to the behaviour of the estimator in a finite sample situation - this will be illustrated with data from a standard normal distribution later in this section. As a consequence they are also useful for deriving optimal smoothing parameters by assuming that f is of a particular form. Silverman (1986) provides a full discussion of this strategy. Similar expressions for multivariate data were derived by Cacoullos (1966).

In the study of the simple properties of adaptive estimators smoothing parameters of the form

$$h/f(X_i)^{\alpha} \quad , \quad 0 < \alpha < 1$$

will be investigated. In particular the values $\alpha = 0, 1/2$ and $1/p$ representing fixed kernels and the adaptive methods of Abramson and Breiman et al will be considered. Silverman (1986) discusses choice of α and shows that when $\alpha = 1/p$ each kernel approximately "catches" the same number of observations in all parts of the density. Because of integration and other problems associated with the nearest neighbour approach only those adaptive methods whose smoothing parameters involve $f(X_i)$, as opposed to $f(x)$, will be considered further.

The true value of the density f is to be used in studying these smoothing parameters. This is of course unrealistic because in practice f is unknown and a pilot estimate such as a fixed kernel estimate is used. An analysis based on using f is then ignoring the extra variability which will be incurred when estimating the smoothing parameter and so comparisons with $\alpha = 0$ should be favourable to those

methods with $\alpha > 0$. However, a study of the properties of adaptive methods using the true f in the smoothing parameter is instructive as it indicates the heights to which a method which uses a pilot estimate might aspire.

The notation $\hat{f}_\alpha(x)$ will be used to refer to the estimator

$$\hat{f}_\alpha(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)^\alpha}{h} .K((x-X_i).f(X_i)^\alpha/h) \qquad (2.2.5)$$

The exact mean and variance can be expressed in integral forms as:

$$E(\hat{f}_\alpha(x)) = \int \frac{f(y)^\alpha}{h}.K((x-y).f(y)^\alpha/h)f(y)dy \qquad (2.2.6)$$

$$V(\hat{f}_\alpha(x)) = \frac{1}{n}\left[\int \frac{f(y)^{2\alpha}}{h^2} K((x-y).f(y)^\alpha/h)^2.f(y)dy - \left\{E(\hat{f}_\alpha(x))\right\}^2\right] \qquad (2.2.7)$$

If we make a change of variable to $t = (x-y)f(x)^\alpha/h$ and use Taylor series approximations then:

$$V(\hat{f}_\alpha(x)) = \frac{f(x)^{\alpha+1}}{nh} \int K(t)^2dt + o((nh)^{-1}) \qquad (2.2.8)$$

$$E(\hat{f}_1(x)) = f(x) + h^2 \frac{\{f^{(1)}(x)^2 - 1/2.f(x).f^{(2)}(x)\}}{f(x)^3} \int t^2K(t)dt + o(h^2) \qquad (2.2.9)$$

$$E(\hat{f}_{1/2}(x)) = f(x) + \frac{h^4.A(t)}{24.f(t)} \int t^4K(t)dt + o(h^4) \qquad (2.2.10)$$

where,

$$A(t) = \frac{-f^{(4)}(t)}{f(t)} + \frac{8.f^{(3)}(t).f^{(1)}(t)}{f(t)^2} + \frac{6.(f^{(2)}(t))^2}{f(t)^2} - \frac{36.f^{(2)}(t).f^{(1)}(t)^2}{f(t)^3} + \frac{24.(f^{(1)}(t))^4}{f(t)^4} \qquad (2.2.11)$$

That choosing $\alpha = 1/2$ reduces the asymptotic bias to $o(h^4)$ was first shown by Abramson (1982) who also demonstrated that no other

value of α will give this result and that the same results hold in the multivariate case. The term $A(t)$ in the coefficient of the h^4 term in (2.2.10) given by (2.2.11) was derived by Silverman (1986) using a computer algebraic manipulation package.

If f is assumed to be the standard normal density then these expressions simplify to :

$$V(\hat{f}_\alpha(x)) = \frac{f(x)^{\alpha+1}}{nh} \int K(t)^2 dt + o((nh)^{-1}) \qquad (2.2.12)$$

$$E(\hat{f}_1(x)) = f(x) + \frac{h^2}{2f(x)} .(x^2+1) + o(h^2) \qquad (2.2.13)$$

$$E(\hat{f}_{1/2}(x)) = f(x) + \frac{h^4}{8.f(x)} .(x^4+6.x^2+3) + o(h^4) \qquad (2.2.14)$$

These expressions for the mean indicate that when $f(x) \longrightarrow 0$, for example in the tails of the density, the mean and hence the bias will tend to infinity. It is of course unreasonable to expect asymptotic expressions to provide good approximations to finite sample behaviour in regions where the density is low because there will be a scarcity of data. However, the extreme nature of these results also raises the question of how good these approximations might be in regions where the density is high. To further examine this question numerical integration will be used to obtain exact results to compare with the asymptotic expressions. The composite trapezoidal rule (Burden et al (1981)) was used for these and subsequent one dimensional integrals in this chapter with the number of subintervals chosen to ensure a relative accuracy of at least 10^{-3}. The following discussion refers to the estimation of a standard normal density.

Figure 2.1 shows in the case $\alpha = 0$ that the exact bias is well approximated by the asymptotic bias over the whole range of x while

the exact standard deviation is well approximated by the asymptotic expression for $|x| > 2$ but over estimated for x in $(-2,2)$. For $\alpha = 1/2$ and 1 the asymptotic approximations of standard deviation are qualitatively similar to when $\alpha = 0$ except for a slightly more marked underestimation in the tails (see figures 2.2 and 2.3). However, for the bias, in both cases the asymptotic estimates are very poor approximations. They are most effective near the mode at $x = 0$ but rapidly deteriorate as $|x|$ becomes larger. In contrast, the exact bias approaches zero in the tails. The flattest exact bias curve is that for $\alpha = 1/2$ (figure 2.2) followed by $\alpha = 1$ (figure 2.3) while $\alpha = 1$ has the lowest standard deviation curve followed by that for $\alpha = 1/2$.

Figures 2.1, 2.2 and 2.3 are based on using an optimal smoothing parameter which minimises the exact mean integrated squared error (see Section 3.1 for more details) and thus balances the integrated squared bias with the integrated variance.

For $n = 50$, when h is reduced below its optimal value the exact bias curves for each method become much closer to the x-axis as expected. For $\alpha = 0$ the asymptotic bias curves provide good approximations over all x. However, for $\alpha = 1/2$ and 1 they become better approximations over a broader range of x as h decreases, this being more marked for $\alpha = 1/2$ than for $\alpha = 1$, but are still inadequate in the tails. Also, as h decreases the standard deviation curves, both exact and asymptotic, become more peaked. If, instead, h is increased above its optimal value then the peaks and troughs in the exact bias curves for each method become much higher. For $\alpha = 0$ the asymptotic approximations are reasonable for a large range of h but the quality of those for $\alpha = 1/2$ and 1

rapidly deteriorate. The exact and asymptotic standard deviation curves become much flatter as h increases.

When the sample size n increases then plots of bias and standard deviation at the optimal h values show that for each method the asymptotic approximations become much better as expected. However, even for very large n values the asymptotic biases for $\alpha = 1/2$ and $\alpha = 1$ are still poor approximations in the tails but those for $\alpha = 1/2$ tend to deteriorate at larger values of $|x|$ than for $\alpha = 1$.

To illustrate the effect of varying h, figures 2.4, 2.5 and 2.6 show the exact and asymptotic biases and standard deviations at $x = 0$ over a broad range of h values for $\alpha = 0, 1/2$ and 1 respectively. The decreasing nature of the standard deviation as h increases is imitated by the asymptotic expressions but at a higher level for each value of α. When $\alpha = 0$ the asymptotic bias follows the exact but tends to decrease at a faster rate as h increases. In contrast, for $\alpha = 1/2$ and 1, the asymptotic expressions only provide reasonable approximations for small h until near where the exact curve redescends - this turning point feature of the exact curve is not captured at all by the asymptoptic expressions.

Similar plots were obtained but are not included for a range of x-values between 0 and 2. For $\alpha = 0$ the correspondences between the exact and asymptotic values of bias and standard deviation were generally good for each x considered. As h increases the bias approximation worsens while that for the standard deviation improves with the opposite effect as h decreases, as might be expected. However, this was not very marked except when h is large for the bias and very small for the standard deviation. For $\alpha = 1/2$ and

1 the bias approximations get much worse for all h as x increases. The asymptotic bias curves for these methods always increase from zero for each x with the rate of increase becoming higher as |x| gets larger due to the presence of f(x) in the denominators (see expressions (2.2.13) and (2.2.14)) and therefore they never capture the turning points characteristic of the exact curves. In contrast, the standard deviation approximations were found to be quite reasonable.

To examine the nature of the Taylor Series approximations in the integrals for the expectation of density estimates with $\alpha = 0$ and 1 plots were obtained of the exact functions together with their estimates at $x = 0$.

For $\alpha = 0$ figure 2.7 shows the function f(x-ht) with $h = 0.5$, which is an N(0,4) density, along with its quadratic approximation

$$f(x) - h.t.f^{(1)}(x) + 1/2.h^2.t^2.f^{(2)}(x). \qquad (2.2.15)$$

The true function is estimated accurately about zero. In the exact integrand f(x-h.t) is multiplied by the kernel function K(t), an N(0,1) density function, and similarly when using the Taylor Series approximation, (2.2.15) is also multiplied by K(t). Figure 2.8 shows the estimated integrand to be a very good approximation to the exact one. For values of $h < 0.5$ plots of the exact and estimated integrands show them to be almost indistinguishable. As h increases from 0.5 the exact integrand is well approximated in the centre but negative sidelobes start to appear in the tails which become more pronounced with larger h. Figure 2.9 illustrates this effect for $h = 2$. These negative parts occur because the quadratic approximation to f(x-h.t) becomes narrower and more pointed as h increases and it

is not until $|t|$ becomes large that multiplying by $K(t)$ can bring it back to zero. Evaluating the integral using the estimated integrand then results in a value less than the exact one which eventually becomes negative as h increases and hence produces the increasing disparity shown in figure 2.4.

For $\alpha = 1$ at $x = 0$ the function $f(x-ht/f(x))$ is a $N(0,(f(0)/h)^2)$ density and is well approximated by the quadratic expansion

$$f(x) - \frac{h.t}{f(x)} \cdot f^{(2)}(x) + \frac{1}{2} \frac{h^2}{f(x)^2} t^2 f^{(2)}(x) \qquad (2.2.16)$$

in the centre. As $h \longrightarrow 0$ the approximation is good for an increasingly broad range of t while as $h \longrightarrow \infty$ the range becomes much narrower.

Figures 2.10, 2.11 and 2.12 show the exact integrand and its approximation with the function f and the kernel K expanded as Taylor Series up to terms in h^2 but without performing any multiplications. When $h = 0.1$ the two are very close while for $h = 0.3$ the approximation is good in the centre but a large peak has appeared in each tail of the estimate. For $h = 0.5$ this effect is more pronounced and in fact these peaks in the tails rapidly increase in height as h gets larger. This then explains why the estimated value of $E[\hat{f}_1(0)]$ just becomes increasingly large and is only a good approximation for small h as illustrated in figure 2.6. The reason for this is that f also occurs in the function K so that when Taylor Series expansions are substituted into

$$\frac{f(x-ht/f(x))}{f(x)} \cdot K\left[\frac{t}{f(x)} \cdot f(x-ht/f(x))\right] \qquad (2.2.17)$$

the result is a function with large negative troughs in the tails.

This is then multiplied by the quadratic estimate of $f(x-ht/f(x))$ which gives rise to the large positive peaks.

2.3. Optimal Smoothing for Adaptive Estimators of Univariate Densities.

The most common measure of the global accuracy of $\hat{f}$ as an estimate of f is the mean integrated squared error (MISE). This was first used by Rosenblatt (1956) and is defined as

$$E \int (\hat{f}(x)-f(x))^2 dx = \int \text{bias}^2(\hat{f}(x))dx + \int V(\hat{f}(x))dx \qquad (2.3.1)$$

For reasons discussed in Section 2.2, when $\alpha = 0$, the true MISE can be estimated well by substituting asymptotic expressions for the bias and variance instead of the true values. Silverman (1986, Section 3.4) shows that if f is a normal density then the value of h which minimises the MISE is

$$h_{opt} = (4/3)^{1/5} \sigma n^{-1/5} \qquad (2.3.2)$$

Such a value ensures that, asymptotically, the squared bias and variance converge to zero at the same rate and that the MISE converges to zero at rate $n^{-4/5}$. Fryer (1976) considered the special case when the true density is normal and the kernel is a standard normal density so that (2.3.1) can be evaluated exactly and hence minimised over h. It turns out that the results obtained are very similar to those given by (2.3.2). However, for the adaptive methods exact calculations of mean and variance need to be made. Evaluation of (2.3.1) therefore requires two levels of numerical integration. Once evaluated, the MISE can be plotted as a function of h and the optimal value for a specific sample size determined. This can be carried out for different distributions and sample sizes.

Firstly, optimal smoothing will be carried out for the standard

normal distribution. Bowman (1985) showed in an extensive simulation study that normal optimal smoothing is effective for recovering the shapes of a wide range of densities when fixed kernels are used. It is hoped that this will also carry over to adaptive smoothing. Using normality as a criteria for smoothing parameter choice is also motivated by the goodness-of-fit problem to be described in chapter 5.

The MISE (2.3.1) was calculated for a number of sample sizes between 25 and 6400 when the smoothing parameters $h/f(x_i)^{\alpha}$ with $\alpha = 0$, $1/2$ and 1 were applied to univariate standard normal data to find optimal values of h. For $\alpha = 1$ there is a sudden shift in these values near $n = 400$. This occurs as a result of the local minimum in the integrated squared bias (ISB) curve as illustrated in figure 2.13. For a given value of h integrated variance decreases as a function of n whereas integrated squared bias does not. As the sample size increases there comes a point at which the MISE is minimised further by switching from a value of h near the local minimum of the ISB curve to a value nearer zero. In fact when $n = 383$ the MISE takes the same value, 0.00420, for $h = 0.2045$ and $h = 0.0669$. This feature of the ISB curve is not present when $\alpha = 0$ or $1/2$.

Approximation formulae for the optimal smoothing parameters, obtained by regressing $\log(h_{opt})$ on $\log(n)$, are given by :

$$\begin{aligned}
&\text{Fixed } (\alpha = 0) && : 1.198.n^{-0.214} && , \quad 25 < n < 6400 \\
&\text{Adaptive } (\alpha = 1/2) && : 0.896.n^{-0.235} && , \quad 25 < n < 6400 \\
&\text{Adaptive } (\alpha = 1) && \begin{cases} : 0.260.n^{-0.042} & , \quad 25 < n < 400 \\ : 0.580.n^{-0.364} & , \quad 400 < n < 6400 \end{cases}
\end{aligned} \qquad (2.3.3)$$

The asymptotic expressions for bias and variance can be combined to

give an expression for the MSE which can then be minimised algebraically with respect to h. The results from this suggest rates proportional to $n^{-1/5}$ for $\alpha = 0$ and 1 and $n^{-1/9}$ for $\alpha = 1/2$. Clearly these are only appropriate for the case $\alpha = 0$ despite the large sample sizes considered. The asymptotic suggestion that when $\alpha = 1/2$ the bias will be lower is supported though by the plot of the exact bias at h_{opt} for normal data (figure 2.2) which is flatter than for $\alpha = 0$ and 1 (figures 2.1 and 2.3).

The minimised MISE's are plotted in figure 2.14 where it is clear that $\alpha = 1/2$ does indeed produce the best performance at all sample sizes considered. The kink in the $\alpha = 1$ curve occurs at the sample size $n = 383$ after which MISE is further minimised by switching to smaller h values.

This process of finding optimal smoothing parameters was repeated for two other distributions, namely a Gamma$(2, \sqrt{2})$ representing skewness and the bimodal mixture $0.5.N(0.866, 0.5^2) + 0.5.N(-0.866, 0.5^2)$. The parameters of both distributions have been chosen to give unit variance so that direct comparisons with the standard normal results may be made.

The results for the $Ga(2, \sqrt{2})$ distribution are plotted in figure 2.15. This shows that $\alpha = 1/2$ performed best at virtually all sample sizes considered. $\alpha = 0$ produced a poor performance for $n < 400$ but is virtually indistinguishable from $\alpha = 1/2$ thereafter. $\alpha = 1$ is slightly inferior to $\alpha = 1/2$ even for large sample sizes. The minimised MISE's for each method and at each sample size are all higher than the corresponding results for the standard normal. For $\alpha = 0$ and 1/2 the h_{opt} values are proportional to $n^{-0.29}$ and $n^{-0.34}$ respectively and so again only broadly in line with the

asymptotic suggestion when $\alpha = 0$. When $\alpha = 1$ h_{opt} is again proportional to two different powers of n according to the sample size due to a local minimum in the integrand squared bias curve.

For the bimodal distribution the results are qualitatively similar to those from the Gamma (see figure 2.16). However, larger sample sizes are required for $\alpha = 0$ to approach the performance of $\alpha = 1/2$ and for $\alpha = 1/2$ to be superior to $\alpha = 1$. The minimum MISE's for each method and sample size are all greater than those for the standard normal but less than the corresponding results for the Gamma. This time, when $\alpha = 0$, h_{opt} is proportional to $n^{-0.24}$ and so once more is in line with the asymptotic results whereas for $\alpha = 1/2$ it is $n^{-0.20}$ and for $\alpha = 1$ it has two different values depending on sample size.

In order to assess the effectiveness of applying normal optimal smoothing to non-normal data the MISE's were calculated when the formulae (2.3.3) were applied to data from the Gamma and bimodal distributions. These results are illustrated in figures 2.17 and 2.18. With data from a Gamma distribution the relative performances of the three methods are very similar to the Gamma optimal case. With the bimodal distribution it is difficult to identify a markedly superior performance by any of the three methods when the normal optimal formulae are used.

For the three underlying distributions considered, the best overall performance is achieved by the adaptive method with $\alpha = 1/2$ as implied by the asymptotic theory. The results also indicate that the use of normal optimal smoothing provides a simple and reasonably effective technique. As expected this is least effective with the bimodal distribution. Also, when the sample size is large, there is

little loss in MISE but much gain in computational simplicity by using the fixed kernel method.

From a practical point of view the use of $\alpha = 1$ has the great advantage of producing an estimator whose smoothing parameter h is scale invariant thus removing the need to provide an estimate of scale.

Suppose that the unknown density f is to be estimated from a random sample of size n from X using the adaptive estimator $\hat{f}_\alpha$ with smoothing parameter h. Then the MISE can be calculated from the integrated squared bias and variance based on (2.2.6) and (2.2.7).

If the X_i's are transformed to $Z_i = k X_i$, $i = 1,\ldots,n$ for some constant $k > 0$ then

$$g(z) = \frac{1}{k} f(z/k) \tag{2.3.4}$$

where g is the density function for the random variable $Z = kX$.

If the same smoothing parameter, h, is used in an adaptive estim

$$E(\hat{g}_\alpha(z)) = \int \frac{g(y)^\alpha}{h} K\left[\frac{(z-y)g(y)^\alpha}{h}\right] g(y)dy$$

$$E(\hat{g}_\alpha(z)) = \int \frac{g(y)^\alpha}{h} K\left[\frac{(z-y)g(y)^\alpha}{h}\right] g(y)dy \tag{2.3.5}$$

$$= \int \frac{f(y/k)^\alpha}{k^\alpha h} .K\left[\frac{(z-y)f(y/k)^\alpha}{k^\alpha . h}\right] \frac{f(y/k)}{k} dy \tag{2.3.6}$$

Now let $w = y/k$ so that $dy = kdw$ and

$$E(\hat{g}_\alpha(z)) = \int \frac{f(w)^\alpha}{k^\alpha h} .K\left[\frac{(z-kw)f(w)^\alpha}{k^\alpha h}\right] f(w)dw. \tag{2.3.7}$$

Hence,

$$\int \text{bias}^2(z)\, dz =$$

$$\int \left\{\frac{1}{k^\alpha} \int \frac{f(w)^\alpha}{h} K\left[\frac{(z-kw)f(w)^\alpha}{k^\alpha h}\right] f(w)dw - \frac{f(z/k)}{k}\right\}^2 dz \tag{2.3.8}$$

But $z = kx$ which gives $dz = kdx$ and hence

$$\int \text{bias}^2(z)\, dz =$$

$$\frac{1}{k} \int \left\{ \frac{1}{k^{\alpha-1}} \int \frac{f(w)^{\alpha}}{h} K\left[\frac{(x-w)f(w)}{k^{\alpha-1}h}\right] f(w)dw - f(x) \right\}^2 dx \qquad (2.3.9)$$

$$= \frac{1}{k} \int \text{bias}^2(x)dx \qquad (2.3.10)$$

if and only if $\alpha = 1$. Similarly,

$$E(\hat{g}_{\alpha}(z)^2) = \int \frac{g(y)^{2\alpha}}{h^2} K^2\left[\frac{(z-y)g(y)^{\alpha}}{h}\right] g(y)dy \qquad (2.3.11)$$

$$= \int \frac{f(y/k)^{2\alpha}}{k^{2\alpha}\, h^2} K^2\left[\frac{(z-y)f(y/k)^{\alpha}}{k^{\alpha}h}\right] \frac{f(y/k)}{k}\, dy \qquad (2.3.12)$$

Again let $w = y/k$ which gives:

$$E(\hat{g}_{\alpha}(z)^2) = \int \frac{f(w)^{2\alpha}}{k^{2\alpha}\, h^2} K^2\left[\frac{(z-kw)f(w)^{\alpha}}{k^{\alpha}h}\right] f(w)dw \qquad (2.3.13)$$

Hence,

$$\int V(\hat{g}_{\alpha}(x))dz = \int \left[E(\hat{g}_{\alpha}(z)^2) - \left[E(g_{\alpha}(z))\right]^2\right] dz \qquad (2.3.14)$$

$$\int \left[\int \frac{f(w)^{2\alpha}}{k^{2\alpha}h^2} K^2\left[\frac{(z-kw)f(w)^{\alpha}}{k^{\alpha}h}\right] f(w)\, dw \right.$$

$$\left. - \left[\int \frac{f(w)^{\alpha}}{k^{\alpha}h} K\left[\frac{(z-kw)f(w)^{\alpha}}{k^{\alpha}h}\right] f(w)\, dw\right]^2 dz \right. \qquad (2.3.15)$$

which on substituting $z = kx$, gives

$$\int V(g_{\alpha}(z))dz = \frac{1}{k} \int V(\hat{f}_{\alpha}(x))dx \qquad (2.3.16)$$

if and only if $\alpha = 1$.

Therefore,

$$\text{MISE}(\hat{g}_1(z)) = \frac{1}{k} \text{MISE}\ (\hat{f}_1(x)) \qquad (2.3.17)$$

so that the optimal h for estimating g is the same as that for

estimating f and is thus independent of the scale of the data provided $\alpha = 1$.

The scalings of h for other values of α when estimating a density with non-unit variance are as follows :

	α	Scaling
	0	$\hat{\sigma}$
	1/2	$\hat{\sigma}^{1/2}$
general	α	$\hat{\sigma}^{(1-\alpha)}$

where $\hat{\sigma}$ is an estimate of the standard deviation of the data. In practice it is probably better to use a robust estimate of scale such as the median of the absolute deviations from the median divided by 0.6745 (MAD/0.6745) (Hogg (1979)) rather than the empirical standard deviation.

Silverman (1986, 5.3) suggests the practical solution of multiplying the h derived for data with unit variance by the geometric mean of the $\{\tilde{f}(x_i)\}$, where $\tilde{f}(\cdot)$ is the pilot estimate employed in the smoothing parameter, to free h from the scale of the data.

2.4. Optimal Smoothing for Adaptive Estimators of Multivariate Densities.

In order to assess the performance of the three smoothing techniques (i.e. $\alpha = 0$, 1/2 and 1/p) in higher dimensions a standard p-variate normal distribution $N_p(\underline{0}, I_p)$ and a long tailed p-variate normal mixture $0.219.N_p(\underline{0}, 4I_p) + 0.871.N_p(\underline{0}, 0.16I_p)$ which also has a unit covariance matrix, were considered. I_p denotes the identity matrix of order p. The choice of non-normal distribution is rather limited because of the considerable computational

advantages of retaining radial symmetry as discussed below. Also radial symmetry allows the use of the same smoothing parameter in each co-ordinate direction so that for example when $\alpha = 0$ the kernel function over point $\underline{x}_i$ is a $N_p(\underline{x}_i, hI_p)$ density.

As in the univariate case it is necessary to use numerical integration to evaluate the mean integrated squared errors $\int E\{\hat{f}(\underline{x})-f(\underline{x})\}^2 d\underline{x}$. If the underlying densities are chosen to be radially symmetric then it will follow that the mean squared error will also be radially symmetric as a function of $\underline{x}$. Therefore, if a transformation to polar co-ordinates is made i.e.

$$
\begin{aligned}
x_1 &= r_1.\cos\theta_1.\sin\theta_2.....\sin\theta_{p-1} \\
x_2 &= r_1.\sin\theta_1.\sin\theta_2.....\sin\theta_{p-1} \\
x_3 &= r_1 \quad .\cos\theta_2.\sin\theta_3.....\sin\theta_{p-1} \\
x_4 &= r_1 \quad .\cos\theta_3.\sin\theta_4.....\sin\theta_{p-1} \\
&\vdots \\
x_p &= r_1 \quad \cos\theta_{p-1}
\end{aligned}
\qquad (2.4.1)
$$

where $r_1 > 0$, $\theta_1 \in [0,2\pi]$ and $\theta_j \in [o,\pi]$, $j = 2,.....,p-1$. The Jacobian of this transformation is :

$$
J = \begin{cases} r_1 & , \quad p = 2 \\ r_1^{p-1} \prod_{j=2}^{p-1} \sin^{j-1}\theta_j & , \quad p = 3,4,5,..... \end{cases} \qquad (2.4.2)
$$

The MISE can be written as :

$$
\int E\{\hat{f}(\underline{x})-f(\underline{x})\}^2 d\underline{x} = C_p.\int_0^\infty r_1^{p-1}.E\{\hat{f}(r_1)-f(r_1)\}^2 dr_1
$$

where

$$C_p = 2\pi, \quad n = 2,$$

$$C_p = \int_0^{2\pi} d\theta_1 \cdot \int_0^{\pi} \sin\theta_2 d\theta_2 \cdot \int_0^{\pi} \sin^2\theta_3 d\theta_3 \ldots\ldots \int_0^{\pi} \sin^{p-2}\theta_{p-1} d\theta_{p-1}$$

$$, \; n = 3,4,5,\ldots\ldots \qquad (2.4.4)$$

In fact, C_p = p.volume of a unit sphere = $p \dfrac{\pi^{p/2}}{\Gamma(p/2+1)}$ where $\Gamma(\cdot)$ indicates the Gamma function.

The MISE can be written as:

$$Cp \int_0^{\infty} r_1^{p-1} \, (B^2(r_1)+V(r_1))dr_1 \qquad (2.4.5)$$

where $B(r_1)$ and $V(r_1)$ denote the bias and variance terms at radial distance r_1. The dimensionality of the integral of the MSE has therefore been reduced from p to 1.

Now let $\underline{r}$ denote the vector $(0,0,\ldots\ldots,0,r_1)^T$ so that $B(r_1) = E(\hat{f}(\underline{r}))-f(\underline{r})$. Expression (2.2.6) shows that this p-dimensional integral may be written as

$$B(r_1) = \int b(\|\underline{r}\| \, , \|\underline{y}\|, \|\underline{r}-\underline{y}\|)dy \qquad (2.4.6)$$

for a suitably chosen function b and where $\|\cdot\|$ denotes the Euclidean norm.

Another polar coordinate transformation allows us to write

$$\|\underline{y}\| = r_2 \qquad (2.4.7)$$

and

$$\|\underline{r}-\underline{y}\|^2 = \|\underline{r}\|^2 + \|\underline{y}\|^2 - 2\|\underline{r}\|.\|\underline{y}\|.\cos\theta$$

$$= r_1^2 + r_2^2 - 2r_1.r_2.\cos\theta \qquad (2.4.8)$$

where θ is the angle between vectors $\underline{r}$ and $\underline{y}$.

The p-dimensional integral for the bias has therefore been reduced to only two i.e.

$$B(r_1) = C_{p-1} \int_0^{\infty} r_2^{p-1} \left[\int_0^{\pi} b(r_1, r_2, r_1^2 + r_2^2 - 2.r_1.r_2.\cos\theta).(\sin\theta)^{p-2} d\theta \right] dr_2 \qquad (2.4.9)$$

For example, if both f and the kernel K are p-dimensional standard normal then the integral will depend on $\|\underline{y}\|$ in $f(\underline{y})$ and on $\|\underline{x}-\underline{y}\|$ in $K(\cdot)$. The argument for evaluating the variance term $V(r_1)$ follows that for $B(r_1)$ using the appropriate functions.

This shows that when both the underlying density and kernel function are radially symmetric and are functions of $\|\underline{y}\|$ and $\|\underline{x}-\underline{y}\|$ respectively the original p^2 dimensional integral can be reduced to two two dimensional integrals for the MSE followed by a one dimensional integral to calculate the MISE.

This then makes numerical integration a feasible tool for the calculations involved for multivariate density estimators. As mentioned earlier it does restrict the choice of underlying density though. For example, to construct a long tailed alternative, a p-dimensional density may be formed by the product of p 1-dimensional t-distributions on 3 degrees of freedom scaled to have unit variance. The dimension reduction technique described above cannot then be used because

$$f(\underline{x}) = (2/\pi)^p.(1+x_1^2)^{-2} \ldots\ldots (1+x_p^2)^{-2} \qquad (2.4.10)$$

cannot be expressed as a function of $\|\underline{x}\|$ as in the normal case.

To overcome this problem, consider a random vector $\underline{X}$ with density

$$f(\underline{x}) = a.N_p(\underline{x},\sigma_1^2 .I_p) + (1-a).N_p(\underline{x},\sigma_2^2 .I_p),\ 0 < a < 1. \quad (2.4.11)$$

In this case $E(\underline{X}) = 0$ and $V(\underline{X}) = (a\sigma_1^2 + (1-a)\sigma_2^2).I_p$. To obtain a unit covariance matrix set

$$a\sigma_1^2 + (1-a)\sigma_2^2 = 1. \quad (2.4.12)$$

Hence,

$$a = \frac{1-\sigma_2^2}{\sigma_1^2 -\sigma_2^2} \quad (2.4.13)$$

In order to keep $0 < a < 1$ values of σ_1 and σ_2 within a region bounded by $\sigma_1 = \sigma_2$, $\sigma_2 = 1$, $\sigma_2 = 0$ and $\sigma_1 > \sigma_2$ are allowable. Choosing $\sigma_1 = 2$ and $\sigma_2 = 0.4$ gives a value of $a = 0.219$ and results in a long tailed alternative for which the dimension reduction technique can be used.

The first distribution to be considered is the standard p-dimensional normal distribution. The h-values which minimise MISE were found for each of the three methods at several sample sizes ($25 \leqslant n \leqslant 6400$) for dimensions two to six and also for dimension ten. The two-dimensional integrals were evaluated by NAG subroutine D01FCF (NAG (1988)) which uses an adaptive subdivision strategy. Plots were obtained of the minimum MISE vs log(n) for each of these dimensions and are illustrated in figures 2.19-2.24.

In dimension 2 the two adaptive methods are equivalent and are consistently better than $\alpha = 0$ for each n. In dimension 3 $\alpha = 1/p$ achieves the smallest MISE at each n followed by $\alpha = 1/2$. $\alpha = 1/p$ is also the best in dimension 4 and the difference between $\alpha = 1/2$ and $\alpha = 0$ is greatly reduced especially for large sample sizes ($n > 1600$). When the dimension is 5 $\alpha = 1/p$ again achieves

the smallest MISE's at each n while $\alpha = 1/2$ and $\alpha = 0$ are very similar in performance. In dimension 6 $\alpha = {}^{1}/_{p}$ again performs best but $\alpha = 0$ is better than $\alpha = {}^{1}/_{2}$ at each n. The relative positions are the same in dimension 10 but now $\alpha = 1/2$ performs very much worse than the other two.

To illustrate the bias and standard deviation these were plotted as a function of x_1 only - any cross section through the origin will have the same profile because of radial symmetry. Plots were obtained for each dimension considered previously and for $n = 50$ only. The optimal values of the smoothing parameter were used in each case. See figures 2.25-2.28 for dimensions 3 and 6.

In each dimension the maximum bias occurs at $\underline{x} = \underline{0}$ where the curvature in the density is greatest for $\alpha = 0$ and $1/p$ but at $x = \pm (1,.....,1)^T$ for $\alpha = 1/2$ except in dimension 10 when it is also at $\underline{x} = \underline{0}$. For each dimension and method the bias approaches zero in the tails and the positive part of the bias curves also become much closer to zero as the dimension increases. The flattest bias curve in each dimension is for $\alpha = 1/2$ followed by $\alpha = 1/p$. However by dimension 10 the differences are much less marked.

The standard deviation curves for each method have a mode at $\underline{x} = \underline{0}$ and approach zero in the tails for each dimension. However, in contrast to the bias curves, $\alpha = 0$ has the lowest standard deviation curve in each dimension followed by $\alpha = 1/p$ with that for $\alpha = 1/2$ being most peaked. As dimension increases the relative differences increase, especially between $\alpha = 1/2$ and the other two.

In conclusion, for multivariate standard normal data, by using an adaptive method a reduction in bias is achieved over the fixed but at

the expense of increasing the variance of the estimated density. For $\alpha = 1/p$ this trade-off still results in an optimal MISE smaller than for the other two but the increasingly poor performance in terms of standard deviation for $\alpha = 1/2$ explains why it does increasingly worse in MISE than the other two.

For the long tailed multivariate normal mixture the optimal h-values were determined together with the corresponding MISE's at several sample sizes for dimensions two to six. The plots of minimum MISE vs log(n) are shown in figures 2.29-233. The minimum MISE's are much higher but the overall pattern of results is very similar to the multivariate standard normal case with the adaptive $\alpha = 1/p$ method increasingly achieving smaller MISE's than the other two methods. Also by dimension 6, $\alpha = 0$ is generally doing a little better than $\alpha = 1/2$ at each n. The increasingly poor performance of $\alpha = 1/2$ is again due to the large increase in variance, relative to the other two methods, as dimension increases.

There are strong similarities in the results for these two underlying distributions which indicate that in general the adaptive method with $\alpha = 1/p$ would make a good choice for multivariate data. The asymptotic argument in favour of $\alpha = 1/2$ becomes increasingly less effective as dimensionality increases.

As in the univariate case the normal optimal h-values were used to calculate the MISE when estimating the long tailed density by the $\alpha = 1/p$ method. Results for sample sizes 100 and 1600 are given in table 2.1

Table 2.1. Values of MISE incurred when estimating the long tailed Normal mixture density by the $\alpha = 1/p$ method using Normal optimal h-values.

	MISE (% increase from minimum value)	
Dimension	n = 100	n = 1600
2	0.01282 (1%)	0.00240 (2%)
3	0.01849 (2%)	0.00373 (2%)
4	0.02168 (1%)	0.00557 (1%)
5	0.02281 (0.5%)	0.00724 (2%)
6	0.02191 (1%)	0.00808 (4%)

The increases are generally only by a few percent thus giving support to a strategy of using the normal optimal h-values to smooth non-normal multivariate data. These values can be approximated by the simple formulae given in the following table.

Table 2.2. Approximation formulae for the normal optimal global smoothing parameters when using the adaptive $(\alpha = 1/p)$ method.

Dimension	Optimal h.
2	$0.393.n^{-0.185}$
3	$0.336.n^{-0.143}$
4	$0.331.n^{-0.131}$
5	$0.331.n^{-0.120}$
6	$0.333.n^{-0.112}$

The scale invariance of property of the $\alpha = 1/p$ method as discussed for univariate data also carries over to the multivariate case with vectors replacing single variables in the proof of Section 2.3.

Figure 2.1. Exact and asymptotic bias and standard deviation incurred by the fixed kernel method when estimating an N(0,1) density, n = 50, h = 0.52.

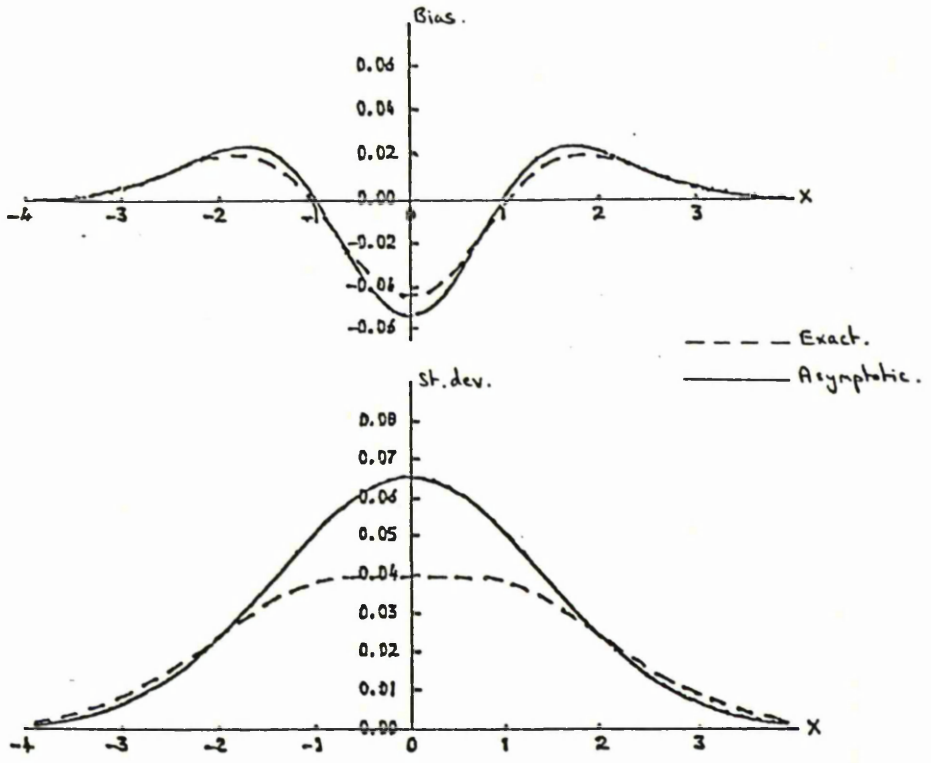


Figure 2.2. Exact and asymptotic bias and standard deviation incurred by the adaptive method with $\alpha = 0.5$ when estimating an N(0,1) density, n = 50, h = 0.351.

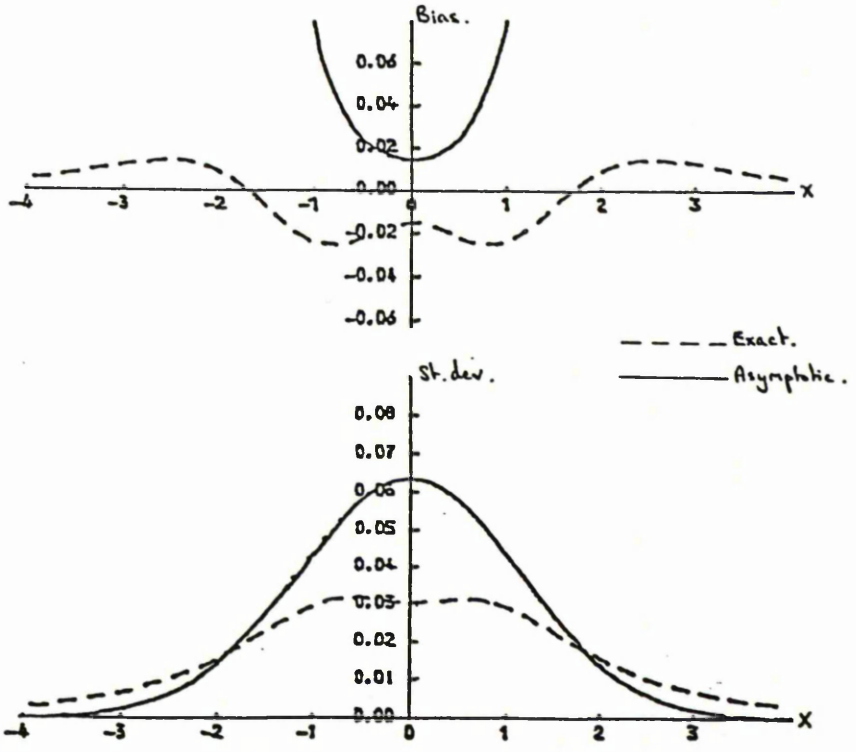


Figure 2.3. Exact and asymptotic bias and standard deviation incurred by the adaptive method with $\alpha = 1$ when estimating an N(0,1) density, $n = 50$, $h = 0.219$.

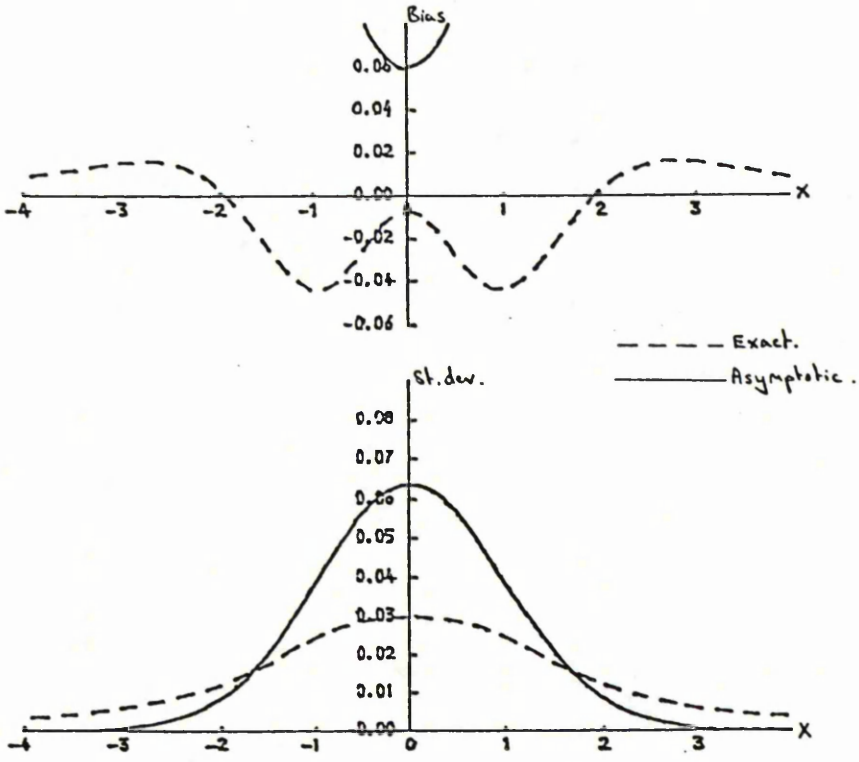


Figure 2.4. Exact and asymptotic bias and standard deviation incurred by the fixed kernel method as a function of h when estimating an N(0,1) density, $x = 0$, $n = 50$.

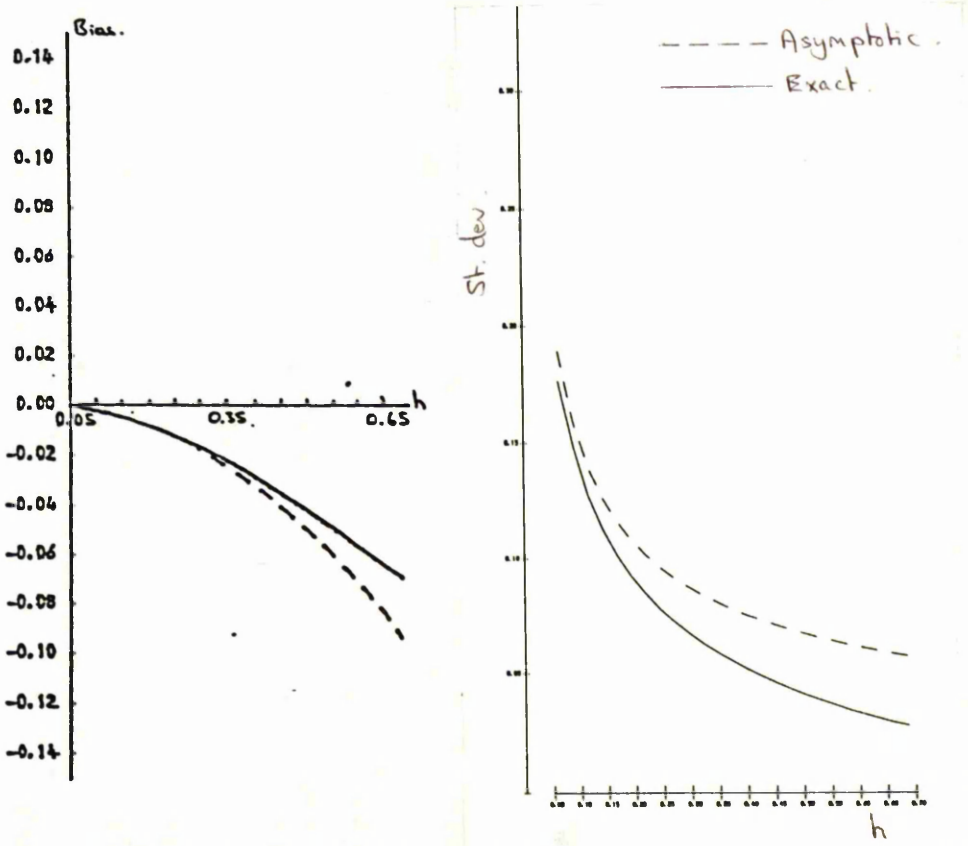


Figure 2.5. Exact and asymptotic bias and standard deviation incurred by the adaptive method with $\alpha = 0.5$ as a function of h when estimating an N(0,1) density, $x = 0$, $n = 0.50$.

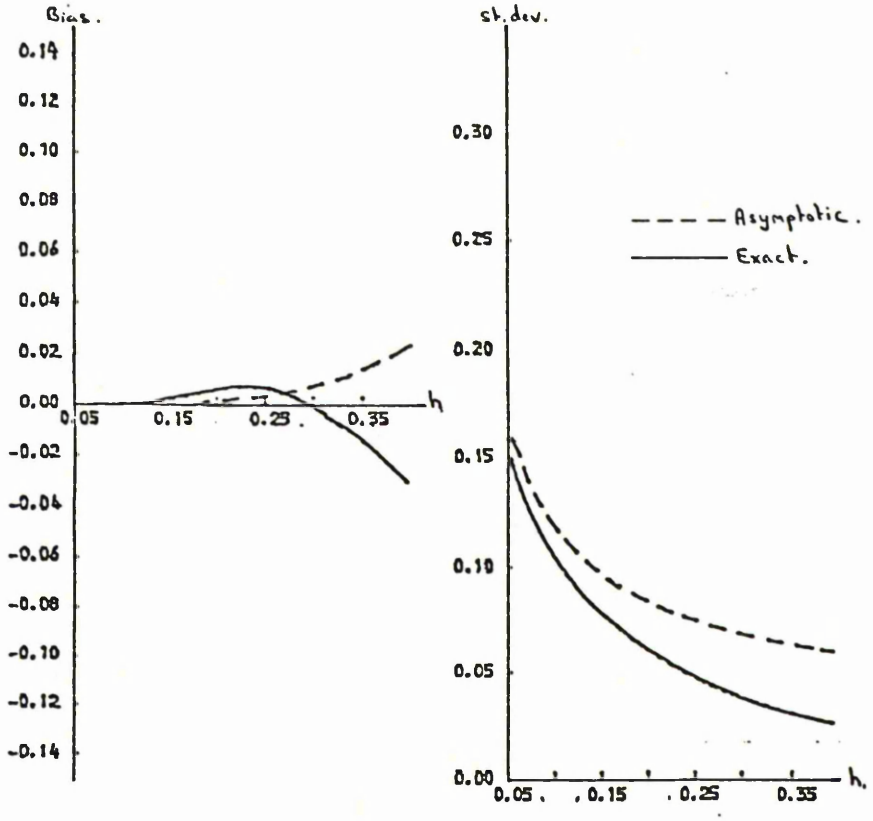


Figure 2.6. Exact and asymptotic bias and standard deviation incurred by the adaptive method with $\alpha = 1$ as a function of h when estimating an N(0,1) density, $x = 0$, $n = 50$.

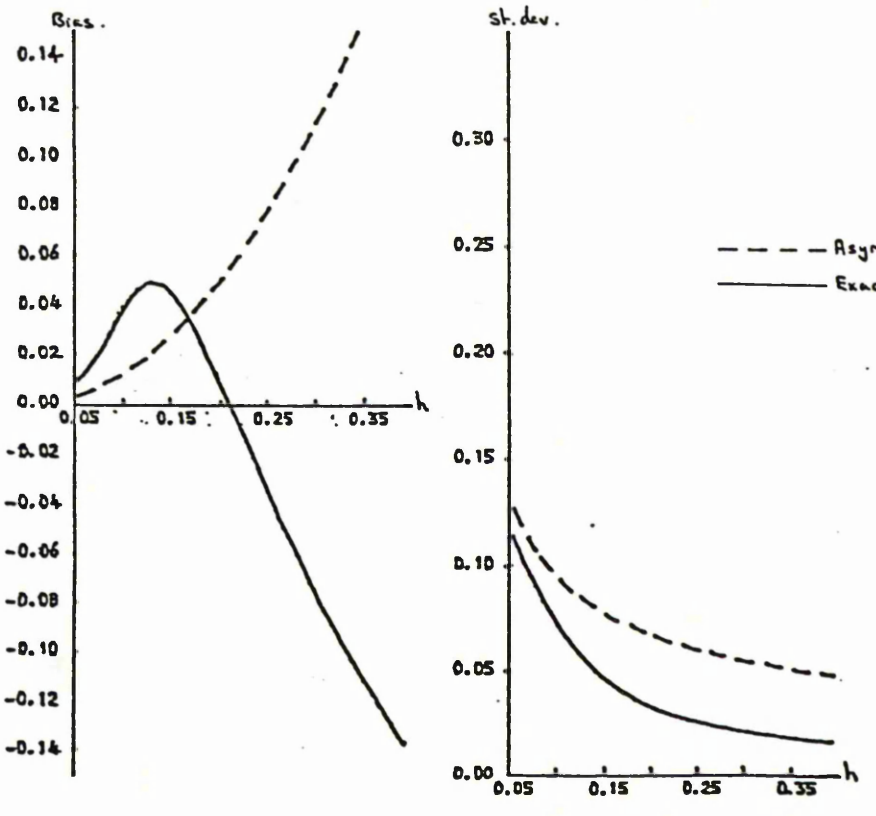


Figure 2.7. f(x-ht) and its quadratic approximation when f is an N(0,1) density function, x = 0, h = 0.5.

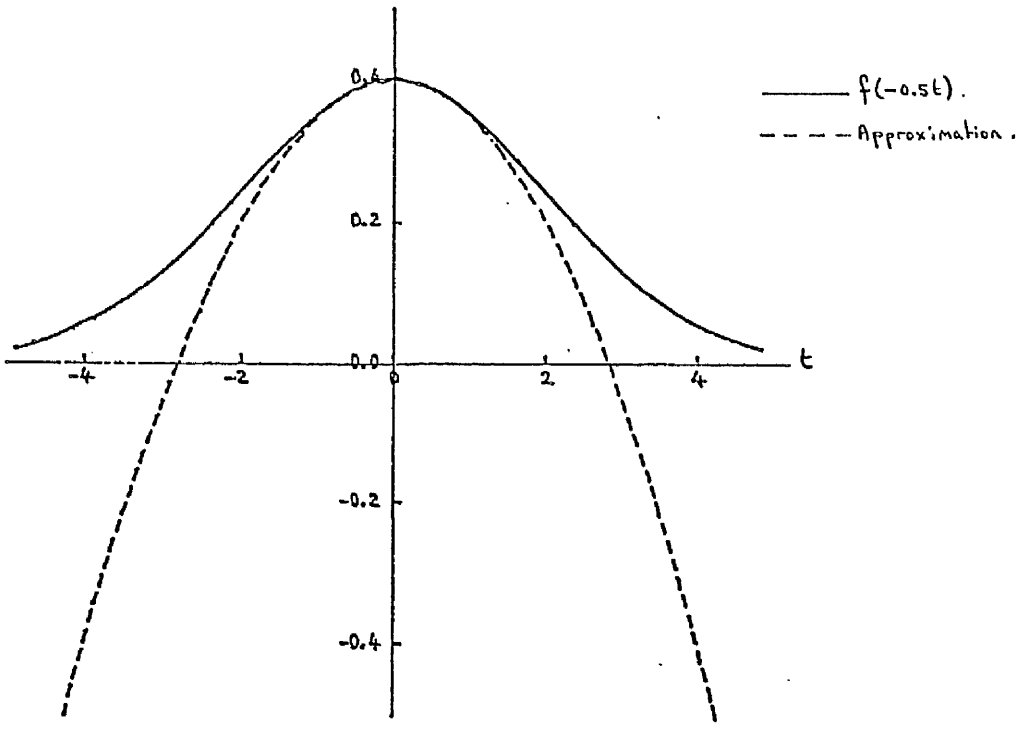


Figure 2.8. f(x-ht).K(t) and its approximation when f and K are N(0,1) density functions, x = 0, h = 0.5.

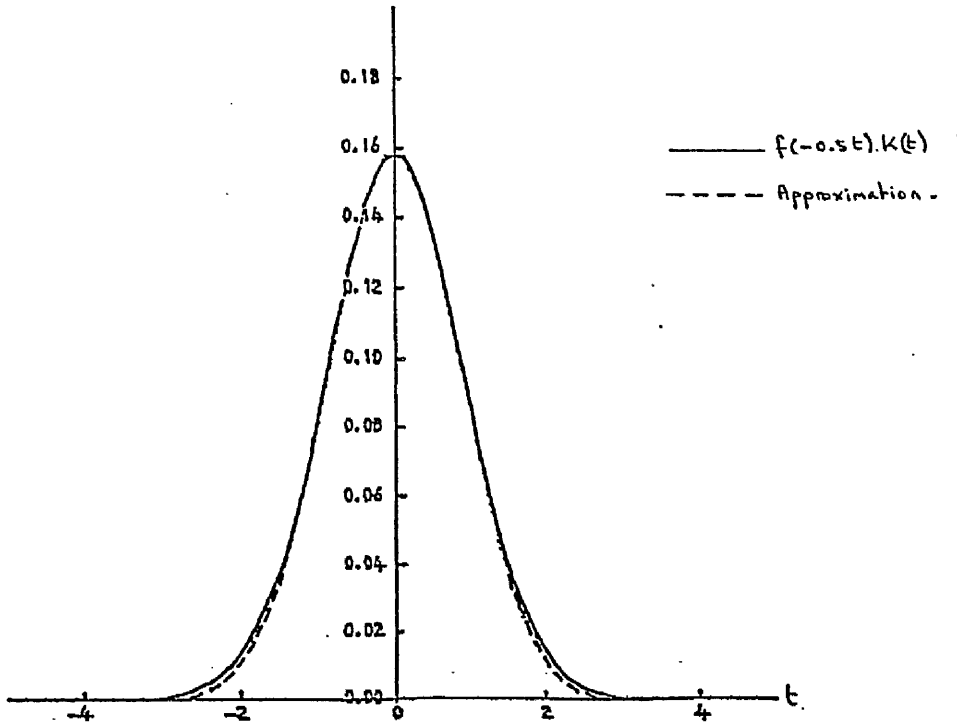


Figure 2.9. f(x-ht).K(t) and its approximation when f and K are N(0,1) density functions, x = 0, h = 2.

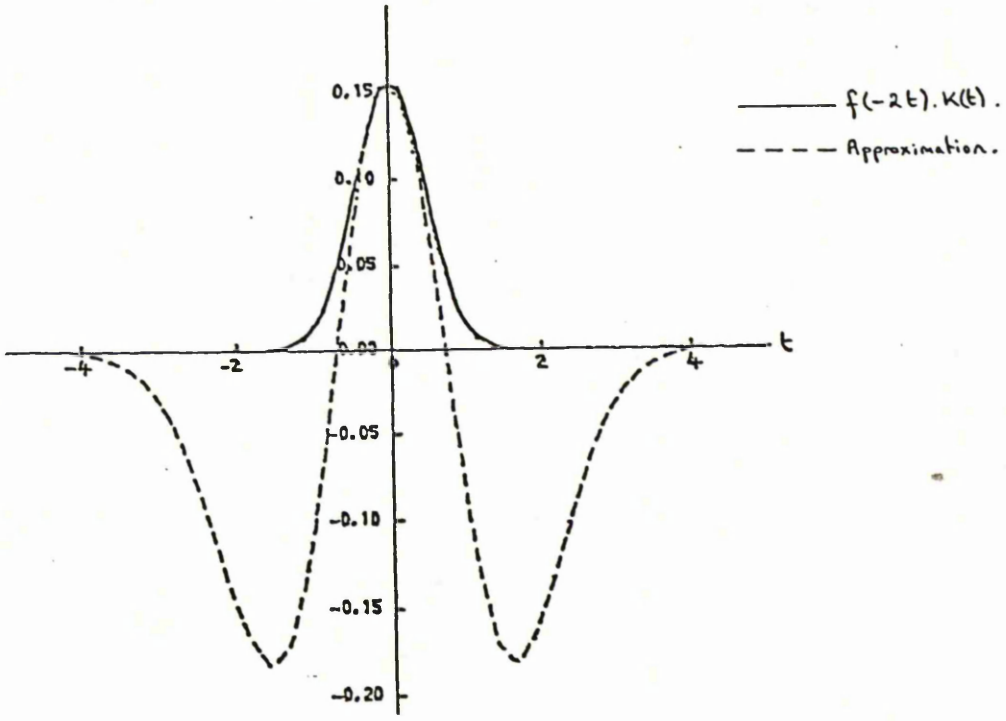


Figure 2.10. $f(x-ht/f(x))^2.K(t.f(x-ht/f(x))/f(x))/f(x)$ and its approximation when f and K are N(0,1) density functions, x = 0, h = 0.1.

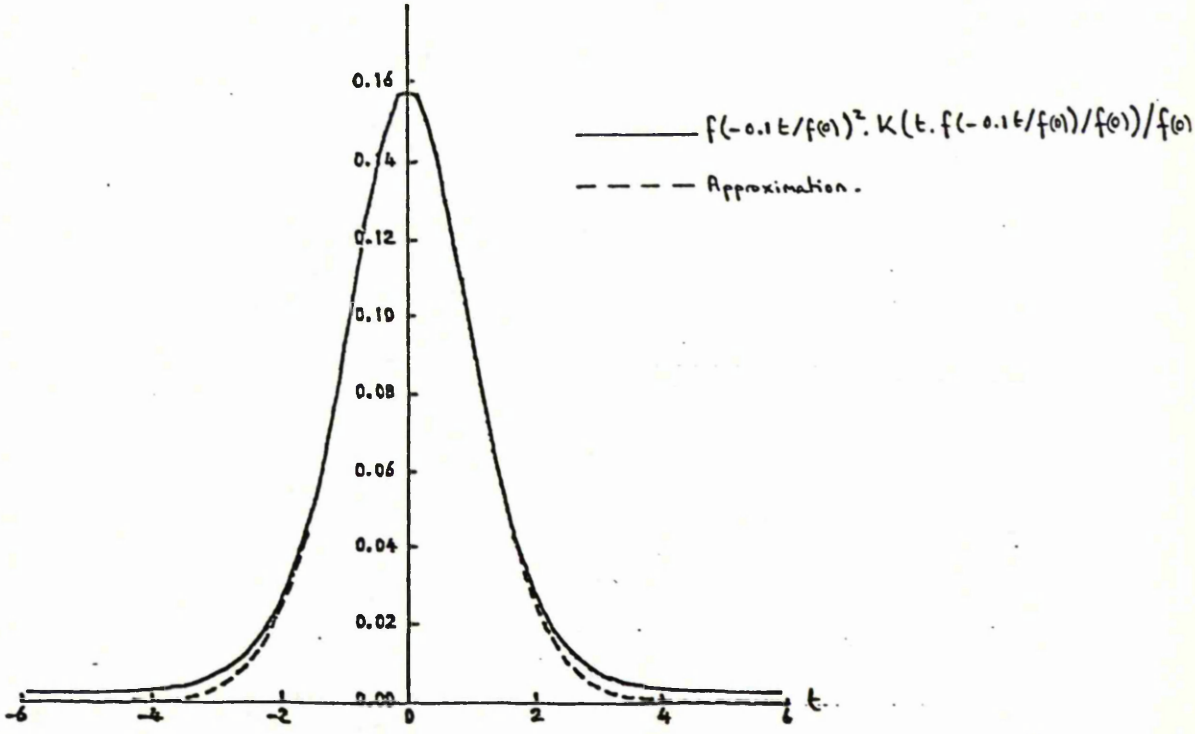


Figure 2.11. $f(x-ht/f(x))^2.K(t.f(x-ht/f(x))/f(x))/f(x)$ and its approximation when f and K are $N(0,1)$ density functions, $x = 0$, $h = 0.3$.

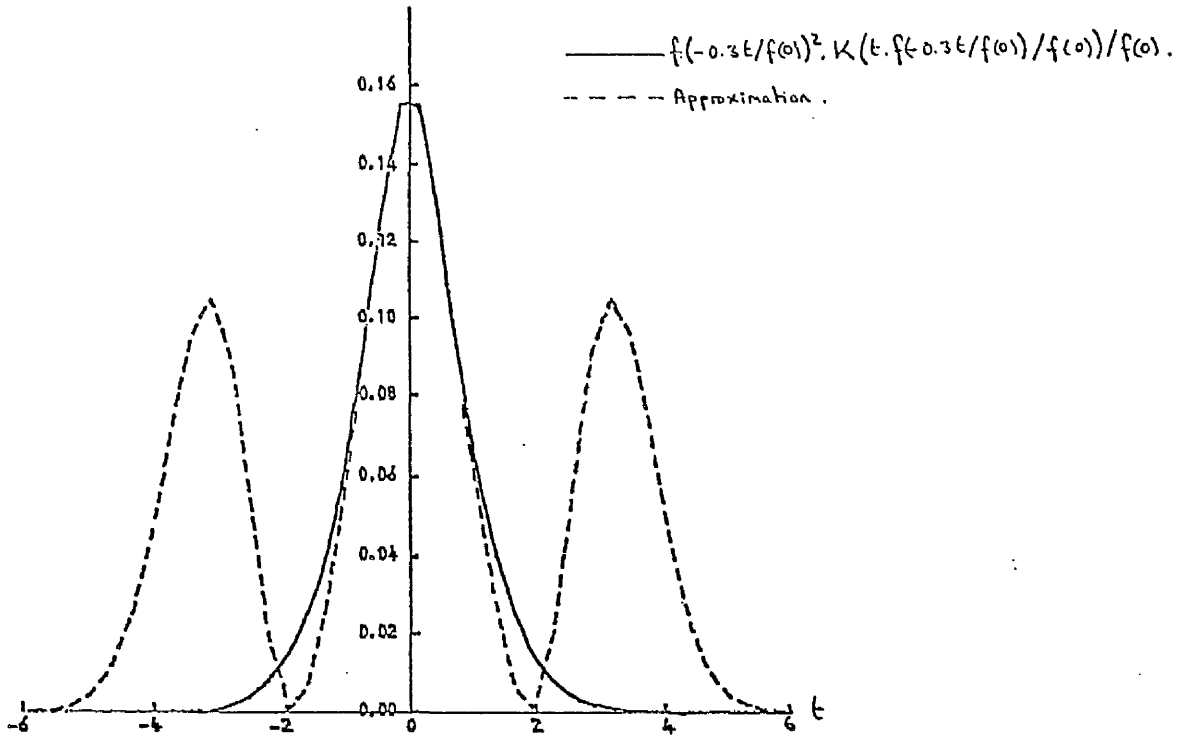


Figure 2.12. $f(x-ht/f(x))^2.K(t.f(x-ht/f(x))/f(x))/f(x)$ and its approximation when f and K are $N(0,1)$ density functions, $x = 0$, $h = 0.5$.

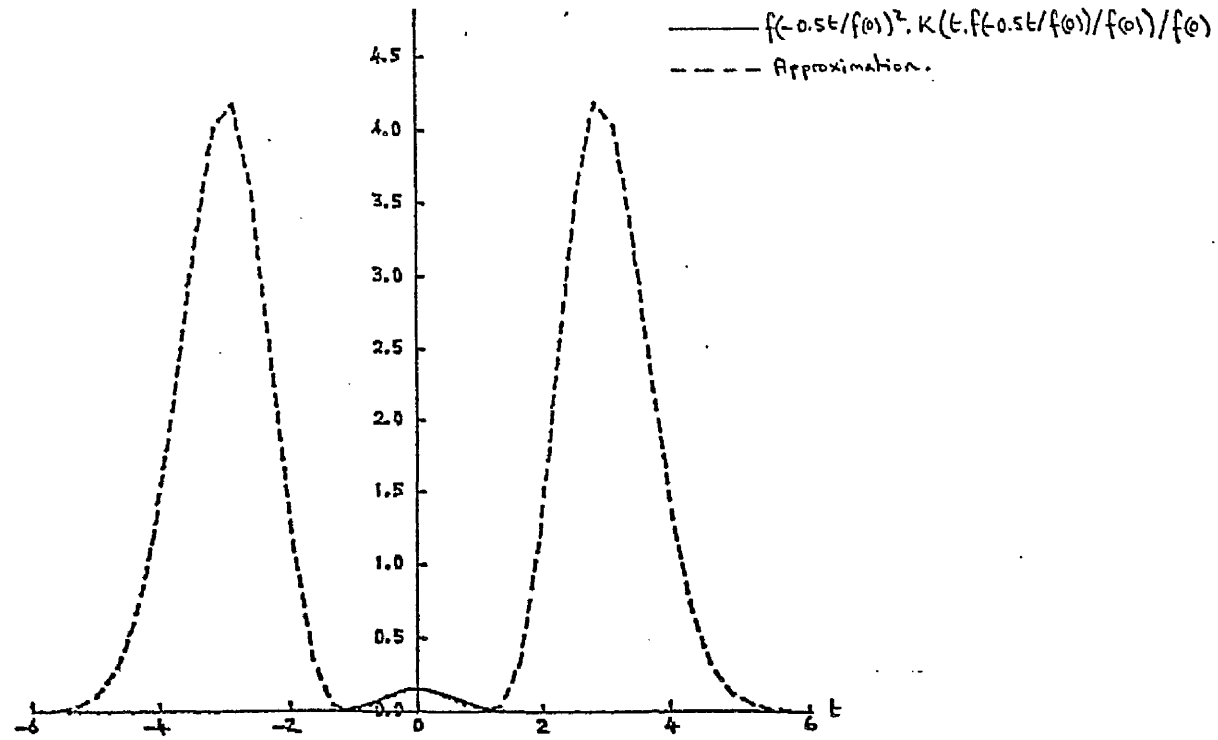


Figure 2.13. Exact integrated squared bias incurred by the adaptive method with $\alpha = 1$ as a function of the smoothing parameter h when estimating an N(0,1) density.

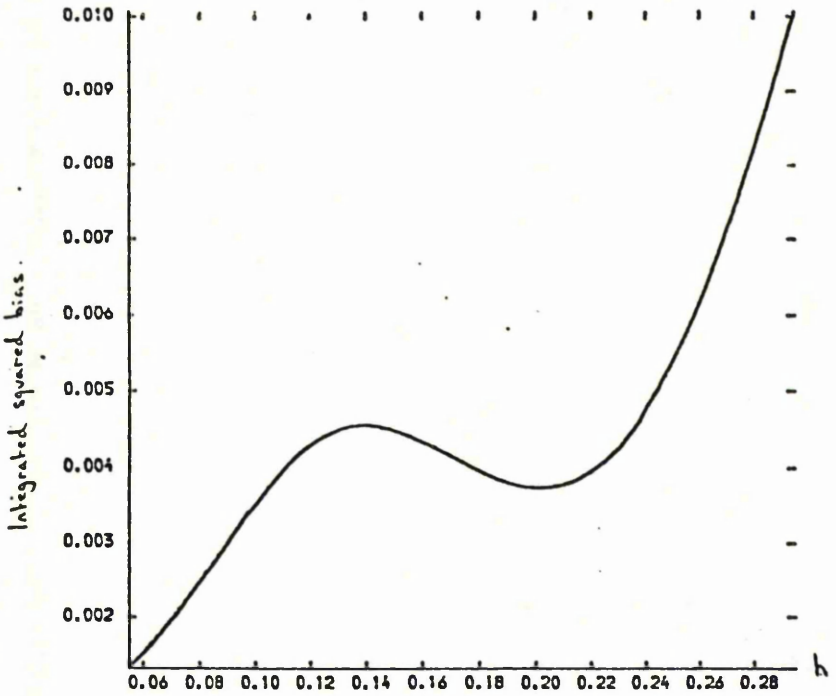


Figure 2.14. Minimum MISE as a function of log(n) for the methods $\alpha = 0$, $1/2$ and 1 when the underlying density is N(0,1).

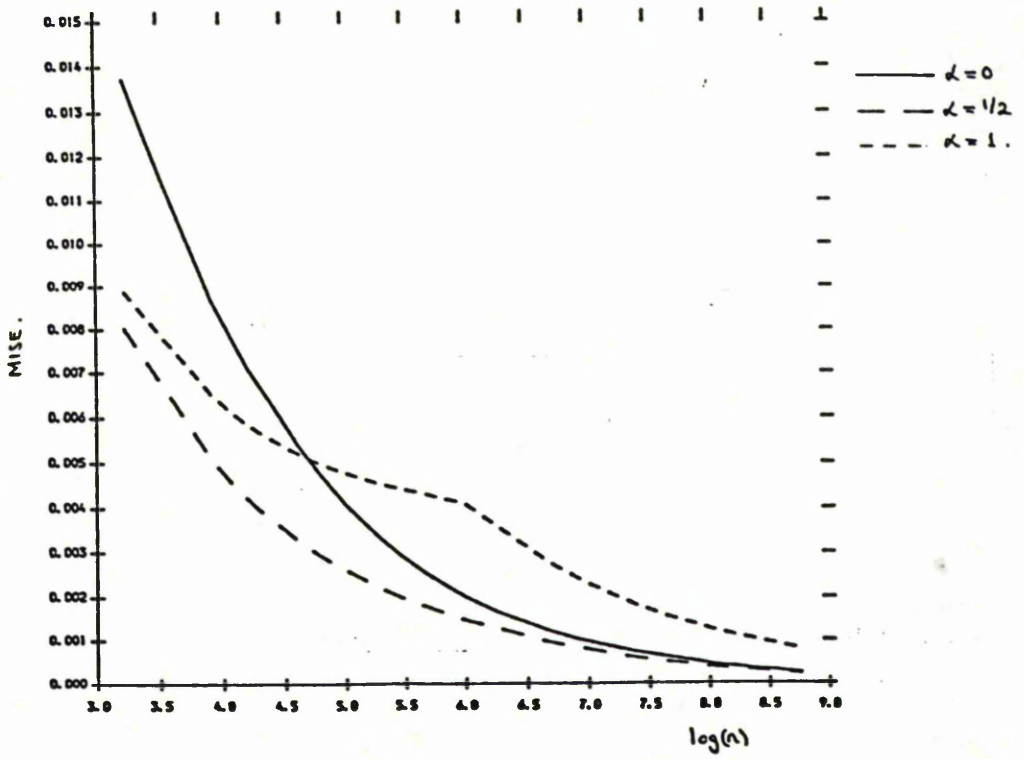


Figure 2.15. Minimum MISE as a function of log(n) for the methods $\alpha = 0, \frac{1}{2}$ and 1 when the underlying density is Gamma $(2, \sqrt{2})$.

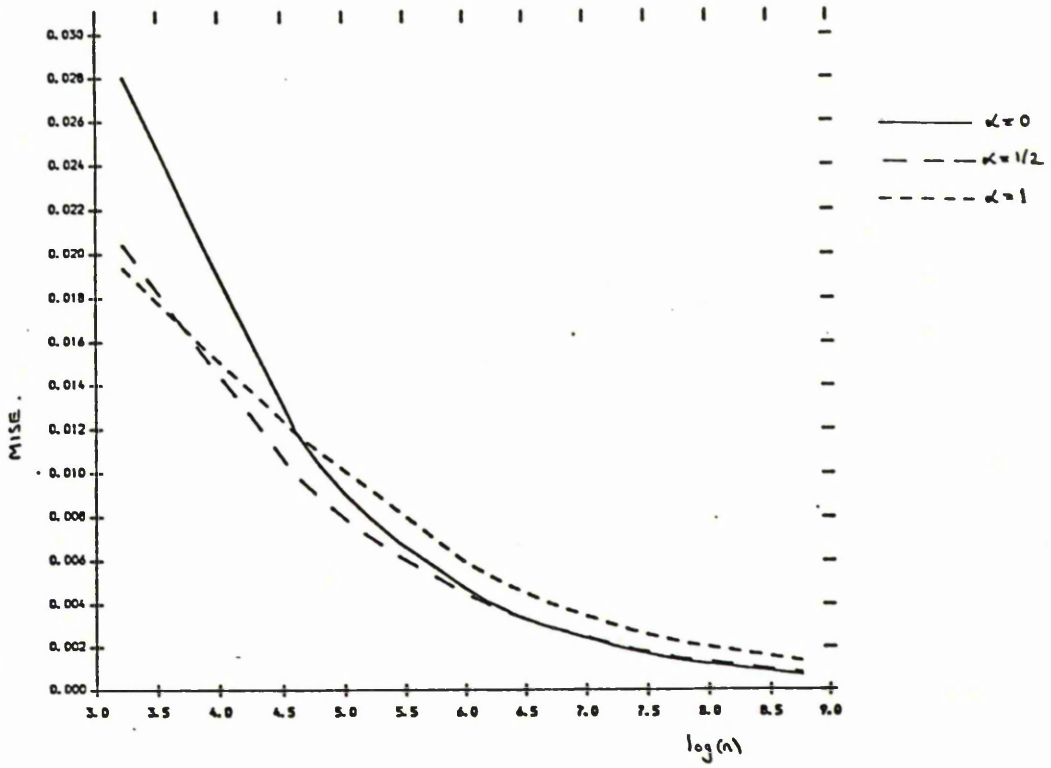


Figure 2.16. Minimum MISE as a function of log(n) for the methods $\alpha = 0, \frac{1}{2}$ and 1 when the underlying density is $0.5N(-0.866, 0.5^2) + 0.5N(0.866, 0.5^2)$.

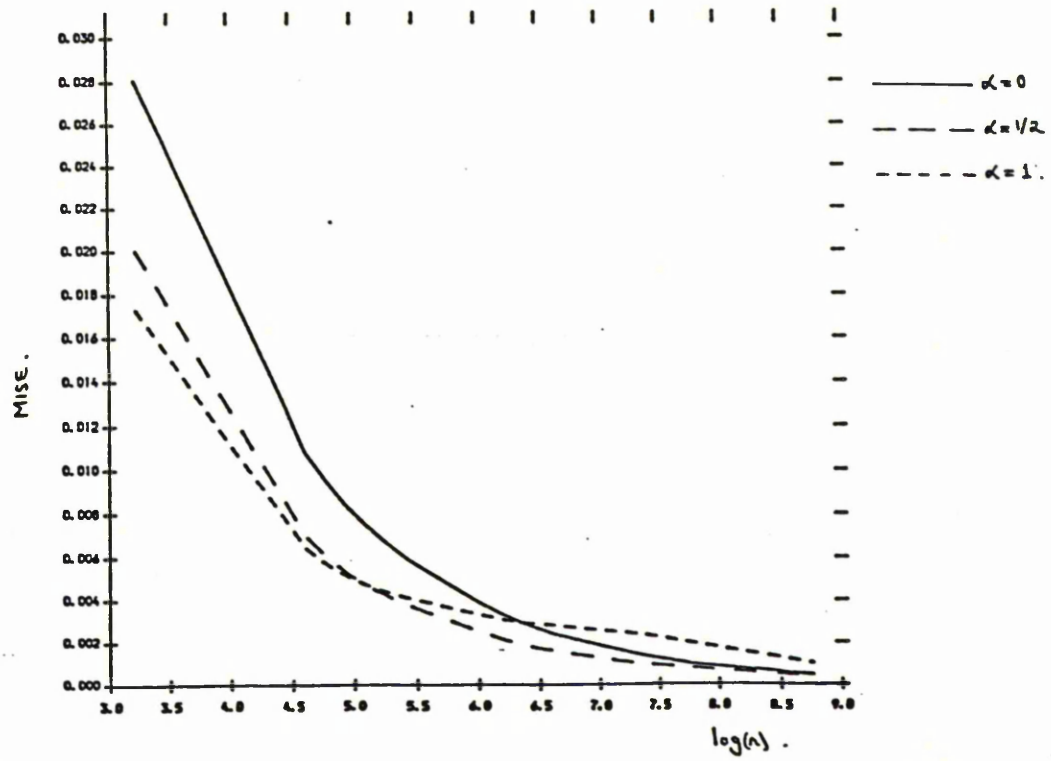


Figure 2.17. The MISE when using normal optimal smoothing parameters, as a function of log(n), for the methods $\alpha = 0, \frac{1}{2}$ and 1 when the underlying density is Gamma $(2, \sqrt{2})$.

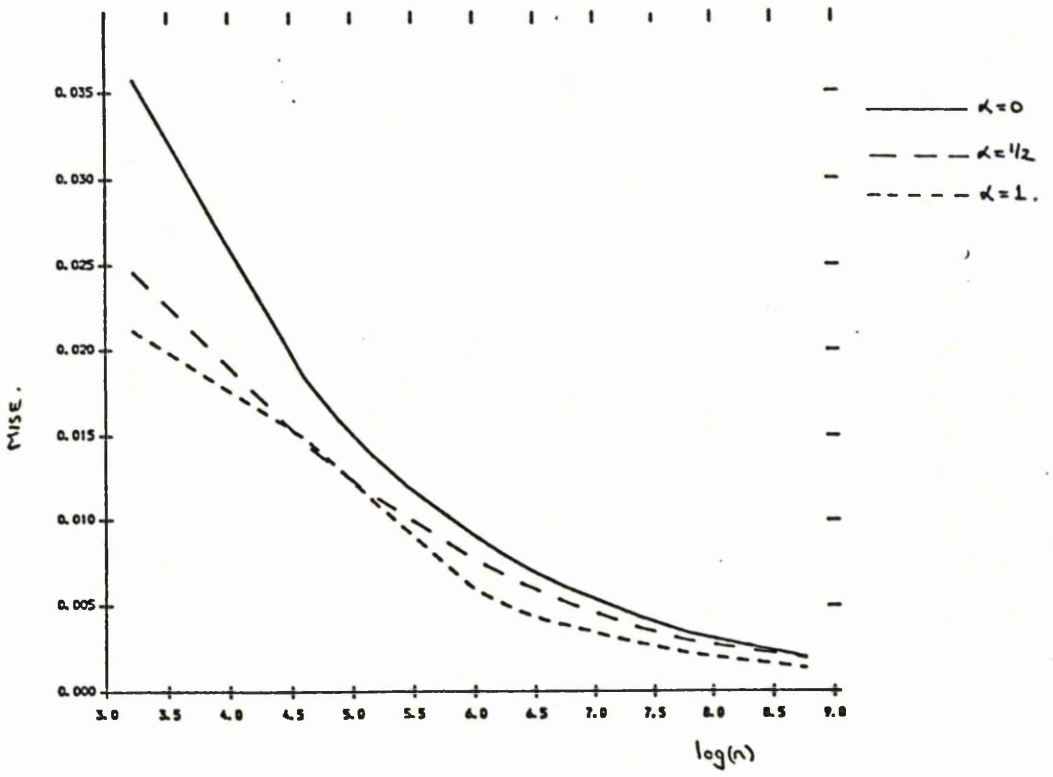


Figure 2.18. The MISE when using normal optimal smoothing parameters, as a function of log(n), for the methods $\alpha = 0, \frac{1}{2}$ and 1 when the underlying density is $0.5N(-0.866, 0.5^2) + 0.5N(0.866, 0.5^2)$

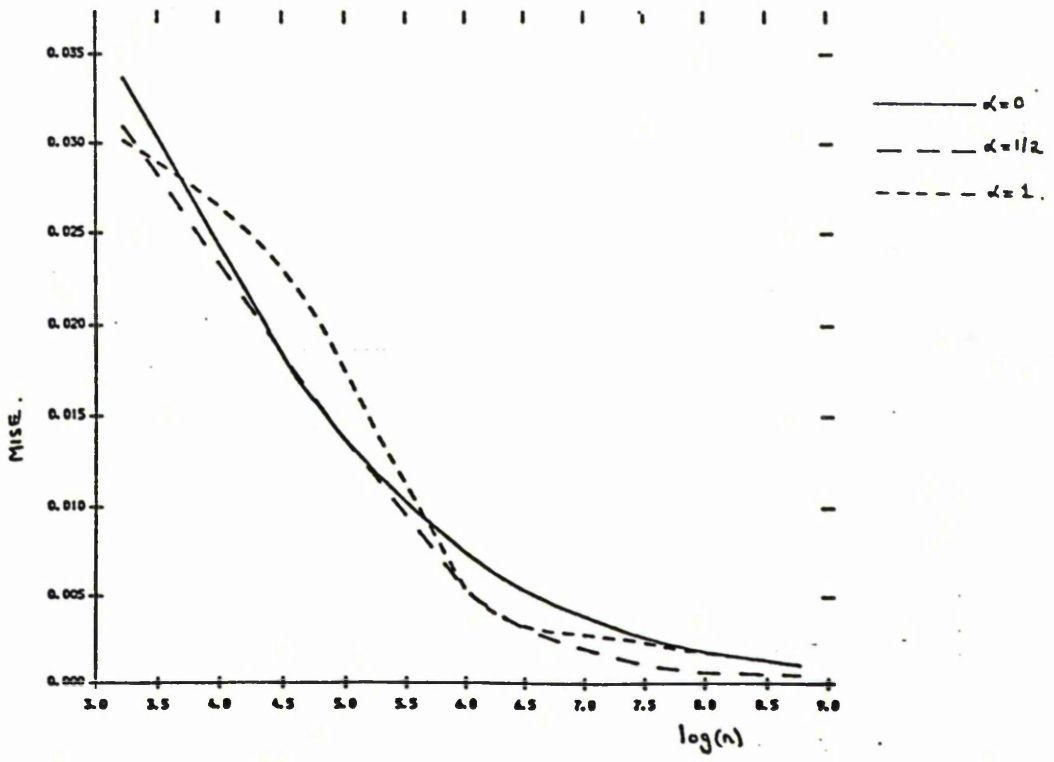


Figure 2.19. Minimum MISE as a function of log(n) for the methods $\alpha = 0$ and $1/2$ when the underlying density is $N_2(\underline{0}, I_2)$.

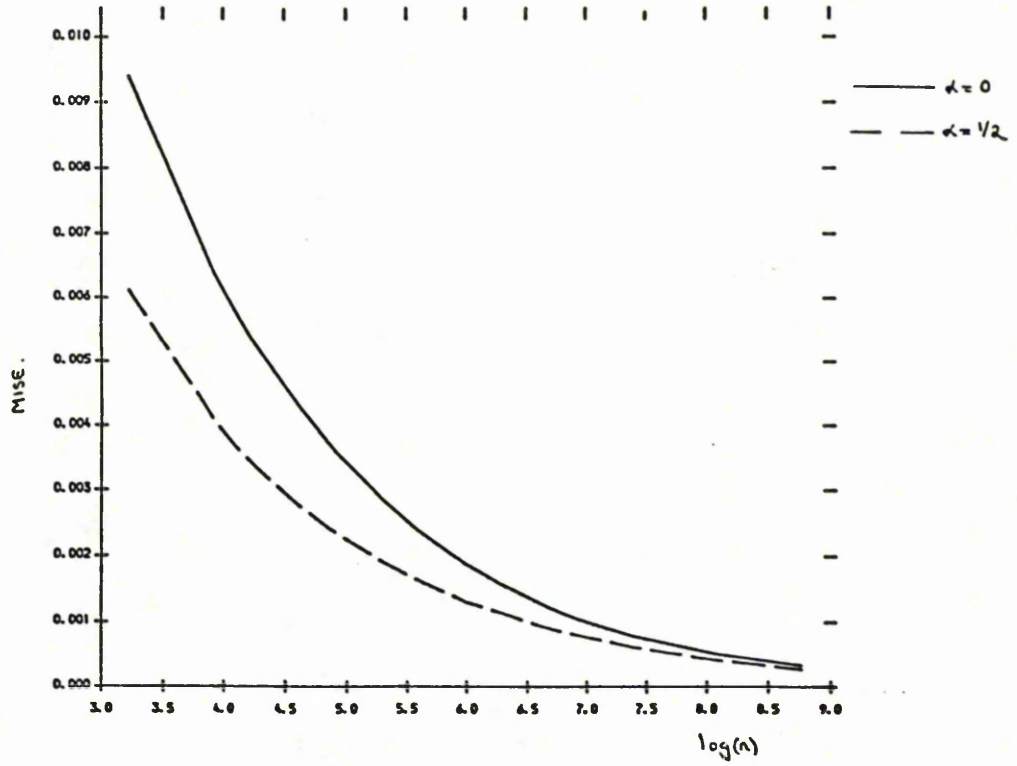


Figure 2.20. Minimum MISE as a function of log(n) for the methods $\alpha = 0$, $1/2$ and $1/3$ when the underlying density is $N_3(\underline{0}, I_3)$.

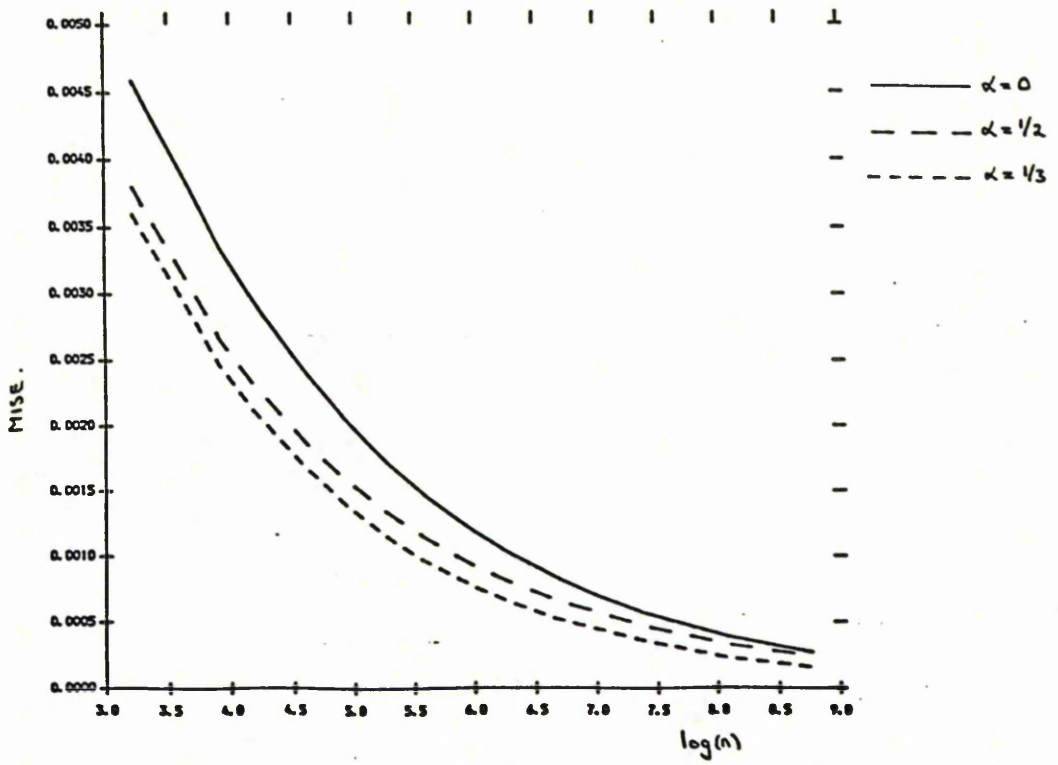


Figure 2.21. Minimum MISE as a function of log(n) for the methods $\alpha = 0, 1/2$ and $1/4$ when the underlying density is $N_4(\Omega, I_4)$.

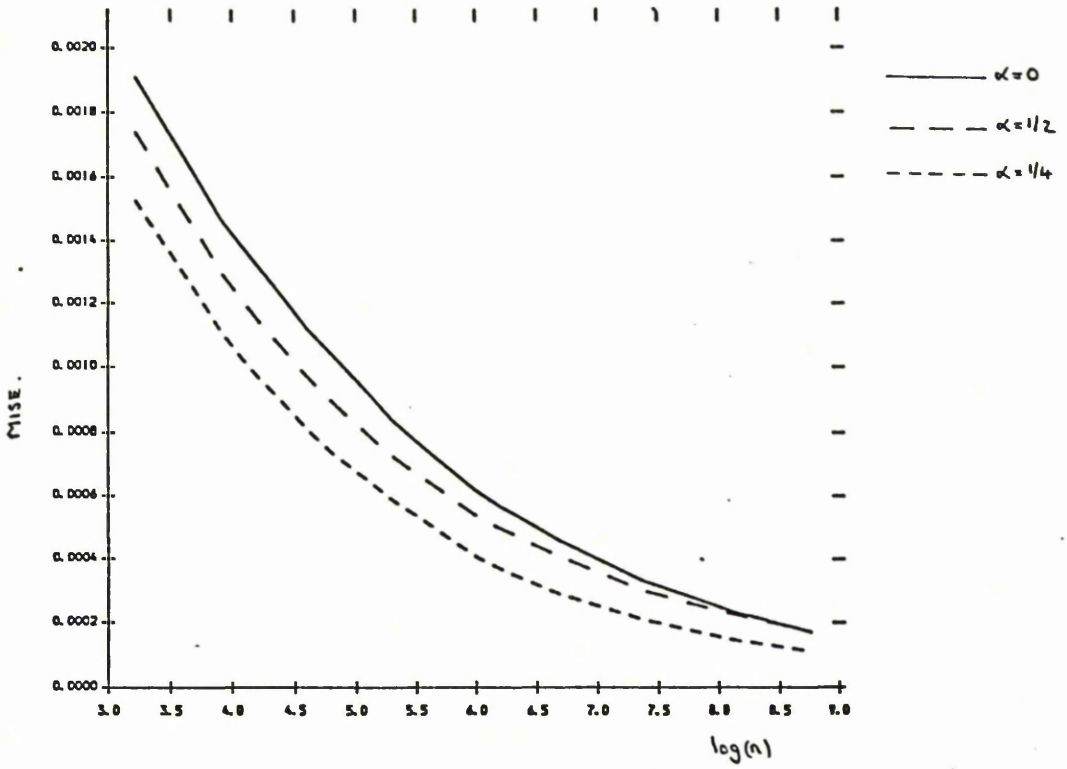


Figure 2.22. Minimum MISE as a function of log(n) for the methods $\alpha = 0, 1/2$ and $1/5$ when the underlying density is $N_5(\Omega, I_5)$.

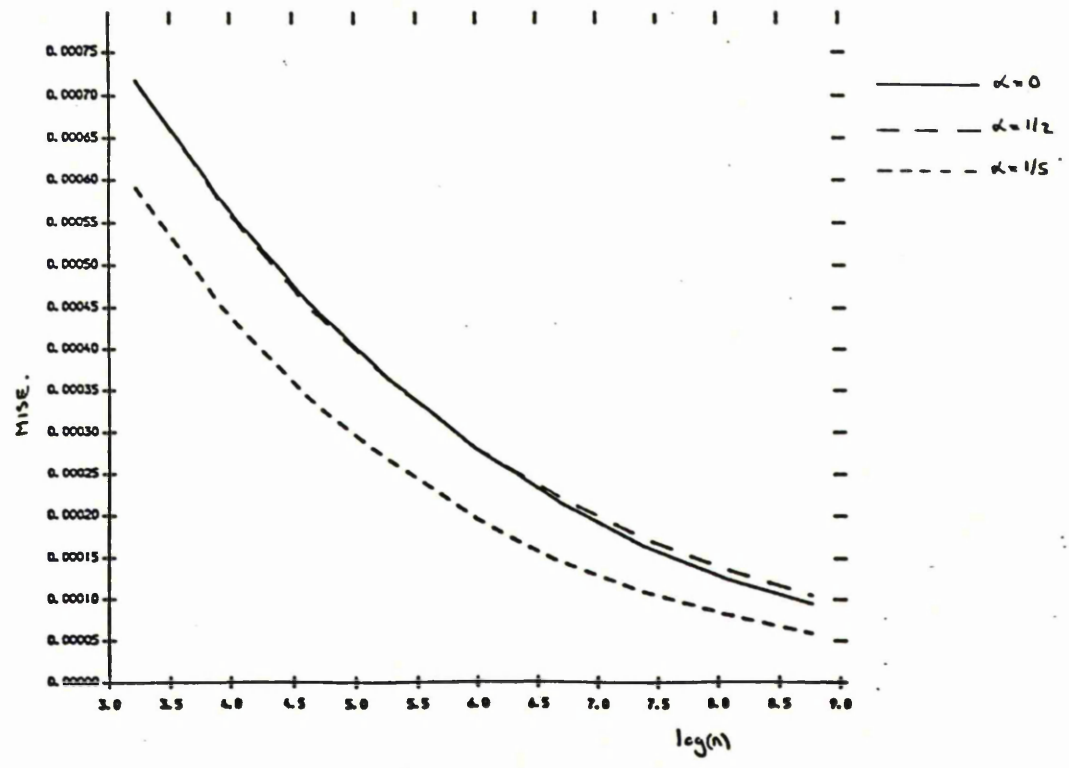


Figure 2.23. Minimum MISE as a function of log(n) for the methods $\alpha = 0,\ 1/2$ and $1/6$ when the underlying density is $N_6(\underline{0}, I_6)$.

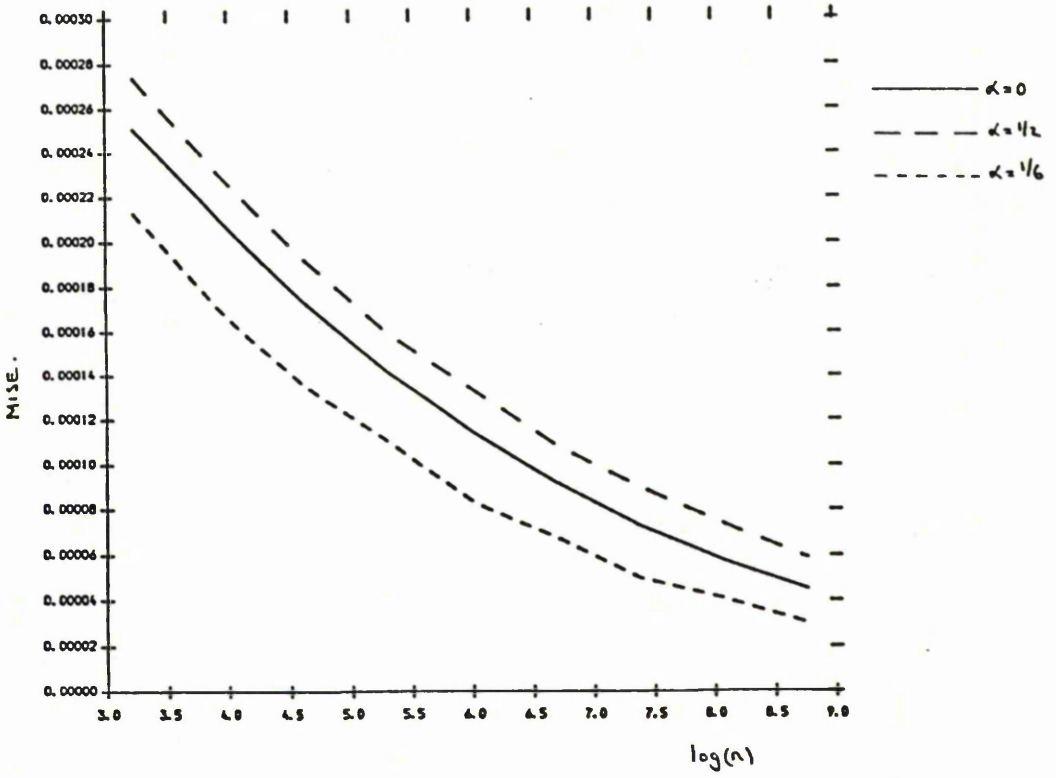


Figure 2.24. Minimum MISE as a function of log(n) for the methods $\alpha = 0,\ 1/2$ and $1/10$ when the underlying density is $N_{10}(\underline{0}, I_{10})$.

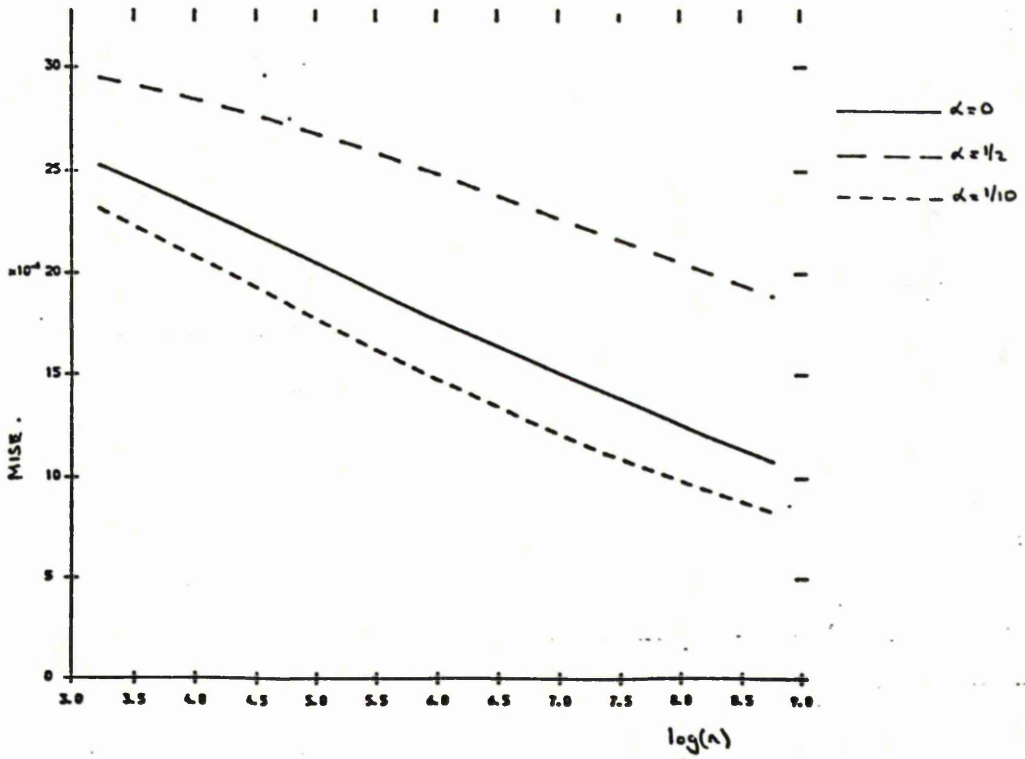


Figure 2.25. The exact bias at the optimal h-values for methods $\alpha = 0$, $1/2$ and $1/3$ as a function of x_1 for a sample of size 50 from a $N_3(\underline{0}, I_3)$ distribution.

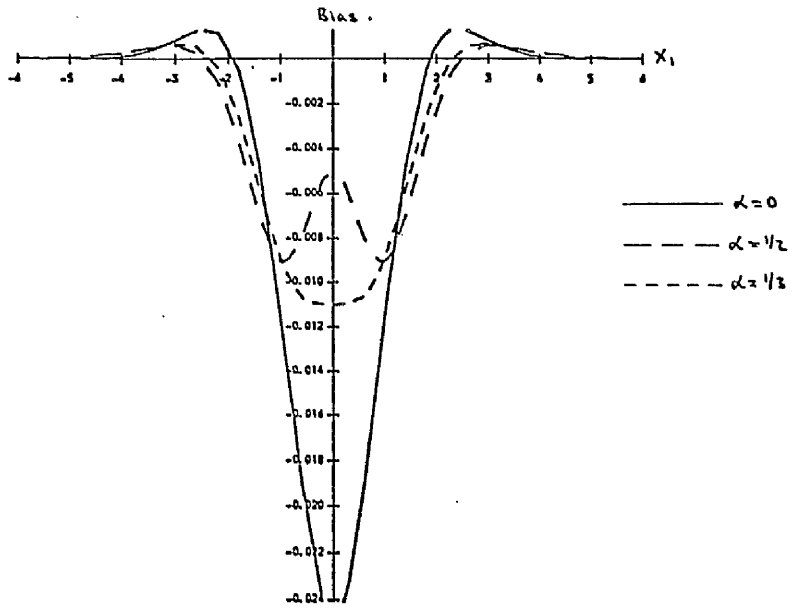


Figure 2.26. The exact standard deviation at the optimal h-values for methods $\alpha = 0$, $1/2$ and $1/3$ as a function of x_1 for a sample of size 50 from a $N_3(\underline{0}, I_3)$ distribution.

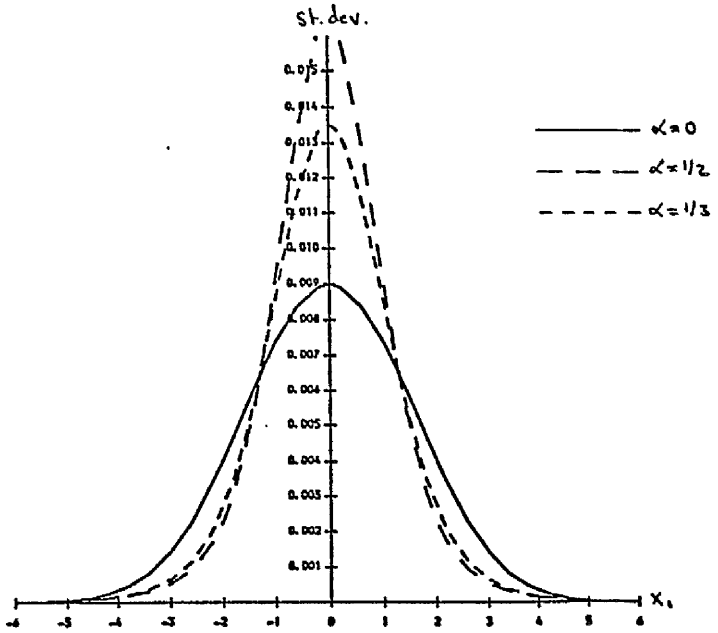


Figure 2.27. The exact bias at the optimal h-values for methods $\alpha = 0,\ 1/2$ and $1/6$ as a function of x_1 for a sample of size 50 from a $N_6(\underline{0}, I_6)$ distribution.

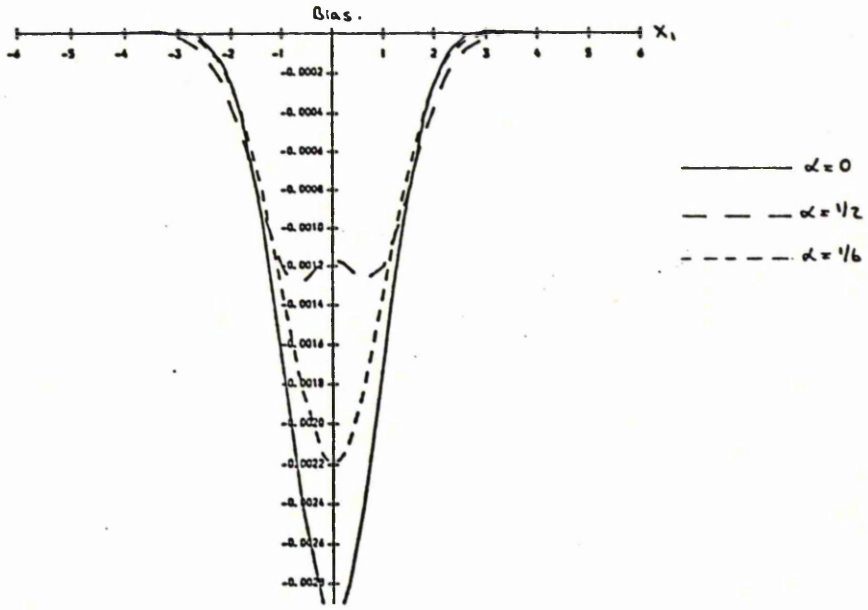


Figure 2.28. The exact standard deviation at the optimal h-values for methods $\alpha = 0,\ 1/2$ and $1/6$ as a function of x_1 for a sample of size 50 from a $N_6(\underline{0}, I_6)$ distribution.

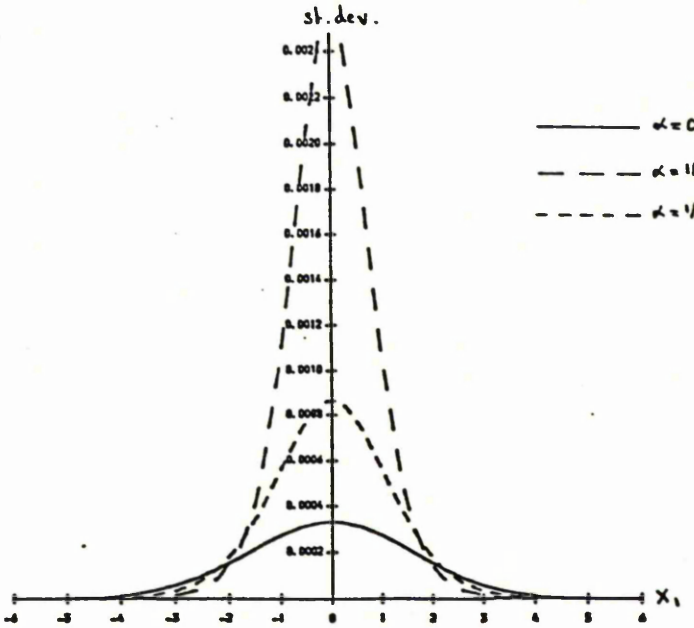


Figure 2.29. Minimum MISE as a function of log(n) for the methods $\alpha = 0$ and $1/2$ when the underlying density is $0.219N_2(\underline{0},4.I_2) + 0.781N_2(\underline{0},0.16\ I_2)$.

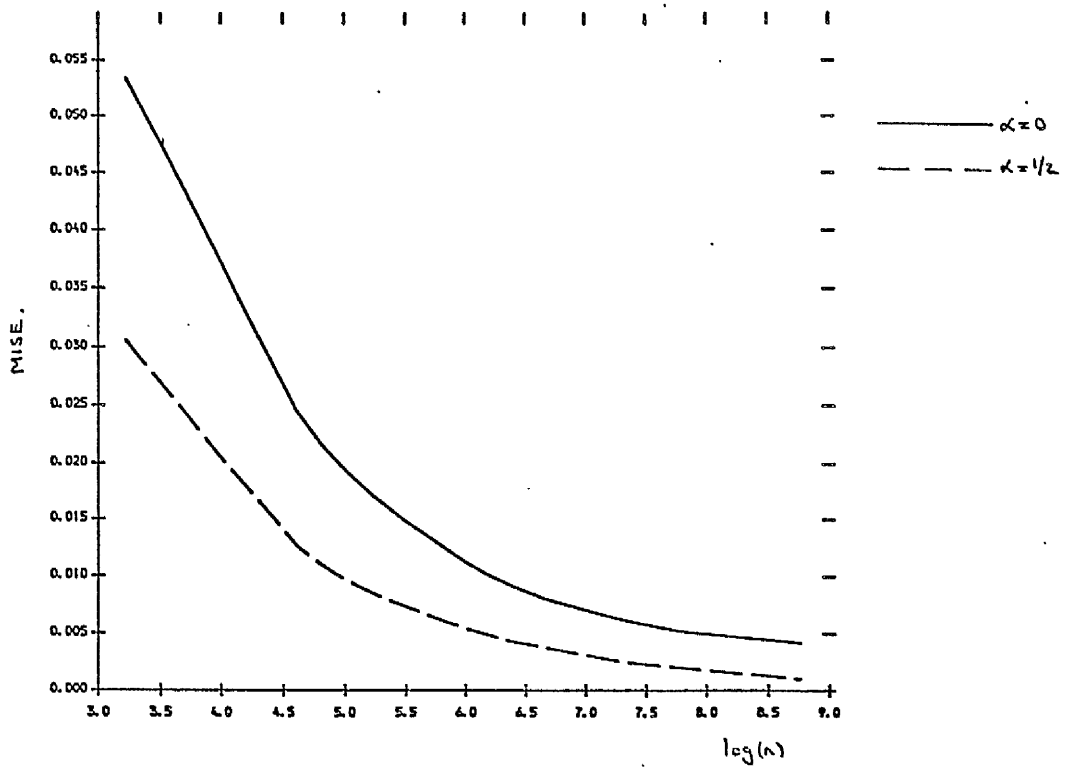


Figure 2.30. Minimum MISE as a function of log(n) for the methods $\alpha = 0$, $1/2$ and $1/3$ when the underlying density is $0.219N_3(\underline{0},4\ I_3) + 0.781N_3(\underline{0},0.16\ I_3)$.

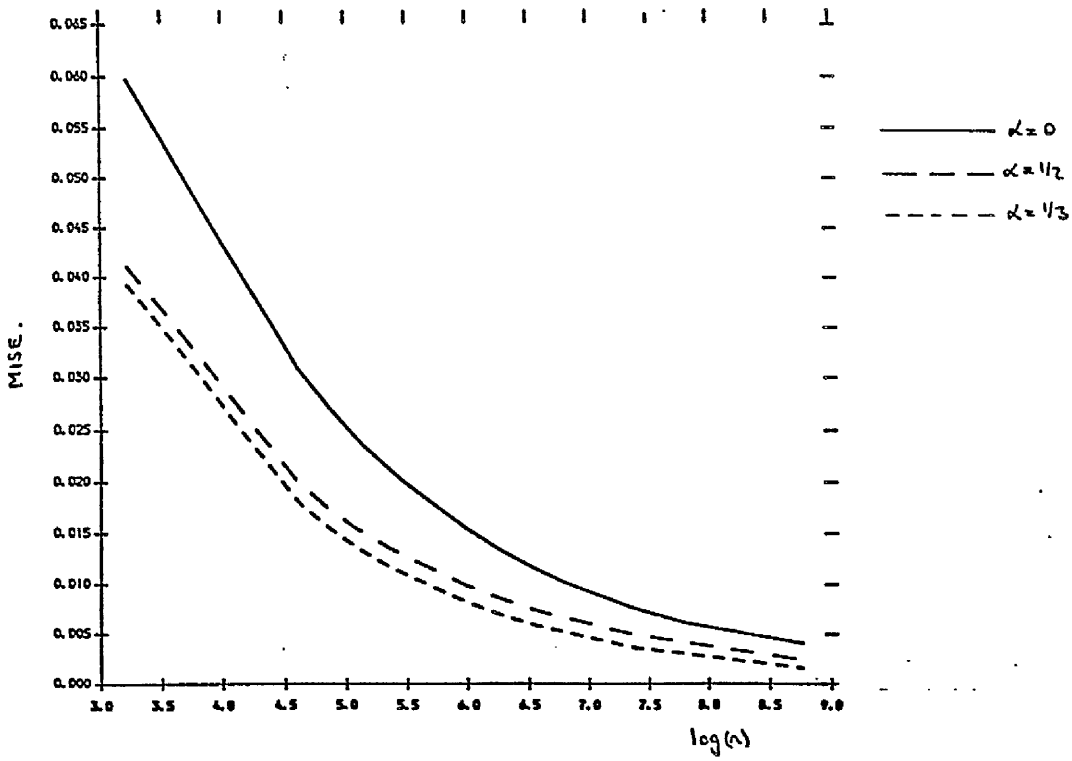


Figure 2.31. Minimum MISE as a function of log(n) for the methods $\alpha = 0$, $1/2$ and $1/4$ when the underlying density is $0.219N_4(\underline{0}, 4\ I_4) + 0.781N_4(\underline{0}, 0.16\ I_4)$.

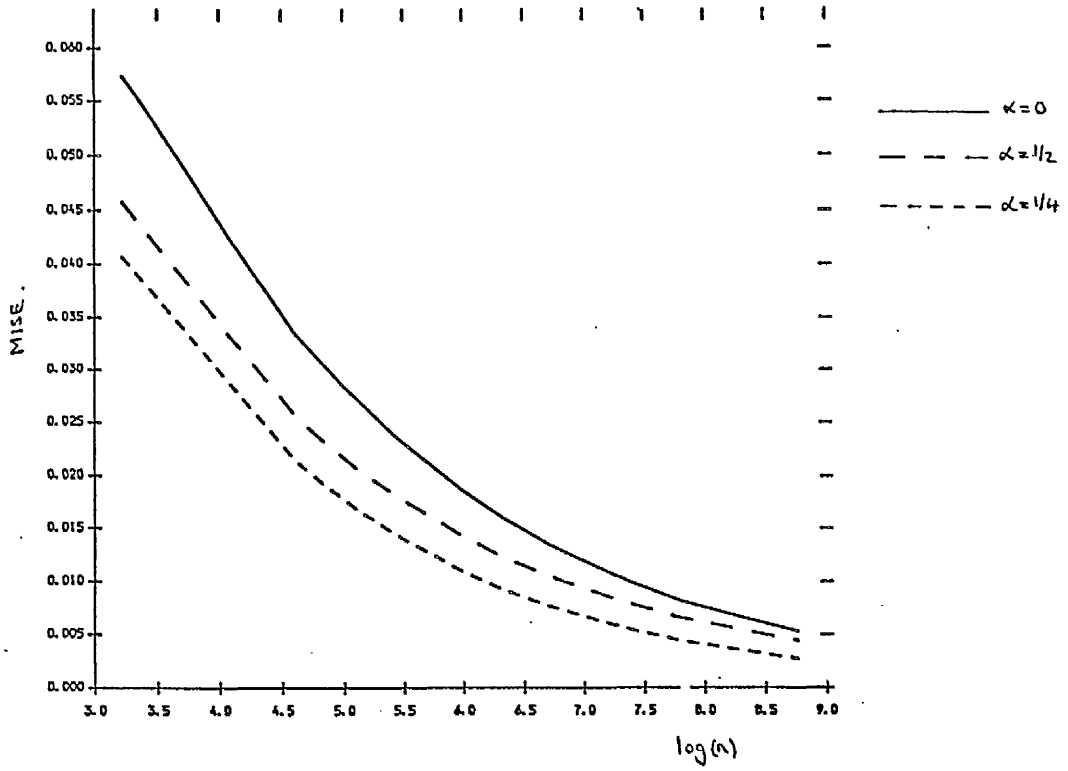


Figure 2.32. Minimum MISE as a function of log(n) for the methods $\alpha = 0$, $1/2$ and $1/5$ when the underlying density is $0.219N_5(\underline{0}, 4\ I_5) + 0.781N_5(\underline{0}, 0.16\ I_5)$.

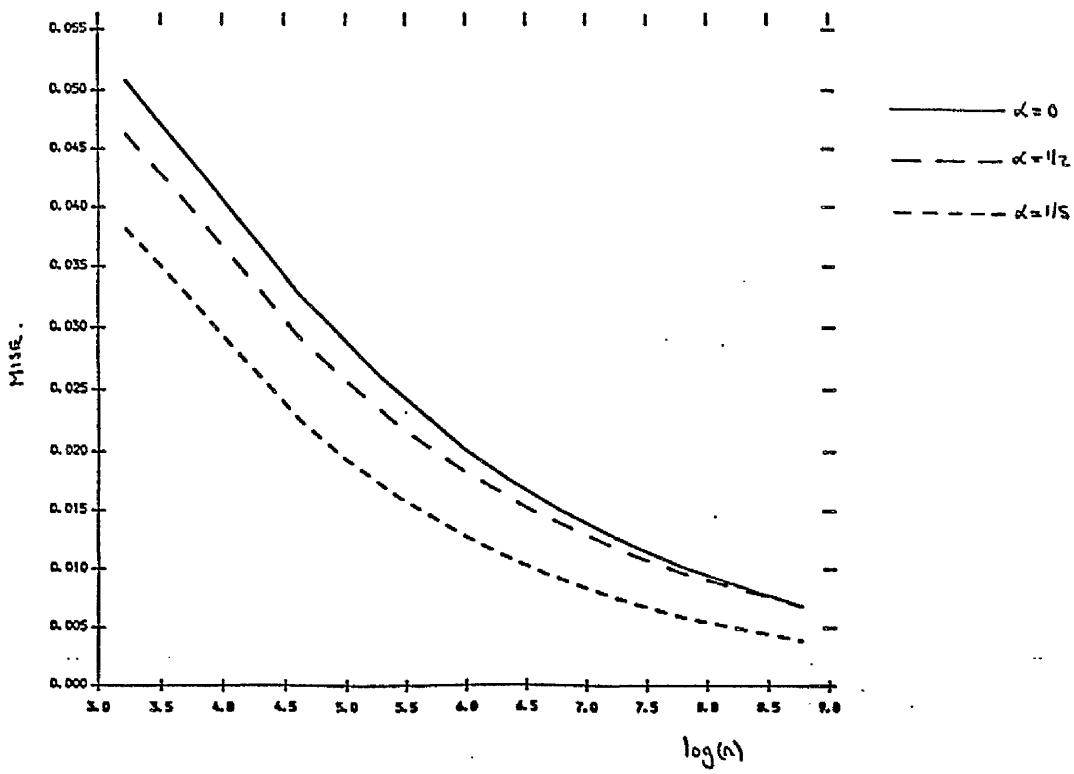
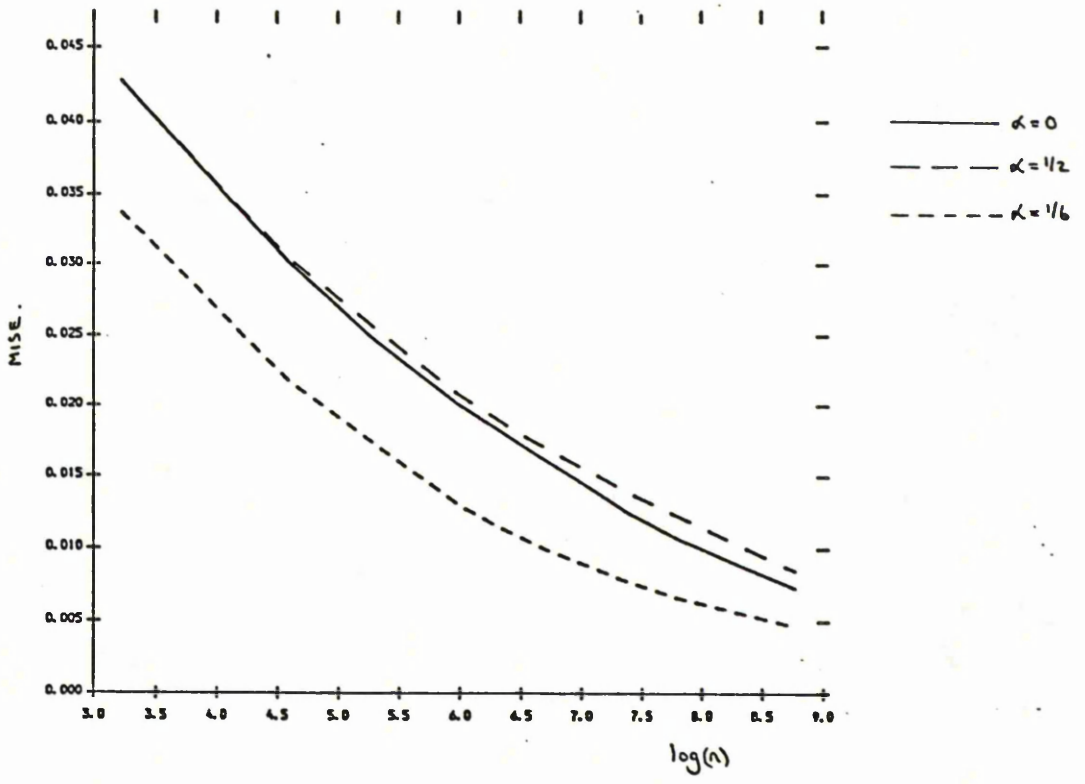


Figure 2.33. Minimum MISE as a function of log(n) for the methods $\alpha = 0, \frac{1}{2}$ and $\frac{1}{6}$ when the underlying density is $0.219N_6(\underline{0}, 4\ I_6) + 0.781N_6(\underline{0}, 0.16\ I_6)$.



Chapter 3. Bias reduction for Nonparametric Estimators.

A. Density Estimators.

3.1. Introduction.

As in chapter 2 we have an independent random sample $\{X_1,\ldots,X_n\}$ from an unknown univariate probability density function, f, for which an estimate is required. Let $\hat{f}(x)$ denote the estimator of f at the point x.

In chapter 2 the exact properties of adaptive kernel density estimators were studied and compared with those based on using a fixed kernel. Global comparisons were made by evaluating the exact MISE for each of the estimators at different sample sizes and for different underlying distributions which showed that greater accuracy could often be obtained by using an adaptive method. For standard normal data it was also shown that the ideal versions of the adaptive estimators, and in particular $\alpha = 1/2$, are effective at reducing bias but at the expense of some increase in variance.

In this chapter particular attention will be paid to the bias of estimators. Rosenblatt (1956) proves that if $\hat{f}(x)$ is symmetric and jointly Borel measurable in $\{X_1,\ldots,X_n\}$ then $\hat{f}(x)$ is not unbiased for $f(x)$. The proof is valid for estimators which are allowed to take negative values as well as those restricted to being non-negative. Yamoto (1972) considers the special case of non-negative kernel estimators which integrate to one and proves that for finite samples these are biased. We therefore cannot construct estimators completely without bias. The aim in this chapter then is to find kernel estimators which have less bias than a simple fixed kernel estimator based on using a standard normal kernel. It would also be

desirable if such estimators had reduced MISE. These properties will be particularly useful when the estimate based on actual observed sample data is to be used in another statistical procedure such as the construction of a confidence interval for $f(x)$ when correct centring is very important. This will be considered further in Chapter 4.

When $\hat{f}(x)$ is a fixed kernel density estimator its exact expectation is given by

$$E[\hat{f}(x)] = \int_{-\infty}^{\infty} \frac{1}{h}.K\left[\frac{x-y}{h}\right].f(y).dy \qquad (3.1.1)$$

where the kernel K is a symmetric function satisfying the conditions (2.1.2) but with the more general moment condition

$$m_j = \int_{-\infty}^{\infty} t^j.K(t).dt = \begin{cases} 1 & , j = 0 \\ 0 & , j = 1,\ldots,k-1 \\ \beta_k \neq 0, & j = k \end{cases} \qquad (3.1.2)$$

If K is a symmetric non-negative density function with finite variance then (2.1.2) and (3.1.2) will hold for $k = 2$ - for example when using a standard normal density.

This expected value is a convolution of the underlying density with the kernel scaled by the smoothing parameter h and is a smoothed version of f. The bias, $E[\hat{f}(x)]-f(x)$, depends explicitly on both K and h but not on the sample size n. However, if h is chosen as a function of n then the bias will depend implicitly on n. Indeed, it is usually assumed that h is a positive function of n such that

$$\lim_{n\to\infty} h(n) = 0 \quad \text{and} \quad \lim_{n\to\infty} n.h(n) = \infty \qquad (3.1.3)$$

which ensures that the bias (and variance) tend to zero as n gets larger so that the estimator is consistent. The dependence of the value of h on n will be assumed but suppressed in the subsequent notation.

As discussed in chapter 2, Rosenblatt (1956) and Parzen (1962) used Taylor series expansions to obtain asymptotic expressions for the bias and variance of a fixed kernel estimator with $k = 2$ which provide a good approximation to the exact small sample bias and variance at the point x. These are given by

$$\text{bias}(\hat{f}(x)) = (1/2).\ h^2.f^{(2)}(x).\beta_2 + o(h^2) \qquad (3.1.4)$$

$$\text{var}(\hat{f}(x)) = f(x).(n.h)^{-1}\int K(t)^2 dt + o((nh)^{-1}) \qquad (3.1.5)$$

The optimal h for minimising the MSE (MISE) based on these results is proportional to $n^{-1/5}$. This means that the bias and variance reduce at the same rate and results in an optimal rate of convergence for the MSE (MISE) of $n^{-4/5}$. In fact, if h is chosen to be proportional to n^{-a}, where $a > 0$, then provided $a < 1$, the estimator is consistent in MSE (MISE). Choosing $a > \frac{1}{5}$ will result in an estimator with smaller bias than for $a = \frac{1}{5}$ but at the expense of increased variance and MSE (MISE). Rather than simply adjusting the value of the smoothing parameter to reduce bias we require methods that will not only reduce bias but also MSE (MISE). In this chapter three main approaches will be described and discussed.

The first is to use 'higher order' kernels which will eliminate terms up to $O(h^r)$ in the asymptotic expansion

$$\text{bias}(\hat{f}(x)) = \sum_{j=1}^{r} \frac{(-1)^j}{j!} f^{(j)}(x).m_j.h^j + O(h^{r+1}) \qquad (3.1.6)$$

which was first proposed by Bartlett (1963). The resulting estimators will have lower order bias, faster optimal rates of convergence in MSE (MISE) but require the relaxation of the non-negativity constraint for the kernel. In particular, the developments of Gasser et al (1985) and Shucany and Sommers (1977) are discussed in Sections 3.2 and 3.3 respectively.

Secondly, it is proposed to construct an estimator for the bias and subtract it from the original density estimator. In Section 3.4 the principal asymptotic bias term when $k = 2$ (i.e. $(1/2)\, h^2 f^{(2)}(x)$) is estimated using the second derivative of the density estimator. This approach is shown to be equivalent to using the 'higher order' weight function $W(t) = K(t) - (1/2)\, K^{(2)}(t)$. In Section 3.6 the second derivative is estimated separately from the density while in Section 3.5 an estimator of the exact bias based on (3.1.1) is considered.

Finally, the approach of Terrell and Scott (1980) who use a multiplicative correction factor and relax the integral constraint is discussed.

Their effectiveness in a finite sample situation is assessed through a simulation study for different shapes of underlying density. This study also involves comparisons with adaptive methods which have been shown in chapter 2 to be effective in reducing bias. In the initial discussion of each of the methods the emphasis will be on their asymptotic properties.

3.2. Minimum Variance and Optimal Kernels.

The MISE is an appropriate loss function when the shape of the underlying density is of principal interest. For a kernel with $k = 2$ the asymptotic value of the MISE when h is chosen optimally is

$$(5/4).\beta_2^{\,2/5}.(\int K(t)^2dt)^{4/5}.(\int f^{(2)}(x)^2dx)^{1/5}.n^{-4/5} \qquad (3.2.1)$$

which can in turn be made as small as possible, provided h is chosen optimally, by minimising

$$\int K(t)^2dt \qquad (3.2.2)$$

with respect to the function K subject to the constraints that

$$\int K(t)dt = 1$$
$$\text{and} \qquad \int t^2.K(t)dt = 1. \qquad (3.2.3)$$

The solution to this problem in this context was given by Epanechnikov (1969) and the so called "Epanechnikov kernel" is the quadratic function

$$K(t) = \begin{cases} (3/(4.\sqrt{5})).(1-t^2/5), & -\sqrt{5} \leqslant t \leqslant \sqrt{5} \\ 0 \quad , & \text{o.w.} \end{cases} \qquad (3.2.4)$$

This kernel among those with $k = 2$ is optimal then in the sense that it minimises the asymptotic MISE. However, Table 3.1 of Silverman (1986) shows that there is little asymptotic loss in using some other suboptimal kernel with $k = 2$.

Gasser et al (1985) further generalise this problem by considering two classes of kernel which satisfy moment conditions of various orders for estimating a density and its derivatives. These are

"minimum variance" and "optimal" kernels which minimise the asymptotic variance and MISE respectively.

A kernel K_ν of order k for estimating the ν^{th} derivative ($\nu = 0,1,2....,$) is defined as

$$\int_{-\tau}^{\tau} t^j K_\nu(t)dt = \begin{cases} 0 & , \ j = 0,\ldots,\nu-1,\nu+1,\ldots,k-1 \\ (-1)^\nu.\nu! & , \ j = \nu \\ \beta_k \neq 0 & , \ j = k. \end{cases} \qquad (3.2.5)$$

It is assumed that $0 \leqslant \nu \leqslant k-2$ and that ν and k are either both even or both odd. As discussed earlier use of a kernel satisfying (3.2.5) will result in an asymptotic bias of $O(h^k)$.

The minimum variance kernels of order (ν,k) are solutions to the variational problem:

$$\text{Minimise} \int_{-\tau}^{\tau} K_\nu(t)^2 dt \qquad (3.2.6)$$

subject to the moment conditions (3.2.5).

The support is taken to be $[-1,1]$ so that $\tau = 1$. They show that the minimum variance kernels are symmetric when ν is even and antisymmetric when ν is odd and that those of order (ν,k) on $[-1,1]$ are uniquely defined polynomials of degree $k-2$ with $k-2$ real roots in $(-1,1)$.

Optimal kernels of order (ν,k) are again symmetric for ν even and antisymmetric for ν odd and minimise the functional

$$\left| \int_{-1}^{1} t^k.K_\nu(t)dt \right|^{(2\nu+1)} . \left[\int_{-1}^{1} K_\nu(t)^2 dt \right]^{(k-\nu)} \qquad (3.2.7)$$

subject to conditions (3.2.5).

To avoid degeneracy, the kernel order (ν,k) defined on $[-1,1]$ must have at least $k-2$ changes of sign on $(-1,1)$.

Table 1 below gives the functional form of the minimum variance and optimal kernels of various orders while Table 2 gives values of B, V and T which are defined by

$$B = \int_{-1}^{1} t^k . K_\nu(t)dt$$
$$V = \int_{-1}^{1} K_\nu(t)^2 dt \qquad (3.2.8)$$

and $$T = \left[V^{(k-\nu)} \cdot B^{(2\nu+1)}\right]^2/(2k+1)$$

and are the components of the expressions for the asymptotic bias, variance and MISE which depend on the kernel. V and T are the functionals minimised above. These tables are reproduced from Gasser et al (1985).

Table 3.1. Examples of minimum variance and optimal kernels.

Order	Minimum variance	Optimal
(0,2)	$^1/_2$	$(^3/_4).(-x^2+1)$
(0,4)	$(^3/_8)(-5x^2+3)$	$(^{15}/_{32}).(7x^2-10x^2+3)$
(0,6)	$(^{15}/_{128}).(63x^4-70x^2+15)$	$(^{35}/_{256}).(-99x^6+189x^4-105x^2+15)$
(1,3)	$(-^3/_2).x$	$(^{15}/_4).(x^3-x)$
(1,5)	$(^{15}/_8)(7x^3-5x)$	$(^{105}/_{32}).(-9x^5+14x^3-5x)$
(2,4)	$(^{105}/_{32}).(-45x^4+42x^2-5)$	$(^{105}/_{16})(-5x^4+6x^2-1)$
(2,6)	$(^{105}/_{32}).(-45x^4+42x^2-5)$	$(^{315}/_{64})(77x^6-135x^4+63x^2-5)$

Table 3.2. Asymptotic bias, variance and MISE for optimal and minimum variance kernels, standard and one order higher ($k=\nu+2$ and $k=\nu+4$).

ν	k	kernel	B	V	T
0	2	opt.	0.2000	0.6000	0.3491
		min. V.	0.3333	0.5000	0.3701
	4	opt.	-0.0476	1.250	0.6199
		min. V.	-0.0857	1.125	0.6432
1	3	opt.	-0.4286	2.143	0.7477
		min. V.	-0.6000	1.500	0.8137
	5	opt.	0.1515	11.93	2.168
		min. V.	0.2381	9.375	2.328
2	4	opt.	1.333	35.00	6.685
		min. V.	1.714	22.50	7.262
	6	opt.	-0.6293	381.6	27.16
		min. V.	-0.90091	275.6	29.50

The values of T in Table 3.2 show that the minimum variance kernels are all suboptimal by less than 10% with respect to asymptotic MISE.

One of the anomalies with the kernels when $k-v \geq 4$ is that the theory requires that the underlying density should have at least k-v continuous derivatives but the kernels themselves are discontinuous at + and - 1 which leads to estimates which are also discontinuous. To partially circumvent this problem Muller (1984) constructs kernels with support [-1,1] of order (ν,k) which are u times differentiable $(u > 0)$ and minimise the variance of the u^{th} derivative of the estimate. These kernels are polynomials of degree $(k + 2u - 2)$ and are tabulated in Table 1 of that paper for $u = 2$ and 3 and for important values of ν and k.

When using the optimal kernel of order (0,4) to estimate the density i.e.

$$K(t) = (^{15}/_{32}).(7t^4 - 10t^2 + 3) \qquad (3.2.9)$$

a practical formula for choosing the value of the smoothing parameter can be obtained by assuming that f is a particular distribution such as $N(0,1)$. The asymptotic MISE (Silverman (1986, p.67)) can then be minimised with respect to h to give:

$$h_{opt} = 3.243 \left[\int f^{(4)}(x)^2 dx \right]^{-1/9} . n^{-1/9} \qquad (3.2.10)$$

and using the result that $\int f^{(4)}(x)^2 \, dx = 1.85125$ for a standard normal distribution gives

$$h_{opt} = 3.029.n^{-1/9} \qquad (3.2.11)$$

for $N(0,1)$ data. If the variance is not one then (3.2.11) should be multiplied by a (robust) measure of scale.

Alternatively, the exact procedure described in Section 3.1 of Chapter 2 can be used by again assuming the underlying distribution is $N(0,1)$. This was carried out for nine sample sizes between $n = 25$ and $n = 6400$ and resulted in the formula

$$h_{opt} = 3.904.n^{-0.134} \qquad (3.2.12)$$

which is not too dissimilar from (3.2.11). Use of (3.2.12) in fact gives slightly larger values of h than does (3.2.11) but these differences lessen as n increases.

A problem when estimating the density function ($\nu = 0$) using a kernel with $k \geq 4$ is that negative estimates of the density can be obtained especially in the tails where there is little data. Such negative values should therefore be reset to zero but the resulting estimate will then not be a bona fide density in that it will not integrate to one.

A simple way to overcome this is by then rescaling the estimate by $\left[\int \hat{f}(x)dx \right]^{-1}$. The result of this however, is that anomalies in

parts of the distribution with little data affect the estimate in the more important main body of the distribution. An alternative iterative correction procedure which converges to a bona fide estimate was suggested by Gajek (1986) which has an optimality property over any other correction procedure with respect to weighted MISE i.e. $E\int(\hat{f}(x) - f(x))^2.W(x)dx$. A necessary condition on the weight function w in the MISE though is that $\int w(x)\ dt$ is finite which means that the weight w should increase for "large" arguments which thereby increases the importance of the tails. No guidance is given in the paper as to a more specific choice of weight function and the effect different choices have on the resulting estimate nor to the speed of convergence of the algorithm.

In the derivation of formula (3.2.12), which involved numerical integration, no rescaling was carried out which removed the influence of the tails as discussed above. This is justified by the average areas based on 25 simulations of data from four different distributions given in the following table.

Table 3.3. The mean and standard deviation of the areas of density estimators using the kernel $K(t) = (^{15}/_{32}).(7t^4-10t^2+3)$ with no rescaling based on 25 simulations.

Distribution	n = 50		N = 100	
	Mean	St. deviation	Mean	St. deviation
N(0,1)	1.015	0.000066	1.008	0.000025
Gamma (2, $\sqrt{2}$)	1.003	0.000007	1.001	0.000001
0.5[N(x-0.866,0.5) + N(x+0.866,0.5)]	1.033	0.000044	1.032	0.000035
t(3)	1.006	0.000020	1.002	0.000012

The largest mean errors occur for the bimodal normal mixture but these are only 3.3% and 3.2% for $n = 50$ and 100 respectively. All the others are 1.5% or less. There is also little variability in the results for each of the sets of 25 samples. For larger sample sizes the errors will be even less. Hence, resetting negative estimates to zero but not rescaling results in a density estimate which is very close to being bona fide.

On the other hand, these results could be used to argue that if rescaling is carried out the influence of the tails on the main part of the distribution will be quite small. Therefore, the decision on whether to rescale or not really depends on what the estimate is required for. If only a simple pictorial representation of the density estimate is required then doing no rescaling simplifies the computations and should be quite adequate. However, if the estimate is to be used in another statistical procedure then it may be better to carry out some form of rescaling to obtain a true density.

3.3 Jacknifing.

This is based on the generalised jacknife method of Schucany, Gray and Own (1971). In this context let $\{X_1,\ldots,X_n\}$ be n independent and identically distributed observations from the distribution $F(\theta)$ where θ is an unknown parameter and also let t_1 and t_2 be two estimators of θ which are biased.

i.e. $$E[t_r(X_1,\ldots,X_n)] - \theta = b_r(n,\theta) \neq 0, \quad r = 1,2.$$

Let $R = b_1(n,\theta)/b_2(n,\theta)$. If $\hat{\theta}$ is now defined to be $(t_1-Rt_2)/(1-R)$ then $E[\hat{\theta}] = \theta$ assuming that R is known.

For density estimation Schucany and Sommers (1977) define the "jacknife estimator" to be

$$g(x) = [\hat{f}(x;K,h_1) - R.\hat{f}(x;K,h_2)]/(1-R) \qquad (3.3.1)$$

where $R \neq 1$ is a constant, K is the kernel function and h_1 and h_2 are smoothing parameters. The estimators $\hat{f}(\cdot;K,h_1)$ and $\hat{f}(\cdot;K,h_2)$ are both assumed to satisfy (3.1.6) for $k = 2$. They show that if R is set to be the ratio of the principal terms in the asymptotic expansion for the bias which, because the kernel functions are the same for both estimators, is simply h_1^2/h_2^2, then the term in the bias, $E[g(x)] - f(x)$, containing $f^{(2)}(x)$ is removed.

They also show that

$$g(x) = (n.h_1)^{-1} \sum_{i=1}^{n} [K(Z_i) - a^{-3}.\ K((Z_i/a)]/(1-a^{-2}) \qquad (3.3.2)$$

where $a = h_2/h_1$ and $Z_i = (x-X_i)/h_1$ so that the same estimator could have been produced by using the single kernel

$$K^*(t) = [K(t) - a^{-3}.\ K(t/a)]/(1-a^{-2}) \ . \qquad (3.3.3)$$

K^* satisfies the condition (3.1.2) for $k = 4$ and is therefore in the class of kernels suggested by Bartlett with $k = 4$.

Note that the jacknife estimator can be expressed as

$$g(x) = \hat{f}(x;K,h_1) + [1/(a^2-1)].[\hat{f}(x;K,h_1) - \hat{f}(x;K,ah_1)] \qquad (3.3.4)$$

which suggests that $[1/(a^2-1)].[\hat{f}(x;K,h_1) - \hat{f}(x;K,ah_1)]$ is an additive bias reducing corrector factor for $\hat{f}(x;K,h_1)$.

Now let

$$g(x) = (n.h_1)^{-1} \ . \sum_{i=1}^{n} K^*(x-X_i)/h_1) \qquad (3.3.5)$$

so that

$$E[g(x)] = f(x) + (1/24).h_1^4.f^{(4)}(x).\int t^4.K^*(t)dt + o(h_1^4) \tag{3.3.6}$$

and

$$V[g(*)] = f(x)/(nh_1)).\int K^*(t)^2dt + o(nh_1)^{-1} \tag{3.3.7}$$

The quantities $\int t^4.K^*(t)dt$ and $\int K^*(t)^2dt$ can be simplified as follows:

$$\int t^4.K^*(t)dt = \frac{a^2}{(a^2-1)} . \int t^4\, K(t)dt - \frac{a^{-1}}{(a^2-1)} . \int t^4 K(t/a)dt$$

$$= \frac{a^2}{(a^2-1)} .m_4 - \frac{a^{-1}}{(a^2-1)} .a^5.m_4$$

$$= -a^2.m_4 \tag{3.3.8}$$

where m_4 is defined by (3.1.2) and equals 3 if K is the standard normal density.

$$\int K^*(t)^2dt = \int \left[\frac{a^4}{(a^2-1)^2} \left\{K(t)^2 + a^{-6}.K(t/a)^2 \right.\right.$$

$$\left.\left. - 2.a^{-3}.K(t).K(t/a)\right\}\right] dt$$

$$= \frac{a^4}{(a^2-1)^2} . \left[\int K(t)^2dt + a^{-6} \int K(t/a)^2dt - 2a^{-3}.\int K(t)K(t/a)dt\right] \tag{3.3.9}$$

When K is a standard normal density

$$\int K^*(t)^2dt = \frac{a^4}{(a^2-1)^2} . \left[\frac{1}{2\sqrt{\pi}} + \frac{a^{-5}}{2\sqrt{\pi}} - \frac{2\, a^{-2}}{\sqrt{2.\pi.(1+a^2)}}\right] \tag{3.3.10}$$

Schucany and Sommers suggest without justification using a value of a close to 1. This suggestion can be investigated by considering expressions (3.3.8) and (3.3.10). If a standard normal density function is used as the kernel then when $9a^4$ (i.e. the square of (3.3.8)) is plotted against a it can be seen from figure 3.1 that

for $a > 1$ the integrated squared bias would increase very rapidly. The plot of (3.3.10) against a (figure 3.2) shows that the integrated variance would decrease fairly rapidly for $0 < a < 1$ while for $a > 1$ the decrease is quite slow. Combining these in the expression $T = (V^8 \, . \, B^2)^{1/9}$ which is proportional to the asymptotic MISE evaluated at the asymptotically optimal smoothing parameter and again plotting against a (figure 3.3) shows that the MISE would decrease fairly rapidly for $0 < a \leqslant 0.6$, more gently for $0.6 < a < 1$, attain a minimum as $a \longrightarrow 1$ and increase fairly slowly for $a > 1$. This gives credence then to a choice of a near 1 but for a large range of values of a the increased loss in MISE would not be great.

It is of interest to see what form $K^*(t)$ takes in the limit as $a \longrightarrow 1$. This is also not considered by Schucany and Sommers but using L'Hopital's rule gives

$$\lim_{a \to 1} K^*(t) = 0.5.t.K^{(1)}(t) + 1.5.K(t)$$

$$= K^L(t), \quad \text{say.} \qquad (3.3.11)$$

If $K^L(t)$ is a standard normal density then

$$K^L(t) = -0.5t^2.N(t;0,1) + 1.5.N(t;0,1)$$

$$= N(t;0,1) - 0.5.(t^2-1).N(t;0,1).$$

$$= N(t;0,1) - 0.5.N^{(2)}(t;0,1) \qquad (3.3.12)$$

where $N(t;0,1)$ denotes the standard normal density evaluated at t. This is equivalent to estimating the density using a normal kernel and then subtracting an estimate of the principal bias term, $0.5.h^2.f^{(2)}(x)$, based on the second derivative of the estimate. This approach will be discussed further in Section 3.4.

The value of $B = \int t^4.K^L(t)dt = -3$ and $V = \int K^L(t)^2dt = 0.476$

so that $T = 0.660$ which compares with 0.620 for the optimal kernel and 0.643 for the minimum variance kernel of Gasser et al (1985). (See table 3.1).

Jacknifing can also be implemented by using a different kernel for each of the two estimators in (3.3.1) (i.e K_1 and K_2) and Schucany and Sommers give algebraic details based on such an approach. The bias term $h^4.f^{(4)}(x)$ can be eliminated if K_1 and K_2 are chosen such that they differ in the second or fourth moments and h_1 and h_2 are chosen in the ratio

$$\frac{h_1^2}{h_2^2} = \frac{\int t^4 K_2(t)dt \; . \int t^2.K_1(t)dt}{\int t^2 K_2(t)dt \; . \int t^4.K_1(t)dt} \qquad (3.3.13)$$

Bias terms containing higher powers of h can be eliminated if more than two estimators are combined.

In a small simulation study they demonstrate the effectiveness of the jacknife estimator in reducing MSE.

3.4. Using a kernel function $W(t) = K(t) - (1/2).K^{(2)}(t)$.

Consider estimating a density f using a fixed kernel density estimator with $k = 2$ in (3.1.6) and then subtracting an estimate of the principal bias term $(1/2)h^2.f^{(2)}(x)$ based on the second derivative of the estimator.

i.e. $$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} \{h^{-1}K(x-X_i)/h)-(1/2)h^2. \frac{d^2}{dx^2} h^{-1}K((x-X_i)/h)\} \qquad (3.4.1)$$

The term in brackets, after carrying out the differentiation, is

$$h^{-1}.K((x-X_i)/h) - (1/2).h^2/(h^{-3}.K^{(2)}((x-X_i)/h))$$

$$= h^{-1} [K((x-X_i)/h)-(1/2).K^{(2)}((x-X_i)/h)] \qquad (3.4.2)$$

Hence,

$$\hat{f}(x) = n^{-1}.h^{-1} \sum_{i=1}^{n} W((x-X_i)/h) \qquad (3.4.3)$$

where

$$W(t) = K(t) - (1/2).K^{(2)}(t) \qquad (3.4.4)$$

Assume that the kernel K integrates to one and satisfies the moment condition (3.1.2) with k = 2. Then

$$\int W(t)dt = \int \{K(t) - (1/2).K^{(2)}\}dt = 1 \quad \text{provided} \int K^{(2)}(t)dt = 0 \qquad (3.4.5)$$

$$\int t.W(t)\,dt = 0 \quad \text{provided} \quad \int tK^{(2)}(t)dt = 0 \qquad (3.4.6)$$

$$\int t^2.W(t)\,dt = 0 \quad \text{provided} \quad \int t^2.K^{(2)}(t)dt = 2.m_2 \qquad (3.4.7)$$

$$\int t^3.W(t)\,dt = 0 \quad \text{provided} \quad \int t^3.K^{(2)}(t)dt = 0 \qquad (3.4.8)$$

$$\int t^4.W(t)\,dt = m_4 - (1/2) \int t^4.K^{(2)}(t)dt \neq 0. \qquad (3.4.9)$$

Clearly W(t) is a kernel function satisfying the moment conditions (3.1.2) with k = 4 so that the asymptotic bias will be $O(h^4)$.

If K is chosen to be the standard normal density then the above conditions ((3.4.5-(3.4.9)) on the kernel and it's second derivative are satisfied.

i.e.
$$\begin{aligned} W(t) &= N(t;0,1) - 0.5.N^{(2)}(t;0,1) \\ &= N(t;0,1) - 0.5(t^2-1).N(t;0,1) \\ &= 0.5(3-t^2).N(t;0,1) \end{aligned} \qquad (3.4.10)$$

This is the same weight function that was found by jacknifing using N(0,1) kernels and letting $a \longrightarrow 1$.

It is of interest to find other functions K which satisfy (3.4.5)-(3.4.8). Suppose then that K is a polynomial of degree m defined on $[-1,1]$ and zero otherwise. K needs to be a symmetric function so it is only necessary to consider terms involving even powers of t.

i.e. $K(t) = a_m t^m + a_{m-2}t^{m-2}+\ldots+a_0$, m even (3.4.11)

where the $\{a_i\}$ are constants. Therefore,

$$K^{(2)}t = m.(m-1)a_m t^{m-2} +(m-2)(m-3).a_{m-2}t^{m-4} +\ldots+ 2.1.a_2 \quad (3.4.12)$$

Now,

$$\int_{-1}^{1} K(t)dt = 2.\sum_{i=0}^{m} a_i/(i+1) = 1, \quad i, m \text{ even} \quad (3.4.13)$$

$$\int_{-1}^{1} K^{(2)}(t)dt = 2\sum_{i=0}^{m} a_i.i = 0, \quad i, m \text{ even} \quad (3.4.14)$$

$$\int_{-1}^{1} t^2K(t)dt = 2\sum_{i=0}^{m} a_i/(i+3) \quad i, m \text{ even} \quad (3.4.15)$$

and

$$\int_{-1}^{1} t^2.K^{(2)}(t)dt = 2.\sum_{i=0}^{m} i.(i-1).a_i/(i+1) \quad i, m \text{ even} \quad (3.4.16)$$

so that to satisfy (3.4.7) it is required that

$$\sum_{i=0}^{m} i.(i-1).a_i(i+1) = 2\sum_{i=0}^{m} a_i/(i+3) \quad (3.4.17)$$

i.e. $$\sum_{i=0}^{m} a_i\{((i^3+2.i^2-5.i-2)/((i+1).(i+3))\} = 0 \quad (3.4.18)$$

For $m = 0$ condition (3.4.18) implies that $-(2/3)a_0 = 0$ so that $a_0 = 0$. For $m = 2$ condition (3.4.14) implies that $0.a_0 + 2a_2 = 0$ so that $a_2 = 0$. Therefore, no polynomials of degree 2 or less satisfy the required conditions. However, when $m = 4$ we have:

$$0.a_0 + 2.a_2 + 4.a_4 = 0 \qquad \text{from (3.4.14)}$$

$$a_0 + (1/3).a_2 + (1/5).a_4 = 1/2 \qquad \text{from (3.4.13)}$$

$$-(2/3).a_0 + (4/15).a_2 + (74/35).a_4 = 0 \qquad \text{from (3.4.18)}$$

Solving this system of linear equations produces the kernel function

$$K(t) = \begin{cases} (21/80).t^4 - (21/40).t^2 + 249/400, & -1 \leqslant t \leqslant 1 \\ 0 & , \quad \text{o.w} \end{cases} \qquad (3.4.19)$$

so that

$$K^{(2)}(t) = \begin{cases} (252/80).t^2 - 42/40, & -1 \leqslant t \leqslant 1 \\ 0 & , \text{ o.w.} \end{cases} \qquad (3.4.20)$$

Hence,

$$W(t) = K(t) - (1/2).K^{(2)}(t)$$

$$= \begin{cases} (21/80)t^4 - (21/10).t^2 + 459/400, & -1 \leqslant t \leqslant 1 \\ 0 & , \quad \text{o.w.} \end{cases} \qquad (3.4.21)$$

Plots of $W(t)$ based on the standard normal kernel and on (3.4.21) together with the optimal kernel of order (0,4), are shown in figure 3.4. Each of these functions take negative values. The one based on the standard normal has a broader shape than the other two, is negative for $|t| > \sqrt{3}$ and asymptotically approaches zero. The optimal polynomial is more peaked than (3.4.21), crosses the t-axis at ± 0.65 and has minima of -0.27 at $t = \pm 0.85$. (3.4.21) crosses the t-axis at $t = \pm 0.77$ and has minima of -0.69 at ± 1. Both the polynomials are discontinuous at ± 1.

The quantities B, V and T (3.2.8) are given in the following table.

Table 3.4. Asymptotic bias, variance and MISE for kernels of the form $W(t) = K(t) - 0.5.K^2(t)$.

W(t)	B	V	T
$0.5(3-t^2)N(t;0,1)$	-3.000	0.4760	0.6600
$\frac{21}{80}\,t^4 - \frac{21}{10}\,t^2 + \frac{459}{400}$	-0.0826	1.1258	0.6384

In terms of asymptotic MISE there is little loss in using one of these two weight functions instead of the asymptotically optimal kernel of order (0,4) which has $T = 0.6199$. The polynomial version performs slightly better than the minimum variance kernel of order (0,4) which has $T = 0.6432$. The normal version has the advantage that the resulting density estimate, before adjustment for any negative values, has continuous derivatives of all orders.

3.5. Estimating the exact bias.

The exact expected value of a fixed kernel estimator, $\hat{f}(x)$, is given by (3.1.1) and involves the true f which in practice is unknown. It can be replaced by a fixed kernel estimate to provide in turn an estimate of the expectation i.e.

$$\int_{-\infty}^{\infty} h^{-1}K((x-y)/h).\ \hat{f}(y)dy = A(x), \text{ say.} \qquad (3.5.1)$$

An estimate of the bias is then

$$A(x) - \hat{f}(x) \qquad (3.5.2)$$

so that a bias corrected estimate of $f(x)$ is

$$\hat{f}^*(x) = \hat{f}(x) - [A(x) - \hat{f}(x)]$$

$$= 2.\hat{f}(x) - A(x) \qquad (3.5.3)$$

To show that this has asymptotically reduced bias consider:

$$E[\hat{f}^*(x)] = 2.E[\hat{f}(x)] - E[A(x)] \qquad (3.5.4)$$

Now,

$$E[\hat{f}(x)] = f(x) + (1/2).h^2.f^{(2)}(x) + O(h^4) \qquad (3.5.5)$$

from (3.1.6) if $\hat{f}(\cdot)$ is based on using a kernel with $k = 2$. Also,

$$E[A(x)] = \int_{-\infty}^{\infty} h^{-1}.K((x-y)/h).\ E[\hat{f}(y)]dy$$

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

$$+ \int_{-\infty}^{\infty} h^{-1}.K((x-y)/h).(1/2).h^2\ f^{(2)}(y)dy + O(h^4) \qquad (3.5.6)$$

By making the change of variable $t = (x-y)/h$ and using the assumptions that K is symmetric and integrates to one

$$E[A(x)] = f(x) + h^2.f^{(2)}(x) + O(h^4) \qquad (3.5.7)$$

Therefore,

$$E[\hat{f}^*(x)] = f(x) + O(h^4). \qquad (3.5.8)$$

If K is assumed to be a standard normal density then $A(x)$ can be evaluated analytically as follows:

$$A(x) = \int_{-\infty}^{\infty} \{h^{-1}.K((x-y)/h).\hat{f}(y)\}dy$$

$$= \int_{-\infty}^{\infty} \{h^{-1}.K((x-y)/h).n^{-1}.h^{-1} \sum_{i=1}^{n} K((y-x_i)/h\}dy$$

$$= n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \{h^{-1}.K((x-y)/h).h^{-1}K((y-x_i)/h)\}dy$$

$$= n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \{N(x;y,h^2).N(y;x_i,h^2)\}dy$$

$$= n^{-1} \sum_{i=1}^{n} N(x;x_i,2h^2) \qquad (3.5.9)$$

$$= \hat{f}(x;\sqrt{2h}) \quad \text{based on an } N(0,1) \text{ kernel.}$$

Therefore, the bias corrected estimator based on a standard normal kernel is

$$2.\hat{f}(x;h) - \hat{f}(x;\sqrt{2}h) \qquad (3.5.10)$$

which is just a jacknife estimator with $a = \sqrt{2}$. When $a = \sqrt{2}$ $B = 6.000$, $V = 0.4065$ and $T = 0.6690$ and so it is slightly inferior in terms of asymptotic MISE to those kernels with $k = 4$ considered in Section 3.2 - 3.4 - see tables 3.2 and 3.4.

3.6. Subtracting an estimate of $(1/2).h^2.f^{(2)}(x)$.

In the asymptotic expression for the bias of a fixed kernel estimator the principal term when $k = 2$ is $(1/2).h^2.f^{(2)}(x)$. (3.1.6) In this section it is proposed to construct a kernel estimator for this quantity and subtract it from the original estimator.

In order to implement this method a choice of kernel and appropriate smoothing parameter needs to be made for estimating the second derivative. A fixed kernel estimator for $f^{(2)}(x)$ is defined to be

$$\hat{f}^{(2)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_2^3} \cdot K_2\left[\frac{x-X_i}{h_2}\right] \qquad (3.6.1)$$

where K_2 is derived from a twice diffentiable function and satisfies the conditions

$$\int t^j.K_2(t)dt = \begin{cases} 0 & , j = 0,1,3,\ldots,k-1 \\ 2! & , j = 2 \\ \beta_{2,k} \neq 0 & , j = k \end{cases} \qquad (3.6.2)$$

so that the order of the kernel is (2,k) and h_2 is the smoothing parameter (Muller and Gasser (1979)).

If it is assumed that f is four times differentiable then

$$E[\hat{f}^{(2)}(x)] = \frac{1}{h_2^2} \cdot \int K_2(t) \cdot f(x-h_2 t)dt \qquad (3.6.3)$$

and

$$Var(\hat{f}^{(2)}(x)) = \frac{1}{n.h_2^5} \cdot \int K_2(t)^2 \cdot f(x-h_2 t)dt$$

$$- \frac{1}{nh_2^4} \left\{ \int K_2(t) \cdot f(x-h_2)dt \right\}^2 \qquad (3.6.4)$$

Muller and Gasser (1979).

The asymptotic MISE is then

$$\frac{1}{n.h_2^5} \cdot \int K_2(t)^2 dt + \frac{1}{(2!)^2} \cdot h_2^{2(k-2)} \cdot \int f^{(k)}(x)^2 dx$$

$$\cdot \left[\int t^k K_2(t)dt \right]^2 + o\left[nh_2^{-5} + h_2^{2(k-2)}\right] \qquad (3.6.5)$$

The estimator is therefore consistent in MISE if

$$\lim_{n\to\infty} h_2 = 0 \quad \text{and} \quad \lim_{n\to\infty} n.h_2^5 = \infty \qquad (3.6.6)$$

and $f^{(2)}(\cdot)$ is continuous in x. (Muller and Gasser (1979)).

Therefore, if h_2 is proportional to n^{-a}, provided $0 < a < 1/5$, $\hat{f}^{(2)}(x)$ will be consistent for $f^{(2)}(x)$. Hence, differentiating $\hat{f}(x)$ twice will not provide a consistent estimator if the asymptotically optimal value of h, which is proportional to $n^{-1/5}$, is used.

The value of h_2 which minimises the asymptotic MISE (3.6.5) is

$$h_2 = \left[\frac{5}{2.(k-2)} \cdot \frac{V_{2,k}}{B_{2,k}^2} \cdot \frac{1}{\int f^{(k)}(t)^2 dt} \cdot \frac{1}{n} \right]^{\frac{1}{(2k+1)}} \qquad (3.6.7)$$

where

$$V_{2,k} = \int K_2(t)^2 dt$$

and

$$B_{2,k} = \frac{1}{k!} \int t^k K_2(t)dt$$

so that if $k = 4$ h_2 should be chosen proportional to $n^{-1/9}$.

That the method of this section will asymptotically reduce the bias to $O(h^4)$ can be seen as follows:

$$\hat{f}^*(x) = \hat{f}(x) - (1/2) .h^2 \hat{f}^{(2)}(x) \qquad (3.6.8)$$

where h and h_2 are the smoothing parameters used in the kernel estimators, based on the kernel functions K and K_2, of the density and second derivative respectively. Now,

$$E[\hat{f}^*(x)] = E[\hat{f}(x)] - (1/2)h^2 \, . \, E[\hat{f}^{(2)}(x)]$$

$$= f(x) + (1/2) \, h^2 \, f^{(2)}(x) + O(h^4)$$

$$- (1/2) \, h^2 \left\{ f^{(2)}(x) + (h_2^4/4!) \int t^4 K_2(t)dt \, . f^{(4)}(x) + O(h_2^6) \right\}$$

$$= f(x) + O(h^4), \text{ provided } \hat{f}^{(2)}(x) \text{ is consistent} \qquad (3.6.9)$$

Also,

$$V[\hat{f}^*(x)] = V[\hat{f}(x)] + (1/4)h^4 \, . \, V[\hat{f}^{(2)}(x)]$$

$$- h^2 \, Cov\left[\hat{f}(x), \, \hat{f}^{(2)}(x)\right]$$

Now,

$$Cov\left[\hat{f}(x), \, \hat{f}^{(2)}(x)\right]$$

$$= Cov\left[\frac{1}{nh} \sum_{i=1}^{n} K\left[\frac{x-X_i}{h}\right], \, \frac{1}{n \, h_2^3} \sum_{j=1}^{n} K_2\left[\frac{x-X_j}{h_2}\right]\right]$$

$$= E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x-X_i}{h}\right] . \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_2^3} K_2\left[\frac{x-X_i}{h_2}\right]\right]$$

$$- E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left[\frac{x-X_i}{h}\right]\right] E\left[\frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_2^3} K_2\left[\frac{x-X_j}{h_2}\right]\right]$$

Consider the first term:

$$\frac{1}{n^2}\ E\Big[\sum_{i=1}^{n}\frac{1}{h}K\Big[\frac{x-X_i}{h}\Big]\ .\ \sum_{j=1}^{n}\frac{1}{h_2^3}K_2\Big[\frac{x-X_j}{h_2}\Big]\Big]$$

$$=\frac{1}{n^2}\ E\Big[\sum_{i=1}^{n}\frac{1}{h}K\Big[\frac{x-X_i}{h}\Big]\ .\ \frac{1}{h_2^3}K_2\Big[\frac{x-X_i}{h_2}\Big]+\sum_{i\neq j}\frac{1}{h}K\Big[\frac{x-X_i}{h}\Big]\ .\ \frac{1}{h_2^3}K_2\Big[\frac{x-X_j}{h_2}\Big]\Big]$$

$$=\frac{1}{n}E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\ \frac{1}{h_2^3}K_2\Big[\frac{x-Y}{h_2}\Big]\Big]+\frac{(n-1)}{n}\ .\ E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\Big]E\Big[\frac{1}{h_2^3}K_2\Big[\frac{x-Y'}{h_2}\Big]\Big],\ Y\neq Y'.$$

Therefore,

$$\mathrm{Cov}(\hat{f}(x),\ \hat{f}^{(2)}(x))$$

$$=\frac{1}{n}E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\ .\ \frac{1}{h_2^3}K_2\Big[\frac{x-Y}{h_2}\Big]\Big]-\frac{1}{n}E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\Big].E\Big[\frac{1}{h_2^3}K_2\Big[\frac{x-Y'}{h_2}\Big]\Big]$$

$$=\frac{f(x)}{nh_2^3}\int K(t).K_2(ht/h_2)dt+O(n^{-1})$$

$$-\frac{1}{n}\Big[\Big[f(x)+\frac{h^2}{2}f^{(2)}(x)+O(h^4)\Big].\Big[f^{(2)}(x)+\frac{h_2^4}{24}\int t^4K_2(t)dt\ f^{(4)}(x)$$

$$+O(h_2^6)\Big]\Big]$$

using the separate changes of variable $t = (x-y)/h$ and $t = (x-y')/h_2$

$$=\frac{f(x)}{nh_2^3}\int K(t).K_2(ht/h_2)dt+O(n^{-1})$$

provided $h = O(n^{-1/5})$ and $h_2 = O(n^{-1/9})$. Hence,

$$V\big[\hat{f}^*(x)\big]=\frac{f(x)}{nh}\int K(t)^2dt+\frac{h^4}{4}\ .\ \frac{f(x)}{nh_2^5}\int K_2(t)^2dt$$

$$-h^2\ .\ \frac{f(x)}{nh_2^3}\int K(t).K_2(ht/h_2)dt+O(n^{-1}) \qquad (3.6.10)$$

If $h = O(n^{-1/5})$ and $h_2 = O(n^{-1/9})$ then asymptotically

$$V[\hat{f}^*(x)]=\frac{f(x)}{nh}\int K(t)^2dt \qquad (3.6.11)$$

which is the same as for $\hat{f}(x)$ when using a kernel of order $(0,2)$.

The ratio of the asymptotically optimal smoothing parameters for estimating the density and it's second derivative is

$$\frac{h}{h_2} = \frac{\left\{\frac{1}{4} \cdot \frac{V_{0,2}}{B_{0,2}^2} \cdot \frac{1}{\int f^2(t)^2 dt}\right\}^{1/5}}{\left\{\frac{5}{4} \cdot \frac{V_{2,4}}{B_{2,4}^2} \cdot \frac{1}{\int f^4(t)^2 dt}\right\}^{1/9}} \cdot n^{-4/45} \qquad (3.6.12)$$

$$= c.n^{-4/45} \quad , \text{ say.}$$

Therefore,

$$h_2 = c^{-1}.n^{4/45}.h \qquad (3.6.13)$$

but c depends on the unknown quantity

$$\frac{\left[\int f^{(4)}(t)^2 dt\right]^{1/9}}{\left[\int f^{(2)}(t)^2 dt\right]^{1/5}} = r, \text{ say.} \qquad (3.6.13)$$

If it is assumed that both f and K are standard normal densities then

$$r = \frac{(1.85125)^{1/9}}{(0.21157)^{1/5}} = 1.46092.$$

Also,

$$V_{0,2} = \int N(t,1)^2 dt = \frac{1}{2\sqrt{\pi}}$$

$$B_{0,2} = \frac{1}{2!} \cdot \int t^2 N(t,1) dt = \frac{1}{2}$$

$$V_{2,4} = \int N^2(t,1)^2 \, dt = \frac{3}{8\sqrt{\pi}}$$

$$B_{2,4} = \frac{1}{4!} \cdot \int t^4 N^{(2)}(t,1) \, dt = \frac{1}{2} .$$

Hence, for standard normal data choose

$$h_2 = 0.887.n^{4/45}.h \tag{3.6.15}$$

so that more smoothing will be appropriate for estimating $N^{(2)}(x;0,1)$ provided $n > 4$. The asymptotically optimal h value for estimating the standard normal density itself is $1.059.n^{-1/5}$ and substituting this into (3.6.15) gives

$$h_2 = 0.940.n^{-1/9} \tag{3.6.16}$$

If, however an optimal polynomial kernel of order (2,4) is used for estimating the second derivative such as

$$K_2(t) = \begin{cases} (105/16).(-5t^4+6t^2-1), & |t| \leq 1 \\ 0 & , \text{ o.w.} \end{cases} \tag{3.6.17}$$

(see table 3.1) then the optimal h_2 value is

$$h_2 = \left\{\frac{5}{4} \cdot \frac{35}{324^{-1}} \cdot \frac{1}{\int f^{(4)}(x)^2 dx} \cdot \frac{1}{n}\right\}^{1/9} \tag{3.6.18}$$

Also, if a standard normal kernel is used in the estimator of the density then

$$\frac{h}{h_2} = 0.26841 \cdot \frac{\left[\int f^4(x)^2 dx\right]^{1/9}}{\left[\int f^2(x)^2 dx\right]^{1/5}} \cdot n^{-4/45} \tag{3.6.19}$$

If we again assume that f is an $N(0,1)$ density then

$$h_2 = 2.550 \cdot n^{-4/45}.h \tag{3.6.20}$$

$$= 2.701 \cdot n^{-1/9} \tag{3.6.21}$$

if $h = 1.059 \cdot n^{-1/5}$.

Three other underlying distributions each with unit variance but of different shape were considered for calculating r. These are as follows:

(i) Normal mixture ; $f(x) = 0.5N(x;-0.866,0.5)$
$+ 0.5N(x;0.866,0.5)$ - bimodal

(ii) t(3) scaled to have unit variance:

$$f(x) = \frac{2}{\pi} \cdot \frac{1}{(1+x^2)^2} \text{ - long tailed.}$$

(iii) Gamma $(2, \sqrt{2})$: $f(x) = 2.x.e^{-\sqrt{2}\,x}$ - skewed.

The values of $\int f^{(2)}(x)^2 dx$, $\int f^{(4)}(x)^2 dx$ and r are given in table 3.5.

Table 3.5 The values of $\int f^{(2)}(x)^2 dx$, $\int f^{(4)}(x)^2 dx$ and $r = \left[\int f^{(4)}(x)^2 dx\right]^{1/9} \cdot \left[\int f^{(2)}(x)^2 dx\right]^{1/5}$ for various underlying distributions.

Distribution	$\int f^{(2)}(x)^2 dx$	$\int f^{(4)}(x)^2 dx$	r
N(0,1)	0.2116	1.8513	1.4609
Bimodal	3.5536	562.2524	1.5682
Long tailed	3.2229	814.6744	1.6665
Skewed	16.0208	11.3969	0.7525

Despite the integral quantities varying considerably the ratios r are fairly similar, except in the case of the highly skewed Gamma, with the average value being 1.3620. Substituting this value for r into the ratio of smoothing parameters when $N^{(2)}(t;0,1)$ is used as a

kernel for estimating $f^{(2)}(x)$, (3.6.12), gives

$$h_2 = 0.95.n^{4/45}.h \qquad (3.6.22)$$

When the optimal polynomial (3.6.17) is used for K_2 and $r = 1.3620$ is substituted into the ratio of optimal smoothing parameters (3.6.12) it is found that

$$h_2 = 2.74.n^{4/45}.h \qquad (3.6.23)$$

The slightly simpler

$$h_2 = 2.74.n^{1/11}.h \qquad (3.6.24)$$

will be used in practice. This result is very similar to the result when the data is assumed to have come from an $N(0,1)$ distribution (3.6.21).

These formulae based on an average value of r should provide a simple but reasonably effective guide to the amount of smoothing required for estimating $f^{(2)}(x)$ for a variety of underlying distributions.

As regards the best choice of kernel for estimating $f^{(2)}(x)$ the asymptotic MISE's evaluated at the asymptotically optimal h_2 values i.e. $T = (V^4.B^{10})^{1/9}$ can be calculated for the kernel function $N^{(2)}(t;0,1)$ and the optimal polynomial of order (2,4). The ratio of these values gives the asymptotic efficiency. The values of T are

$$N^{(2)}(t;0,1) \quad : T = 7.930 \quad (V = {}^{3}/_{(8\sqrt{\pi})}, \qquad B = 12)$$

Optimal polynomial : $T = 6.685$ (Table 3.2)

so that $T(\text{optimal})/T(\text{Normal}) = 0.843$, i.e. the normal kernel is only 84.3% as efficient as the optimal one. This is in contrast to the very high efficiencies of suboptimal kernels for estimating the density

itself.

Consider now the situation when the variance is not equal to one. Let the random variable X have unit variance and pdf $f(x)$ and let $Y = \alpha.X$ where α is a scalar.

Then, the pdf of Y, g say, is

$$g(y) = \alpha^{-1}.f(\alpha^{-1}.y)$$

with derivatives:

$$g^{(1)}(y) = \alpha^{-2}.f^{(1)}(\alpha^{-1}.y)$$

$$g^{(2)}(y) = \alpha^{-3}.f^{(2)}(\alpha^{-1}.y)$$

$$g^{(3)}(y) = \alpha^{-4}.f^{(3)}(\alpha^{-1}.y)$$

$$g^{(4)}(y) = \alpha^{-5}.f^{(4)}(\alpha^{-1}.y).$$

Hence,

$$\int g^{(2)}(y)^2 dy = \int \alpha^{-6} f^{(2)}(\alpha^{-1}.y)dy$$

$$= \alpha^{-5} \int f^{(2)}(t)^2 dt$$

and

$$\int g^{(4)}(y)^2 dy = \int \alpha^{-10} .f^{(4)}(\alpha^{-1}.y)dy$$

$$= \alpha^{-9} \int f^{(4)}(t)^2 dt$$

where $t = \alpha^{-1}.y$.

Therefore,

$$\left[\int g^{(4)}(y)^2 dy\right]^{1/9} . \left[\int g^{(2)}(y)^2 dy\right]^{-1/5}$$

$$= \left[\alpha^{-9}. \int f^{(4)}(t)^2 dt\right]^{1/9} . \left[\alpha^{-5} \int f^{(2)}(t)^2 dy\right]^{-1/5}$$

$$= \left[\int f^{(4)}(t)^2 dt\right]^{1/9} . \left[\int f^{(2)}(t)^2 dt\right]^{-1/5} = r.$$

Therefore when the variance is not one the smoothing parameter for the density, which will itself involve an estimate of the standard deviation, can be multiplied by the same factor as in the unit variance case.

Muller et al (1987) describe a number of methods of smoothing parameter choice for derivatives in a nonparametric regression context. Their "factor method" is the one that might best also be applied to derivatives of density functions. If we choose kernels K and K_ν for estimating the density and its ν^{th} derivative, respectively with both having the same k value in (3.2.5) then the ratio of optimal smoothing parameters is

$$\frac{h}{h_\nu} = \left[\frac{(2\nu+1)k}{k-\nu} \cdot \frac{V_{\nu,k} \cdot B_{o,k}^2}{V_{o.k.} \cdot B_{\nu,k}^2} \right]^{\frac{1}{2k+1}}$$

$$= d_{\nu,k} \qquad (3.6.25)$$

So, by using the same value of k the terms in the ratio dependent on derivatives of the unknown density cancel out leaving a constant which depends only on the kernels K and K_ν.

For the case $\nu = 2$ it is necessary to use

$$K(t) = {}^{15}/_{32}.(7t^4 - 10t^2 + 3), \quad |t| < 1 \qquad (3.6.26)$$

and

$$K_2(t) = ({}^{105}/_{16})(-5t^4 - 6t^2 + 1), \quad |t| < 1 \qquad (3.6.27)$$

which both have $k = 4$ and result in the factor $d_{2,4} = 0.8919$. Table 1 of their paper lists the factors $d_{\nu,k}$ for various values of ν and k.

Hence, if using this method to find a value for h_2 a choice of h for estimating the density using a kernel with $k = 4$ needs to be made. As discussed in Section 3.2. an estimator of the density based on a kernel with $k = 4$ will have bias = $O(h^4)$ which is the same order as when an estimator of $(1/2)h^2f^{(2)}(x)$ is subtracted from a kernel estimator with $k = 2$. Also, because (3.6.26) has discontinuities at ± 1 the resulting estimate will itself be discontinuous. Therefore, as we are trying to bias correct an estimator with bias of order h^2 it is perhaps better to base the choice of smoothing parameter for use with K_2 (3.6.27) on the degree of smoothing employed with a kernel having $k = 2$ and continuous derivatives such as a standard normal kernel. The small sample behaviour of subtracting bias and using a kernel with $k = 4$ will be investigated further in the simulation study of Section 3.8.

Using the factor method also requires more computation as smoothing parameters have to be chosen for kernels with $k = 2$ and $k = 4$. Muller et al (1987) suggest using cross-validation. On the other hand use of a formula such as (3.6.24) requires only one smoothing parameter choice for estimation of the density. However, similar values of h_2 are in fact obtained if instead of cross-validation the normal optimal formulae derived by finding those h-values which minimise the exact MISE for a variety of sample sizes are used i.e.

$$\text{N(0,1) kernel} : h = 1.2.n^{-0.214} \cdot \hat{\sigma} \qquad (3.6.28)$$

$$\text{Optimal polynomial } (k = 4) : h = 3.904.n^{-0.134} \cdot \hat{\sigma} \qquad (3.6.29)$$

When the optimal polynomial (3.6.27) is used for K_2 the factor method gives

$$h_2 = (0.8919)\ (3.904.n^{-0.134}.\ \hat{\sigma})$$
$$= 3.382.n^{-0.134}.\ \hat{\sigma} \qquad (3.6.30)$$

whereas using (3.6.24) results in

$$h_2 = (2.7.n^{1/11}).(1.2.n^{-0.214}).\ \hat{\sigma}$$
$$= 3.24.n^{-0.123}.\ \hat{\sigma} \qquad (3.6.31)$$

Using the factor method in the form (3.6.30) results in smaller h_2 values than found using (3.6.31) for all $n \geq 45$. The differences however, are not large. For example if $n = 100$ $h_2 = 1.825$ when the factor method (3.6.30) is used and 1.839 if (3.6.31) is used. When the sample size is much larger the differences are still small - at $n = 1000$ $h_2 = 1.340$ using (3.6.30) and 1.385 when using (3.6.31). These small differences should have little effect in practice.

A problem with subtracting an estimate of the bias from a non-negative estimate is that it may result in a negative estimate of density as is the case with the optimal kernels, which have negative sidelobes, of Section 3.2. Such an estimate must be reset to zero but the results of the simulation study given in table 3.6 indicate that if no rescaling takes place this will make little difference. The details of the study are as described in Section 3.2 and the same seed was used for the random number generator so that results can be directly compared. The simulated data sets were smoothed using smoothing parameters calculated from the exact optimal formula for standard normal data.

Table 3.6. The mean and standard deviation of the areas of density estimates based on subtracting an estimate of the asymptotic bias from a fixed normal kernel estimate with no rescaling and for 25 simulations.

Distribution	N = 50		N = 100	
	Mean	St. dev.	Mean	St. dev.
N(0,1)	1.009	0.000024	1.004	0.000007
Gamma $(2, \sqrt{2})$	1.002	0.000003	1.001	0.000000
0.5N(x;-0.866,0.5) +0.5N(x;0.866,0.5)]	1.020	0.000032	1.018	0.000021
t(3)	1.003	0.000014	1.001	0.000009

These means and standard deviations are all less than the corresponding results when using the optimal kernel of order (0,4). (Table 3.3). The largest mean errors occur again when estimating the bimodal normal mixture but at only 2% for $n = 50$ and 1.8% for $n = 100$ these are still very small.

3.7. Relaxing the integral constraint.

In order to improve the rate of convergence in MISE of the kernel mmethod from $O(n^{-4/5})$ to rates like $O(n^{-8/9})$ Terrell and Scott (1980) consider relaxing the constraint that the density estimate should integrate to one rather than allowing the kernel function to take negative values. They limit the choice of kernel to be symmetric and non-negative so that the odd moments in the Taylor series expansion of $E[\hat{f}(x)]$ are all zero. Equation(3.1.6) may then be written in the form:

$$E[\hat{f}(x)] = f(x).\left[1 + \frac{a_2}{f(x)} . h^2 + \frac{a_4}{f(x)} .h^4 +...\right] \qquad (3.7.1)$$

where $a_i = (-1)^i . f^{(i)} \left[\int t^i . K(t) dt \right] / i!$.

Taking logarithms, applying the series expansion for natural logarithms and ignoring higher order terms in h then h^4 gives:

$$\log\{E[\hat{f}(x)]\} = \log(f(x)) . + \frac{a_2}{f(x)} . h^2 + \frac{\left[a_4 f(x) - (1/2) a_2^2\right] h^4}{f(x)^2} \qquad (3.7.2)$$

The term in h^2 can be eliminated by considering the linear combination:

$$(4/3) . \log\{E[\hat{f}(x;h)]\} - (1/3) . \log\{E[\hat{f}(x;2h)]\}$$

$$= \log f(x) - \frac{\left[4a_4 f(x) - 2a_2^2\right] h^4}{f(x)^2} \qquad (3.7.3)$$

Taking exponentials and using a series expansion for the exponential function then gives:

$$E[\hat{f}(x;h)]^{4/3} . E[\hat{f}(x;2h)^{-1/3}$$

$$= f(x) + \frac{\left[2a_2^2 - 4a_4 f(x)\right]}{f(x)} . h^4 \qquad (3.7.4)$$

The estimator they propose is then the ratio of two non-negative kernel estimators i.e.

$$\hat{f}^*(x) = \hat{f}(x;h)^{4/3} . \hat{f}(x;2h)^{-1/3}$$

$$= \hat{f}(x;h) . \left\{\hat{f}(x;h)/\hat{f}(x;2h)\right\}^{2/3} \qquad (3.7.5)$$

so that the term in brackets is a multiplicative correction factor for $\hat{f}(x;h)$. The value of $\hat{f}^*(x)$ is taken to be zero if $\hat{f}(x;h)$ is zero.

A fixed kernel estimator based on smoothing parameter h tends to systematically overestimate in the tails and underestimate in the

main body of the distribution. Increasing the value of h tends to exaggerate this effect. Therefore, the effect of the correction factor should be to increase $\hat{f}(x;h)$ in the main body and decrease it in the tails and so thereby reducing bias.

Terrell and Scott show that $\hat{f}^*(x)$ does indeed have bias of order h^4 and variance of order $(nh)^{-1}$. In fact,

$$E[\hat{f}^*(x)] = f(x) + \frac{2a_2^2 - 4a_4 f(x)}{f(x)} \cdot h^4 + o(h^4) \qquad (3.7.6)$$

and

$$V[\hat{f}^*(x)] = V\left[\frac{4}{3}\hat{f}(x;h) - \frac{1}{3}\hat{f}(x;2h)\right] + O(\frac{1}{n})$$

$$= \frac{16}{9} V[\hat{f}(x;h)] + \frac{1}{9} V\left[\hat{f}(x;2h)\right] - \frac{8}{9} \mathrm{Cov}\left[\hat{f}(x;h),\ \hat{f}(x;2h)\right] + O(\frac{1}{n})$$

Now, for any non-zero constant k,

$$V[\hat{f}(x;kh)] = \frac{f(x)}{knh} \int K(t)^2 dt + O(\frac{1}{n}) \qquad (3.7.7)$$

$$\mathrm{Cov}\left[\hat{f}(x;h), \hat{f}(x;kh)\right] = \frac{1}{n}\left[\int \frac{1}{kh^2} \cdot K\left[\frac{x-y}{h}\right] K\left[\frac{x-y}{kh}\right] f(y)dy\right.$$

$$\left. - \int \frac{1}{h} K\left[\frac{x-y}{h}\right] f(y)dy \cdot \int \frac{1}{kh} K\left[\frac{x-y}{kh}\right] f(y)dy\right]$$

Making the change of variable $t = (x-y)/h$ and expanding as Taylor series gives:

$$\mathrm{Cov}[\hat{f}(x;h), \hat{f}(x;2h]$$

$$= \frac{1}{n}\left[\int \frac{1}{kh} K(t).K(t/k) \left\{f(x) - ht\, f^{(1)}(x) + \frac{h^2t^2}{2} f^{(2)}(x)\right.\right.$$

$$\left.\left. + o(h^2)\right\}dt\right] - \frac{1}{n}\left[((f(x) + o(h^2))(f(x) + o(h^2))\right]$$

$$= \frac{1}{knh} \cdot f(x).\int K(t).K(t/k)dt + O(\frac{1}{n}) \qquad (3.7.8)$$

Hence,

$$V[\hat{f}^*(x)] = \frac{16}{9} \cdot \frac{f(x)}{nh} \cdot \int K(t)^2 dt$$

$$+ \frac{1}{9} \cdot \frac{f(x)}{2nh} \cdot \int K(t)^2 dt - \frac{8}{9} \cdot \frac{f(x)}{2nh} \cdot \int K(t)\, K(t/2) dt$$

$$+ O(\frac{1}{n}) \qquad (3.7.9)$$

These results can be used to derive an asymptotically optimal smoothing strategy when using a standard normal kernel and also assuming that the underlying distribution is standard normal.

For $K(t) = N(t;0,1)$ we have $a_2 = (1/2).f^{(2)}(x)$ and $a_4 = (1/8).f^{(4)}(x)$ so that

$$\text{bias } \hat{f}^*(x) = \frac{(1/2).\left\{N^{(2)}(x;0,1)^2 - N^{(4)}(x;0,1).N(x;0,1)\right\} h^4}{N(x;0,1)}$$

$$= N(x;0,1).(2x^2-1).h^4 \qquad (3.7.10)$$

Therefore,

$$\text{bias}^2(x) = N(x;0,0.5).\frac{1}{2\sqrt{\pi}}.(4x^4-4x^2+1)h^8 \qquad (3.7.11)$$

so that

$$\int \text{bias}^2(x)dx = \frac{2.h^8}{2\sqrt{\pi}} = \frac{h^8}{\sqrt{\pi}} \qquad (3.7.12)$$

Also,

$$V[\hat{f}^*(x)] = f(x).\left[\frac{16}{9nh} \cdot \int N(t;0,1)^2 dt + \frac{1}{18nh} \int N(t;0,1)^2 dt\right.$$

$$\left. - \frac{8}{18nh} \cdot \int N(t;0,1).2.N(t;0,4)dt\right] \qquad (3.7.13)$$

$$[\text{Nb. } K(t) = N(t;0,1) \Rightarrow K(t/2) = 2.N(t;0,4)].$$

Hence,

$$V[\hat{f}^*(x)] = f(x).\left[\frac{0.3586}{nh}\right] \qquad (3.7.14)$$

and

$$\int V[\hat{f}^*(x)]dx = \frac{0.3586}{nh} \qquad (3.7.15)$$

Combining the expressions for the integrated squared bias (3.7.12) and integrated variance (3.7.15) we have:

$$\text{MISE} = \frac{h^8}{\sqrt{\pi}} + \frac{0.3586}{nh} \qquad (3.7.16)$$

Differentiating (3.7.16) with respect to h, setting the result equal to zero and then solving for h gives:

$$h = 0.7547.n^{-1/9} \qquad (3.7.17)$$

Substituting this optimal h-value into the MISE (3.7.16) gives the minimum MISE of $0.418.n^{-8/9}$.

In their paper Terrell and Scott consider devising a smooth strategy for standard normal data but using a uniform kernel i.e.

$$K(t) = \begin{cases} 1, & |t| < \frac{1}{2} \\ 0, & \text{O.W.} \end{cases} \qquad (3.7.18)$$

However, there are some errors in their calculations. For the bias we have $a_2 = f^{(2)}(x)/_{24}$ and $a_4 = f^{(4)}(x)/_{1920}$. Using (3.7.4) then gives:

$$\text{bias } \hat{f}^*(x) = \frac{1}{720}(x^4 + 4x^2 - 2).\ N(x;0,1)h^4 \qquad (3.7.19)$$

By (3.7.9) the variance is:

$$\text{Var}\{\hat{f}^*(x)\} = 25.N(x;0,1)/(18nh) \qquad (3.7.20)$$

A straightforward calculation then shows that the MISE is given by

$$\text{MISE} = \frac{25}{18nh} + \frac{17}{663552\sqrt{\pi}}.\ h^8 \qquad (3.7.21)$$

which is minimised by

$$h = 2.840.n^{-1/9} \qquad (3.7.22)$$

with a resulting minimum MISE = $0.611.n^{-8/9}$ which is 46% larger then the value obtained when a standard normal kernel is used - a marked drop in performance.

Because of the multiplicative correction factor, use of the estimator $\hat{f}^*(x)$ (3.7.5) will result in an estimate which does not integrate to one. In fact Terrell and Scott remark that the area will always converge to one from above for any sampling density. To check on the size of error the simulation study carried out in Sections 3.2 and 3.6 was again undertaken using the estimator (3.7.5) with the smoothing parameter chosen by (3.7.17) multiplied by a robust measure of scale. The results are given in table 3.7.

Table 3.7. The mean and standard deviation of the areas of density estimates found using the estimator $\hat{f}^*(x)$ with a standard normal kernel. 25 simulations were carried out.

Distribution	N = 50		N = 100	
	Mean	St. dev.	Mean	St. dev.
N(0,1)	1.026	0.000049	1.017	0.000014
Gamma $(2, \sqrt{2})$	1.010	0.000015	1.008	0.000006
0.5N(x;-0.866,0.5) + 0.5N(x;0.866,0.5)	1.039	0.000023	1.034	0.000017
t(3)	1.021	0.000016	1.015	0.000012

The mean errors are all at least 0.8% with the largest at over 3% again for the bimodal normal mixtures. These means and standard

deviations are consistently higher than those found in Section 3.2 and 3.6.

The approach of this section may be generalised by considering a sequence of non-negative and symmetric kernel density estimates $\{\hat{f}(x;ih);\ i=1,\ldots,s\}$ and taking a multiplicative combination of these with i^{th} exponent

$$(-1)^{(i-1)} \cdot \frac{2.s.(s-1)....(s-i+1)}{(s+1)(s+2)....(s+i)} . \qquad (3.7.23)$$

The resulting non-negative estimator will have asymptotic MISE = $O(n^{-4s/(4s-1)})$.

3.8. Simulation study.

In Sections 3.2-3.7 six methods for constucting density estimators which reduce asymptotic bias and MISE have been discussed. These each fall into at least one of three approaches to this problem which are:

(i) Subtracting a bias reducing correction factor.

(ii) Using a kernel which can take negative values so that certain higher order moments than the second are zero. (i.e. $k > 2$ in (3.2.5)).

(iii) Using a multiplicative correction factor.

A number of the six methods clearly only fall into one of these three categories. The first is when subtracting an estimate of the principal asymptotic bias term $(1/2)h^2f^{(2)}(x)$ where the estimate of $f^{(2)}(x)$ utilises a different degree of smoothing to that used for estimating the density itself (Section 3.6). Secondly, there are the minimum variance and optimal kernels of Gasser et al (1985) and thirdly the method of Terrell and Scott (1980) which falls into category (iii).

However, each of the other three methods (Jacknifing, using the weight function $W(t) = K(t) - (1/2)\, K^{(2)}(t)$ and estimating the exact bias) can be placed into either categories (i) or (ii). Also, each of these three methods were shown to be equivalent to another of these three methods for certain choices of kernel and/or relevant parameters. Subtracting an estimate of the exact bias is equivalent to Jacknifing using normal kernels and the suboptimal value of the parameter $a = \sqrt{2}$. Jacknifing was shown to be equivalent to using a kernel of order 4 but when again using normal kernels and optimally letting $a \longrightarrow 1$ it is also equivalent to using the weight function $W(t) = K(t) - 1/2K^{(2)}(t)$ with $K(\cdot)$ taken to be the standard normal density. $W(t)$ constructed in this way performs slightly worse asymptotically than the minimum variance and optimal kernels of order 4. However, the polynomial version of $W(t)$ (3.4.21) does perform asymptotically better than the minimum variance kernel but is still slightly inferior to the optimal kernel.

In the simulation study then, which is intended to examine small sample performance, three of the methods discussed in this chapter are included. These are either the only method or those which have the best asymptotic properties in each of the above three categories and are:

1. Subtracting an estimate of $(1/2)h^2f^{(2)}(x)$ from an estimate based on an N(0,1) kernel. The estimate of the second derivative will use the optimal kernel of order (2,4) i.e.

$$K(t) = \begin{cases} (105/6)(-5t^4+6t^2-1) , & |t| < 1 \\ 0 , & \text{otherwise} . \end{cases}$$

2. Using the optimal kernel of order 4, i.e.

$$K(t) = \begin{cases} (15/32)(7t^4-10t^2-3) , & |t| < 1 \\ 0 , & \text{otherwise} . \end{cases}$$

3. The method of Terrell and Scott (1980) with the two estimates both using an N(0,1) kernel i.e.

$$\hat{f}^*(x) = \hat{f}(x,h) \cdot \left[\hat{f}(x,h)/\hat{f}(x,2h)\right]^{2/3}$$

In addition, the following three methods are also included.:

4. Using a simple fixed N(0,1) kernel estimate.

5. The adaptive method with $\alpha = \frac{1}{2}$ based on an N(0,1) kernel and a pilot estimate constructed using method 4.

6. The adaptive method with $\alpha = 1$ which is again based on an N(0,1) kernel and a method 4 pilot estimate.

The fixed kernel method 4 is included so that the performances of the other asymptotically superior methods can be directly compared. The two adaptive methods are also included because it was demonstrated in chapter 2 that their ideal versions can reduce both bias and MISE. In this study the performances of their feasible versions will be assessed.

Data were simulated from 4 distributions each having a different density shape. These are:

1. Standard normal.
2. Gamma $(2, \sqrt{2})$ - skewed.
3. 0.5.N(-0.866,0.5) + 0.5N(0.866,0.5) - bimodal normal mixture.
4. Student's t(3) - long tailed.

In chapter 2, method performance was assessed via exact calculations of MISE using numerical integration. The aim of the simulation study

techniques in this chapter are to try and reflect the performances as maybe realised in actual real applications of the methods. To this end empirical loss functions and an easily implemented practical smoothing parameter choice based on normality are used. Such an approach also enables feasible versions of the adaptive methods to be easily included.

Samples of size 50 and 100 were used. For a particular random sample the underlying density is estimated at x-values equally spaced at intervals of 0.2. For the distributions 1-4 the densities were estimated in (-4,4), (0,10), (-5,5) and (-6,6) respectively and therefore resulted in estimates at 40, 50, 50 and 60 points respectively. This was repeated for 1000 random samples in each case thus producing 1000 estimates of the density at each x-value. These were used to construct empirical estimates of $E[\hat{f}(x_k)]$ and $E[\hat{f}(x_k)^2]$, $k = 1,\ldots,M$ where M denotes the number of distinct x-values at which the density is estimated for a given distribution. The next step was to use these in estimates of the loss functions average squared bias (ASB) and average variance (AVAR) which can be combined to estimate the average mean squared error (AMSE), i.e.

$$\text{ASB} = \frac{1}{M} \sum_{k=1}^{M} \left[\frac{1}{1000} \sum_{j=1}^{1000} \hat{f}_j(x_k) - f(x_k)\right]^2 \qquad (3.8.1)$$

$$\text{AVAR} = \frac{1}{M} \sum_{k=1}^{M} \left[\frac{1}{1000} \sum_{j=1}^{1000} \hat{f}_j(x_k)^2 - \left[\frac{1}{1000} \sum_{j=1}^{1000} \hat{f}_j(x_k)\right]^2\right] \qquad (3.8.2)$$

$$\text{ASME} = \text{ASB} + \text{AVAR} \qquad (3.8.3)$$

Here $f(\cdot)$ denotes the true density and $\hat{f}_j(\cdot)$ the estimate based on the j^{th} random sample.

This procedure was then in turn repeated 10 times when $n = 50$ and 5 times when $n = 100$. The averages of the resulting 10 (or 5) ASB's, AVAR's and AMSE's, to be denoted AASB, AAVAR and AAMSE respectively, were used to compare performances with the standard errors providing a measure of stability. The main computational effort in this whole procedure has gone into obtaining good estimates of $E[\hat{f}(x_k)]$ and $E[\hat{f}(x_k)^2]$.

The smoothing strategy employed for each method is based on the appropriate optimal formula for smoothing data from a standard normal distribution. These are as follows:

Method.

1. $h_{Nopt} = 1.2n^{-0.214}\,\hat{\sigma}$ with $h_2 = 2.7h_{Nopt}$
2. $h_{Nopt} = 3.904\,n^{-0.134}\,\hat{\sigma}$
3. $h_{Nopt} = 0.75\,n^{-1/9}\,\hat{\sigma}$
4. $h_{Nopt} = 1.2\,n^{-0.214}\,\hat{\sigma}$
5. $h_{Nopt} = 0.9\,n^{-0.235}\,\sqrt{\hat{\sigma}}$
6. $h_{Nopt} = 0.26\,n^{-0.042}$

For methods 1-5 $\hat{\sigma}$ represents a robust measure of scale. An estimate based on the median of the absolute deviation (MAD) was used. (Hogg (1979)), i.e.

$$\hat{\sigma} = \text{median}\,|\,x_i - \text{median}\,(x_i)\,|/0.6745 \qquad (3.8.4)$$

Method 6 is scale invariant as discussed in chapter 2.

As remarked in chapter 2, the simulation study of Bowman (1985) shows that normal optimal smoothing provides a simple but effective

guide. It is most effective for unimodal densities but also has some success in the presence of bimodality as long as the modes are not too highly separated. In addition though, results were also obtained when using the above formulae scaled by the factors $^2/_3$ and $^4/_3$.

Data samples from the Gamma $(2, \sqrt{2})$ distribution only take positive values so that the density function is zero for negative x. Hence, we also require $\hat{f}(x)$ to be zero for all negative x. The density should therefore only be estimated at positive x-values but, for example, if a fixed normal kernel is used an estimate which integrates to one can only be obtained by constructing the estimate for x-values greater than about -4h. Solutions to this problem are discussed by Silverman (1986, p.29-32). The one chosen to be used here is to reflect the data in the origin and then estimate the density for $x > 0$ using the data set of size $2n$ but still with the value $^1/_n$ in the estimator. The effect of this is to add the size of the invalid positive estimate for certain negative x-values to the estimate at the corresponding positive x's. This then results in an estimate which does integrate to one over positive x and satisfies $\hat{f}^{(1)}(0+) = 0$.

The seeds for the random number generator were chosen so that for a particular sample size n, and distribution the seed is the same. This results in the same data sets being sampled and enables direct comparisons to be made both between the methods as well as within methods when using different amounts of smoothing. The full results are contained in tables 3.8-3.15. Also, for sample size 50 they are illustrated in figures 3.5-3.16.

The AASB results for each distribution, except the highly skewed Gamma $(2, \sqrt{2})$, indicate the effectiveness of subtracting an estimate

Table 3.8. Values of AASB, AAVAR and AAMSE for samples of size 50 from an N(0,1) distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	4.52×10^{-6} (6.4×10^{-7})	2.13×10^{-3} (1.9×10^{-5})	2.13×10^{-3}
	2	1.15×10^{-5} (1.0×10^{-6})	1.52×10^{-3} (1.5×10^{-5})	1.53×10^{-3}
	3	7.76×10^{-6} (6.7×10^{-7})	2.26×10^{-3} (2.0×10^{-5})	2.26×10^{-3}
	4	6.35×10^{-5} (2.4×10^{-6})	1.58×10^{-3} (1.5×10^{-5})	1.64×10^{-3}
	5	6.86×10^{-5} (1.8×10^{-6})	1.33×10^{-3} (1.4×10^{-5})	1.41×10^{-3}
	6	3.11×10^{-4} (5.8×10^{-6})	1.51×10^{-3} (1.5×10^{-5})	1.82×10^{-3}
1	1	2.34×10^{-5} (1.5×10^{-6})	1.30×10^{-3} (1.4×10^{-5})	1.33×10^{-3}
	2	1.27×10^{-4} (3.5×10^{-6})	9.54×10^{-4} (1.1×10^{-5})	1.08×10^{-3}
	3	3.32×10^{-5} (1.5×10^{-6})	1.39×10^{-3} (1.4×10^{-5})	1.43×10^{-3}
	4	2.86×10^{-4} (4.8×10^{-6})	9.74×10^{-4} (1.1×10^{-5})	1.26×10^{-3}
	5	2.89×10^{-4} (3.4×10^{-6})	8.99×10^{-4} (1.0×10^{-5})	1.19×10^{-3}
	6	4.65×10^{-4} (3.2×10^{-6})	9.39×10^{-4} (1.0×10^{-5})	1.40×10^{-3}
4/3	1	1.45×10^{-4} (3.8×10^{-6})	9.45×10^{-4} (1.1×10^{-5})	1.09×10^{-3}
	2	6.85×10^{-4} (8.1×10^{-6})	7.44×10^{-4} (1.0×10^{-5})	1.43×10^{-3}
	3	1.42×10^{-4} (3.4×10^{-6})	1.00×10^{-3} (1.1×10^{-5})	1.15×10^{-3}
	4	7.66×10^{-4} (7.6×10^{-6})	7.06×10^{-4} (8.8×10^{-6})	1.47×10^{-3}
	5	9.70×10^{-4} (7.8×10^{-6})	6.53×10^{-4} (8.2×10^{-6})	1.62×10^{-3}
	6	1.47×10^{-3} (8.2×10^{-6})	6.05×10^{-4} (6.8×10^{-6})	2.08×10^{-3}

Table 3.9. Values of AASB, AAVAR and AAMSE for samples of size 50 from a Gamma $(2, \sqrt{2})$ distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	1.20×10^{-2} (3.5×10^{-5})	1.04×10^{-3} (6.9×10^{-6})	1.31×10^{-2}
	2	1.23×10^{-2} (3.5×10^{-5})	7.95×10^{-4} (6.5×10^{-6})	1.31×10^{-2}
	3	1.19×10^{-2} (3.5×10^{-5})	1.09×10^{-3} (6.9×10^{-6})	1.30×10^{-2}
	4	1.20×10^{-2} (3.2×10^{-5})	7.78×10^{-4} (5.7×10^{-6})	1.27×10^{-2}
	5	1.19×10^{-2} (3.3×10^{-5})	6.82×10^{-4} (6.0×10^{-6})	1.26×10^{-2}
	6	1.11×10^{-2} (3.5×10^{-5})	7.73×10^{-4} (7.3×10^{-6})	1.19×10^{-2}
1	1	1.21×10^{-2} (3.3×10^{-5})	6.85×10^{-4} (6.0×10^{-6})	1.28×10^{-2}
	2	1.20×10^{-2} (2.8×10^{-5})	5.04×10^{-4} (5.2×10^{-6})	1.25×10^{-2}
	3	1.20×10^{-2} (3.3×10^{-5})	7.03×10^{-4} (6.0×10^{-6})	1.27×10^{-2}
	4	1.21×10^{-2} (2.9×10^{-5})	4.90×10^{-4} (4.5×10^{-6})	1.26×10^{-2}
	5	1.16×10^{-2} (2.9×10^{-5})	4.54×10^{-4} (4.4×10^{-6})	1.20×10^{-2}
	6	1.02×10^{-2} (3.0×10^{-5})	5.10×10^{-4} (5.2×10^{-6})	1.07×10^{-2}
4/3	1	1.20×10^{-2} (2.9×10^{-5})	4.85×10^{-4} (5.0×10^{-6})	1.24×10^{-2}
	2	1.17×10^{-2} (2.4×10^{-5})	3.29×10^{-4} (3.6×10^{-6})	1.20×10^{-2}
	3	1.20×10^{-2} (3.0×10^{-5})	5.02×10^{-4} (4.9×10^{-6})	1.25×10^{-2}
	4	1.22×10^{-2} (2.5×10^{-5})	3.41×10^{-4} (3.6×10^{-6})	1.26×10^{-2}
	5	1.12×10^{-2} (2.5×10^{-5})	3.15×10^{-4} (3.2×10^{-6})	1.15×10^{-2}
	6	1.01×10^{-2} (2.5×10^{-5})	3.35×10^{-4} (3.7×10^{-6})	1.04×10^{-2}

Table 3.10. Values of AASB, AAVAR and AAMSE for samples of size 50 from a $[0.5N(-0.866,0.5^2) + 0.5N(0.866,0.5^2)]$ distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	6.18×10^{-4} (4.3×10^{-6})	1.30×10^{-3} (7.9×10^{-6})	1.92×10^{-3}
	2	1.94×10^{-3} (6.5×10^{-6})	9.52×10^{-4} (6.4×10^{-6})	2.89×10^{-3}
	3	4.96×10^{-4} (3.9×10^{-6})	1.40×10^{-3} (8.2×10^{-6})	1.90×10^{-3}
	4	1.13×10^{-3} (4.5×10^{-6})	9.42×10^{-3} (5.6×10^{-6})	2.07×10^{-3}
	5	1.75×10^{-3} (5.2×10^{-6})	9.17×10^{-4} (5.7×10^{-6})	2.67×10^{-3}
	6	2.21×10^{-3} (6.3×10^{-6})	1.14×10^{-3} (7.3×10^{-6})	3.35×10^{-3}
1	1	2.24×10^{-3} (5.5×10^{-6})	7.87×10^{-4} (5.5×10^{-6})	3.03×10^{-3}
	2	3.78×10^{-3} (4.2×10^{-6})	5.43×10^{-4} (4.3×10^{-6})	4.33×10^{-3}
	3	1.73×10^{-3} (5.0×10^{-6})	8.61×10^{-4} (5.7×10^{-6})	2.59×10^{-3}
	4	2.56×10^{-3} (5.1×10^{-6})	5.59×10^{-4} (3.8×10^{-6})	3.12×10^{-3}
	5	3.42×10^{-3} (4.3×10^{-6})	6.04×10^{-4} (4.3×10^{-6})	4.03×10^{-3}
	6	3.77×10^{-3} (3.7×10^{-6})	7.84×10^{-4} (5.6×10^{-6})	4.56×10^{-3}
4/3	1	3.23×10^{-3} (3.7×10^{-6})	5.26×10^{-4} (4.1×10^{-5})	3.75×10^{-3}
	2	4.16×10^{-3} (5.0×10^{-6})	3.74×10^{-4} (3.1×10^{-6})	4.53×10^{-3}
	3	2.86×10^{-3} (4.5×10^{-6})	5.93×10^{-4} (4.4×10^{-6})	3.45×10^{-3}
	4	3.74×10^{-3} (5.9×10^{-6})	3.81×10^{-4} (2.9×10^{-6})	4.12×10^{-3}
	5	4.56×10^{-3} (5.5×10^{-6})	4.18×10^{-4} (3.1×10^{-6})	4.98×10^{-3}
	6	4.91×10^{-3} (5.1×10^{-6})	5.15×10^{-4} (3.8×10^{-6})	5.42×10^{-3}

Table 3.11. Values of AASB, AAVAR and AAMSE for samples of size 50 from a t(3) distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	4.51×10^{-6} (9.9×10^{-7})	1.28×10^{-3} (7.5×10^{-6})	1.29×10^{-3}
	2	1.88×10^{-5} (2.0×10^{-6})	9.49×10^{-4} (5.7×10^{-6})	9.67×10^{-4}
	3	6.19×10^{-6} (1.0×10^{-6})	1.36×10^{-3} (7.8×10^{-6})	1.36×10^{-3}
	4	5.93×10^{-5} (3.7×10^{-6})	9.72×10^{-4} (5.5×10^{-6})	1.03×10^{-3}
	5	1.72×10^{-5} (1.8×10^{-6})	7.84×10^{-4} (4.6×10^{-6})	8.01×10^{-4}
	6	4.61×10^{-5} (1.8×10^{-6})	8.39×10^{-4} (5.2×10^{-6})	8.85×10^{-4}
1	1	4.43×10^{-5} (3.0×10^{-6})	8.27×10^{-4} (5.0×10^{-6})	8.71×10^{-4}
	2	1.99×10^{-4} (5.6×10^{-6})	6.37×10^{-4} (4.2×10^{-6})	8.35×10^{-4}
	3	3.95×10^{-5} (2.8×10^{-6})	8.67×10^{-4} (5.2×10^{-6})	9.07×10^{-4}
	4	2.42×10^{-4} (6.6×10^{-6})	6.25×10^{-4} (3.8×10^{-6})	8.68×10^{-4}
	5	1.63×10^{-4} (5.2×10^{-6})	5.37×10^{-4} (3.4×10^{-6})	7.00×10^{-4}
	6	2.15×10^{-4} (5.1×10^{-6})	5.21×10^{-4} (3.4×10^{-6})	7.37×10^{-4}
4/3	1	2.04×10^{-4} (5.8×10^{-6})	6.25×10^{-4} (4.1×10^{-6})	8.29×10^{-4}
	2	7.09×10^{-4} (9.6×10^{-6})	5.12×10^{-4} (3.8×10^{-6})	1.22×10^{-3}
	3	1.55×10^{-4} (5.2×10^{-6})	6.45×10^{-4} (4.1×10^{-6})	8.00×10^{-4}
	4	5.91×10^{-4} (9.4×10^{-6})	4.66×10^{-4} (3.1×10^{-6})	1.06×10^{-3}
	5	6.33×10^{-4} (9.5×10^{-6})	3.97×10^{-4} (2.7×10^{-6})	1.03×10^{-3}
	6	9.59×10^{-4} (1.1×10^{-5})	3.47×10^{-4} (2.4×10^{-6})	1.31×10^{-3}

Table 3.12. Values of AASB, AAVAR and AAMSE for samples of size 100 from an N(0,1) distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	2.53×10^{-6} (3.9×10^{-7})	1.19×10^{-3} (1.5×10^{-5})	1.19×10^{-3}
	2	7.60×10^{-6} (8.0×10^{-7})	8.33×10^{-4} (9.3×10^{-6})	8.41×10^{-4}
	3	4.58×10^{-6} (5.6×10^{-7})	1.21×10^{-3} (1.6×10^{-5})	1.21×10^{-3}
	4	4.48×10^{-5} (2.5×10^{-6})	9.19×10^{-4} (1.0×10^{-5})	9.64×10^{-4}
	5	3.53×10^{-5} (1.3×10^{-6})	8.12×10^{-4} (9.7×10^{-6})	8.47×10^{-4}
	6	3.46×10^{-4} (5.5×10^{-6})	8.54×10^{-4} (1.1×10^{-5})	1.20×10^{-3}
1	1	1.66×10^{-5} (1.3×10^{-6})	7.29×10^{-4} (7.7×10^{-6})	7.46×10^{-4}
	2	9.32×10^{-5} (3.3×10^{-6})	5.17×10^{-4} (5.9×10^{-6})	6.11×10^{-4}
	3	2.64×10^{-5} (1.6×10^{-6})	7.44×10^{-4} (3.3×10^{-6})	7.71×10^{-4}
	4	1.95×10^{-4} (5.1×10^{-6})	5.65×10^{-4} (5.7×10^{-6})	7.60×10^{-4}
	5	1.49×10^{-4} (2.9×10^{-6})	5.52×10^{-4} (6.2×10^{-6})	7.01×10^{-4}
	6	4.27×10^{-4} (3.6×10^{-6})	5.33×10^{-4} (6.9×10^{-6})	9.60×10^{-4}
4/3	1	9.83×10^{-5} (3.4×10^{-6})	5.23×10^{-4} (5.8×10^{-6})	6.21×10^{-4}
	2	5.02×10^{-4} (7.4×10^{-6})	3.97×10^{-4} (4.0×10^{-6})	8.98×10^{-4}
	3	1.17×10^{-4} (3.6×10^{-6})	5.34×10^{-4} (5.8×10^{-6})	6.51×10^{-4}
	4	5.26×10^{-4} (7.9×10^{-6})	4.05×10^{-4} (4.1×10^{-6})	9.31×10^{-2}
	5	4.88×10^{-4} (6.4×10^{-6})	4.07×10^{-4} (4.6×10^{-6})	8.96×10^{-4}
	6	1.29×10^{-3} (1.0×10^{-5})	3.40×10^{-4} (4.4×10^{-6})	1.63×10^{-3}

Table 3.13. Values of AASB, AAVAR and AAMSE for samples of size 100 from a Gamma $(2, \sqrt{2})$ distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	1.19×10^{-2} (3.5×10^{-5})	5.73×10^{-4} (5.8×10^{-6})	1.25×10^{-2}
	2	1.23×10^{-2} (2.9×10^{-5})	4.27×10^{-4} (5.1×10^{-6})	1.27×10^{-2}
	3	1.18×10^{-2} (3.6×10^{-5})	5.75×10^{-4} (5.6×10^{-6})	1.24×10^{-2}
	4	1.19×10^{-2} (3.2×10^{-5})	4.45×10^{-4} (4.3×10^{-6})	1.23×10^{-2}
	5*	1.20×10^{-2} (4.0×10^{-5})	4.01×10^{-4} (4.3×10^{-6})	1.24×10^{-2}
	6	1.12×10^{-2} (3.7×10^{-5})	4.25×10^{-4} (5.9×10^{-6})	1.16×10^{-2}
1	1	1.21×10^{-2} (2.8×10^{-5})	3.76×10^{-4} (4.3×10^{-6})	1.25×10^{-2}
	2	1.22×10^{-2} (2.5×10^{-5})	2.77×10^{-4} (2.6×10^{-6})	1.24×10^{-2}
	3	1.20×10^{-2} (2.9×10^{-5})	3.72×10^{-4} (4.1×10^{-6})	1.24×10^{-2}
	4	1.21×10^{-2} (2.5×10^{-5})	2.82×10^{-4} (2.8×10^{-6})	1.24×10^{-2}
	5*	1.18×10^{-2} (3.3×10^{-5})	2.74×10^{-4} (2.8×10^{-5})	1.20×10^{-2}
	6	1.02×10^{-2} (3.0×10^{-5})	2.83×10^{-4} (4.0×10^{-6})	1.05×10^{-2}
4/3	1	1.20×10^{-2} (2.5×10^{-5})	2.71×10^{-4} (2.6×10^{-6})	1.25×10^{-2}
	2	1.18×10^{-2} (2.2×10^{-5})	1.82×10^{-4} (1.7×10^{-6})	1.20×10^{-2}
	3	1.20×10^{-2} (2.5×10^{-5})	2.67×10^{-4} (2.7×10^{-6})	1.23×10^{-2}
	4	1.22×10^{-2} (2.1×10^{-5})	1.99×10^{-4} (1.8×10^{-6})	1.24×10^{-2}
	5*	1.14×10^{-2} (2.9×10^{-5})	1.97×10^{-4} (1.8×10^{-6})	1.16×10^{-2}
	6	1.00×10^{-2} (2.4×10^{-5})	1.85×10^{-4} (2.6×10^{-6})	1.02×10^{-2}

* Results based on only 4 runs.

Table 3.14. Values of AASB, AAVAR and AAMSE for samples of size 100 from a $[0.5N(-0.866,0.5^2) + 0.5N(0.866,0.5^2)]$ distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	4.25×10^{-4} (5.8×10^{-6})	6.91×10^{-4} (3.5×10^{-6})	1.12×10^{-3}
	2	1.60×10^{-3} (6.0×10^{-6})	4.80×10^{-4} (2.6×10^{-6})	2.08×10^{-3}
	3	4.12×10^{-4} (6.0×10^{-6})	7.14×10^{-4} (3.0×10^{-6})	1.13×10^{-3}
	4	8.39×10^{-4} (6.8×10^{-6})	5.21×10^{-4} (2.1×10^{-6})	1.36×10^{-3}
	5	1.13×10^{-3} (7.6×10^{-6})	5.31×10^{-4} (3.0×10^{-6})	1.66×10^{-3}
	6	1.89×10^{-3} (1.03×10^{-6})	6.09×10^{-4} (4.5×10^{-6})	2.50×10^{-3}
1	1	1.86×10^{-3} (5.7×10^{-6})	4.07×10^{-4} (3.0×10^{-6})	2.27×10^{-3}
	2	3.75×10^{-3} (6.8×10^{-6})	2.75×10^{-4} (2.6×10^{-6})	4.02×10^{-3}
	3	1.57×10^{-3} (5.9×10^{-6})	4.27×10^{-4} (3.0×10^{-6})	2.00×10^{-3}
	4	2.13×10^{-3} (6.4×10^{-6})	3.04×10^{-4} (2.1×10^{-6})	2.43×10^{-3}
	5	2.75×10^{-3} (5.6×10^{-6})	3.54×10^{-4} (3.0×10^{-6})	3.11×10^{-3}
	6	3.56×10^{-3} (5.4×10^{-6})	4.27×10^{-4} (4.5×10^{-6})	3.99×10^{-3}
4/3	1	2.96×10^{-3} (4.9×10^{-6})	2.72×10^{-4} (2.5×10^{-6})	3.24×10^{-3}
	2	4.06×10^{-3} (5.8×10^{-6})	1.84×10^{-4} (1.6×10^{-6})	4.24×10^{-3}
	3	2.72×10^{-3} (5.4×10^{-6})	2.91×10^{-4} (2.7×10^{-6})	3.01×10^{-3}
	4	3.30×10^{-3} (7.9×10^{-6})	2.05×10^{-4} (1.7×10^{-6})	3.50×10^{-3}
	5	3.94×10^{-3} (6.8×10^{-6})	2.52×10^{-4} (2.5×10^{-6})	4.19×10^{-3}
	6	4.70×10^{-3} (7.7×10^{-6})	2.84×10^{-4} (3.3×10^{-6})	4.99×10^{-3}

Table 3.15. Values of AASB, AAVAR and AAMSE for samples of size 100 from a t(3) distribution.

Scaling factor for h_{Nopt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	3.25×10^{-6} (3.1×10^{-7})	7.07×10^{-4} (9.2×10^{-6})	7.10×10^{-4}
	2	1.56×10^{-5} (7.5×10^{-7})	5.07×10^{-4} (7.2×10^{-6})	5.23×10^{-4}
	3	4.41×10^{-6} (3.3×10^{-7})	7.13×10^{-4} (9.2×10^{-6})	7.17×10^{-4}
	4	4.17×10^{-5} (1.1×10^{-6})	5.52×10^{-4} (7.2×10^{-6})	5.94×10^{-4}
	5	5.24×10^{-6} (2.8×10^{-7})	4.57×10^{-4} (6.8×10^{-6})	4.62×10^{-4}
	6	7.51×10^{-5} (1.3×10^{-6})	4.52×10^{-4} (6.9×10^{-6})	5.27×10^{-4}
1	1	3.27×10^{-5} (1.1×10^{-6})	4.49×10^{-4} (6.5×10^{-6})	4.82×10^{-4}
	2	1.57×10^{-4} (2.6×10^{-6})	3.35×10^{-4} (5.2×10^{-6})	4.92×10^{-4}
	3	3.39×10^{-5} (1.0×10^{-6})	4.51×10^{-4} (6.6×10^{-6})	4.85×10^{-4}
	4	1.70×10^{-4} (2.1×10^{-6})	3.51×10^{-4} (5.0×10^{-6})	5.21×10^{-4}
	5	5.68×10^{-5} (1.1×10^{-6})	3.14×10^{-4} (4.9×10^{-6})	3.70×10^{-4}
	6	1.72×10^{-4} (1.8×10^{-6})	2.77×10^{-4} (4.3×10^{-6})	4.49×10^{-4}
4/3	1	1.48×10^{-4} (2.4×10^{-6})	3.35×10^{-4} (5.2×10^{-6})	4.83×10^{-4}
	2	5.71×10^{-4} (4.6×10^{-6})	2.65×10^{-4} (4.3×10^{-6})	8.36×10^{-4}
	3	1.33×10^{-4} (2.1×10^{-6})	3.33×10^{-4} (5.1×10^{-6})	4.66×10^{-4}
	4	4.23×10^{-4} (3.3×10^{-6})	2.59×10^{-4} (3.9×10^{-6})	6.81×10^{-4}
	5	2.97×10^{-4} (2.6×10^{-6})	2.34×10^{-4} (3.8×10^{-6})	5.31×10^{-4}
	6	8.37×10^{-4} (4.2×10^{-6})	1.82×10^{-4} (2.9×10^{-6})	1.02×10^{-3}

of $(h^2/2)\, f^{(2)}(x)$ (method 1) and the multiplicative correction factor (method 3) in achieving low bias. For data from distributions 1, 3 and 4 these two methods generally have by far the lowest AASB with the performance of Method 3 perhaps slightly superior overall to that of 1. For the two unimodal distributions their AASB at $1.33\, h_{Nopt}$ is still less than that based on the fixed normal kernel (method 4) at h_{Nopt}. When the underlying density is highly skewed their performance is similar to that of the other non-adaptive methods.

The two adaptive methods, $\alpha = 1/2$ (method 5) and $\alpha = 1$ (method 6) attain lower bias than the other methods for the Gamma $(2, \sqrt{2})$ data. In particular $\alpha = 1$ is markedly superior to the others in this case. For the other distributions $\alpha = 1/2$ generally has lower AASB than $\alpha = 1$ in line with both the asymptotic and exact results for their ideal versions described in chapter 2. Also, for the unimodal distributions 1, 2 and 4 $\alpha = 1/2$ generally has similar or lower AASB than method 4 $(\alpha = 0)$ when $n = 50$ but when $n = 100$ the differences are more marked in favour of $\alpha = 1/2$. Both $\alpha = 1/2$ and $\alpha = 1$ are least effective in terms of bias when the distribution is bimodal when their AASB's are much higher than those for method 4.

The estimator based on the optimal kernel of order 4 (method 2) is on the whole superior to method 4 in terms of AASB for data from the two symmetric unimodal distributions. For the highly skewed Gamma $(2, \sqrt{2})$ it has lower AASB than method 4 for larger smoothing parameters but for data from the bimodal normal mixture it consistently has much higher AASB. In comparison with the two adaptive methods it achieves lower bias for $N(0,1)$ data but its performance for distributions 3 and 4 generally lies between the two except for larger amounts of smoothing when it is particularly superior to $\alpha = 1$.

In contrast, the AAVAR results show methods 1 and 3 to have much higher variance than each of the other methods for each distribution and amount of smoothing. The only exception is for distribution 3 when $\alpha = 1$ also has high AAVAR. The values for method 3 are consistently at a slightly higher level than those of 1. On the other hand the two adaptive methods achieve between them many of the lowest AAVAR values and generally outperform method 4 for all but the bimodal distribution. Method 2 is fairly similar overall to 4 when $n = 50$ but is almost consistently a little superior when $n = 100$.

When the bias and variance results are combined to give the AAMSE the values for methods 1 and 3 decrease as h increases except for the bimodal distribution 3 when they increase. For this particular distribution they have much the lowest AAMSE of all the methods for both sample sizes and each value of h_{Nopt}. For the symmetric, unimodal distributions their performance are poor at $0.667\ h_{Nopt}$ due to high variance but clearly the best at $1.333\ h_{Nopt}$ when they have lower values than method 4 at h_{Nopt}. The two adaptive methods are overall most successful in terms of AAMSE for unimodal densities and in particular for the highly skewed distribution 2 when $\alpha = 1$ is superior to $\alpha = 1/2$. For the symmetric 1 and 4 though, $\alpha = 1/2$ is generally better than both $\alpha = 1$ and method 4. $\alpha = 1$ is markedly inferior to method 4 for $N(0,1)$ data in contrast to the exact MISE results of chapter 2. Finally, method 2 generally has similar or lower AAMSE than method 4 except again for the bimodal distribution 3 when it is far inferior.

The results of this study indicate that the choice of a method

in practice depends to some extent on the properties required. If particularly low bias is desirable then either method 1 or 3 should be used. When using a Normal optimal smoothing parameter the bias should be much lower but the resulting MISE is likely to be similar of slightly higher than that based on the unadjusted fixed normal kernel due to the increased variance. However, if the amount of smoothing is increased the bias will still be low but the MISE should fall below that based on the fixed kernel. The exception in the study was for the bimodal density when methods 1 and 3 were both better than 4 in terms of both AASB and AAMSE at each level of smoothing considered. If the underlying density is unimodal, and in particular long tailed or skewed, then $\alpha = {}^{1}/_{2}$ would make a good all round choice with Normal optimal smoothing again providing a good guide. For such distributions and choice of h-value it should be superior to the fixed normal kernel estimator on all counts.

3.9 Examples.

The data considered in this Section consist of the survival times for two groups of rats in an investigation of the toxicity of cytoxan, a chemical agent used for chemotherapy. The first group of 40 rats was given half the dosage twice weekly while the second group of 44 was given the full dosage once weekly.

These two data sets were investigated by McLachlan et al (1982). They assumed that death was attributable either to the regrowth of the tumour or to the toxicity of the cytoxan with toxic death usually preceding the former. Hence, they considered the failure time density for each group to be a mixture of two densities corresponding to each of the causes. Toxicity at the two dosage levels was then compared via a likelihood ratio test of the homo-

geneity of the mixing proportions. In constructing the test they assumed the mixing densities to be normal with different means and variances.

For group 1 density estimates were obtained using methods 1, 4 and 5 with the normal optimal formulae given in Section 3.8 being used to choose h. These estimates are illustrated in figures 3.17-3.19. They are each unimodal with mode at about 12 weeks but differ in their tail behaviour. That based on method 1, which subtracts an estimate of the asymptotic bias of method 4, has tails which go to zero quite abruptly. For method 4 the upper tail still goes to zero quite quickly but in a smoother manner than for method 1 while the lower tail is positive at zero. For the adaptive method 5 with $\alpha = {}^1/_2$ the upper tail approaches zero at a slower rate than for 4 while its lower tail takes a slightly higher positive value. These plots then are not inconsistent with the mixture assumption of McLachlan et al. However, if the normal optimal h-values are scaled by 0.5 than when using methods 1 and 4 the resulting density estimates are trimodal (see figures 3.20, 3.21) with modes at about 8, 12.5 and 17.5 weeks and suggests a mixture of three densities in the approximate ratio 3:2:1. Scaling h_{Nopt} for method 5 by 0.5 still results in a unimodal density (figure 3.22) but the change in gradient of the density at about 8 weeks suggests that if h is further reduced the density will be at least bimodal.

For the second group of rats who were given the full dosage once weekly density estimates were again obtained by methods 1, 4 and 5 using Normal optimal smoothing and are illustrated in figures 3.23-3.25. Each estimate is markedly skewed with a large mode at about 7.5 weeks. Those based on methods 1 and 4 are noisy in the right

hand tail which arises from small clusters of a few higher valued observations. These two estimates are very similar with perhaps the bias corrected method 1 appearing slightly more data responsive. The estimate based on the adaptive method 5, on the other hand, has a right hand tail which goes smoothly to zero and also a left hand tail which goes to zero at a slower rate than in the case of methods 1 and 4. Again, each of these estimates is not incompatible with the hypothesis of a mixture of two densities but with a quite different mixing proportion to the first group.

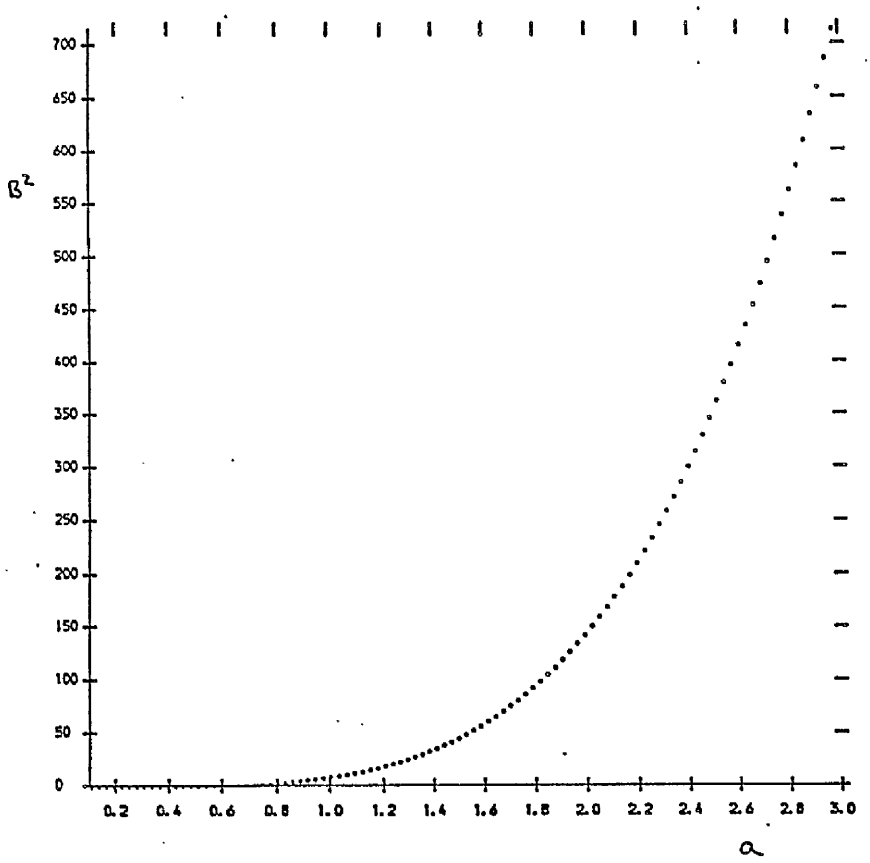


Figure 3.1. $\left[\int t^4.K^*(t)dt\right]^2$ (i.e. B^2) when K^* is based on standard normal kernels, as a function of the parameter α for the jacknife estimator.

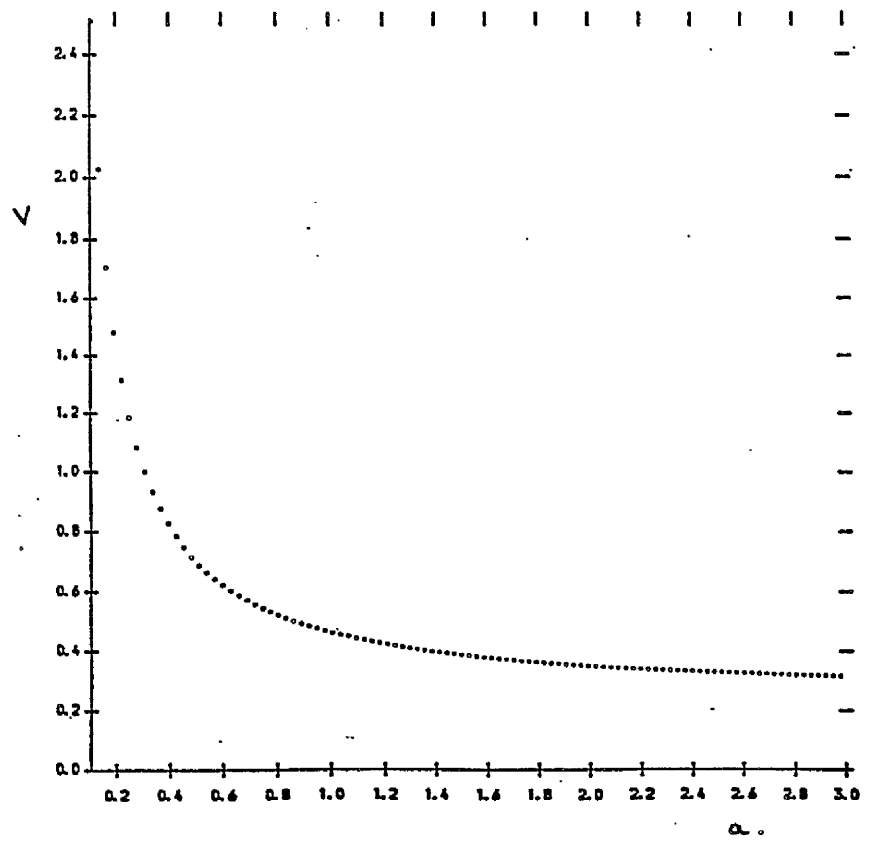


Figure 3.2. $\int K^*(t)^2dt$ (i.e. V) when K^* is based on standard normal kernels, as a function of the parameter α for the jacknife estimator.

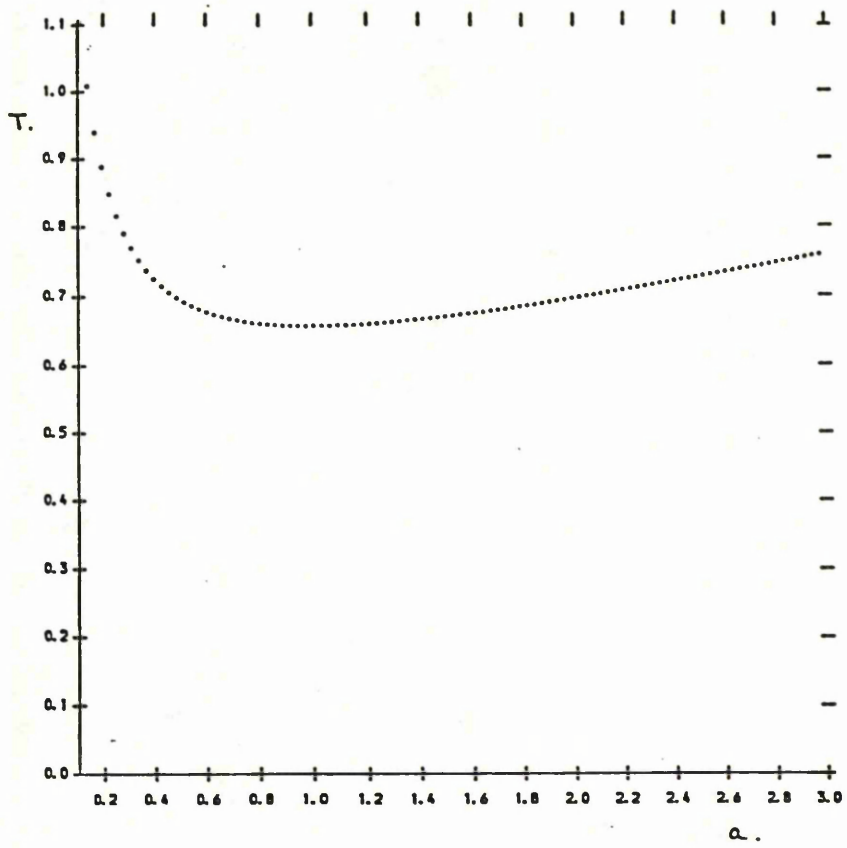


Figure 3.3. $T = \left[v^8, B^2\right]^{1/4}$ as a function of a for the jacknife estimator with kernel function K^* based on standard normal kernels.

Figure 3.4. Three higher-order kernel functions.

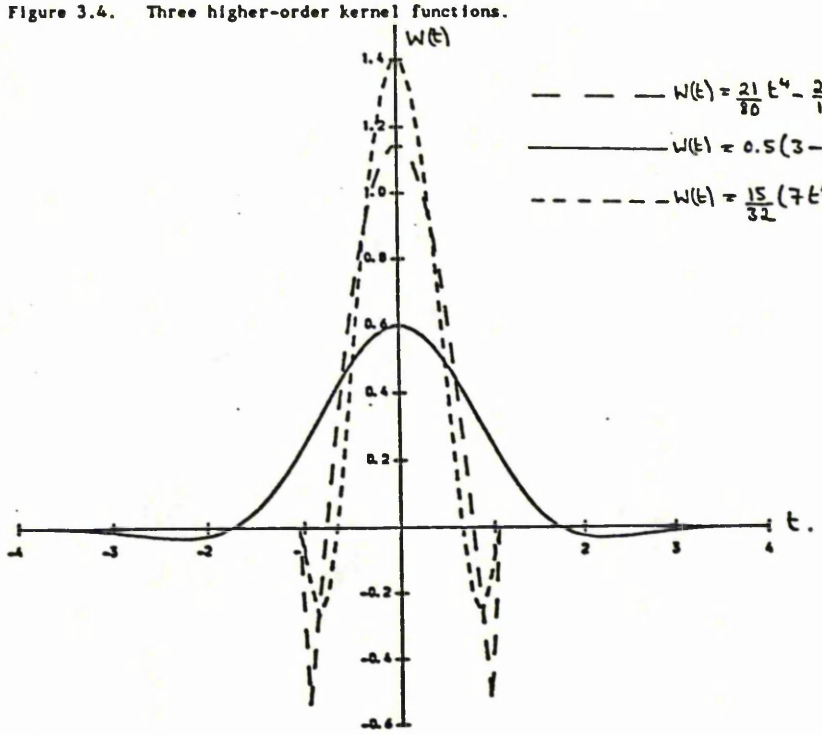


Figure 3.5. AASB incurred by the six estimators for samples of size 50 from an N(0,1) distribution.

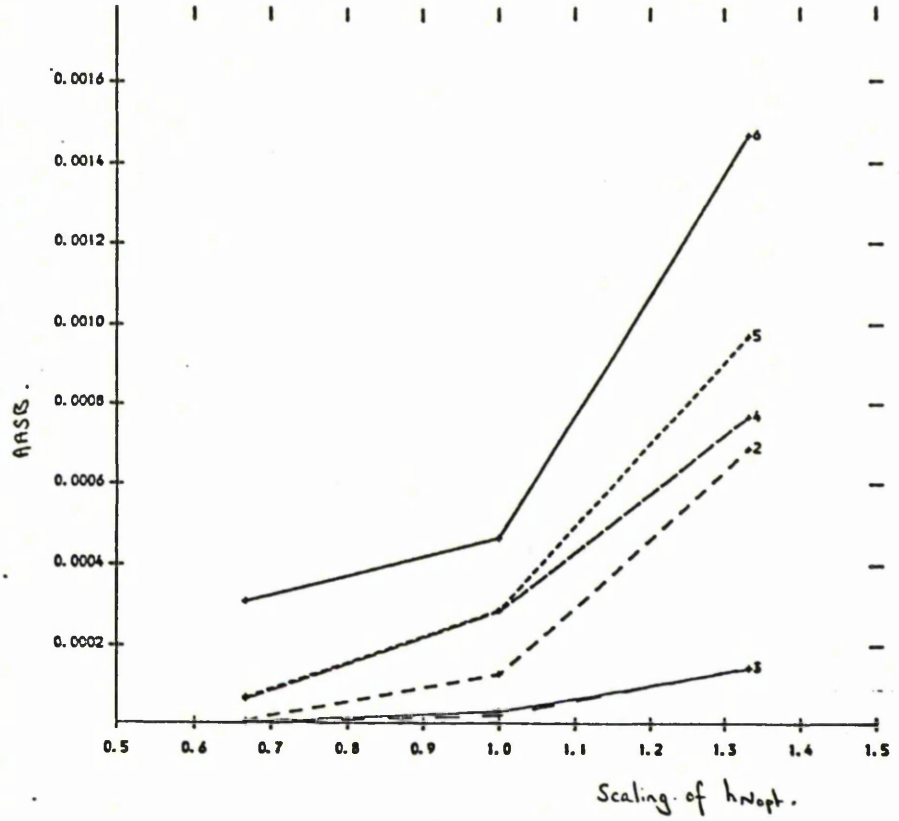


Figure 3.6. AASB incurred by the six estimators for samples of size 50 from a Gamma $(2, \sqrt{2})$ distribution.

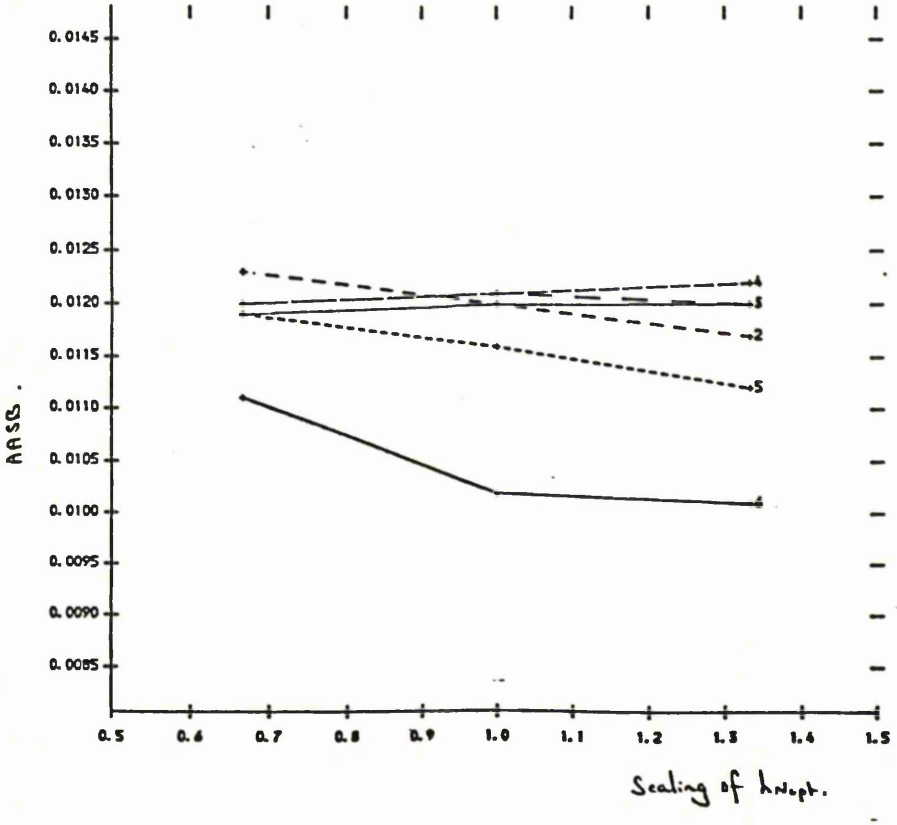


Figure 3.7. AASB incurred by the six estimators for samples of size 50 from a $0.5N(-0.866, 0.5^2) + 0.5N(0.866, 0.5^2)$ distribution.

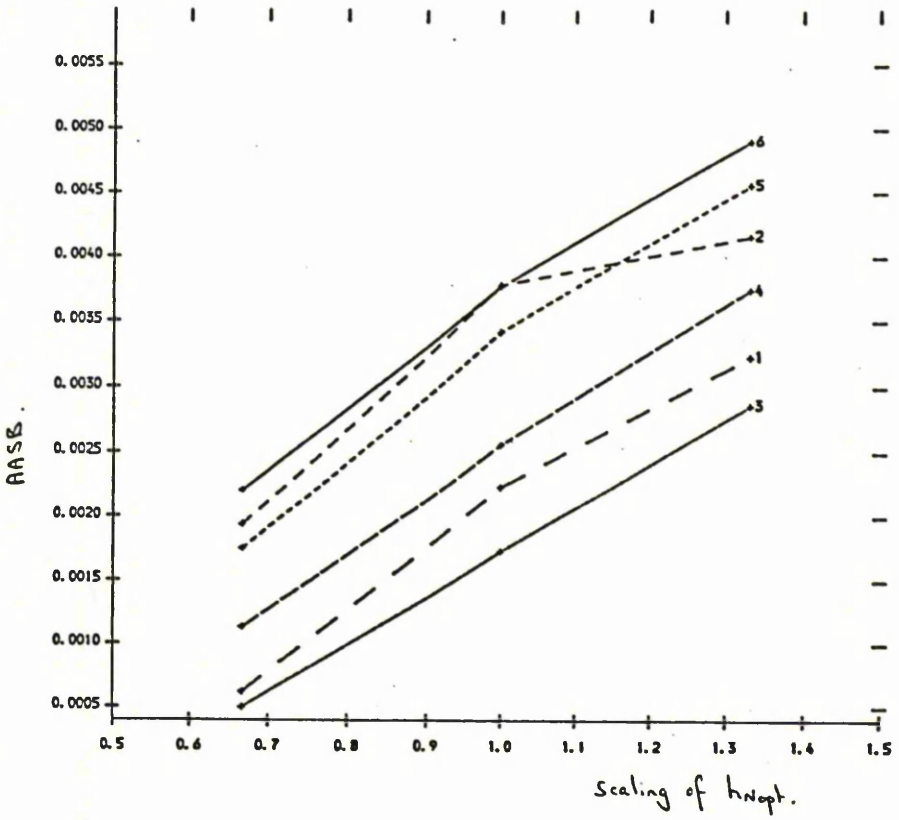


Figure 3.8. AASB incurred by the six estimators for samples of size 50 from a t(3) distribution.

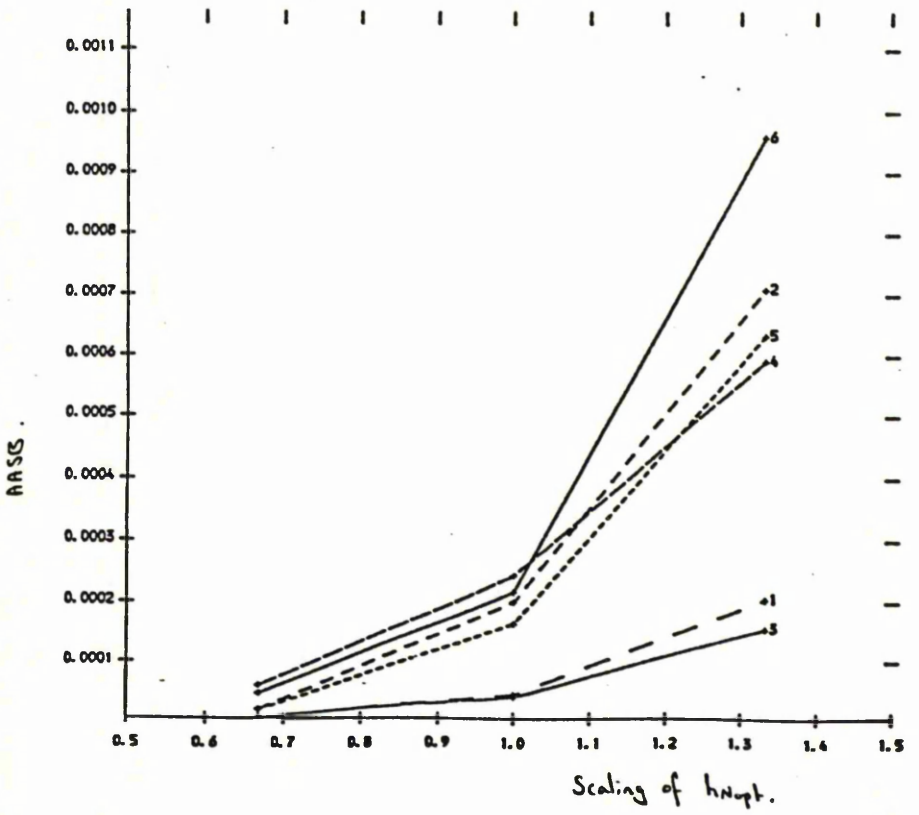


Figure 3.9. AAVAR incurred by the six estimators for samples of size 50 from an N(0,1) distribution.

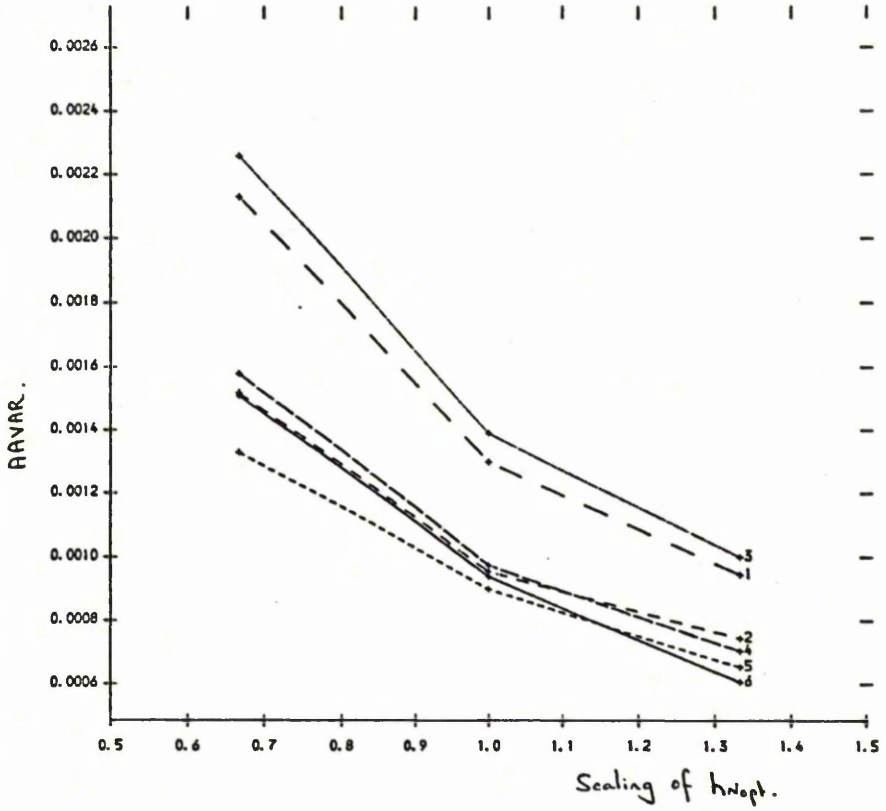


Figure 3.10. AAVAR incurred by the six estimators for samples of size 50 from a Gamma $(2, \sqrt{2})$ distribution.

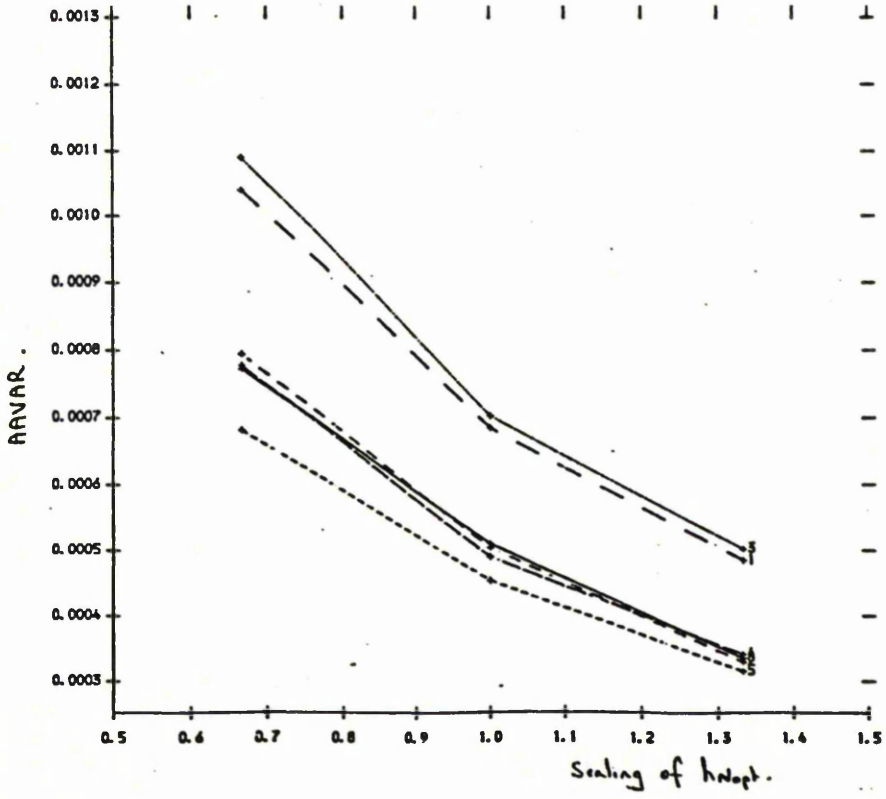


Figure 3.11. AAVAR incurred by the six estimators for samples of size 50 from a $0.5N(-0.866, 0.5^2) + 0.5N(0.866, 0.5^2)$ distribution.

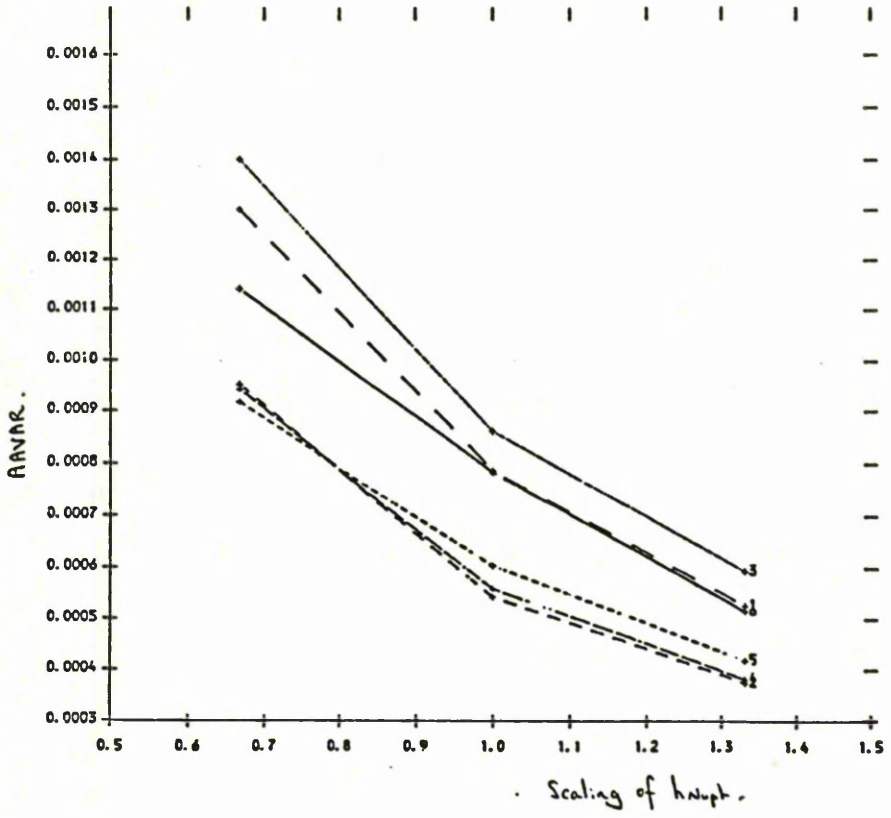


Figure 3.12. AAVAR incurred by the six estimators for samples of size 50 from a t(3) distribution.

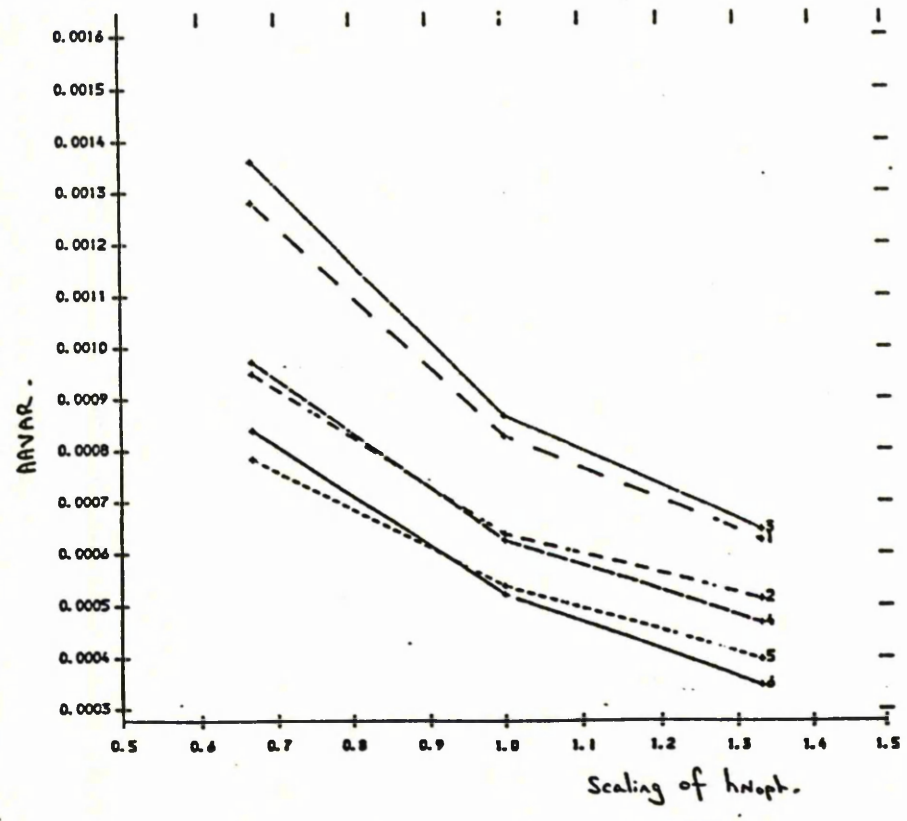


Figure 3.13. AAMSE incurred by the six estimators for samples of size 50 from an N(0,1) distribution.

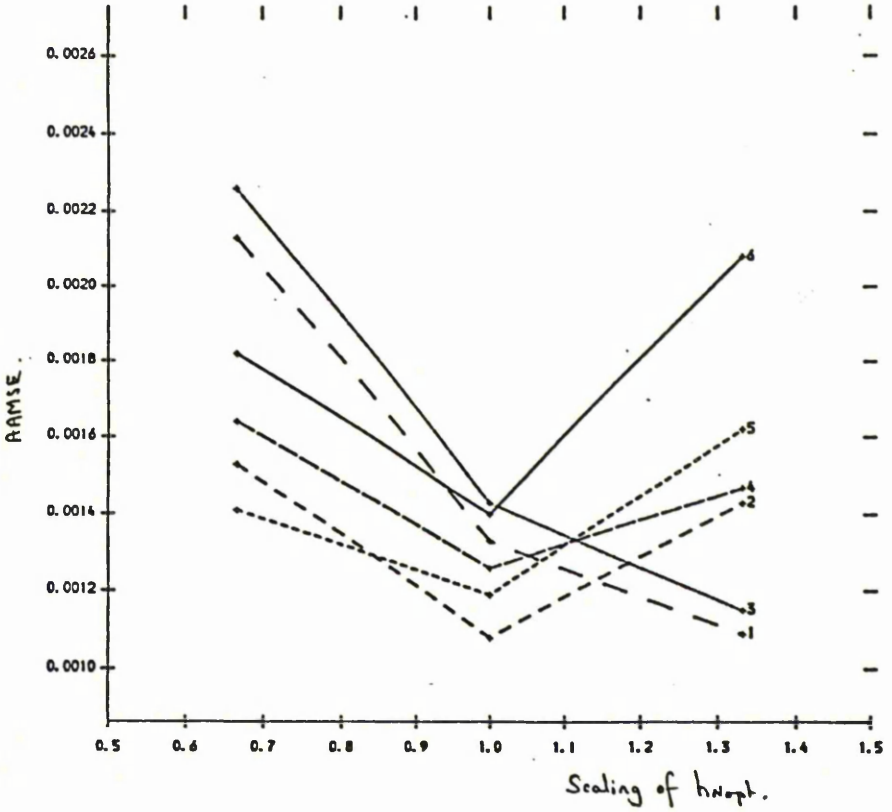


Figure 3.14. AAMSE incurred by the six estimators for samples of size 50 from a Gamma $(2, \sqrt{2})$ distribution.

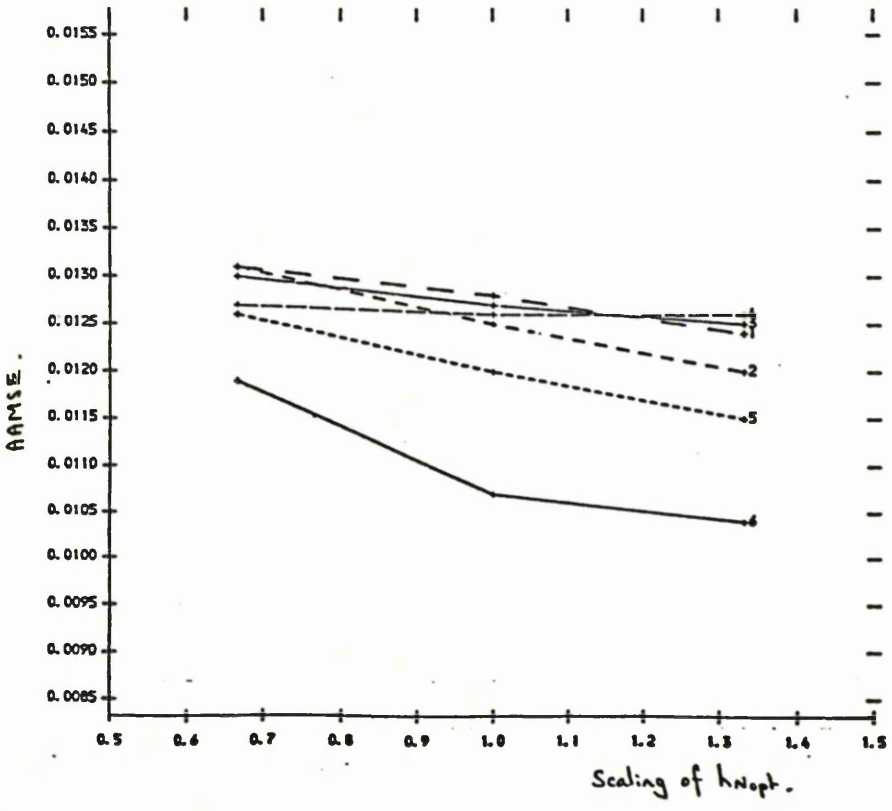


Figure 3.15. AAMSE incurred by the six estimators for samples of size 50 from a $0.5N(-0.866, 0.5^2) + 0.5N(0.866, 0.5^2)$ distribution.

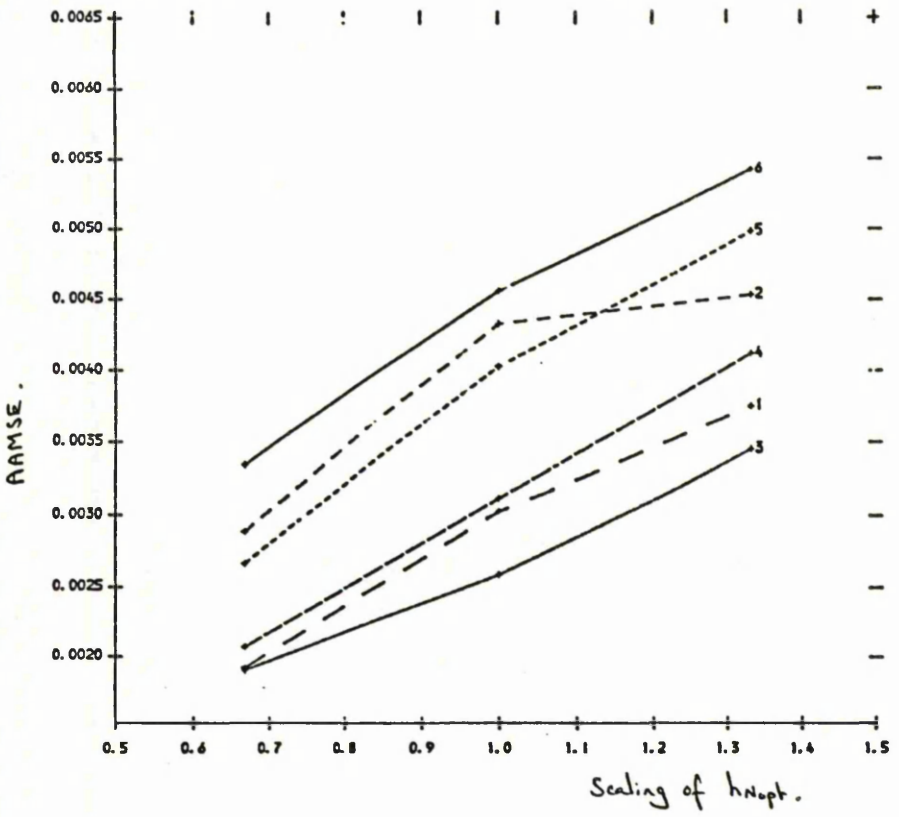


Figure 3.16. AAMSE incurred by the six estimators for samples of size 50 from a t(3) distribution.

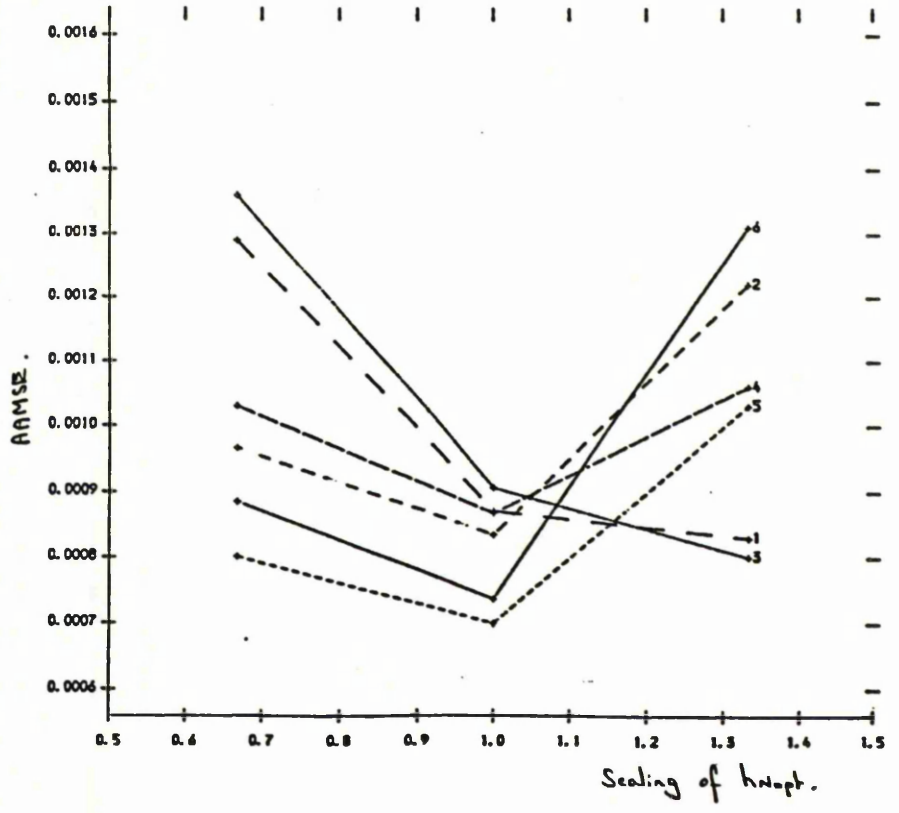


Figure 3.17. Bias corrected fixed kernel estimate (method 1) for the first group of rats given half the full dose of cytoxan twice weekly, h = 2.424.

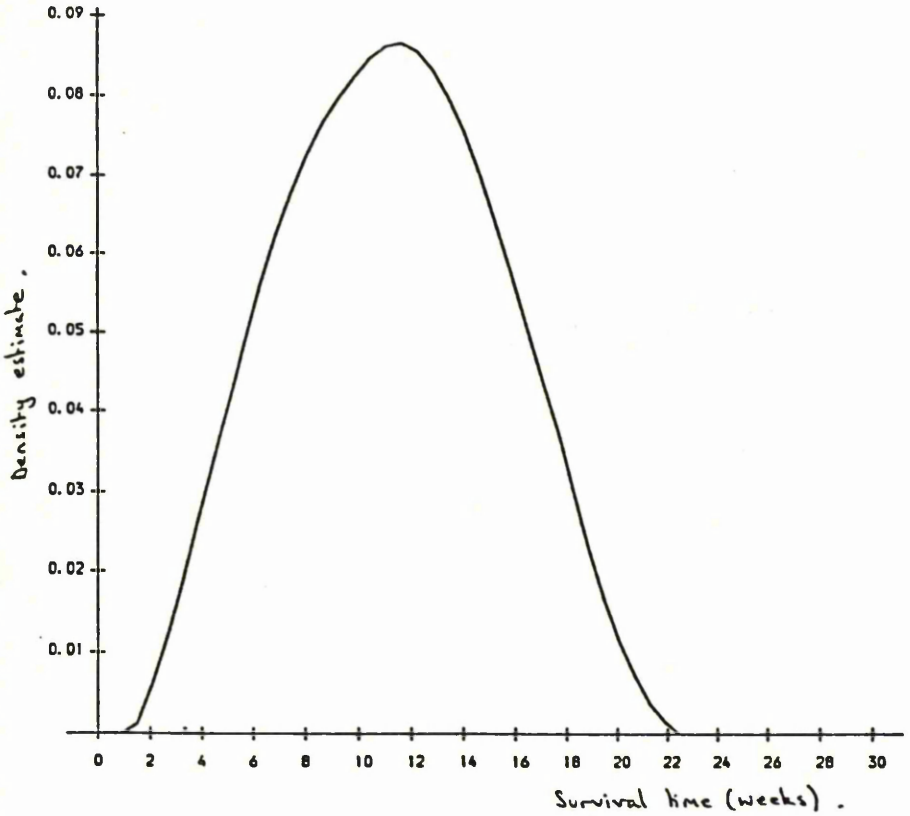


Figure 3.18. Fixed kernel estimate (method 4) for the first group of rats given half the full dose of cytoxan twice weekly, h = 2.424.

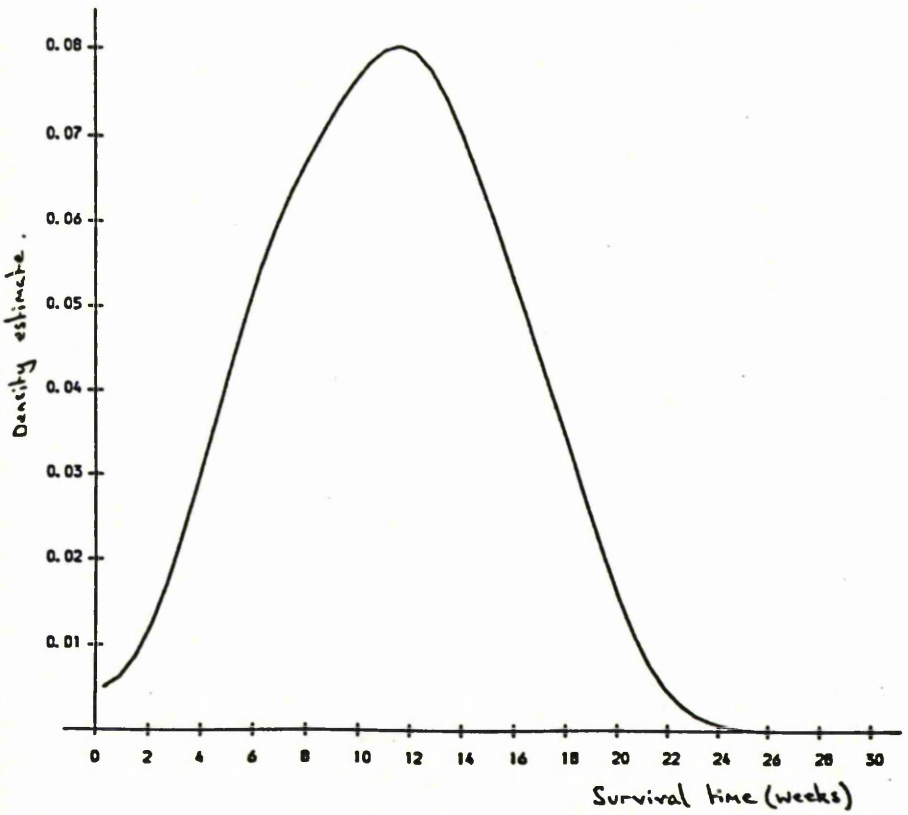


Figure 3.19. Adaptive kernel estimate with $\alpha = 1/2$ (method 5) for the first group of rats given half the full dose of cytoxan twice weekly, $h = 0.978$.

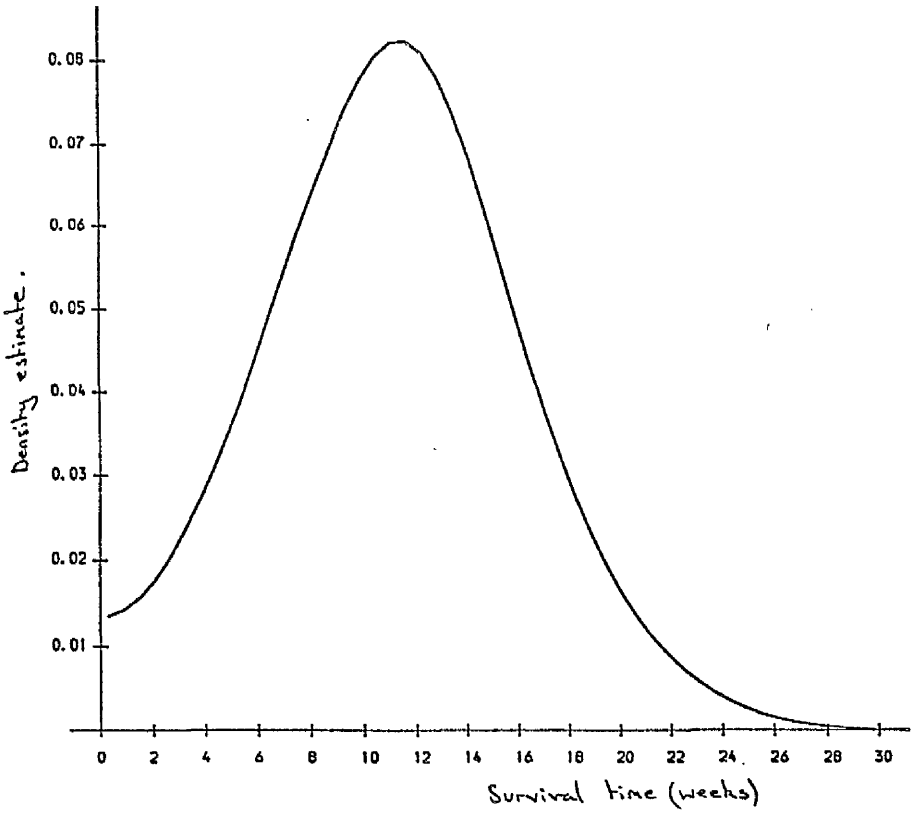


Figure 3.20. Bias corrected fixed kernel estimate (method 1) for the first group of rats given half the full dose of cytoxan twice weekly, $h = 1.212$.

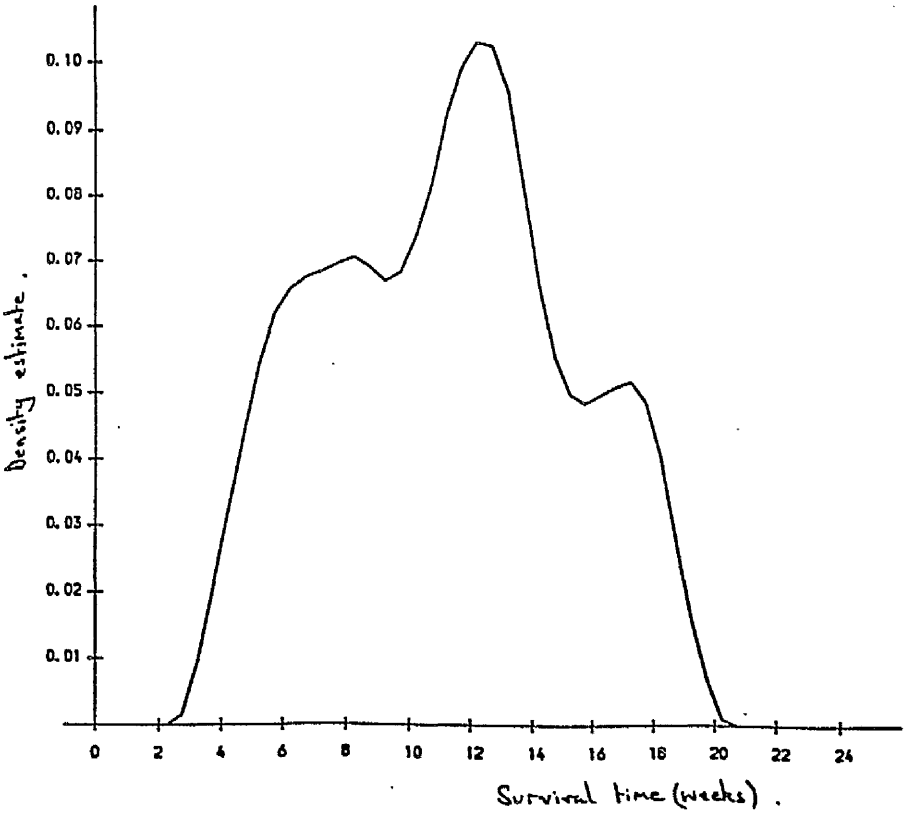


Figure 3.21. Fixed kernel estimate (method 4) for the first group of rats given half the full dose of cytoxan twice weekly, h = 1.212.

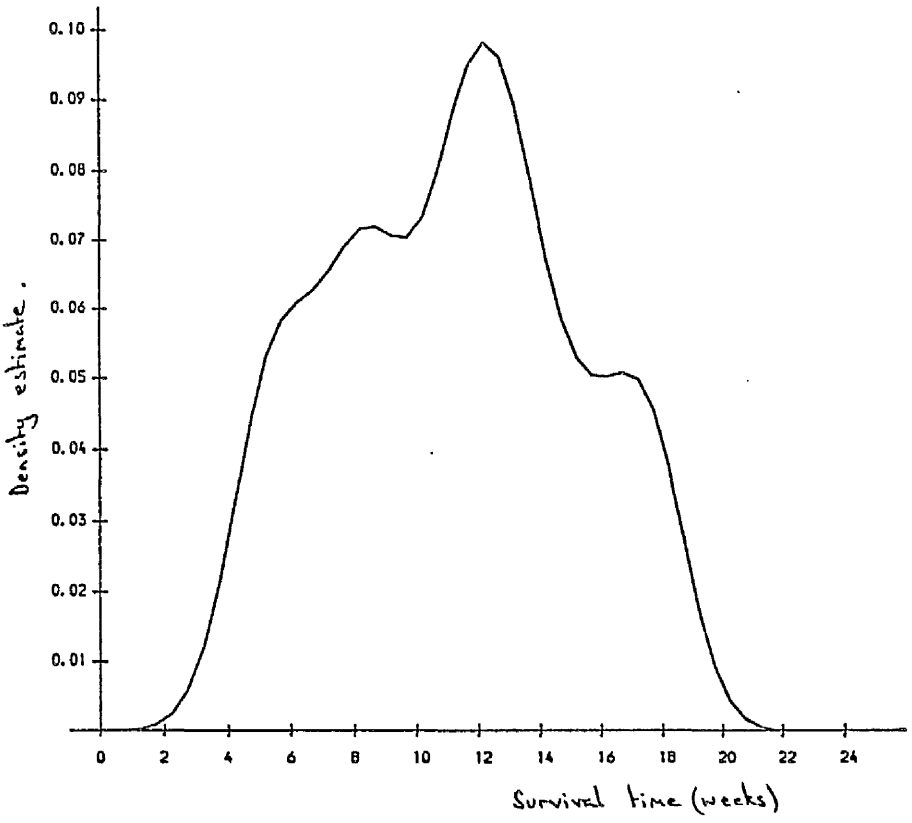


Figure 3.22. Adaptive kernel estimate with $\alpha = 1/2$ (method 5) for the first group of rats given half the full dose of cytoxan twice weekly, h = 0.399.

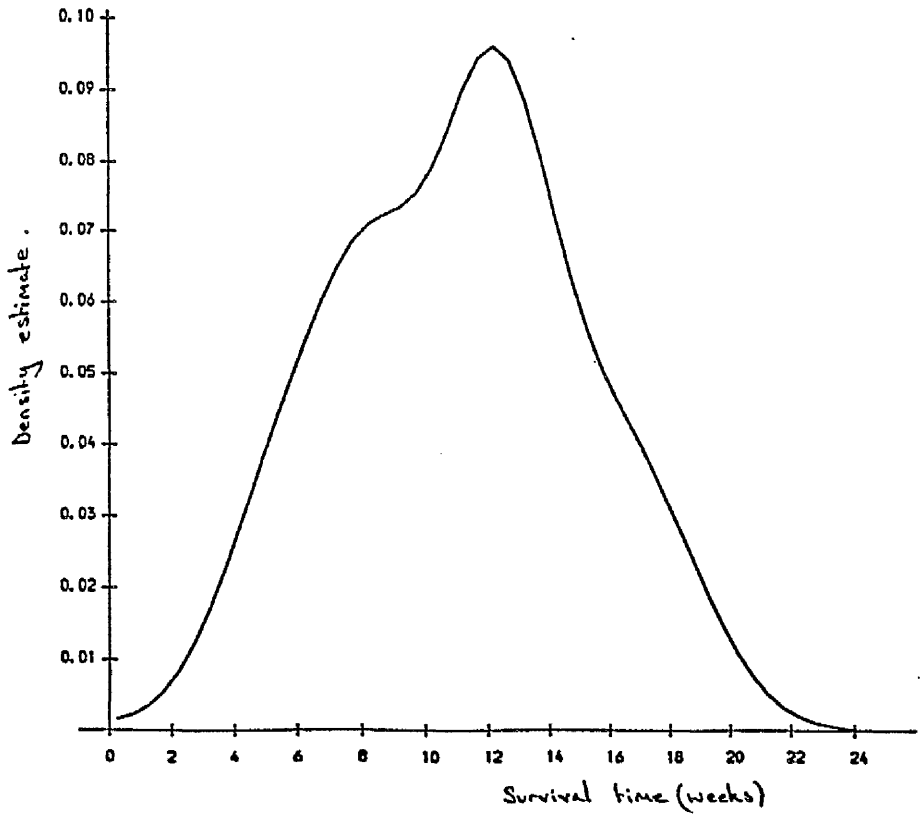


Figure 3.23. Bias corrected fixed kernel estimate (method 1) for the second group of rats given the full dose of cytoxan once weekly, h = 0.594.

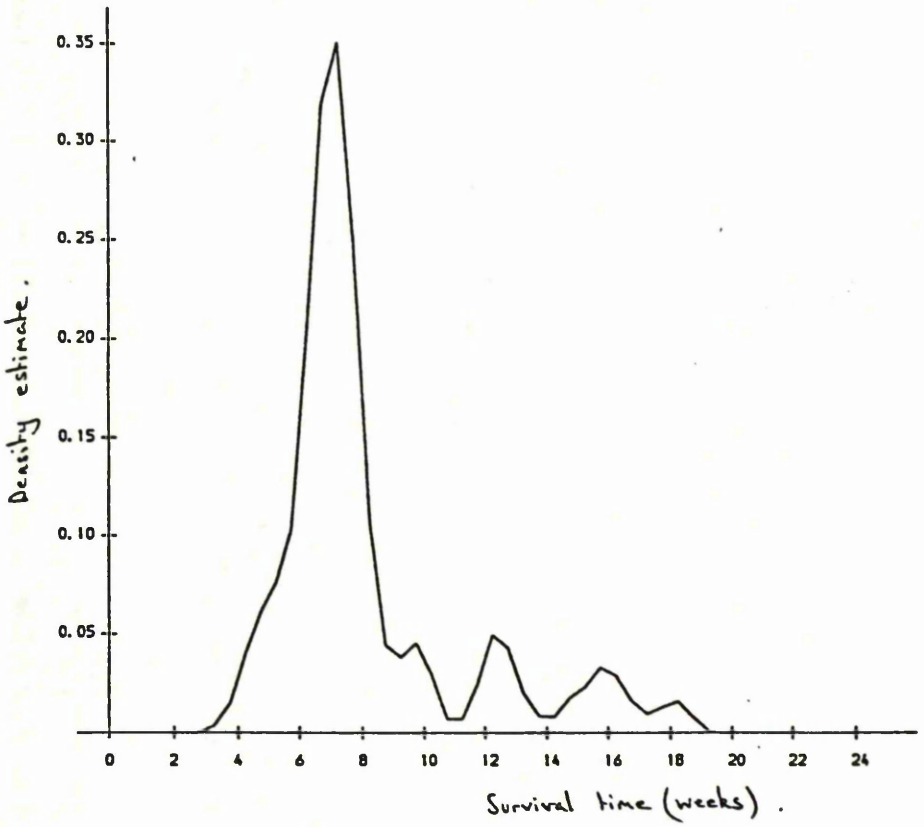


Figure 3.24. Fixed kernel estimate (method 4) for the second group of rats given the full dose of cytoxan once weekly, h = 0.594.

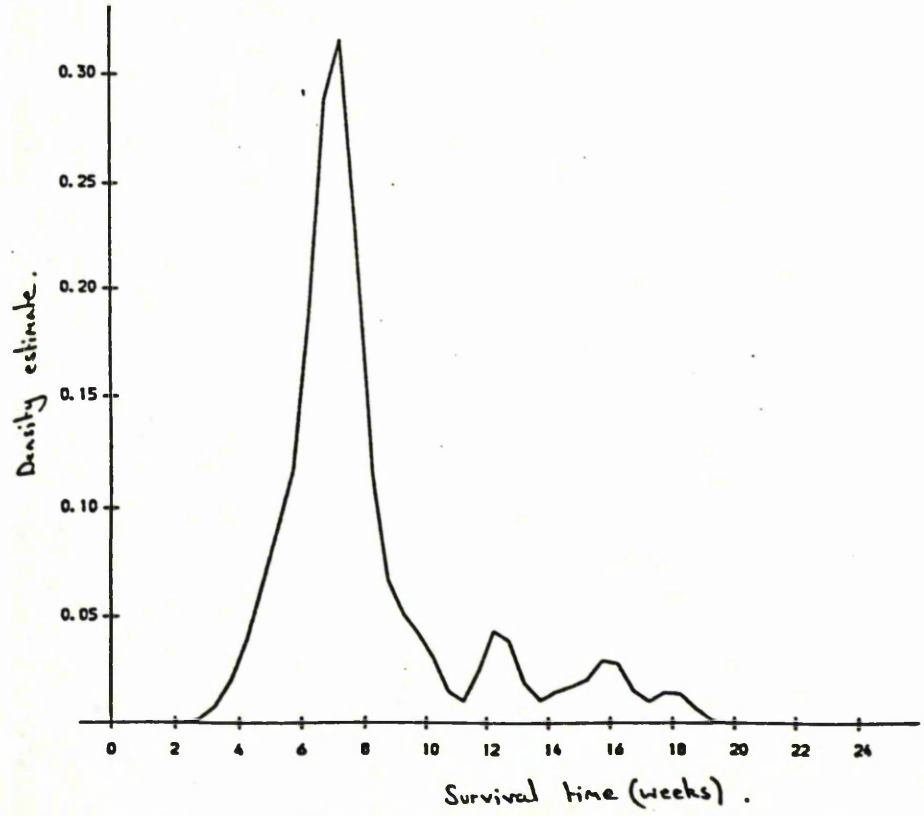


Figure 3.25. Adaptive kernel estimate with $\alpha = 1/2$ (method 5) for the second group of rats given the full dose of cytoxan once weekly, h = 0.390.

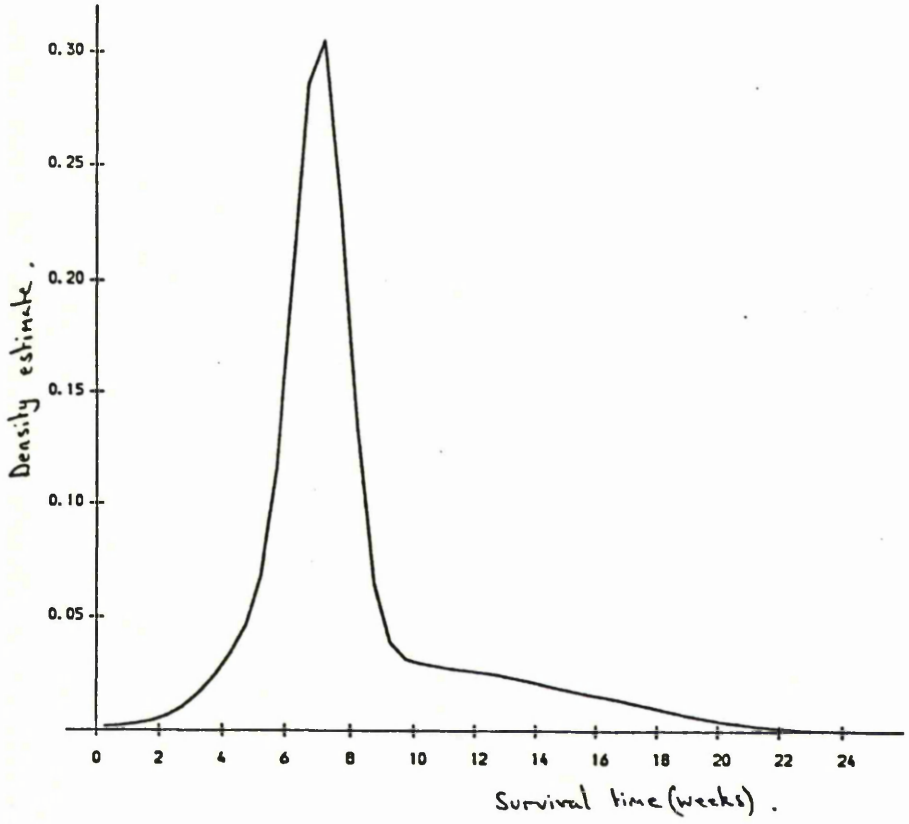
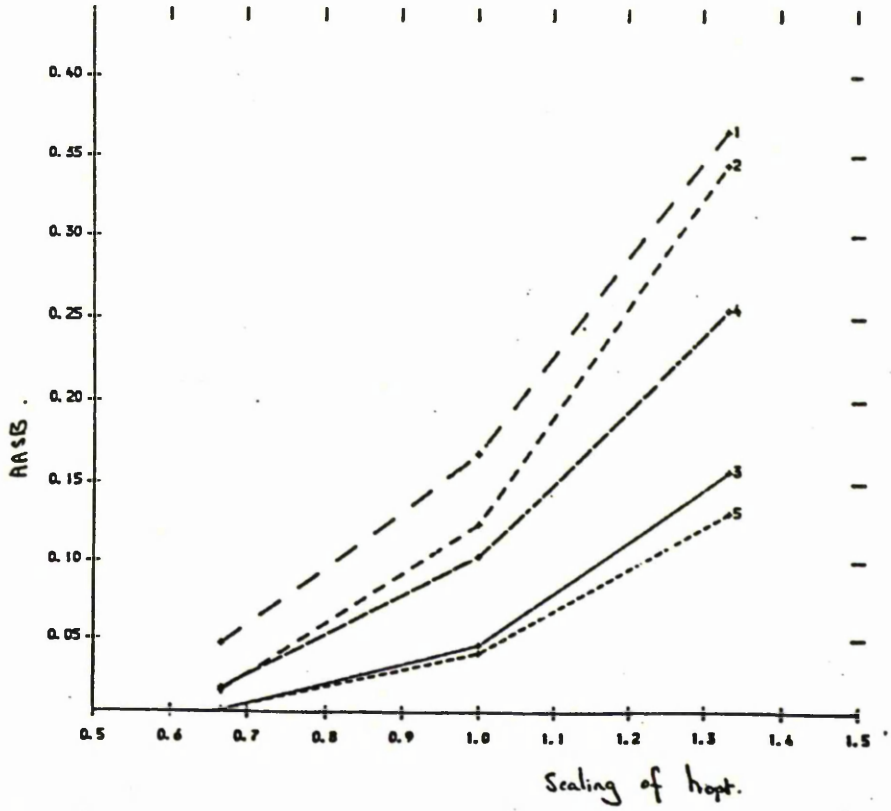


Figure 3.26. AASB incurred by the 5 estimators in estimating $g_1(x)$ on (0,1), n = 50, $\sigma^2 = 3.713$.



B. Nonparametric Kernel Regression.

3.10. Introduction.

In the following section of this chapter we will consider the univariate regression problem. It is assumed that we have observations (Y_i, x_i), $i = 1,\ldots,n$ which are described by the model

$$Y_i = g(x_i) + e_i \tag{3.10.1}$$

where $g(\cdot)$ is an unknown function having $k \geq 2$ continuous derivatives on $[a,b]$ and the errors $\{e_1,\ldots,e_n\}$ are uncorrelated with zero mean and constant variance, σ^2. It will also be assumed that the design variables $\{x_1,\ldots,x_n\}$ are equally spaced in $[a,b]$ so that

$$x_i = a + (i-0.5).\delta, \quad i = 1,\ldots,n \tag{3.10.2}$$

where

$$\delta = (b-a)/n. \tag{3.10.3}$$

It is required to estimate g on the basis of these observations without making any a priori assumptions as to its particular form, for example linear or quadratic. A number of nonparametric fixed kernel estimators have been proposed in the literature (Watson (1964), Gasser and Muller (1979)) but the one which will be studied here is that proposed by Priestley and Chao (1972), namely

$$\hat{g}(x) = \frac{\delta}{h} \sum_{i=1}^{n} K\left[\frac{x-x_i}{h}\right].Y_i \tag{3.10.4}$$

$K(\cdot)$ is the kernel function of order $(0,k)$ satisfying conditions (2.1.2) and (3.1.2) and h is the smoothing parameter assumed to be a function of n with

$$\lim_{n\to\infty} h \longrightarrow 0 \quad \text{and} \quad \lim_{n\to\infty} nh \longrightarrow \infty \tag{3.10.5}$$

If $[a,b]$ represents a finite interval than a particular problem in estimating g using a kernel with compact support $[-\tau,\tau]$ is increased bias at x-values in the boundary region $[a,a+h\tau) \cup (b-h\tau,b]$. When estimating g at x the x_i's in the interval $[x-h\tau,x+h\tau]$ will be used. If x lies near the boundary then this interval is not completely inside $[a,b]$ so that for equally spaced x_i's the estimate will be based on more data to one side of x than the other. If a kernel with infinite support such as the standard normal is used then it is effectively truncated in practice and so these effects will still be present in a boundary region rather than the whole of $[a,b]$. The only circumstances under which boundary effects will not occur is when g is a periodic function. Gasser and Muller (1979) and Rice (1984) discuss this problem in detail for particular estimators and describe modifications to the kernel function which reduce the boundary bias. To avoid having to consider the boundary in the following discussion in the rest of this Section and Sections 3.11-3.13 it will be assumed that g and its first k derivatives can be continuously periodically extended outside $[a,b]$. It will also be assumed, without loss of generality, that $a = 0$ and $b = 1$.

The exact mean and variance of $\hat{g}(x)$ are:

$$E[\hat{g}(x)] = \frac{1}{n} \sum_{i=1}^{n} K\left[\frac{x-x_i}{h}\right] g(x_i) \qquad (3.10.6)$$

and

$$V[\hat{g}(x)] = \frac{\delta^2\sigma^2}{n^2h^2} \sum_{i=1}^{n} K\left[\frac{x-x_i}{h}\right]^2 \qquad (3.10.7)$$

Calculations involving a Taylor series expansion follow in an analogous way to those for fixed kernel density estimators and yield

the following asymptotic expressions when K is of order 2:

$$E[\hat{g}(x)] = g(x) + \frac{h^2}{2} g^{(2)}(x) + o(h^2) \qquad (3.10.8)$$

and

$$V[\hat{g}(x)] = \frac{\sigma^2}{nh} \int_{-\tau}^{\tau} K(t)^2 dt + o((nh)^{-1}) \qquad (3.10.9)$$

(Priestley and Chao (1972)).

Integrating the squared bias and variance then results in the following expression for the asymptotic MISE:

$$MISE(\hat{g}) = \frac{h^4}{4} \int_0^1 g^{(2)}(x)^2 dx + \frac{\sigma^2}{nh} \int_{-\tau}^{\tau} K(t)^2 dt \qquad (3.10.10)$$

This is a function of the kernel K and the unknown quantities σ^2 and the second derivative of g. Benedetti (1977) shows that the Epanechnikov kernel is locally optimal for MSE when using the Priestley and Chao estimator. Expression (3.10.10) can be minimised with respect to h and results in an optimal h value proportional to $n^{-1/5}$ and a consequent convergence rate for the estimator of $n^{-4/5}$. If h satisfies conditions (3.10.5) then the estimator will be consistent in MISE.

Expression (3.10.8) indicates that the bias will be large when $|g^{(2)}(x)|$ is large which will occur at the peaks and troughs of g. In fact, when using h_{opt} the peaks of g are underestimated and the location of an asymmetric peak is biased towards the less steeply rising side of the peak, (Muller (1985)). A similar result will also hold for the troughs of g. As for density estimation, reducing h to overcome this problem simply increases the variance and hence MISE. We therefore seek methods which will reduce the bias of the estimator

(3.10.4) and also the MISE. Three approaches are considered which are:

a) By using one of the higher order kernels of Gasser et al (1985).

b) Subtracting an estimator of $(1/2)\, h^2\, g^{(2)}(x)$ from the original estimator based on a kernel of order 2.

c) Twicing as described by Stuetzle and Mittal (1979).

A simulation study carried out to examine finite sample performance for a variety of known functions g will also be described.

3.11. Minimum Variance and Optimal Kernels.

The discussion follows that for density estimators (Section 3.2) because the functionals to be minimised and hence the resulting kernels are the same.

3.12. Subtracting an estimate of $1/2\, h^2 g^{(2)}(x)$.

The regression function, g, will firstly be estimated using a fixed standard normal kernel which is of order 2 and then bias corrected by subtracting an estimate of the principal asymptotic bias term. This requires estimating $g^{(2)}(x)$ which will be carried out in two ways using a different kernel in each.

The first is by using the optimal kernel of order (2,4) of Gasser et al (1985). i.e.

$$K_2(t) = \begin{cases} (105/16)(-5t^4+6t^2-1) , & |t| < 1 \\ 0 , & \text{otherwise.} \end{cases} \qquad (3.12.1)$$

The optimal smoothing parameter, $h_{\nu opt}$, for estimating the ν^{th} derivative of a regression function with a kernel of order

(ν,k), in the sense that it minimises the expression for the asymptotic MISE, is

$$h_{\nu opt} = \left[\frac{2\nu+1}{2(k-\nu)} \cdot \frac{V_{\nu,k}}{B_{\nu,k}^2} \cdot \frac{\sigma^2}{\int g^{(k)}(t)^2 dt} \cdot \frac{1}{n}\right]^{\frac{1}{2k+1}} \qquad (3.12.2)$$

where

$$V_{\nu,k} = \int K_\nu(t)^2 dt \qquad (3.12.3)$$

and

$$B_{\nu,k} = \frac{(-1)^k}{k!} \int t^k K_\nu(t) dt \qquad (3.12.4)$$

The smoothing parameter choice for K_ν will be obtained using the factor method of Muller et al (1987) and described in the density estimation context in Section 3.6. This involves considering the ratio of optimal smoothing parameters for estimating the function and its second derivative using kernels of the same order k. The quantities depending on the unknown σ^2 and $g^{(k)}(t)$ then cancel out leaving the ratio as a known constant depending on the kernel functions. In general this constant, $d_{\nu,k}$, is given by:

$$d_{\nu,k} = \frac{h_{\nu,k\ opt}}{h_{o,k\ opt}} = \left[\frac{(2\nu+1)k}{(k-\nu)} \cdot \frac{V_{\nu,k}\, B_{o,k}^2}{V_{o,k}\, B_{\nu,k}^2}\right]^{1/2k+1} \qquad (3.12.5)$$

More specifically, when $\nu = 2$ choose

$$K(t) = \begin{cases} (15/32)(7t^4-10t^2+3), & |t| < 1 \\ 0, & \text{otherwise.} \end{cases} \qquad (3.12.6)$$

which is of order (0,4) so that when using $K_2(t)$ given by (3.12.1) the factor $d_{2,4} = 0.8919$ and

$$h_2 = 0.8919\ h \qquad (3.12.7)$$

where h is based on using (3.12.6).

$g^{(2)}(x)$ can also be estimated by using the second derivative of a standard normal density as a kernel function as carried out by Hardle and Bowman (1988), i.e.

$$K_2(t) = N^{(2)}(t;0,1) = (t^2-1)N(t;0,1) \qquad (3.12.8)$$

When the regression function is estimated using a standard normal kernel then the ratio of optimal smoothing parameters gives:

$$h_2 = 1.3r\sigma^{-4/45}.n^{4/45}\ .\ h \qquad (3.12.9)$$

This expression depends on the unknown error variance. The unbiased estimator used by Rice (1986), based on taking first differences, will be used in practice, i.e.

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{j=1}^{n-1} (Y_{j+1} - Y_j)^2 \qquad (3.12.10)$$

Also, the constant r is given by:

$$r = \frac{\left[\int g^{(4)}(t)^2 dt\right]^{-1/9}}{\left[\int g^{(2)}(t)^2 dt\right]^{-1/5}} \qquad (3.12.11)$$

The discussion in Hardle and Bowman (1988) for estimators on $[0,1]$ implies a choice of $r = 1.15$ which gives

$$h_2 = 1.5\ \sigma^{-1/11}\ .\ n^{1/11}\ .\ h \qquad (3.12.12)$$

where the fraction $4/45$ has been simplified to $1/11$.

3.13. Twicing.

This procedure, originally suggested by Tukey (1977), involves obtaining an estimate of g by the following steps:

(i) Calculate $\hat{g}(x_i)$, $i = 1,\ldots,n$.

(ii) Obtain residuals $r_i = y_i - \hat{g}(x_i)$, $i = 1,..,n$.

(iii) Smooth (r_i,x_i) using the same procedure as in (i) to obtain corrections c_i, $i = 1,\ldots,n$.

(iv) Define $\hat{g}(x_i)^c = \hat{g}(x_i) + c_i$, $i = 1,\ldots,n$ as the final estimate

Stuetzle and Mittal (1979) discuss the asymptotic effect of twicing for kernel regression estimators when the $\{x_i\}$ are equi-spaced and show that it is equivalent to using the kernel

$$W = 2.K - K * K \tag{3.13.1}$$

instead of K where $K * K$ is the convolution of K with itself. If K is a standard normal density then $K * K = N(t;0,2)$ so that

$$W(t) = 2.N(t;0,1) - N(t;0,2). \tag{3.13.2}$$

For (3.13.2) we have:

$$\int t^2W(t)dt = 2\int t^2N(t;0,1)dt - \int t^2N(t;0,2) = 0 \tag{3.13.3}$$

so that the bias term in h^2 is zero. This makes the bias $O(h^4)$ since the term in h^4 is non-zero, i.e.

$$\int t^4W(t)dt = 2\int t^4N(t;0,1)dt - \int t^4N(t;0,2) = -6 \tag{3.13.4}$$

Also,

$$\int W(t)^2dt = 2\int \{2.N(t;0,1)-N(t;0,2)\}^2dt$$

$$= 4\int N(t;0,1)^2dt - 4\int N(t,1)N(t,2)dt + \int N(t;0,2)^2dt = 0.41 \tag{3.13.5}$$

whereas,

$$\int K(t)^2dt = \int N(t;0,1)^2dt \int = 0.28 \qquad (3.13.6)$$

Hence, asymptotically, when $K(t) = N(t;0,1)$, twicing reduces the bias to $O(h^4)$ but increases the variance by nearly 50% for the same choice of smoothing parameter as when using $K(t)$ only.

3.14. Simulation Study.

In this section the finite sample performance of the estimators described in Sections 3.10 - 3.13 will be assessed. These are:

1. Using $K(t) = N(t;0,1)$

2. Using $K(t) = \begin{cases} (15/32)(7t^4-10t^2+3) & , |t| < 1 \\ 0 & , \text{otherwise.} \end{cases}$

3. Estimating the curve using method 1 and then subtracting a an estimate of $(h^2/2).g^{(2)}(x)$. The estimate of the second derivative is based on the kernel

$$K(t) = \begin{cases} (105/16)(-5t^4-6t^2-1) & , |t| < 1 \\ 0 & , \text{otherwise.} \end{cases}$$

with smoothing parameter computed by the factor method of Muller et al (1987) (i.e. by (3.12.7)).

4. The same as method 3 except that the kernel used in the estimation of the second derivative is the second derivative of the standard normal with h_2 calculated by (3.12.12).

5. Twicing based on a standard normal kernel estimate.

Methods 3, 4 and 5 therefore make adjustments to the estimate based on method 1 alone.

The simulations were based on the following five curves of different type:

1. $g_1(x) = N(x;0.5,0,1) + 16x(1-x)$

2. $g_2(x) = N(x;0.25,0.05) + N(x;0.5,0,1)$

3. $g_3(x) = g_2(x) + 1-2x-\log(x+0.6))$

4. $$g_4(x) = \begin{cases} 8x+4 & , \ -0.50 \leqslant x < 0.25 \\ -20x+11 & , \ 0.25 \leqslant x < 0.50 \\ 12x-5 & , \ 0.50 \leqslant x < 0.75 \\ -24x+22 & , \ 0.75 \leqslant x \leqslant 1.50 \end{cases}$$

5. $g_5(x) = 4\sin(2\pi x)$.

These curves were used in the simulation study of Muller and Stadtmuller (1987) in their assessment of an estimator based on a variable smoothing parameter.

Observations, y_i, were generated by considering equispaced points on a particular curve and adding noise sampled from a normal distribution with zero mean and

$$\sigma = 0.1 \left[\max \{g(x_i)\} - \min \{g(x_i)\} \right] \qquad (3.14.1)$$

where the $x_i \in (-0.5,1.5)$. Results were calculated for both 50 and 100 observations in this interval so that $\delta = 2/n$. This resulted in the following error variances being used:

Curve	σ^2 when $n = 50$	σ^2 when $n = 100$
1	3.713	3.861
2	0.633	0.665
3	1.287	1.330
4	3.779	3.905
5	0.637	0.640

The curves were only estimated on the interior region

(0,1) though to avoid problems in estimating near the boundary where the bias increases sharply. 26 observations are contained in (0,1) when $n = 50$ and 50 when $n = 100$.

Optimal smoothing parameters for the Priestly and Chao estimator based on the standard normal and the optimal of order 4 kernels were calculated for each curve using the finite evaluation technique described in Gasser et al (1984). This involves finding the value of h which satisfies:

$$\underset{h}{\text{Min}} \sum_{i} \left\{(\text{bias } (x_i))^2 + \text{var}(x_i)\right\} \qquad (3.14.2)$$

where the bias and variance are both based on the known curve i.e.

$$\text{bias}(x_i) = \frac{2}{nh} \sum_{j=1}^{n} K\left[\frac{x_i - x_j}{h}\right] g(x_j) - g(x_i) \qquad (3.14.3)$$

and

$$\text{var}(x_i) = \frac{4}{n^2h^2} \cdot \sigma^2 \cdot \sum_{j=1}^{n} \left\{K\left[\frac{x_i - x_j}{h}\right]\right\}^2 \qquad (3.14.4)$$

for $x_i \in (0,1)$.

The resulting smoothing parameters are given in the following table:

Table 3.16 Optimal smoothing parameters calculated by the finite evaluation technique.

	N = 50		N = 100	
Curve	K(t)=N(t;0,1)	K(t) = $(15/32)(7t^4-10t^2+3)$	K(t)=N(t;0,1)	K(t) = $(15/32(7t^4-10t^2+3)$
1	0.090	0.350	0.075	0.305
2	0.030	0.130	0.025	0.110
3	0.035	0.140	0.030	0.115
4	0.085	0.335	0.070	0.305
5	0.060	0.315	0.050	0.310

When $\delta = 2/n$ $E[\hat{g}(x)]$ is the same as (3.10.8) but the expression (3.10.9) for $V[\hat{g}(x)]$ should be multiplied by the factor 2. Hence, while choosing the value of h_2 by (3.12.12) h should also be scaled by $2^{-1/11} = 0.94$. There is no effect on the factor method though because when using kernels of the same order k (i.e. 4) the terms of $2^{1/9}$ for both $h_{0,4\ opt}$ and $h_{2,4\ opt}$ cancel.

The simulation size was a run of 1000 samples repeated 10 times for both $n = 50$ and $n = 100$. For each run the average squared bias, average variance and average mean squared error were calculated as in the simulations for the density estimators (3.8.1 - 3.8.3). These were then in turn averaged over the 10 runs to give the AASB, AAVAR and the AAMSE with the standard errors of these means measuring stability.

In practice the optimal values of h will clearly be unknown. Hence, a certain degree of either under or over smoothing will be carried out. To reflect this when $n = 50$ simulation results were evaluated for the optimal h-values scaled by the factors $2/3$ and $4/3$.

Seeds for the random number generator were again chosen so that the results for each estimator are directly comparable for a given curve and sample size. They are also such that direct comparisons can be made between the results for a given method when using different amounts of smoothing to estimate a particular curve.

The full results of the simulation study are contained in tables 3.17-3.22. Also, when $n = 50$ and for the three scalings of h_{opt} theory are presented graphically in figures 3.26-3.40.

When using the optimal h values twicing is clearly the most effective method for reducing the bias. It has the lowest AASB values for each curve when $n = 50$ and the lowest for each except curve 3 at $n = 100$ when it has the second lowest. The method with the worst performance in terms of bias is just using the unadjusted standard normal kernel. This has the highest AASB values for each curve at both $n = 50$ and 100. Out of the two methods which subtract estimates of $1/2h^2g^{(2)}(X)$, method 3, which uses the optimal kernel of order $(2,4)$ and selects h_2 using the factor method, reduces bias considerably more than does method 4. They also both achieve much lower AASB's than method 1 for curve 4 which is not differentiable. The AASB results for the optimal kernel of order $(0,4)$ generally rank at about 3rd or 4th but are much higher than for methods 1 and 3.

For the curves and associated error variances based on (3.14.1) used in this study the AAVAR's for each method are much higher than the AASB's. They are in contrast however to the bias results with the rankings of the methods in terms of AAVAR generally the opposite to those for AASB. Using an $N(0,1)$ kernel is therefore the most effective for reducing variance whilst the AAVAR's for twicing are about 50% higher as predicted by the asymptotic theory.

When the results for bias and variance are combined the best performance is by method 2, the optimal kernel of order $(0,4)$. It has the lowest AAMSE results for each curve except 1 at both $n = 50$ and 100. The poorest overall performance is for twicing with its AAMSE results more markedly dominated than for the others by its poor variance. The AAMSE's of the other three methods are

fairly similar but method 4 is generally slightly superior to both methods 1 and 3.

When the data are undersmoothed using $^2/_3$. h_{opt} the bias of each method decreases while the variances increase as expected. Twicing (method 5) still has the lowest AASB for curves 2, 3 and 4 but the results for method 3 are slightly superior for curves 1 and 5. The orderings in terms of AAVAR are almost the same as when using h_{opt}. All the AAMSE's are higher than at h_{opt} and the best overall is now method 1 with the lowest AAMSE values for curves 1, 3 and 4 and second lowest for 2 and 5. Method 2 now has the second best overall performance.

In contrast, oversmoothing using $^4/_3$ h_{opt} increase bias and reduces variance. The most marked changes in AASB are for methods 1, 2 and 4 while twicing (method 5) achieves the lowest AASB for each curve. The AASB values for methods 3 and 5 are in fact fairly similar to those for method 2 at h_{opt}. The orderings in terms of the AAVAR's are again very similar to when using h_{opt}. The AAMSE's for methods 1, 2 and 4 generally increase above their h_{opt} values but those for methods 3 and 5 decrease for each curve. This results in method 5 having the best performance for each curve except 5 when method 3 is slightly superior. Their AAMSE values are fairly similar to those for method 2 at h_{opt} but the fact that they are still decreasing indicates that they may have become superior if a scaling factor larger than $^4/_3$ had been used.

In conclusion then, this simulation study indicates that the best method for reducing bias is twicing (method 5). It is very effective in this at each of the h-values and sample sizes considered but at the expense of increased variance. It can also achieve low mean squared

error when scaling h_{opt} for the N(0,1) kernel alone by $4/3$. For this value of h its level of bias was fairly similar to those for methods 2 and 4 and smaller than for method 1 at h_{opt}. However, method 3 based on subtracting an estimate of $(h^2/2)\, g^{(2)}(x)$ is only slightly inferior to twicing both in terms of reducing bias and low MSE at $4/3\ h_{opt}$.

For a real data set a practical approach would be to obtain a data based choice of h for an estimate based on an N(0,1) kernel, such as by least squares cross-validation. Using twicing or method 3 with this h will generally provide an estimate with much lower bias than the estimate based on just using the N(0,1) kernel. If this h is then scaled by $4/3$ the resulting estimate should still have lower bias than the estimate which uses only the N(0,1) kernel and original h. In addition though, it should now have a lower mean squared error.

Table 3.17. Values of AASB, AAVAR and AAMSE incurred in estimating $g_1(x)$ from samples of size 50. ($\sigma^2 = 3.713$).

Scaling factor for h_{opt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	4.70×10^{-2} (8.9×10^{-4})	7.07×10^{-1} (2.5×10^{-3})	7.51×10^{-1}
	2	1.69×10^{-2} (7.6×10^{-4})	8.05×10^{-1} (3.0×10^{-3})	8.22×10^{-1}
	3	5.12×10^{-3} (5.5×10^{-4})	9.83×10^{-1} (3.2×10^{-3})	9.88×10^{-1}
	4	1.87×10^{-2} (7.6×10^{-4})	8.12×10^{-1} (2.9×10^{-3})	8.31×10^{-1}
	5	5.32×10^{-3} (5.6×10^{-4})	1.01×10^{0} (3.1×10^{-3})	1.02×10^{0}
1	1	1.68×10^{-1} (1.2×10^{-3})	4.71×10^{-1} (2.0×10^{-3})	6.39×10^{-1}
	2	1.24×10^{-1} (5.4×10^{-4})	5.38×10^{-1} (2.5×10^{-3})	6.62×10^{-1}
	3	4.57×10^{-2} (7.3×10^{-4})	6.59×10^{-1} (2.8×10^{-3})	7.04×10^{-1}
	4	1.03×10^{-1} (8.3×10^{-4})	5.43×10^{-1} (2.4×10^{-3})	6.46×10^{-1}
	5	4.07×10^{-2} (7.9×10^{-4})	6.78×10^{-1} (2.8×10^{-3})	7.19×10^{-1}
4/3	1	3.65×10^{-1} (1.9×10^{-3})	3.53×10^{-1} (1.6×10^{-3})	7.19×10^{-1}
	2	3.44×10^{-1} (1.3×10^{-3})	4.03×10^{-1} (1.8×10^{-3})	7.47×10^{-1}
	3	1.57×10^{-1} (7.0×10^{-4})	4.95×10^{-1} (2.3×10^{-3})	6.56×10^{-1}
	4	2.55×10^{-1} (9.8×10^{-4})	4.08×10^{-1} (1.9×10^{-3})	6.62×10^{-1}
	5	1.31×10^{-1} (7.7×10^{-4})	5.10×10^{-1} (2.3×10^{-3})	6.41×10^{-1}

Table 3.18. Values of AASB, AAVAR and AAMSE incurred in estimating $g_2(x)$ from samples of size 50. ($\sigma^2 = 0.633$).

Scaling factor for h_{opt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	1.32×10^{-2} (3.2×10^{-4})	4.18×10^{-1} (1.1×10^{-3})	4.31×10^{-1}
	2	8.96×10^{-3} (1.8×10^{-4})	3.61×10^{-1} (1.0×10^{-3})	3.70×10^{-1}
	3	3.69×10^{-2} (4.4×10^{-4})	5.33×10^{-1} (1.4×10^{-3})	5.69×10^{-1}
	4	4.07×10^{-3} (1.2×10^{-4})	4.62×10^{-1} (1.3×10^{-3})	4.66×10^{-1}
	5	7.58×10^{-4} (6.0×10^{-5})	5.56×10^{-1} (1.6×10^{-3})	5.57×10^{-1}
1	1	9.06×10^{-2} (7.6×10^{-4})	2.41×10^{-1} (6.3×10^{-4})	3.31×10^{-1}
	2	6.26×10^{-2} (5.7×10^{-4})	2.43×10^{-1} (6.6×10^{-4})	3.06×10^{-1}
	3	2.19×10^{-2} (3.5×10^{-4})	3.13×10^{-1} (8.2×10^{-4})	3.35×10^{-1}
	4	5.02×10^{-2} (5.7×10^{-4})	2.68×10^{-1} (7.0×10^{-4})	3.18×10^{-1}
	5	1.07×10^{-2} (2.8×10^{-4})	3.48×10^{-1} (9.5×10^{-4})	3.59×10^{-1}
4/3	1	2.21×10^{-1} (9.5×10^{-4})	1.79×10^{-1} (5.1×10^{-4})	4.00×10^{-1}
	2	2.52×10^{-1} (7.6×10^{-4})	1.84×10^{-1} (6.0×10^{-4})	4.36×10^{-1}
	3	7.82×10^{-2} (6.3×10^{-4})	2.34×10^{-1} (6.3×10^{-4})	3.12×10^{-1}
	4	1.65×10^{-1} (8.5×10^{-4})	2.00×10^{-1} (5.7×10^{-4})	3.65×10^{-1}
	5	5.01×10^{-2} (5.4×10^{-4})	2.58×10^{-1} (6.9×10^{-4})	3.08×10^{-1}

Table 3.19. Values of AASB, AAVAR and AAMSE incurred in estimating $g_3(x)$ from samples of size 50. ($\sigma^2 = 1.287$).

Scaling factor for h_{opt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	3.25×10^{-2} (7.3×10^{-4})	6.80×10^{-1} (1.7×10^{-3})	7.13×10^{-1}
	2	3.57×10^{-2} (6.0×10^{-4})	7.06×10^{-1} (2.0×10^{-3})	7.42×10^{-1}
	3	1.22×10^{-2} (3.8×10^{-4})	9.07×10^{-1} (2.5×10^{-3})	9.19×10^{-1}
	4	1.02×10^{-2} (3.8×10^{-4})	7.63×10^{-1} (2.1×10^{-3})	7.73×10^{-1}
	5	2.38×10^{-3} (1.7×10^{-4})	9.73×10^{-1} (2.7×10^{-3})	9.75×10^{-1}
1	1	1.49×10^{-1} (1.3×10^{-3})	4.17×10^{-1} (1.1×10^{-3})	5.66×10^{-1}
	2	9.05×10^{-2} (9.2×10^{-4})	4.64×10^{-1} (1.3×10^{-3})	5.55×10^{-1}
	3	3.19×10^{-2} (6.0×10^{-4})	5.68×10^{-1} (1.5×10^{-3})	6.00×10^{-1}
	4	9.29×10^{-2} (1.0×10^{-3})	4.70×10^{-1} (1.3×10^{-3})	5.62×10^{-1}
	5	2.56×10^{-2} (6.0×10^{-4})	6.00×10^{-1} (1.6×10^{-3})	6.26×10^{-1}
4/3	1	3.45×10^{-1} (1.5×10^{-3})	3.11×10^{-1} (9.5×10^{-4})	6.56×10^{-1}
	2	3.43×10^{-1} (1.0×10^{-3})	3.47×10^{-1} (1.1×10^{-3})	6.90×10^{-1}
	3	1.31×10^{-1} (1.0×10^{-3})	4.26×10^{-1} (1.2×10^{-3})	5.57×10^{-1}
	4	2.76×10^{-1} (1.3×10^{-3})	3.50×10^{-1} (1.0×10^{-3})	6.27×10^{-1}
	5	1.07×10^{-1} (1.0×10^{-3})	4.47×10^{-1} (1.2×10^{-3})	5.54×10^{-1}

Table 3.20. Values of AASB, AAVAR and AAMSE incurred in estimating $g_4(x)$ from samples of size 50. ($\sigma^2 = 3.713$).

Scaling factor for h_{opt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	6.06×10^{-2} (5.4×10^{-4})	7.54×10^{-1} (2.6×10^{-3})	8.15×10^{-1}
	2	2.64×10^{-2} (4.0×10^{-5})	8.56×10^{-1} (3.1×10^{-3})	8.82×10^{-1}
	3	1.65×10^{-2} (1.6×10^{-4})	1.05×10^{0} (3.3×10^{-3})	1.06×10^{0}
	4	2.62×10^{-2} (3.2×10^{-4})	8.69×10^{-1} (3.0×10^{-3})	8.95×10^{-1}
	5	1.27×10^{-2} (2.2×10^{-4})	1.08×10^{0} (3.3×10^{-3})	1.10×10^{-1}
1	1	1.95×10^{-1} (1.0×10^{-3})	5.07×10^{-1} (2.2×10^{-3})	7.03×10^{-1}
	2	1.25×10^{-1} (6.6×10^{-4})	5.71×10^{-1} (2.7×10^{-3})	6.97×10^{-1}
	3	4.85×10^{-2} (5.1×10^{-4})	7.03×10^{-1} (2.9×10^{-3})	7.51×10^{-1}
	4	1.18×10^{-1} (7.0×10^{-4})	5.85×10^{-1} (2.5×10^{-3})	7.03×10^{-1}
	5	4.62×10^{-2} (4.7×10^{-4})	7.30×10^{-1} (2.9×10^{-3})	7.76×10^{-1}
4/3	1	4.21×10^{-1} (1.4×10^{-3})	3.82×10^{-1} (1.7×10^{-3})	8.03×10^{-1}
	2	4.24×10^{-1} (1.4×10^{-3})	4.28×10^{-1} (1.9×10^{-3})	8.52×10^{-1}
	3	1.70×10^{-1} (7.9×10^{-4})	5.29×10^{-1} (2.5×10^{-3})	6.99×10^{-1}
	4	3.10×10^{-1} (1.3×10^{-3})	4.41×10^{-1} (2.0×10^{-3})	7.51×10^{-1}
	5	1.43×10^{-1} (7.5×10^{-4})	5.51×10^{-1} (2.5×10^{-3})	6.94×10^{-1}

Table 3.21. Values of AASB, AAVAR and AAMSE incurred in estimating $g_5(x)$ from samples of size 50. ($\sigma^2 = 0.637$).

Scaling factor for h_{opt}	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
2/3	1	7.22×10^{-3} (1.9×10^{-4})	1.81×10^{-1} (5.1×10^{-4})	1.88×10^{-1}
	2	3.30×10^{-4} (4.0×10^{-5})	1.53×10^{-1} (5.4×10^{-4})	1.53×10^{-1}
	3	2.01×10^{-4} (3.3×10^{-5})	2.14×10^{-1} (6.0×10^{-4})	2.14×10^{-1}
	4	2.24×10^{-4} (3.3×10^{-5})	2.00×10^{-1} (5.7×10^{-4})	2.00×10^{-1}
	5	2.12×10^{-4} (3.4×10^{-5})	2.60×10^{-1} (7.0×10^{-4})	2.60×10^{-1}
1	1	3.55×10^{-2} (4.1×10^{-4})	1.21×10^{-1} (4.4×10^{-4})	1.56×10^{-1}
	2	6.23×10^{-3} (1.8×10^{-4})	1.03×10^{-1} (4.7×10^{-4})	1.09×10^{-1}
	3	6.88×10^{-4} (5.9×10^{-5})	1.43×10^{-1} (5.1×10^{-4})	1.44×10^{-1}
	4	2.30×10^{-3} (1.1×10^{-4})	1.34×10^{-1} (4.7×10^{-4})	1.36×10^{-1}
	5	2.40×10^{-4} (3.5×10^{-5})	1.74×10^{-1} (5.4×10^{-4})	1.74×10^{-1}
4/3	1	1.07×10^{-1} (7.0×10^{-4})	9.09×10^{-2} (3.8×10^{-4})	1.98×10^{-1}
	2	4.30×10^{-2} (4.7×10^{-4})	7.69×10^{-2} (3.5×10^{-4})	1.20×10^{-1}
	3	4.95×10^{-3} (1.6×10^{-4})	1.08×10^{-1} (4.7×10^{-4})	1.13×10^{-1}
	4	1.79×10^{-2} (2.8×10^{-4})	1.01×10^{-1} (4.1×10^{-4})	1.19×10^{-1}
	5	1.45×10^{-3} (8.6×10^{-5})	1.31×10^{-1} (5.0×10^{-4})	1.33×10^{-1}

Table 3.22. Values of AASB, AAVAR and AAMSE incurred in estimating $g_1(x)$, $g_2(x)$, $g_3(x)$, $g_4(x)$ and $g_5(x)$ from samples of size 100 and using optimal smoothing parameters.

Curve	Method	AASB (s.e.)	AAVAR (s.e.)	AAMSE
1	1	1.03×10^{-1} (8.2×10^{-4})	2.89×10^{-1} (1.6×10^{-3})	3.92×10^{-1}
	2	7.16×10^{-2} (4.7×10^{-4})	3.15×10^{-1} (1.8×10^{-3})	3.87×10^{-1}
	3	2.23×10^{-2} (4.4×10^{-4})	3.92×10^{-1} (1.9×10^{-3})	4.15×10^{-1}
	4	5.99×10^{-2} (6.0×10^{-4})	3.27×10^{-1} (1.8×10^{-3})	3.87×10^{-1}
	5	1.85×10^{-2} (4.4×10^{-4})	4.16×10^{-1} (2.0×10^{-3})	4.35×10^{-1}
2	1	5.07×10^{-2} (4.1×10^{-4})	1.49×10^{-1} (4.4×10^{-4})	2.00×10^{-1}
	2	2.43×10^{-2} (2.5×10^{-4})	1.50×10^{-1} (4.4×10^{-4})	1.74×10^{-1}
	3	6.70×10^{-3} (1.3×10^{-4})	1.93×10^{-1} (5.4×10^{-4})	1.99×10^{-1}
	4	2.56×10^{-2} (2.8×10^{-4})	1.62×10^{-1} (4.7×10^{-4})	1.88×10^{-1}
	5	3.70×10^{-3} (1.3×10^{-4})	2.15×10^{-1} (6.0×10^{-4})	2.19×10^{-1}
3	1	9.39×10^{-2} (7.3×10^{-4})	2.48×10^{-1} (8.2×10^{-4})	3.42×10^{-1}
	2	3.32×10^{-2} (3.5×10^{-4})	2.87×10^{-1} (8.9×10^{-4})	3.21×10^{-1}
	3	1.05×10^{-2} (2.2×10^{-4})	3.51×10^{-1} (9.8×10^{-4})	3.61×10^{-1}
	4	5.50×10^{-2} (5.1×10^{-4})	2.74×10^{-1} (8.9×10^{-4})	3.29×10^{-1}
	5	1.11×10^{-2} (2.5×10^{-4})	3.58×10^{-1} (1.0×10^{-3})	3.69×10^{-1}
4	1	1.18×10^{-1} (9.8×10^{-4})	3.13×10^{-1} (1.6×10^{-3})	4.32×10^{-1}
	2	8.79×10^{-2} (6.3×10^{-4})	3.19×10^{-1} (1.8×10^{-3})	4.07×10^{-1}
	3	3.35×10^{-2} (3.5×10^{-4})	4.07×10^{-1} (2.0×10^{-3})	4.96×10^{-1}
	4	6.61×10^{-2} (6.0×10^{-4})	3.54×10^{-1} (1.8×10^{-3})	4.20×10^{-1}
	5	2.59×10^{-2} (3.2×10^{-4})	4.51×10^{-1} (2.1×10^{-3})	4.77×10^{-1}
5	1	1.87×10^{-2} (3.5×10^{-4})	7.18×10^{-2} (3.2×10^{-4})	9.05×10^{-2}
	2	5.10×10^{-3} (1.9×10^{-4})	5.14×10^{-2} (2.8×10^{-4})	5.65×10^{-2}
	3	4.20×10^{-4} (5.1×10^{-5})	8.03×10^{-2} (3.5×10^{-4})	8.07×10^{-2}
	4	9.87×10^{-4} (8.0×10^{-5})	7.80×10^{-2} (3.5×10^{-4})	7.90×10^{-2}
	5	1.30×10^{-4} (2.2×10^{-5})	1.03×10^{-1} (3.8×10^{-4})	1.03×10^{-1}

Figure 3.27. AASB incurred by the 5 estimators in estimating $g_2(x)$ on (0,1), n = 50, σ^2 = 0.633.

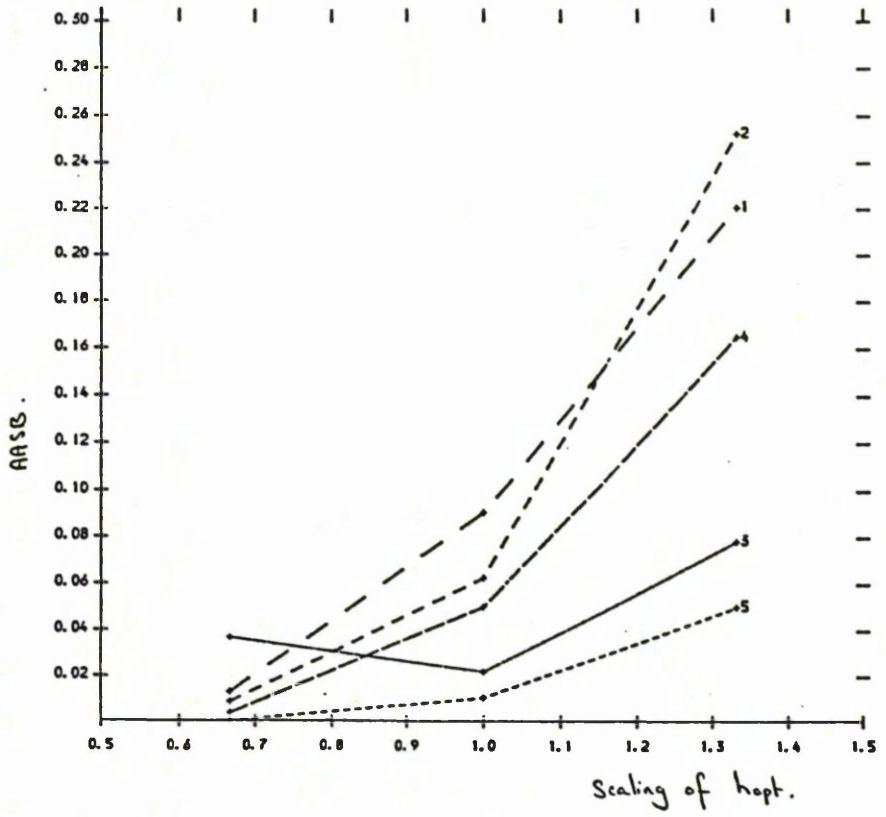


Figure 3.28. AASB incurred by the 5 estimators in estimating $g_3(x)$ on (0,1), n = 50, σ^2 = 1.287.

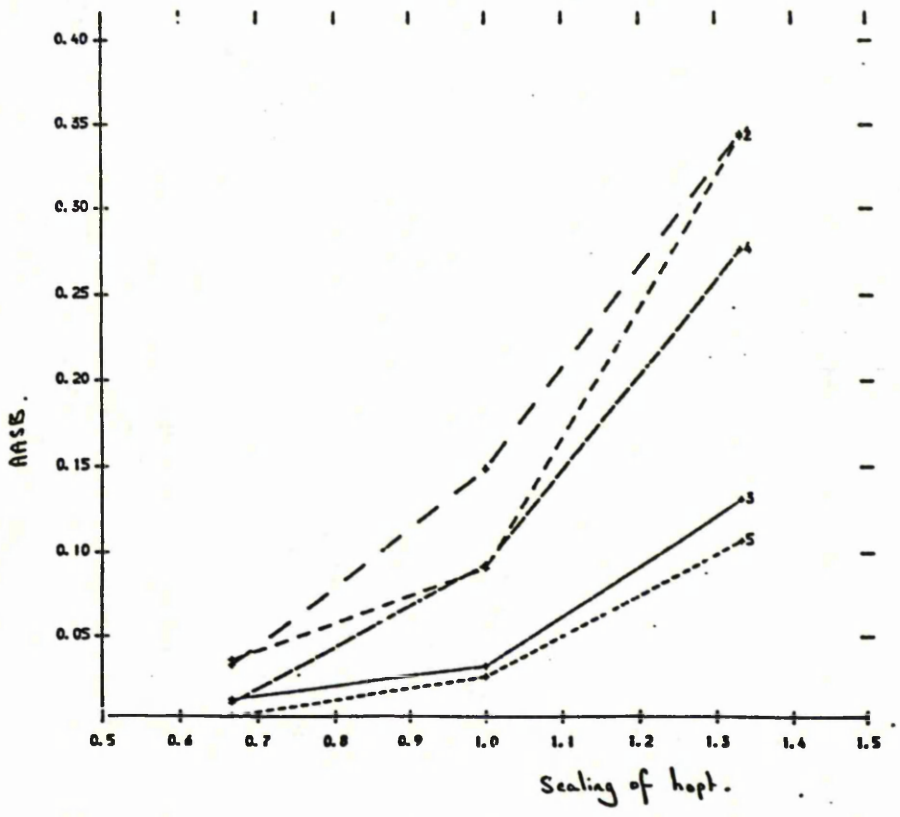


Figure 3.29. AASB incurred by the 5 estimators in estimating $g_4(x)$ on (0,1), n = 50, σ^2 = 3.779.

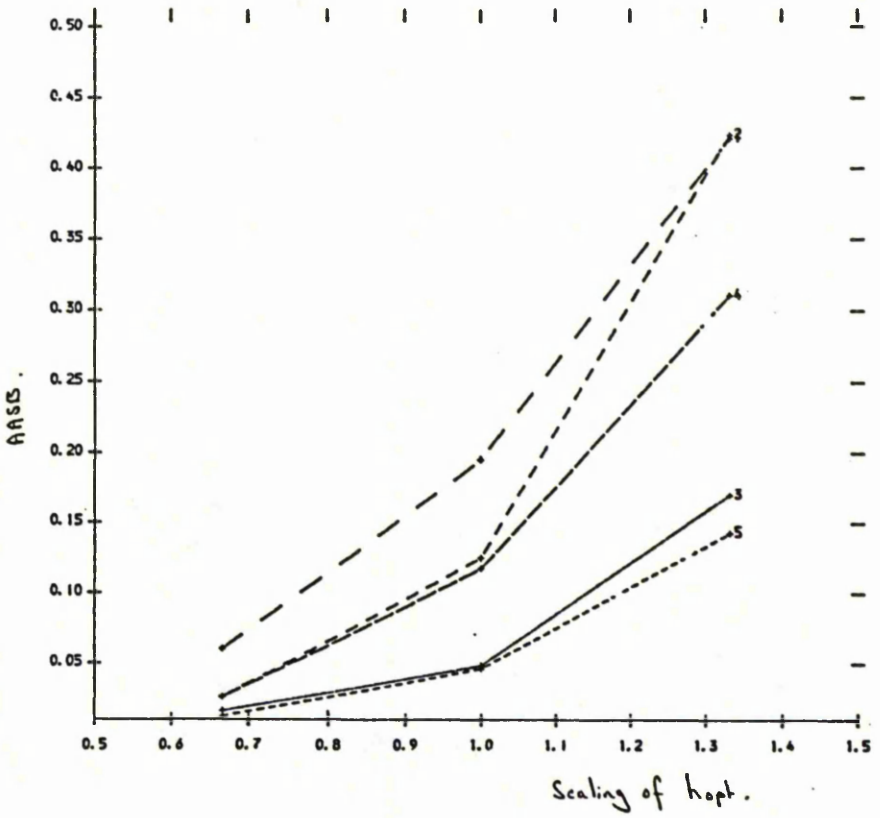


Figure 3.30. AASB incurred by the 5 estimators in estimating $g_5(x)$ on (0,1), n = 50, σ^2 = 0.637.

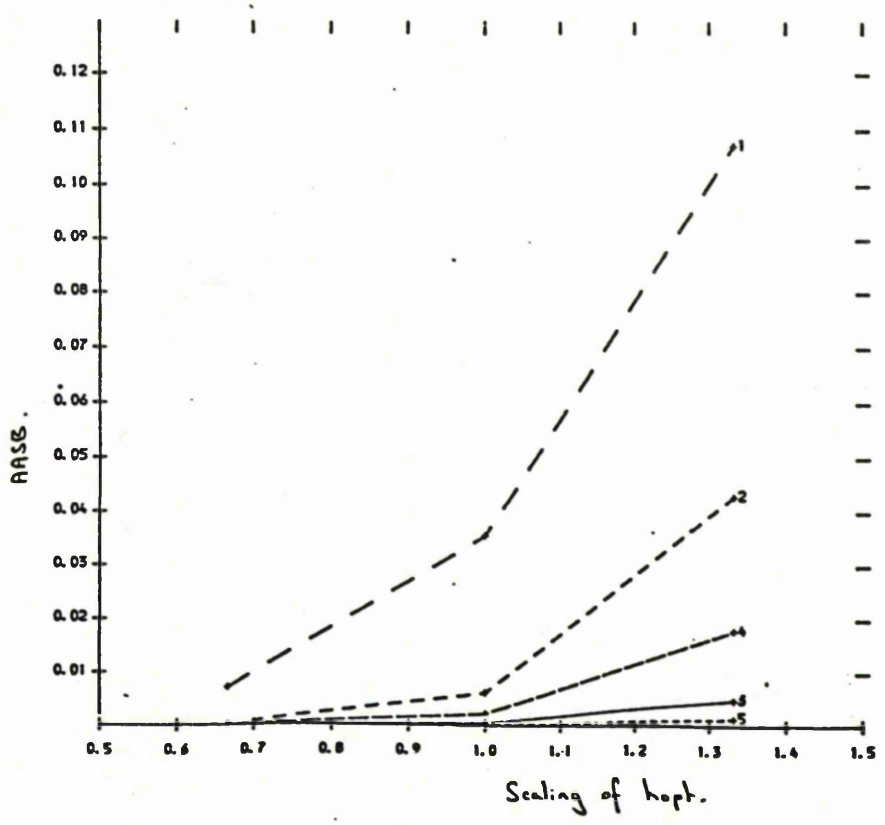


Figure 3.31. AAVAR incurred by the 5 estimators in estimating $g_1(x)$ on (0,1), n = 50, σ^2 = 3.713.

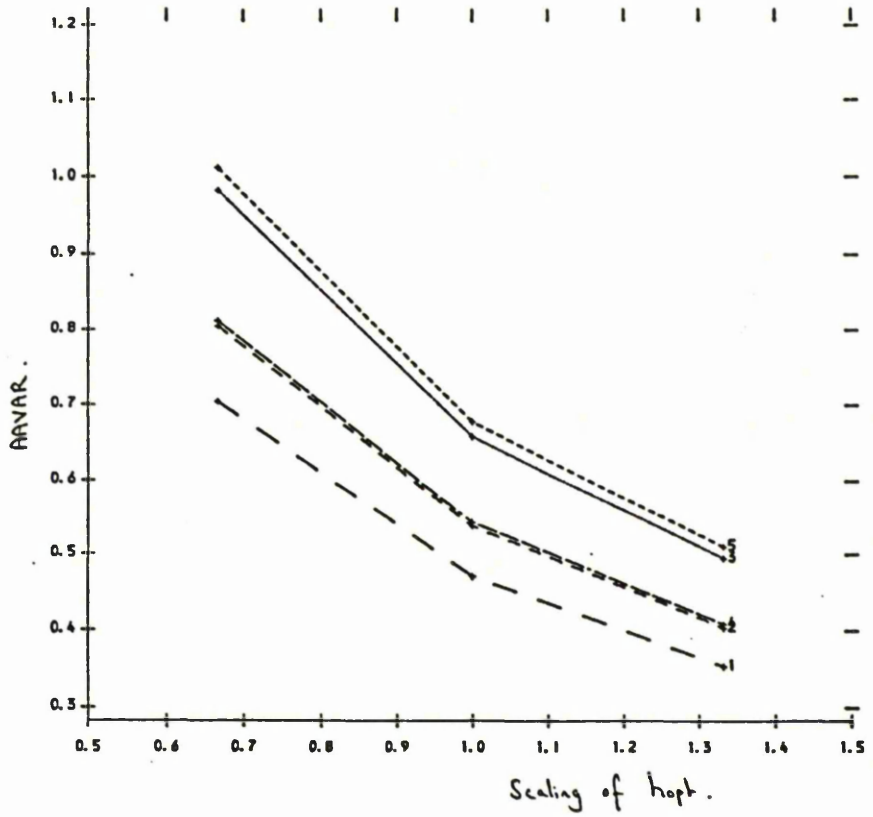


Figure 3.32. AAVAR incurred by the 5 estimators in estimating $g_2(x)$ on (0,1), n = 50, σ^2 = 0.633.

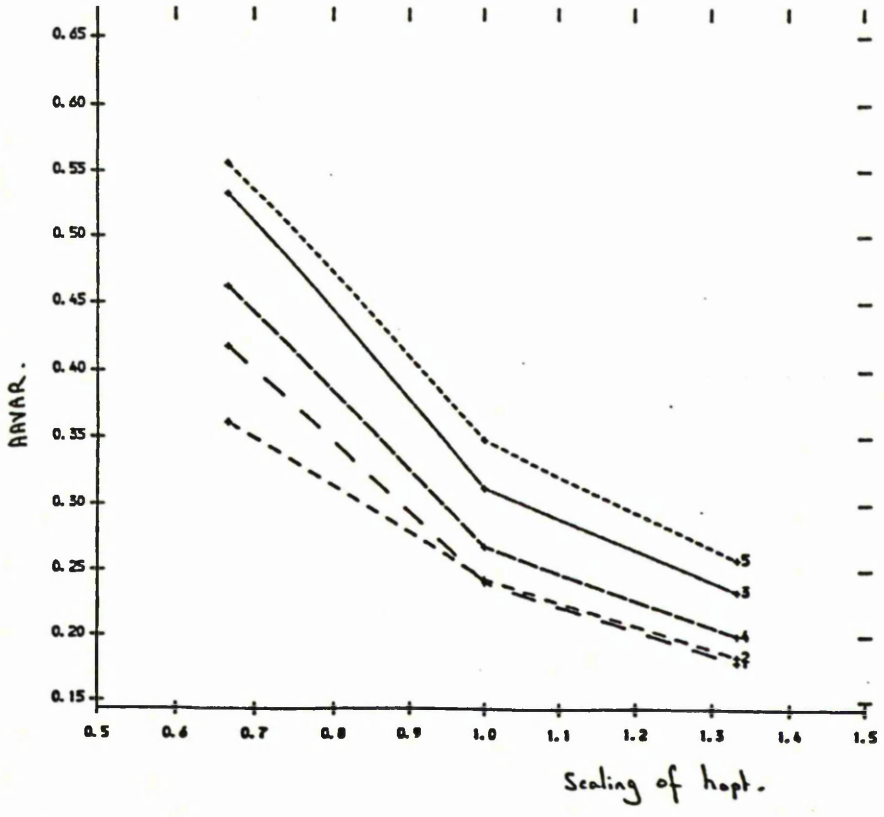


Figure 3.33. AAVAR incurred by the 5 estimators in estimating $g_3(x)$ on (0,1), n = 50, σ^2 = 1.287.

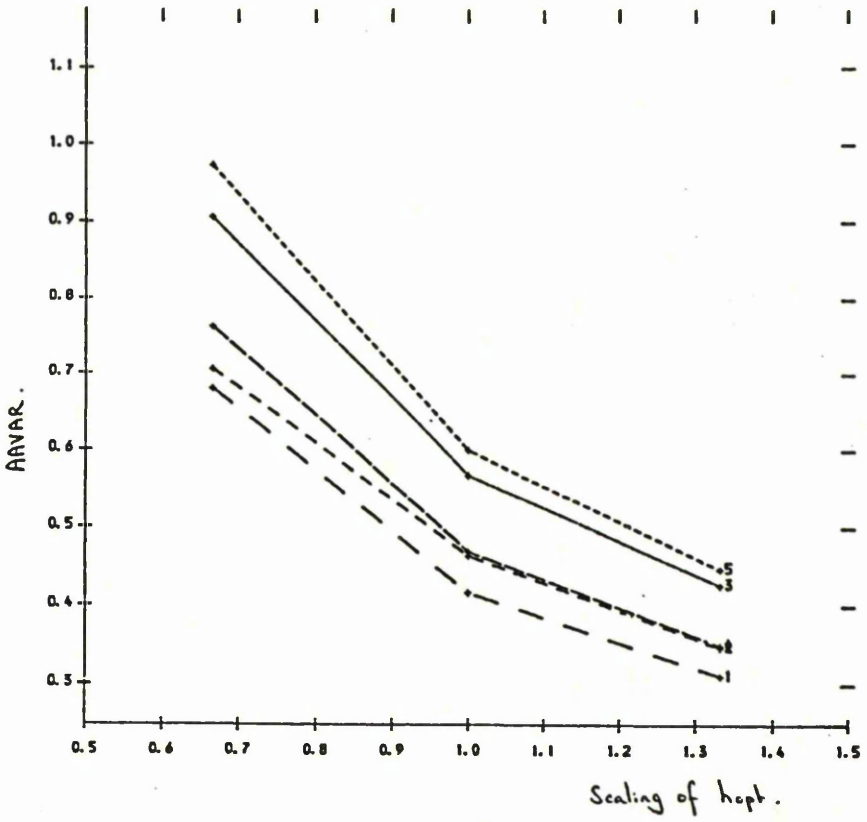


Figure 3.34. AAVAR incurred by the 5 estimators in estimating $g_4(x)$ on (0,1), n = 50, σ^2 = 3.779.

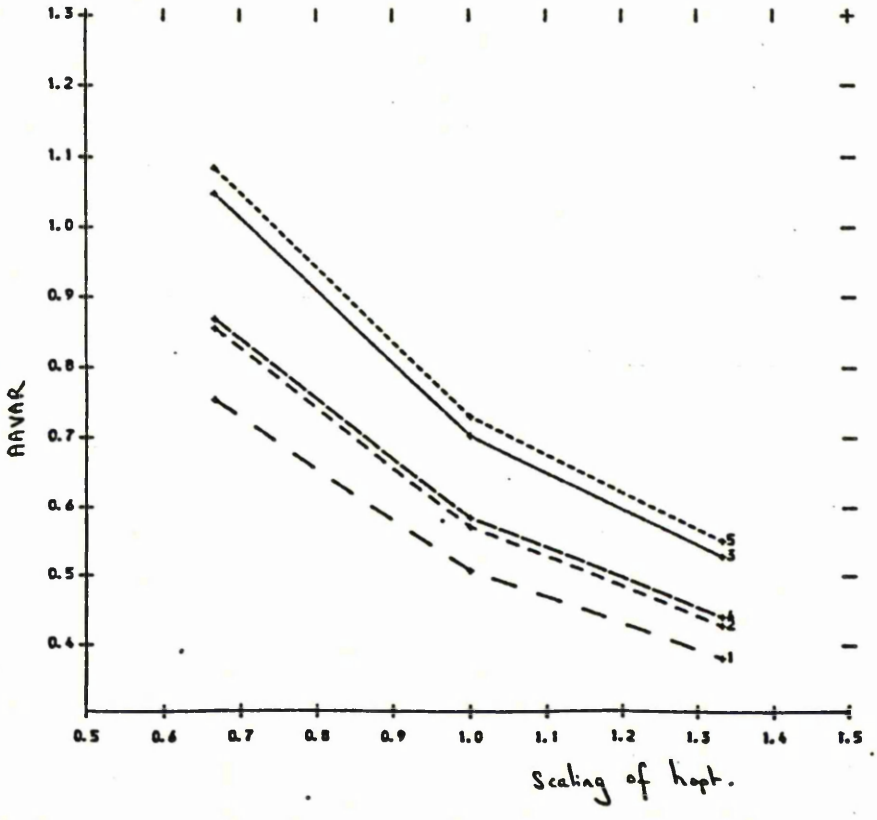


Figure 3.35. AAVAR incurred by the 5 estimators in estimating $g_5(x)$ on (0,1), n = 50, σ^2 = 0.637.

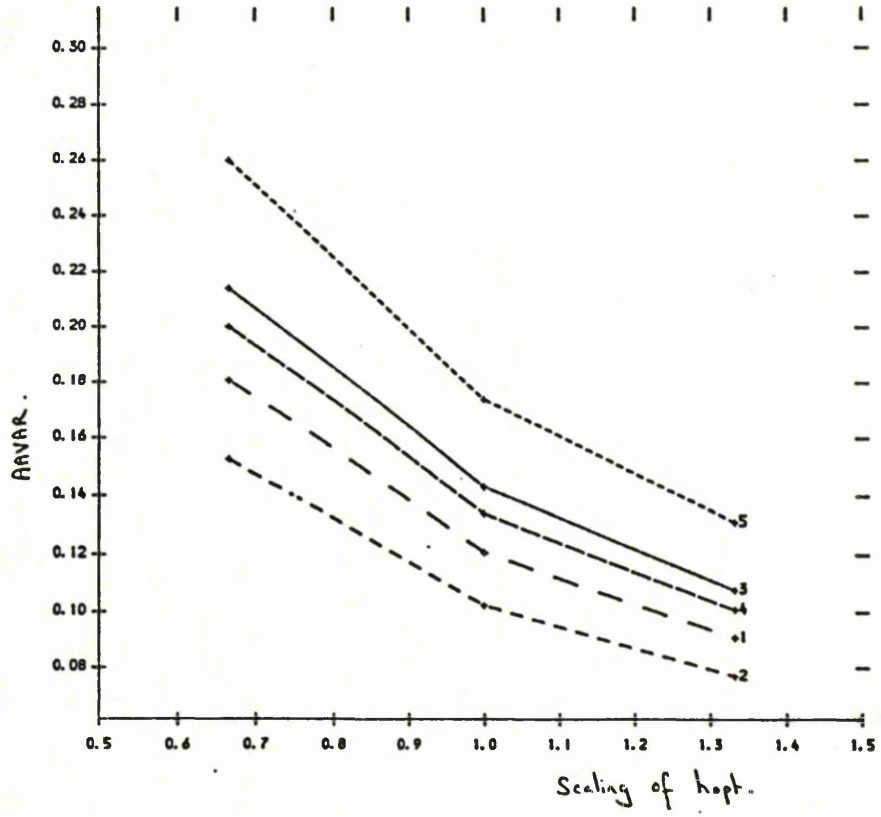


Figure 3.36. AAMSE incurred by the 5 estimators in estimating $g_1(x)$ on (0,1), n = 50, σ^2 = 3.713.

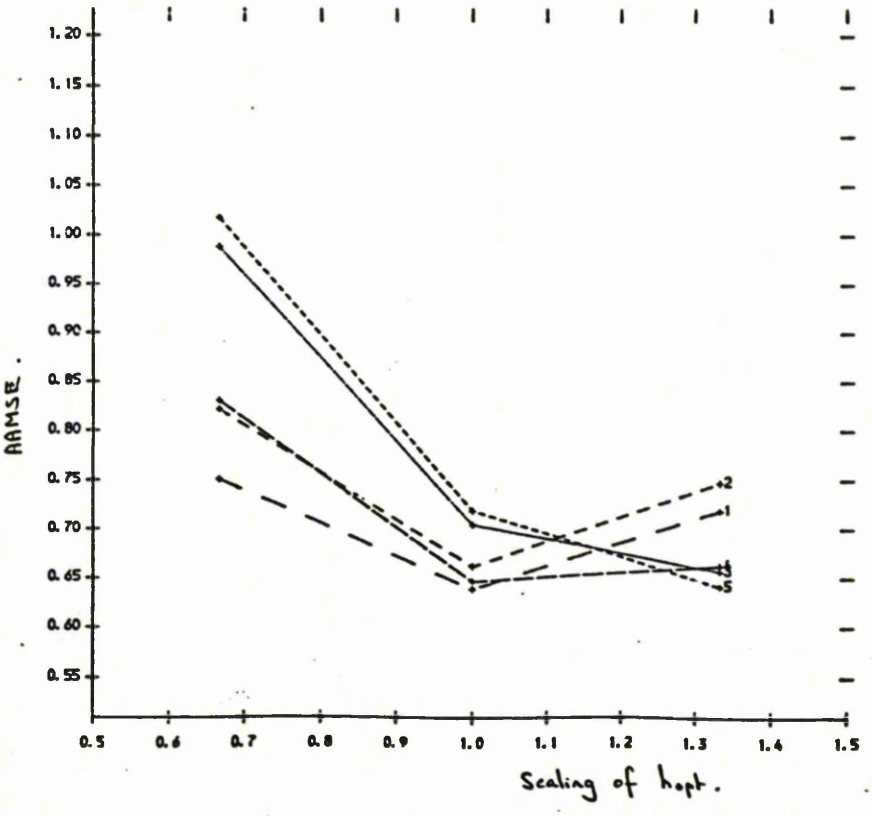


Figure 3.37. AAMSE incurred by the 5 estimators in estimating $g_2(x)$ on (0,1), n = 50, σ^2 = 0.633.

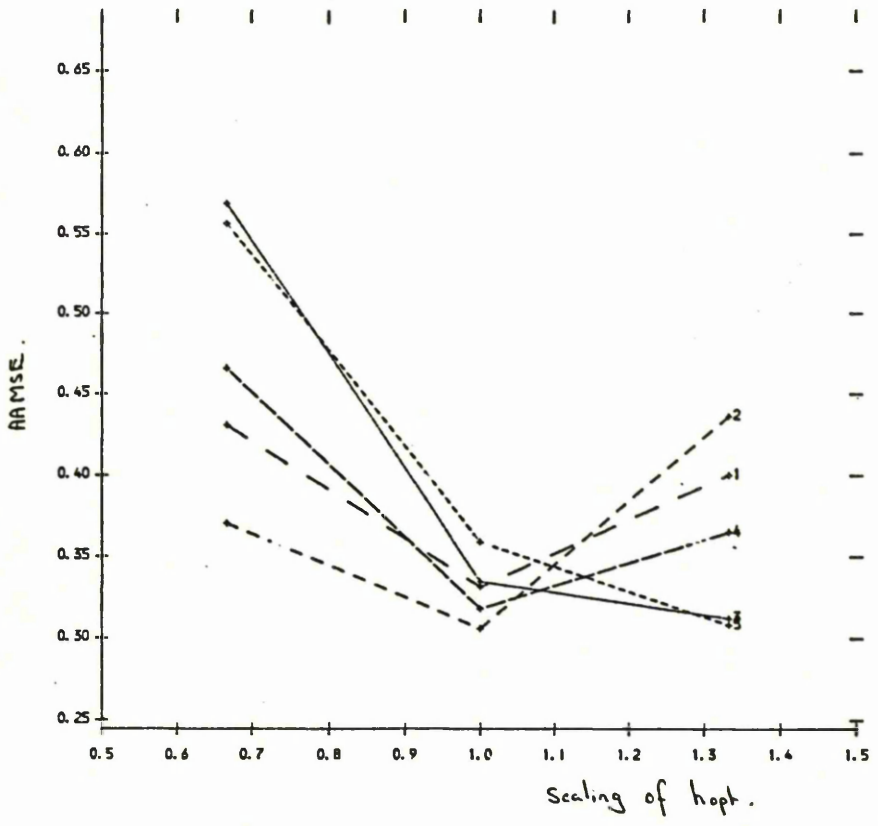


Figure 3.38. AAMSE incurred by the 5 estimators in estimating $g_3(x)$ on (0,1), n = 50, σ^2 = 1.287.

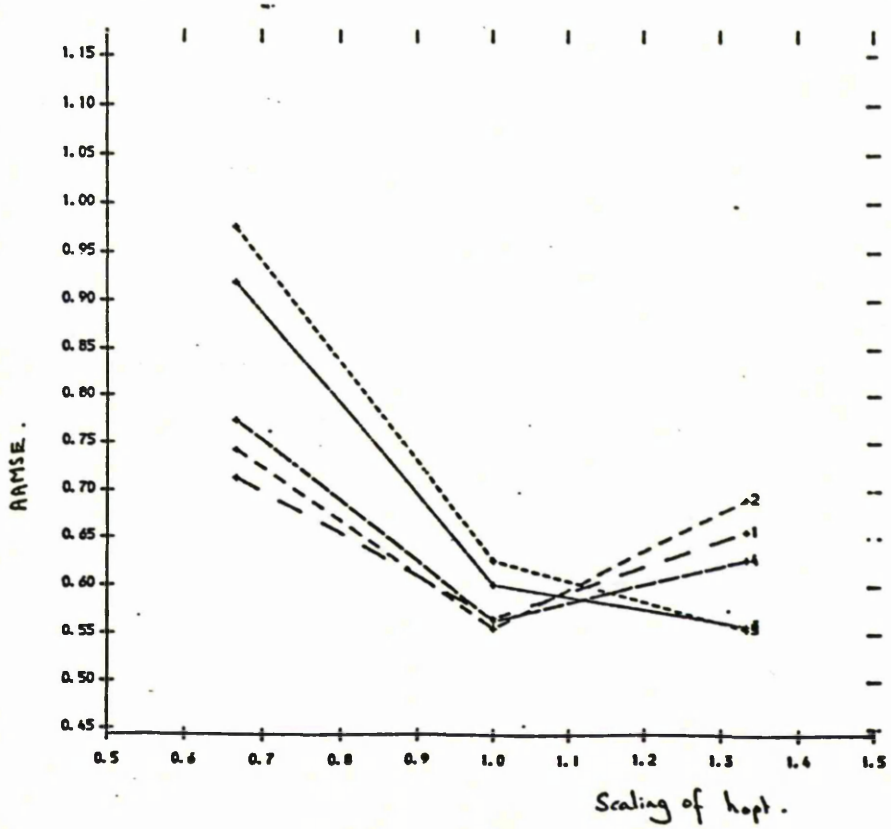


Figure 3.39. AAMSE incurred by the 5 estimators in estimating $g_4(x)$ on (0,1), n = 50, σ^2 = 3.779.

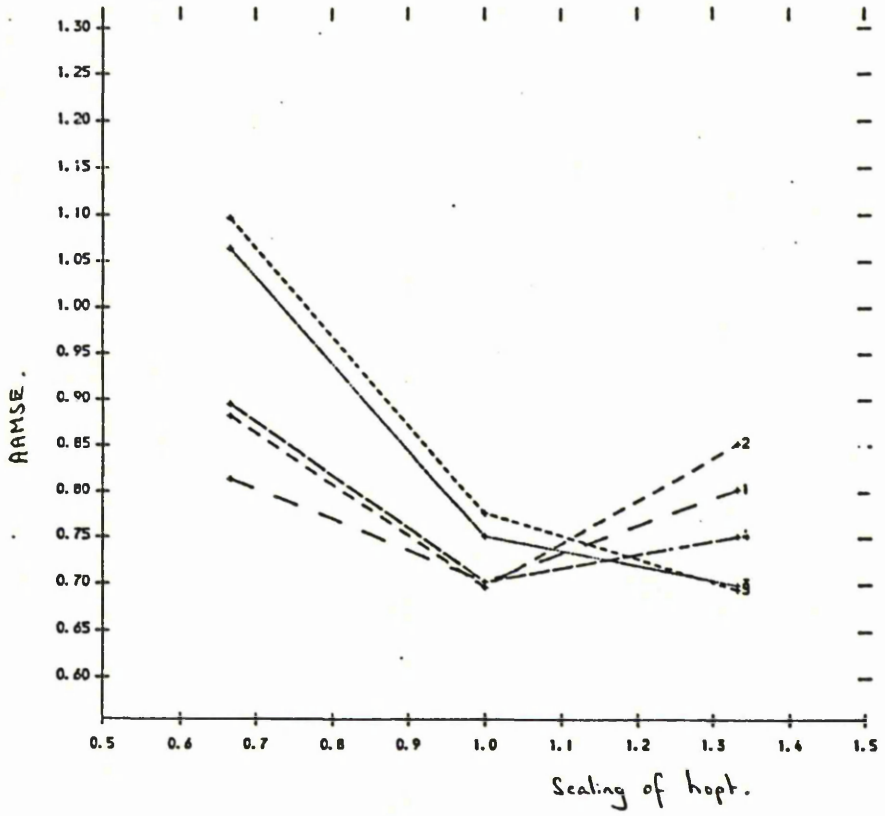
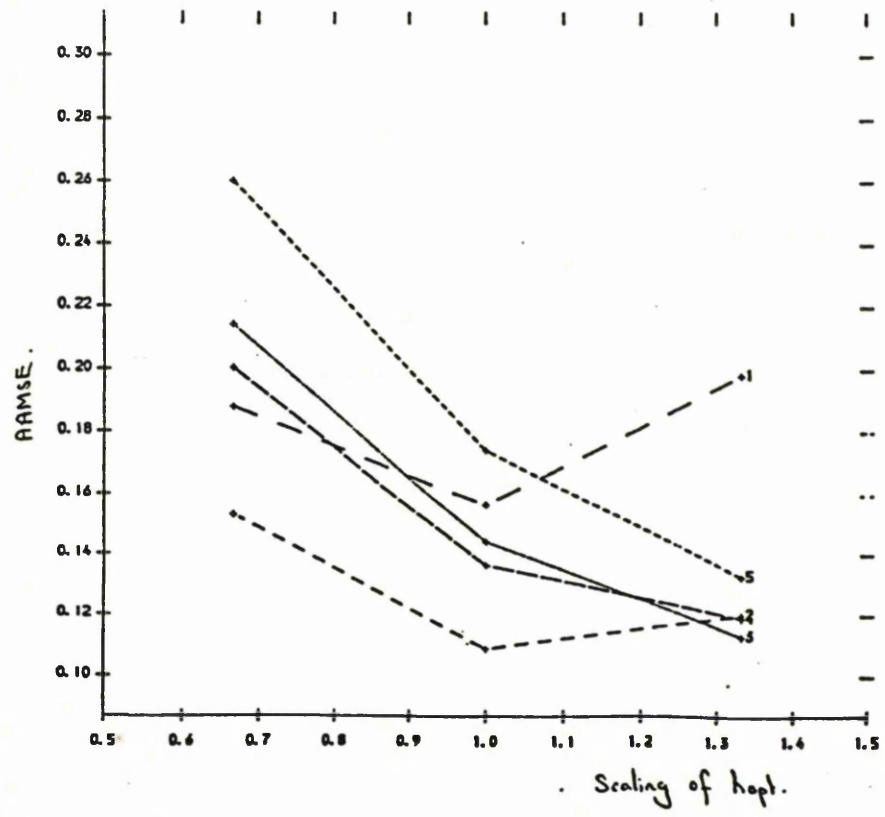


Figure 3.40. AAMSE incurred by the 5 estimators in estimating $g_5(x)$ on (0,1), n = 50, σ^2 = 0.637.



Chapter 4. Pointwise confidence intervals for density functions

4.1. Introduction.

Most of the literature on nonparametric density estimation is concerned with the construction of a point estimate of the underlying density function. However, the estimator, being a function of the sample $d_n = \{X_1, X_2, \ldots, X_n\}$, is a random variable with its own distribution. Therefore, some form of interval estimation is desirable in order to assess its precision and stability as well as to obtain a range of plausible values for the true density function.

In this chapter pointwise confidence intervals for the unknown density $f(x)$, for some fixed x, based on a fixed kernel density estimator will be constructed using two different approaches. Firstly, the central limit theorem will be used to obtain a normal approximation to the distribution of $\hat{f}(x)$. Conditions under which this approximation will hold were first described by Parzen (1962). Secondly, the sampling distribution of $\hat{f}(x)$ will be estimated using bootstrap techniques.

In both cases it is important that the distribution is centred correctly in order to try and obtain the correct coverage probability. The bias inherent in the fixed kernel estimator was discussed in chapter 3 where it was demonstrated that an effective way of reducing it is to subtract an estimate of the principal asymptotic bias term, namely $\frac{1}{2} h^2 \hat{f}^{(2)}(x)$. Such bias corrected estimators will be used in this chapter. Also, comparison will be made with the use of the optimal kernel of order 4, $K(t) = {}^{15}/_{32} (7t^4 - 10t^2 + 3)$, which has some success in reducing bias, particularly for unimodal underlying densities, but also often results in an estimate with lower variance than when subtracting bias or using a fixed Normal kernel.

The effectiveness of the different methods will be assessed through a simulation study using known underlying densities having a variety of different shapes.

Finally, they will also be illustrated in practice through the construction of a pointwise confidence interval for the unknown density of a real data set.

4.2. Using asymptotic normality.

The fixed kernel estimator $\hat{f}_n(x)$, where the subscript n denotes that it is constructed from a sample of size n, can be written as the average

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} W_{nk} \tag{4.2.1}$$

where

$$W_{nk} = \frac{1}{h} K \left[\frac{x-X_k}{h}\right]. \tag{4.2.2}$$

The W_{nk} are independent random variables each identically distributed as the random variable

$$W_n = \frac{1}{h} K \left[\frac{x-Y}{h}\right] \tag{4.2.3}$$

since the X_k are i.i.d. with common distribution F.

Parzen (1962) states conditions under which the sequence of $\hat{f}_n(x)$ is asymptotically normal in the sense that for every real number a

$$\lim_{n\to\infty} \left[\frac{\hat{f}_n(x) - E(\hat{f}_n(x))}{\sqrt{\text{Var}\ (\hat{f}_n(x))}} \leqslant a\right] = \Phi(a) \tag{4.2.4}$$

where $\Phi(\cdot)$ is the c.d.f. of a standard normal distribution.

He also discusses a necessary and sufficient condition for (4.2.4) to hold as well as the Berry-Esseen theorem which provides a bound

for the error in the approximation.

If h is chosen so that $\lim_{n\to\infty} h = 0$ then the estimator $\hat{f}_n(x)$ is asymptotically unbiased. However, in a finite sample situation the bias is well approximated by $\frac{1}{2}.h^2.f^{(2)}(x)$ as discussed in chapter 3, and a good estimator of variance is given by $(nh)^{-1}.f(x).\int K(t)^2dt$. Replacing the unknown $f(x)$ and $f^{(2)}(x)$ by the appropriate kernel estimators results in the root

$$\frac{\left[\hat{f}_n(x) - \frac{1}{2}.h^2\,\hat{f}_n^{(2)}(x)\right] - f(x)}{\sqrt{\left[(nh)^{-1}.\hat{f}_n(x)\int K(t)^2dt\right]}} \qquad (4.2.5)$$

whose distribution is approximately $N(0,1)$. The term root has been used here, as in Beran (1987), to make the distinction from a pivot defined in the classical sense. If $\hat{f}_n(x)$ and $\hat{f}_n^{(2)}(x)$ are consistent estimators of $f(x)$ and $f^{(2)}(x)$ respectively then (4.2.5) is asymptotically pivotal. The introduction of the bias correction in the numerator should make the finite sample distribution of (4.2.5) less dependent on F than if no correction was made.

The root (4.2.5) is appropriate when a kernel such as a standard normal density with $k = 2$ (see equation 3.2.5) is used. If, however, the optimal polynomial $K(t) = \frac{15}{32}(7t^4-10t^2+3)$ is used which has $k = 4$ then the resulting estimator has bias $O(h^4)$ and there is therefore no need to include a bias correction term.

Instead of subtracting an estimate of bias an alternative way of taking it into account is to use the method suggested by Clark (1980) in the related context of kernel regression estimators. This involves basing the confidence intervals on the distribution $N(0,\hat{S})$ where $\hat{S}$ is a consistent estimator of the local mean squared error, S, of the

estimator $\hat{f}_n(x)$. When using a kernel of order 2 a consistent estimator of the mean squared error is given by

$$\hat{S} = \frac{1}{4} h^4 \hat{f}_n^{(2)}(x)^2 + \frac{1}{nh} \hat{f}_n(x) \int K(t)^2 dt \qquad (4.2.6)$$

and the $100(1-\alpha)\%$ confidence interval becomes

$$(\hat{f}_n(x) \pm \Phi^{-1}(1-\alpha/2)\, \hat{S}^{1/2}). \qquad (4.2.7)$$

If the bias and variance of the estimator are denoted by β and ν then intervals based on (4.2.7), as opposed to (4.2.5), are conservative if

$$c^2(\beta^2+\nu) \geq (c\sqrt{\nu} + \beta)^2$$

where $c = \Phi^{-1}(1-\alpha/2)$

i.e. $c^2\beta^2 + c^2\nu \geq c^2\nu + 2c\beta\sqrt{\nu} + \beta^2.$

Dividing through by $\sqrt{\nu}$ and solving for c results in the condition

$$c \geq \frac{\sqrt{1+b^2}+1}{b} \quad \text{or} \quad c \leq \frac{-\sqrt{1+b^2}+1}{b} \qquad (4.2.8)$$

where $b = \beta/\sqrt{\nu}$.

4.3. Using the bootstrap.

In this section let $d_n = (X_1,\ldots,X_n)$ again denote a random sample from an unknown distribution F. The characteristic of F for which a confidence interval is required is $f(x)$, the unknown density which will be denoted by $T(F)$. It is therefore necessary to consider a root $R_n(d_n,T(F))$ which is a functional depending on both the data and the unknown density. The distribution of $R_n(d_n,T(F))$ under F will be denoted by $J_n(F)$. In order to construct a confidence interval

for $T(F) = f(x)$ it is necessary to either know the sampling distribution or estimate the appropriate quantiles of $J_n(F)$. In Section 2 it was assumed that the sampling distribution of the root (4.2.5) was $N(0,1)$. However, in this section $J_n(F)$ will be estimated by $J_n(\hat{F}_n)$ where $\hat{F}_n$ is the empirical distribution function, an unbiased estimator of F. The appropriate quantiles of $J_n(F)$ are then estimated by those of $J_n(\hat{F}_n)$. This is achieved by randomly sampling from $\hat{F}_n$ and then for each sample of size n calculating the root $R_n(\cdot, T(\hat{F}_n))$. The quantiles of $J_n(F)$ are then estimated by the appropriate order statistics of these roots. In this simulation procedure, drawing a random sample from $\hat{F}_n$ means sampling n values from d_n randomly with replacement. The resulting bootstrap $100(1-\alpha)\%$ one-sided confidence interval for $T(F) = f(x)$ then takes the form

$$B_n(\alpha, d_n) = \{t \in T : R_n(d_n, t) \leqslant J_n^{-1}(1-\alpha, \hat{F}_n)\} \tag{4.3.1}$$

or

$$B_n(\alpha, d_n) = \{t \in T : J_n(R_n(d_n, t)) \leqslant 1-\alpha\} \tag{4.3.2}$$

where T is the range space of $T(F)$, $J_n(x,F)$ is the c.d.f. corresponding to $J_n(F)$ evaluated at x and $J_n^{-1}(\alpha,F) = \inf\{x : J_n(x,F) \geqslant \alpha\}$ is an α quantile of $J_n(F)$. The discussion of the bootstrap approaches will refer to one-sided confidence intervals for simplicity of presentation.

Let $d_n^* = \{X_1^*, \ldots, X_n^*\}$ denote a bootstrap sample of size n drawn from $\hat{F}_n$. Then $X_1^*, \ldots, X_n^*$ are conditionally independent given the original sample $d_n = \{X_1, \ldots, X_n\}$. The estimator of the underlying density f based on the bootstrap sample d_n^* is given by

$$\hat{f}_n^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left[\frac{x-X_i^*}{h}\right] \tag{4.3.3}$$

Hence, the expected value of $\hat{f}_n^*(x)$ with respect to the bootstrap sampling is

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

$$= \frac{1}{nh} \cdot n\, E\left\{K\left[\frac{x-X_1^*}{h}\right]\right\}$$

$$= \frac{1}{h}\left[K\left[\frac{x-X_1}{h}\right].\frac{1}{n} + \ldots + K\left[\frac{x-X_n}{h}\right].\frac{1}{n}\right]$$

$$= \frac{1}{nh} \sum_{i=1}^{n} K\left[\frac{x-X_i}{h}\right]$$

$$= \hat{f}_n(x) \qquad (4.3.4)$$

Hence, the distribution of the estimator constructed from the bootstrap samples is centred about $\hat{f}_n(x)$ which is a biased estimator of $f(x)$.

Also, we have

$$V\left[\hat{f}_n^*(x)\right] = E\left[\hat{f}^*(x)^2\right] - \left[E\left[f_n^*(n)\right]\right]^2 .$$

Now,

$$E\left[\hat{f}^*(x)^2\right] = \frac{1}{n^2h^2} E\left[\left[\sum_{i=1}^{n} K\left[\frac{x-X_i^*}{h}\right]\right]^2\right]$$

$$= \frac{1}{n^2h^2} E\left[\sum_{i=1}^{n} K\left[\frac{x-X_i^*}{h}\right]^2 + \sum_{\substack{i,j \\ i\neq j}}^{n} K\left[\frac{x-X_i^*}{h}\right]K\left[\frac{x-X_j^*}{h}\right]\right] .$$

Taking the expectation of each term in the square bracket separately we have firstly that:

$$E\left[\sum_{i=1}^{n} K\left[\frac{x-X_i^*}{h}\right]^2\right] = n.E\left[K\left[\frac{x-X_1^*}{h}\right]^2\right]$$

$$= n\left[K\left[\frac{x-X_1}{h}\right]^2 . \frac{1}{n} + \ldots + K\left[\frac{x-X_n}{h}\right]^2 \frac{1}{n}\right]$$

$$= \sum_{i=1}^{n} K\left[\frac{x-X_i}{h}\right]^2$$

and secondly

$$E\Big[\sum_{\substack{i,j\\ i\neq j}}^{n} K\Big[\frac{x-X_i^*}{h}\Big]\, K\Big[\frac{x-X_j^*}{h}\Big]\Big] = \sum_{\substack{i,j\\ i\neq j}} E\Big[K\Big[\frac{x-X_i^*}{h}\Big]\Big]\, E\Big[K\Big[\frac{x-X_j^*}{h}\Big]\Big]$$

$$= n.(n-1)\, E\Big[K\Big[\frac{x-X_i^*}{h}\Big]\Big].\, E\Big[K\Big[\frac{x-X_j^*}{h}\Big]\Big]$$

$$= n.(n-1)\, .\, \frac{1}{n^2}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]\Big]^2 \quad .$$

Therefore,

$$V\big[\hat{f}_n^*(x)\big] = \frac{1}{n^2h^2}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]^2 + \frac{n-1}{n}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]\Big]^2\Big]$$

$$- \frac{1}{n^2h^2}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]\Big]^2$$

$$= \frac{1}{n^2h^2}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]^2 - \frac{1}{n}\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]\Big]^2\Big] .$$

$$= \frac{1}{n^2}\Big[\sum_{i=1}^{n} \frac{1}{h^2} .\, K\Big[\frac{x-X_i}{h}\Big]^2 - n\Big[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big[\frac{x-X_i}{h}\Big]\Big]^2\Big]$$

(4.3.5)

which is an empirical estimator of

$$\frac{1}{n} .\, V\Big[\frac{1}{h} K\Big[\frac{x-Y}{h}\Big]\Big] = V(\hat{f}_n(x))$$

It can be shown that (4.3.5) is an asymptotically unbiased estimator of $V(\hat{f}_n(x))$ as follows:

$$E\big[V(\hat{f}_n^*(x))\big] = \frac{1}{n^2} E\Big[\sum_{i=1}^{n} \frac{1}{h^2} K\Big[\frac{x-X_i}{h}\Big]^2\Big]$$

$$- \frac{1}{n} E\Big[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big[\frac{x-X_i}{h}\Big]\Big]^2 .$$

Consider the expectation in the second term:

$$\frac{1}{n^2h^2}E\Big[\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]\Big]^2$$

$$= \frac{1}{n^2h^2}\Big[E\sum_{i=1}^{n} K\Big[\frac{x-X_i}{h}\Big]^2 + E\sum_{i\neq j} K\Big[\frac{x-X_i}{h}\Big]K\Big[\frac{x-X_j}{h}\Big]\Big]$$

$$= \frac{1}{n^2h^2}\Big[n.E\,K\Big[\frac{x-Y}{h}\Big]^2 + n(n-1)\Big[EK\Big[\frac{x-Y}{h}\Big]\Big]^2\Big]$$

Hence,

$$E\big[V(\hat{f}_n^*(x))\big] = \frac{1}{n}E\Big[\frac{1}{h^2}K\Big[\frac{x-Y}{h}\Big]^2\Big] - \frac{1}{n^2}E\Big[\frac{1}{h^2}K\Big[\frac{x-Y}{h}\Big]^2\Big] - \frac{(n-1)}{n^2}\Big[E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\Big]\Big]^2$$

$$= \frac{n-1}{n^2}\Big[E\Big[\frac{1}{h^2}K\Big[\frac{x-Y}{h}\Big]^2\Big] - \Big[E\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\Big]\Big]^2\Big]$$

$$\simeq \frac{1}{n}V\Big[\frac{1}{h}K\Big[\frac{x-Y}{h}\Big]\Big] \text{ for large } n$$

$$= V(\hat{f}_n(x)).$$

Therefore, if we consider $(\hat{f}_n(x) - f(x))$ as a root where $\hat{f}_n(x)$ is constructed using a kernel of order 2 the distribution of the bootstrap quantities $f_n^*(x)$ about $\hat{f}_n(x)$ will not mimic the distribution of $f(x)$ about $f(x)$. This is due to the bias of $\hat{f}_n(x)$ as an estimator for $f(x)$ as discussed above.

However, if we consider

$$\Big\{\Big[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\Big] - f(x)\Big\} \tag{4.3.6}$$

then this has an expectation of $0(h^4)$ while the bootstrap version

$$\Big\{\Big[\hat{f}_n^*(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)*}(x)\Big] - \Big[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\Big]\Big\} \tag{4.3.7}$$

has zero expectation using an analogous argument which leads to (4.3.4).

The variance of $\left[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\right]$ which depends on $f(x)$ (and hence F) converges to zero at the rate $n^{-1}h^{-1}$ (see 3.6.11) so as a root we will use

$$\sqrt{nh}\,\left\{\left[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\right] - f(x)\right\} \tag{4.3.8}$$

and estimate the quantiles of its distribution by those of the empirical distribution of

$$\sqrt{nh}\,\left\{\left[\hat{f}_n^*(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)*}(x)\right] - \left[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\right]\right\} \tag{4.3.9}$$

Using (4.3.8) as a root with (4.3.9) for the bootstrap approximation should give more accurate coverage probabilities than using a root with no bias correction term because the distribution of (4.3.8) with lower bias will then be less dependent on F.

If a kernel of order 4 is used then there is no need for the bias correction terms in (4.3.8) and (4.3.9) because an estimator based on such a kernel already has asymptotic bias which is $O(h^4)$.

If we now let

$$R_{n1}(d_n, T(F)) = J_n\{R_n(d_n, T(F)),\ \hat{F}_n\} \tag{4.3.10}$$

then the confidence region given by (4.3.2) can be written as

$$B_n(\alpha, d_n) = \{t \in T : R_{n1}(d_n, t) \leqslant 1 - \alpha\}. \tag{4.3.11}$$

In this construction it is assumed that the distribution $J_{n1}(\cdot, F)$ of $R_{n1}(d_n, T(F))$ is uniform $(0,1)$ which will only be true when the root used is truly a pivotal quantity with a continuous distribution. When this is not the case as with the present problem, Beran (1987) suggests that the coverage probability can be improved

by estimating the distribution $J_{n1}(\cdot,F)$ by $J_{n1}(\cdot,\hat{F}_n)$ to obtain the confidence region

$$B_{n1}(\alpha,d_n) = \{t \in T : R_{n1}(d_n,t) \leqslant J_{n1}^{-1}(1-\alpha,\hat{F}_n)\}. \qquad (4.3.12)$$

The mapping of R_n into R_{n1} using the estimated c.d.f. $J_n(\cdot,\hat{F}_n)$ of R_n is called prepivoting. Beran argues that R_{n1} is closer to being pivotal than R_n is and indeed shows that when beginning with an asymptotically normal root prepivoting is asymptotically equivalent to studentising.

In practice the confidence region B_{n1} is calculated from

$$B_{n1}(\alpha,d_n) = \{t \in T : R_n(d_n,t) \leqslant J_n^{-1}(J_{n1}^{-1}(1-\alpha,\hat{F}_n), \hat{F}_n)\} \qquad (4.3.13)$$

by firstly finding the largest $(1-\alpha)^{th}$ quantile of $J_{n1}(\cdot,\hat{F}_n)$ which we will call C_{n1} and then finding the largest C_{n1}^{th} quantile of $J_n(\cdot,\hat{F}_n)$ to give the required critical value in (4.3.13).

Beran also describes how the prepivoting method can be iterated so that the error in coverage probability decreases as the number of iterations increases.

In general an analytic expression will also not exist for the estimated c.d.f. $J_{n1}(\cdot,\hat{F}_n)$ so it is necessary to use a nested sequence of bootstrap sampling. Beran describes such an approach which leads to the following practical algorithm:

(i) Let $d_1^*,\ldots,d_m^*$ be M bootstrap samples of size n from $\hat{F}_n$.

(ii) Compute $R_n(d_j^*,T(\hat{F}_n))$, $j = 1,\ldots,M$, where $T(\hat{F}_n)$ denotes the bias corrected density estimate. The e.d.f. of the $R_n(d_j^*, T(\hat{F}_n))$ approximates $J_n(\cdot,F)$.

(iii) For every $j = 1,\ldots,M$ let $d^{**}_{j1},\ldots,d^{**}_{jN}$ be N further bootstrap samples of size n each drawn from $\hat{F}_{nj}$.

(iv) Compute $R_n(d^{**}_{jk}, T(\hat{F}_{nj}))$, $k = 1,\ldots,N$, where $T(\hat{F}_{nj})$ denotes the bias corrected density estimate calculated using the bootstrap sample d^{*}_{j}.

(v) Let Z_j be the fraction of the values $R_n(d^{**}_{jk}, T(\hat{F}_{nj}))$, $1 \leqslant k \leqslant N$, computed at step (iv), which are less than

$$R_n(d^{*}_{j}, T(\hat{F}_{n})), \quad j = 1,\ldots,M.$$

The e.d.f. of the $(Z_j: 1 \leqslant j \leqslant M)$ approximates $J_{n1}(\cdot,\hat{F}_n)$ for sufficiently large M and N.

(vi) For a two-sided $100(1-\alpha)\%$ confidence interval obtain the $100(1-\alpha/2)^{th}$ and $100\alpha/2^{th}$ percentiles of the Z_j's and denote these by C_{n1u} and $C_{n1\ell}$ respectively.

(vii) Find the $100C_{n1u}{}^{th}$ and $100C_{n1\ell}{}^{th}$ percentiles of the $R_n(d^{*}_{j}, T(\hat{F}_n))$'s and denote these by b_{nu} and b_{n1} respectively.

(viii) An approximate $10(1-\alpha)\%$ confidence interval for $T(F) = f(x)$ is then given by:

$$\left\{T(\hat{F}_n) - \frac{b_{nu}}{\sqrt{nh}}\ ,\ T(\hat{F}_n) - \frac{b_{n1}}{\sqrt{nh}}\right\}$$

4.4. Simulation study.

In order to study and compare the performance of the various approaches discussed in Sections 4.2 and 4.3 a simulation study was carried out. Firstly, seven methods (four based on asymptotic normality and three on the bootstrap) were used to construct 90% confidence intervals at nine x-values based on a sample size 50 from a standard normal distribution. The empirical confidence levels

were calculated from the results for 100 random samples. As an aid to assessing the performances a 95% prediction interval for the observed coverage probability when the true coverage probability is 0.90 is (0.84, 0.96).

The four methods based on asymptotic normality are as follows:

$$A : \left[\hat{f}_n(x) \pm 1.645 \sqrt{(nh)^{-1} \hat{f}_n(x) \int K(t)^2 dt}\right]$$

where $\hat{f}_n(x)$ is constructed using a fixed standard normal kernel so that $\int K(t)^2 dt = {}^1/_{(2\sqrt{\pi})}$.

$$B : \left[\hat{f}_n(x) \pm 1.645 \sqrt{(nh)^{-1} \hat{f}_n(x) \int K(t)^2 dt}\right]$$

where $\hat{f}_n(x)$ is constructed using the optimal kernel of order 4, i.e. $\int K(t)^2 dt = 1.25$.

$$C : \left[\hat{f}_n(x) - {}^1/_2 h^2 \hat{f}_n^{(2)}(x) \pm 1.645 \sqrt{(nh)^{-1} \hat{f}_n(x) \int K(t)^2 dt}\right]$$

where $\hat{f}_n(x)$ is constructed using a fixed standard normal kernel and $\hat{f}_n^{(2)}(x)$ is constructed using the optimal kernel of order (2,4) i.e. $K(t) = {}^{105}/_{16} \ (-5t^4+6t^2-1)$.

$$D : \left[\hat{f}_n(x) \pm 1.645 \sqrt{\text{Estimated MSE}}\right]$$

where $\hat{f}_n(x)$ is constructed using a fixed standard normal kernel and the estimated MSE, $\hat{S}$, is given by (4.2.6). The optimal kernel of order (2,4) is used to construct an estimate of $\hat{f}_n^{(2)}(x)$ in $\hat{S}$.

Three bootstrap methods were used:

E : Using the root $\sqrt{nh}\left[\hat{f}_n(x) - f(x)\right]$ and the bootstrap approximation $\sqrt{nh}\left[\hat{f}_n^*(x) - \hat{f}_n(x)\right]$ where the density estimates

are constructed using the optimal kernel of order 4. 200 bootstrap samples were taken.

F : Using the root $\sqrt{nh}\left[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x) - f(x)\right]$ and the bootstrap approximation

$$\sqrt{nh}\left[\hat{f}_n^*(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)*}(x) - \left[\hat{f}_n(x) - {}^1/_2\, h^2\, \hat{f}_n^{(2)}(x)\right]\right].$$

A fixed standard normal kernel is used in the estimation of the density and the optimal kernel of order (2,4) in the estimation of the second derivative. 200 bootstrap samples were taken.

G : Using the same root and bootstrap approximation as for method F but prepivoting was also used. 100 first level and 100 second level bootstrap samples were taken.

The amount of smoothing used in estimating the density was controlled by the formulae for global optimal smoothing when the underlying distribution is N(0,1) but scaled by a robust estimate of scale, $\hat{\sigma}$.

These are

$$h = 1.2.n^{-0.214}.\hat{\sigma} \qquad (4.4.1)$$

for the N(0,1) kernel and

$$h = 3.904.n^{-0.134}.\hat{\sigma} \qquad (4.4.2)$$

for the optimal kernel of order 4. the median of the absolute deviations from the median divided by 0.6745, Hogg (1979), was used for $\hat{\sigma}$.

When estimating the second derivative the smoothing parameter given by (4.4.1) was scaled by the factor 2.7. as discussed in chapter 3.6.

Table 4.1. Empirical confidence levels and average lengths of confidence intervals for the N(0,1) density based on methods using asymptotic normality and samples of size 50.

X-value		Method A	Method B	Method C	Method D
-2.00	ECL	0.83	0.79	0.84	0.87
	Avge. length	0.0913	0.0886	0.0886	0.100
-1.00	ECL	0.98	0.99	0.98	0.98
	Avge. length	0.169	0.172	0.169	0.173
-0.50	ECL	0.87	0.93	0.90	0.87
	Avge. length	0.197	0.200	0.197	0.212
-0.25	ECL	0.84	0.91	0.88	0.86
	Avge. length	0.204	0.208	0.204	0.224
0.00	ECL	0.79	0.90	0.91	0.84
	Avge. length	0.206	0.209	0.206	0.228
0.25	ECL	0.84	0.90	0.93	0.84
	Avge. length	0.202	0.206	0.202	0.221
0.50	ECL	0.84	0.94	0.90	0.87
	Avge. length	0.193	0.197	0.193	0.207
1.00	ECL	0.93	0.94	0.92	0.94
	Avge. length	0.165	0.167	0.165	0.170
2.00	ECL	0.80	0.80	0.83	0.86
	Avge. length	0.0915	0.0886	0.0892	0.0999

Table 4.2. Empirical confidence levels and average lengths of confidence intervals for the N(0,1) density based on bootstrap methods and samples of size 50.

X-value		Method		
		E	F	G
-2.00	ECL	0.75	0.78	0.77
	Avge. length	0.083	0.086	0.0902
-1.00	ECL	0.87	0.91	0.94
	Avge. length	0.123	0.159	0.179
-0.50	ECL	0.77	0.84	0.89
	Avge. length	0.119	0.166	0.193
-0.25	ECL	0.69	0.82	0.84
	Avge. length	0.114	0.167	0.194
0.00	ECL	0.68	0.80	0.86
	Avge. length	0.112	0.165	0.191
0.25	ECL	0.68	0.84	0.89
	Avge. length	0.113	0.165	0.192
0.50	ECL	0.72	0.81	0.89
	Avge. length	0.117	0.167	0.190
1.00	ECL	0.84	0.86	0.89
	Avge. length	0.123	0.156	0.183
2.00	ECL	0.78	0.81	0.80
	Avge. length	0.083	0.086	0.0962

Any negative density estimates or confidence bounds obtained were reset to zero. Clearly this will shorten the length of such a confidence interval but the monitoring of the effects of this indicated very little impact on the simulation results.

The results of the simulations are contained in tables 4.1 and 4.2. The same seed for the random number generator was used in the simulations for A-D. Also, the same seed, but different to that used for A-D, was used in the simulations for E and F. Another different seed was used for G. This was done so that a more direct comparison of the methods could be made as they will then be calculating confidence intervals for the same samples. For the bootstrap methods extra levels of simulation are involved in the bootstrap sampling thus necessitating different seeds.

To give an indication of the overall performance of a method the empirical confidence levels (e.c.l's) at each of the x-values can be averaged. For those based on asymptotic normality these averages are 0.86, 0.90, 0.90 and 0.88 for A, B, C, and D respectively. A does not correct for bias and the e.c.l's tend to fall short of the nominal value of 0.90. On the other hand B which is based on the optimal kernel, tends to overestimate this value, most notably when $x = \pm 0.50$ and ± 1.0. C performs well with several e.c.l's close to 0.90 while D again tends to underestimate the nominal level. The lengths of the intervals for A, B and C are all fairly similar with D tending to produce the longest intervals of these four. Also, for these four, the CI's increase in length as x aproaches zero from both sides with the maximum length at the origin - this is because the asymptotic variance is proportional to $f(x)$ which is a maximum at $x = 0$. This is in contrast to E, F and G which each

tend to produce intervals of similar length at each x-value except $x = \pm 2.0$.

For the three bootstrap methods, E, F and G, the average e.c.l's are 0.75, 0.83 and 0.86 respectively. E, which is based on the optimal kernel, consistently underestimates the nominal level and by a large margin for several x-values. The average interval lengths for each x-value are much shorter than for F and G due to its low finite sample variance which was indicated in the simulation results of Chapter 3. There is some improvement in the performance of F over E but it still has a tendency to underestimate the nominal level. The best bootstrap method, in line with the theory, is G which involves prepivoting. Apart from at $x = \pm 2.0$ the e.c.l's for G are all within (0.84, 0.96). It also produces the widest intervals but they are shorter than those produced by A-D.

For each of the seven methods, except perhaps D, there tends to be a drop in the e.c.l's at $x = \pm 2.0$ which is probably due to problems with bias. Indeed, the asymptotic bias at $x = \pm 2.0$ is over 50% larger than the asymptotic standard error for a fixed N(0,1) kernel estimator based on a sample of size 50 and using (4.4.1) to choose h.

For the second stage of the simulation study methods B, C and F were used to construct 90% confidence intervals for pseudo-random samples of size 50 taken from five differently shaped distribution. B and C were chosen because out of the four methods based on asymptotic normality (A-D) these two performed the best in the first stage. This is in terms of accuracy of coverage probability and also the shortness of the average lengths of the intervals. Out of the three bootstrap methods (E-G) the performance of G was far

superior to those of E and F. However, in the more extensive study described below, the long computational time which is required for G meant that it could not be studied further in this way. Method F, which performed better than E, was therefore included.

The five underlying distributions which were used are as follows:

(i) N(0,1).

(ii) Gamma $(2, \sqrt{2})$ - highly skewed.

(iii) 0.5N(-0.866,0.5) + 0.5N(0.866,0.5) - bimodal.

(iv) t(3) - long tailed.

(v) $\chi^2(5)$ - moderately skewed.

The CI's were constructed at 20 equally spaced x-values in the interval (-3,3) for both (iii) and (iv), (0,8) for both (ii) and (v) (-2,2) for (i). The e.c.l.'s at each x-value were based on the results for 100 samples and these were then averaged over the 20 x-values. This was repeated for six different scalings of the appropriate N(0,1) optimal smoothing parameter given by either (4.4.1) or (4.4.2). The data from distributions (ii) and (v) only take positive values and so were reflected in the origin as discussed in chapter 3. For each distribution the same seed was used to obtain the samples for each of the smoothing parameter scalings so that the average e.c.l's for each scaling are directly comparable. The results are given in tables 4.3-4.5.

Table 4.3. Average empirical confidence levels when using the bias corrected normal kernel estimator and asymptotic normality.

	Scaling of the N(0,1) optimal h.					
Distribution	0.4	0.6	0.8	1.0	1.2	1.4
(i)	0.913	0.926	0.931	0.913	0.869	0.808
(ii)	0.520	0.349	0.241	0.170	0.132	0.104
(iii)	0.886	0.886	0.719	0.535	0.383	0.287
(iv)	0.869	0.893	0.898	0.883	0.832	0.773
(v)	0.908	0.866	0.831	0.791	0.748	0.691

Table 4.4. Average empirical confidence levels when using the estimator based on the optimal kernel of order 4 and asymptotic normality.

	Scaling of the N(0,1) optimal h.					
Distribution	0.4	0.6	0.8	1.0	1.2	1.4
(i)	0.907	0.920	0.936	0.921	0.881	0.783
(ii)	0.485	0.347	0.248	0.206	0.183	0.146
(iii)	0.737	0.669	0.473	0.355	0.297	0.250
(iv)	0.868	0.889	0.886	0.861	0.801	0.682
(v)	0.884	0.860	0.818	0.756	0.688	0.608

Table 4.5. Average empirical confidence levels when using the bias corrected normal kernel estimator and the bootstrap.

Distribution	Scaling of the N(0,1) optimal h.					
	0.4	0.6	0.8	1.0	1.2	1.4
(i)	0.841	0.868	0.871	0.859	0.825	0.762
(ii)	0.553	0.448	0.346	0.277	0.234	0.191
(iii)	0.653	0.674	0.562	0.415	0.307	0.223
(iv)	0.757	0.785	0.806	0.800	0.761	0.696
(v)	0.859	0.864	0.830	0.777	0.693	0.615

The main overall feature of the results is that the average e.c.l's generally fall as h gets larger. This is perhaps to be expected because the bias is smaller but the variance larger for small h values with the consequence of more accurately centred but wider CI's. None of the methods have any success for the highly skewed showed Gamma $(2, \sqrt{2})$ distribution. Method F based on the bootstrap also performs poorly for the bimodal normal mixture and the long tailed t-distribution at each h value but has a similar level of success to B and C for the N(0,1) and $\chi^2(5)$ distributions. The two methods based on asymptotic normality have similar levels of success for the N(0,1), t(3) and $\chi^2(5)$ distributions but C performs better for the bimodal normal mixture.

The results of these simulation studies indicate then in practice the best method to use is that based on the bias corrected normal kernel estimator using asymptotic normality. This will be most successful when there is evidence that the underlying distribution

is fairly symmetric and unimodal. If the normal optimal smoothing formula (4.4.1) is being used then this should be scaled by a factor less than one. However, the second part of the simulation study did not include a comparison with the bias corrected normal kernel estimator using the bootstrap with prepivoting which performed very well in the first part. Such a detailed comparison awaits a suitable computational algorithm designed to considerably reduce the number of calculations involved.

4.5. Example.

The data used consists of the annual snowfall (in inches) at Buffalo, New York, for the 63 years from 1910-1972. Silverman (1986, p.44-45) considers these data and shows that by varying the value of the smoothing parameter either a unimodal or a trimodal density estimate is obtained.

Figures 4.1 and 4.2 illustrate a bias corrected estimate with nominal 90% pointwise confidence intervals evaluated at 30 equispaced points. The value 6.378 of the smoothing parameter h is based on formula (4.4.1) scaled by 0.5. (For these data the robust $\hat{\sigma} = 25.797$). The CI's based on asymptotic normality (method C of Section 4.4) in figure 4.1 follow the shape of and are equally spaced about the density estimate. However, in contrast, those based on the bootstrap with prepivoting (method G of Section 4.4) in figure 4.2 contains some small additional structure, in particular for the upper bounds of the two smaller modes. They are also not generally symmetric with the upper bounds in the centre and right of the plot notably much further from the density estimate than the corresponding lower bounds.

Although the two sets of CI's are not simultaneous they do indicate that the true underlying density could be unimodal. (It is likely that simultaneous intervals would be much wider). This is in line with the goodness-of-fit tests carried out by Parzen (1979) who concludes that these data are normally distributed.

Figure 4.1. Bias corrected kernel estimate (h = 6.378) for the Buffalo snowfall data with nominal 90% pointwise confidence intervals based on normality.

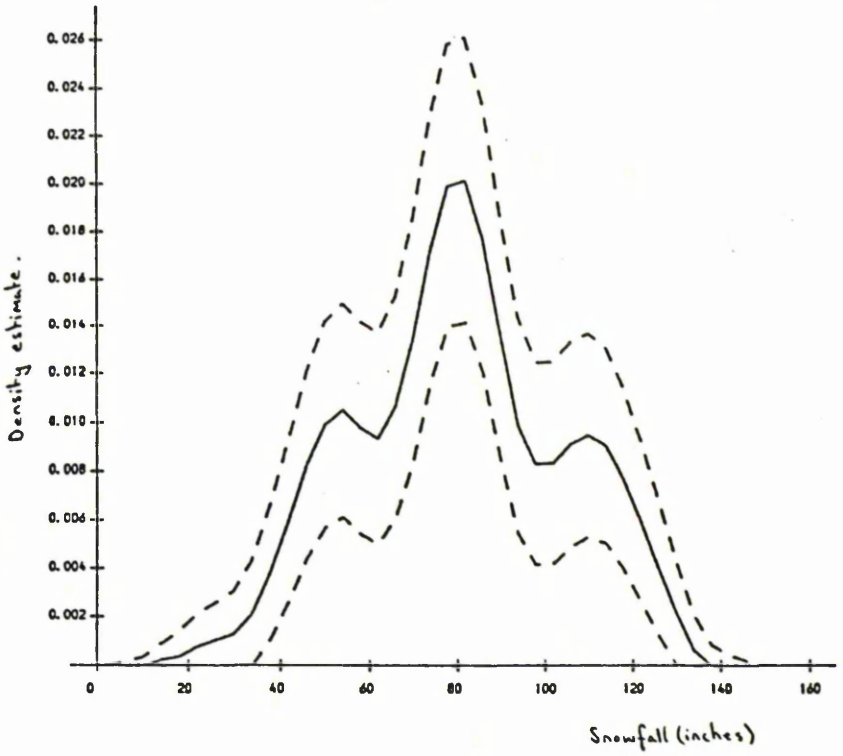
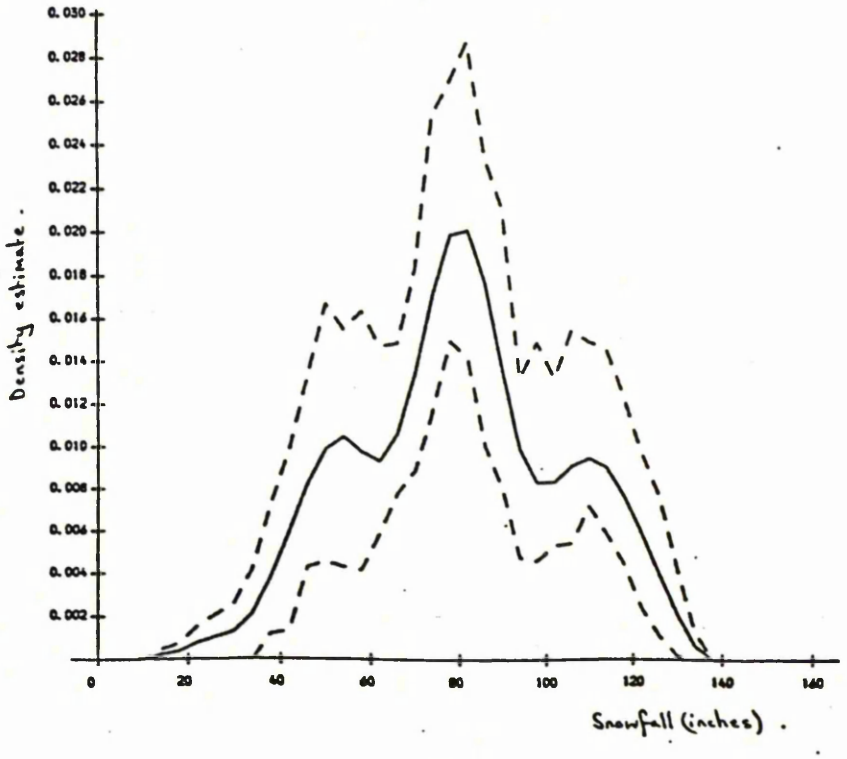


Figure 4.2. Bias corrected kernel estimate (h = 6.378) for the Buffalo snowfall data with nominal 90% pointwise confidence intervals based on the bootstrap with prepivoting.



Chapter 5. Density based Goodness-of-fit tests of Multivariate Normality.

5.1. Introduction.

Given a univariate random sample $\{X_1, X_2, \ldots, X_n\}$ there are many procedures available for testing whether the observed data have come from a normally distributed population.

One class of goodness-of-fit statistics are empirical distribution (EDF) statistics which are based on a comparison between the EDF, $F_n(x)$, and the normal distribution function, $F(x)$. The class includes the Kolmogorov-Smirnov, Cramer von-Mises and Anderson-Darling statistics. Stephens (1974) describes and discusses these in detail and a power study illustrates the effectiveness of the Cramer von-Mises and Anderson-Darling statistics in particular. Another approach is to base a statistic on the distance between empirical and hypothesised characteristic functions as in Koutrouvelis and Kellermeier (1981). More informal methods include quantile-quantile and probability-probability plots.

Bickel and Rosenblatt (1973) derive theoretical results for the maximum of the normalised deviation of a density estimate from its expected value and for quadratic norms of the same quantity. These are used to study the behaviour of goodness-of-fit tests based on these statistics. Penalising departures from normality in this way has strong intuitive appeal but has received little development until recent work by Bowman (1988) who develops two density based tests of normality - one based on integrated squared error and the other on entropy. A power study shows them to be competitive with a number of alternative tests for a wide range of underlying distributions.

If we now have available a random sample $\{\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n\}$ from a p-variate distribution then the assumption of multivariate normality is much more difficult to check. Available procedures concentrate either on combinations of univariate tests of normality or on the geometrical properties in R^p of two or more variates taken together such as the plotting of Mahalanobis distances as in Healy (1968). Cox and Small provide a review of these and also consider departures based on curvature in the variate-variate plots. Paulson, Roohanand and Sullo (1987) consider EDF tests of multivariate normality. The two main ones they consider are those of Anderson-Darling and Cramer von-Mises expressed as functions of Mahalonobis distances. Rosenblatt (1975) examines the asymptotic behaviour of quadratic functions of multivariate density estimates which he suggests will be useful in setting up a goodness-of-fit test of a density function. Koziol (1983) described an omnibus test of multivariate normality based on the "radii and angles" properties of the multivariate normal distribution. He uses the method of Fisher (1958) to combine Rayleigh's test for uniformity of the angles on the p-1 hypersphere with a Cramer von-Mises test for departures of the Mahalanobis distances from a $\chi^2(p)$ distribution. This will be described in more detail in Section 5.2.

There are very few global tests of multivariate normality and these cannot be expected to have the same power as approaches designed to detect specific non-normal features. However, they are particularly useful in that they are able to sweep over the entire distribution for any features which may be missed by a more specific approach.

Much of the standard "classical" multivariate statistical method-

ology depends on the assumption of normality and the effects of departures from this on the methods are generally not clearly or easily understood. It would therefore be useful to have a single effective test for examining this assumption in order to guide the subsequent analysis. Such a test would also be useful prior to carrying out an analysis such as projection pursuit which looks for non-normal features in lower dimensional projections of the data. Such an initial test should help prevent over interpretation of features which may be present even when the data have arisen from a multivariate normal distribution.

If the null hypothesis is rejected then it would be appropriate to carry out other more specific tests and also use graphical methods to determine the causes of the non-normality.

In this chapter it is proposed to extend the univariate density based test statistics of Bowman (1988) to the multivariate case. Critical values will be obtained and their performances assessed via a power study. Comparisons will also be made with the combined approach of Koziol (1983).

In most practical situations the mean vector and covariance matrix are unknown so that the null hypothesis is a composite one. Following usual practice it will be assumed in this chapter that the data have been centred and standardised by the sample mean vector, $n^{-1} \sum_{i=1}^{n} \underline{X}_i$, and covariance matrix, $S = (n-1)^{-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}}).(\underline{X}_i - \bar{\underline{X}})^T$ so that the null distribution is taken to be $N_p(\underline{0}, I_p)$. This standardisation will be reflected in simulation of percentage points and power.

5.2. The omnibus test of Koziol (1983).

It is assumed that we have a random sample from $\underline{X}$ which has a

$N_p(\underline{\mu},\sum)$ distribution. If we make the transformation

$$\underline{Y} = \sum^{-1/2} . (\underline{X}-\underline{\mu}) \qquad (5.2.1)$$

then $\underline{Y} \sim N_p(\underline{0},I)$. If $\underline{Y}$ is transformed to polar co-ordinates, $\underline{Y} \longrightarrow (R,\underline{\theta})$, then

$$R = \underline{Y}^T\underline{Y} \sim \chi^2(p) \qquad (5.2.2)$$

and $\underline{\theta}$ is uniformly distributed on S_{p-1}, the unit hypersphere in R^p.

To derive a test based on the "radii", R, Koziol (1982) uses the empirical process

$$V_n(t) = n^{1/2} \{F_n(t) - G_p(t)\} \qquad (5.2.3)$$

where $F_n(\cdot)$ is the EDF of $R_1,\ldots,R_n$ and $G_p(\cdot)$ is the CDF of the $\chi^2(p)$ distribution. In practice $\underline{\mu}$ and $\sum$ are generally unknown and are replaced by their unbiased estimates. We then consider the empirical process

$$W_n(t) = n^{1/2} \{\hat{F}_n(t) - G_p(t)\} \qquad (5.2.4)$$

where $\hat{F}_n(t)$ is the proportion of $\hat{R}_1,\ldots,\hat{R}_n$ less than or equal to t and

$$\hat{R}_i = (\underline{X}_i-\overline{\underline{X}})^T S^{-1}(\underline{X}_i-\overline{\underline{X}}), \quad i = 1,\ldots,n. \qquad (5.2.5)$$

The Cramer von-Mises statistic

$$J_n = \int_0^\infty W_n^2(t)dG_p(t) \qquad (5.2.6)$$

then measures departures from asymptotic normality and has a limiting distribution which is that of

$$J = \int_0^\infty W^2(\cdot).dG_{(p)} \qquad (5.2.7)$$

where $W(\cdot)$ is a centred Gaussian process whose covariance kernel is given in the paper.

Koziol (1982) determines critical values for J using Pearson curve approximations and suggests that such asymptotic critical values of J_n are reasonable if p is small or n is large.

To implement the statistic J_n practically we put $Z_i = C_p(\hat{R}_i)$, $(i = 1,\ldots,N)$, order the values of Z_i to give $Z_{(1)},\ldots,Z_{(n)}$ and then use the alternative form

$$J_n = \sum_{i=1}^{n} \{Z_{(i)} - (i-1/2)/n\}^2 + (12n)^{-1}. \qquad (5.2.8)$$

To obtain critical values for finite sample sizes it is necessary to use simulation. Koziol (1982) provides a limited table of such values. The biased estimator of $\sum$ with divisor n^{-1} was used in the calculations.

For the angles it is first necessary to scale the $\underline{Y}_i$'s to have unit length i.e.

$$\underline{\ell}_i = \frac{\sum^{-1/2}(\underline{X}_i - \underline{\mu})}{\{(\underline{X}_i - \underline{\mu})^T \sum^{-1} (\underline{X}_i - \underline{\mu})\}^{1/2}}, \quad i = 1,\ldots,n \qquad (5.2.9)$$

so that the elements of $\underline{\ell}_i$ are the direction cosines of the vector $\underline{Y}_i$.

If we let

$$\underline{T} = n^{-1/2} \sum_{i=1}^{n} \underline{\ell}_i \qquad (5.2.10)$$

then Rayleigh's test statistic, $p.\underline{T}^T\underline{T}$, (Mardia (1972)), which is proportional to the squared length of the resultant of the $\underline{\ell}_i$, is asymptotically $\chi^2(p)$ when the data are Normally distributed.

If $\underline{\mu}$ and $\sum$ are unknown the unbiased estimates are used and we have

$$\hat{\underline{\ell}}_i = \frac{S^{-1/2}(\underline{X}_i - \underline{X})}{\{(\underline{X}_i - \bar{\underline{X}})^T S^{-1} (\bar{\underline{X}}_i - \bar{\underline{X}})\}^{1/2}}, \quad i = 1, \ldots, n \qquad (5.2.11)$$

so that

$$\hat{T} = n^{-1/2} \sum_{i=1}^{n} \hat{\ell}_i \; . \qquad (5.2.12)$$

Koziol (1983) uses stochastic integration to show that Rayleigh's statistic should now be expressed as $v^{-1}\hat{\underline{T}}.^T \hat{\underline{T}}$ where

$$v = p^{-1} \left[1 - (2/p) \{\Gamma((p+1)/2)/\Gamma(p/2)\}^2\right] . \qquad (5.2.13)$$

For finite samples a limited simulation study shows that the asymptotic distribution is quite a good approximation for small p or large n.

A test based either on R or on $\underline{\theta}$ will not be consistent because there is a loss of information as R and $\underline{\theta}$ are not minimal sufficient by themselves. However, R and $\underline{\theta}$ are independent and jointly sufficient so Koziol (1983) forms an omnibus test by using Fisher's method.

The p-value of a test statistic is distributed as $U(0,1)$ under the null hypothesis so that minus twice its natural logarithm is easily shown to have a $\chi^2(2)$ distribution. Therefore, if p_1 is the p-value for the Cramer von-Mises test statistic and p_2 the p-value for the independent Rayleigh test statistic

$$-2[\log p_1 + \log p_2] = -2 \log (p_1 . p_2) \sim \chi^2(4). \qquad (5.2.14)$$

Littel and Folks (1971) show that Fisher's method is an optimal procedure for combining independent tests of hypotheses in terms of Bahadur relative efficiency.

5.3. The density based test statistics.

The univariate integrated squared error statistic is derived by analogy with the Cramer von-Mises family for distribution functions which have the form

$$\int_{-\infty}^{\infty} \{F(x) - F_n(x)\}^2 . w(x) . dF(x) \qquad (5.3.1)$$

where F and F_n are $\int \{F(x) - F_n(x)\}^2 . w(x) . dF(x)$ distribution functions and w is a weight function. An effective choice for w is the reciprocal of the variance of $F_n(x)$ as in the Anderson-Darling statistic so that more weight is given to departures at points where F_n is estimated accurately. Using an analogous approach with fixed kernel density estimates leads to the statistic

$$\int \{f(x) - \hat{f}(x)\}^2 dx \qquad (5.3.2)$$

since $dF(x)$ may be written as $f(x).dx$ and the asymptotic variance of the fixed kernel estimator $\hat{f}(x)$ is proportional to $f(x)$.

However, as discussed in chapter 3, the smoothing used in the construction of the density estimate leads to a bias and it is therefore appropriate to modify the test statistics so that the density estimate is compared with the slightly flatter shape of density we would expect if the null hypothesis is true. The appropriate test statistic is then:

$$\int_{-\infty}^{\infty} \{N(x;0,1+h^2) - \hat{f}(x)\}^2 \, dx \qquad (5.3.3)$$

where $N(x;0,1+h^2)$ denotes the normal density in x with zero mean and variance $1+h^2$ and is the expected value of $\hat{f}(x)$ under H_o.

Only the fixed kernel density estimate will be used in this test statistic because of the difficulties in writing an explicit expression for the expected value of an adaptive estimator and also, as discussed in chapter 2, the asymptotic results do not provide a good approximation.

For multivariate data, statistic (5.3.3) easily generalises to

$$\int \{N_p(\underline{x};\underline{0},(1+h^2)I_p) - \hat{f}(\underline{x})\}^2 dx \qquad (5.3.4)$$

If $\hat{f}$ is constructed using standard normal kernels (5.3.4) can be evaluated analytically to give the following simpler computational form:

$$N_p(\underline{0};\underline{0},2(1+h^2)I_p) - (2/n)\sum_i N_p(\underline{x}_i;\underline{0},(1+2h^2)I_p)$$

$$+ (1/n)N_p(\underline{0};\underline{0},2h^2I_p) + (2/n2)\sum_{i<j} N_p(\underline{x}_i;\underline{x}_j,2h^2I_p) \qquad (5.3.5)$$

where n is the sample size. A good choice of smoothing parameter is provided by the value which minimises the asymptotic MISE when estimating a multivariate normal distribution which for a given dimension p and sample size n, is

$$h = \left[\frac{4}{(2p+1)n}\right]^{(1/(p+4))} . \qquad (5.3.6)$$

Bowman (1985) shows that using such an h produces density estimates that recover the shapes of a variety of unimodal densities in the univariate case.

Vasicek (1976) based a test of univariate normality on the property of the normal distribution that its entropy exceeds that of any other distribution that has the same variance. The entropy of a distribution F with a density f is defined as

$$H(f) = -\int_{-\infty}^{\infty} f(x)\log(f(x))dx = -E[\log(f(x)] \qquad (5.3.7)$$

Vasicek's statistic, which estimates $H(f)$, was not derived explicitly in terms of density estimates but is in fact equivalent to $-\frac{1}{n}\sum \log(\tilde{f}(x_i))$ where $\tilde{f}(x_i)$ is a type of nearest neighbour estimate. Bowman (1988) suggests using a fixed kernel approach which leads to the test statistic

$$-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{f}(x_i)) \qquad (5.3.8)$$

with H_o being rejected for small values.

The entropy property of the univariate normal distribution also holds in the multivariate case and so sample entropy

$$-\frac{1}{n}\sum_{i=1}^{n}\log(\hat{f}(\underline{x}_i)) \qquad (5.3.9)$$

provides a goodness-of-fit test statistic for the multivariate normal distribution. In this case adaptive estimators may also be used in constructing $\hat{f}$. As a choice of smoothing parameter formula (5.3.6) may be used for a fixed kernel estimator whereas the normal optimal formulae derived in chapter 2 can be used for the adaptive method. As a result of the comparisons made in chapter 2 when it was found that the adaptive method with $\alpha = 1/p$ was the most effective for multivariate data only this method will be used here. The required pilot estimate will be constructed using fixed kernels smoothed by (5.3.6).

A number of authors, including Huber (1985) and Jones and Sibson (1987), have also used entropy as an index of normality when seeking non-normal one and two dimensional projections in a projection pursuit algorithm.

5.4. A Power Study.

Because of the intractability of the finite sample distributions of each of the three density based statistics, namely integrated squared error, sample entropy using fixed kernels and sample entropy using the adaptive method with $\alpha = 1/p$, simulation was used to calculate 5% critical values. For a given dimension (1,2,...,5 or 6) and sample size (25,50 or 100) a random sample was generated from a standard normal distribution. The data were then centred and standardised and the test statistic evaluated. This was repeated 1000 times. The values of the simulated test statistics were then ordered with the critical value corresponding to the appropriate order statistic.

Simulation was also required to obtain a more extensive set of critical values than those calculated by Koziol (1982, 1983) for the Cramer-von Mises and Rayleigh test statistics described in Section 5.2. This time 5000 simulations were carried out for the three sample sizes (25, 50 and 100) and five dimensions (2,....,6). Nine percentage points of the empirical distributions were evaluated for each dimension and sample size combination and are given in tables 5.1 and 5.2.

Similarly, empirical powers of the tests were calculated by simulating from the following multivariate distributions:

(i) Cauchy (0,1) in each margin - long tailed.

(ii) Lognormal (0,1) in each margin - highly skewed.

(iii) Normal mixture,

$$0.5\ N_p(1.5,\ldots,1.5)^T, I_p)) + 0.5\ N_p((-1.5,\ldots,-1.5)^T, I_p)$$

- bimodal.

(iv) Gamma (2,1) in each margin - moderately skewed.

Sample sizes 25, 50 and 100 were considered for each of dimensions one to six for the density based statistics and dimensions two to six for the others. Results are based on 1000 replications in each case. To evaluate the test statistic for Fisher's method linear interpolation was used in tables 5.1 and 5.2 to obtain the p-values for the Cramer-von Mises and Rayleigh tests. Observed test statistic values greater than the 99.5% point were assigned a p-value of 0.0025 while those less than the 10% point were assigned a p-value of 0.95.

For distributions (i) and (ii), which have densities very unlike the shape of a Normal, all the powers of each test are very close to one which is what one would expect if the tests are at all effective. See tables 5.3 and 5.4.

The powers of the density based tests for distributions (iii) and (iv) are given in table 5.5. These are poor at $n = 25$ which reflects that this is a small sample size at which to estimate a multivariate density. When $n = 100$ all the powers are again close to one with the exception of the entropy statistics for the bimodal normal mixture where there is a decline in power for dimensions 5 and 6.

A more informative comparison can be made when $n = 50$. For distribution (iii) (the bimodal normal mixture) the ISE statistic has the largest power in each dimension except for dimension 1 when the entropy statistic based on the adaptive estimator is marginally superior. There is a marked decrease in power as dimensionality increases which is most marked for the entropy statistics. Of the two entropy statistics the adaptive method has a better power in

each dimension.

For distribution (iv) based on Gamma margins the ISE statistic again has the best power in each dimension except the first when the entropy statistic based on fixed kernels performs slightly better. For this distribution though the adaptive method has no clear advantages over the fixed method in the entropy statistic.

The powers of the non-density based tests for distribution (iii) and (iv) are given in tables 5.5 and 5.6. The results for the bimodal normal mixture show at each sample size a marked decrease in power for each of the tests as the dimension increases. These decreases are most pronounced for the Cramer-von Mises and Fisher test statistics while the Rayleigh statistic has very poor power at each sample size and dimension. The Cramer-von Mises test statistic performs best overall here with Fisher's test being let down by the poor performance of Rayleigh's statistic.

When the underlying distribution is Gamma (2.1) in each margin the powers of the three tests increase as the sample size gets larger but this time there is not the marked decline in performance for increasing dimension. The best now overall is Rayleigh's test with Fisher's method doing slightly worse, again due to the poorer performance of the other test, which this time is based on the Cramer-von Mises statistic.

Comparisons between the density based omnibus tests and Fisher's combined method for distribution (iii) show thast Fisher's method performs worse than the other three at all sample sizes and dimensions considered. The differences are most marked when $n = 25$ and $n = 50$ - the results at $n = 50$ are plotted in figure 5.1.

For distribution (iv) at $n = 25$ Fisher's method has lower power than for each of the density based statistics but the differences are generally not great. The results for $n = 50$ are plotted in figure 5.2. and show the performance of the ISE statistic to be again the best for dimensions 2-6. They also show a marked increase in power for Fisher's method with increasing dimension so that for dimensions 5 and 6 it is performing better than the two entropy statistics. When $n = 100$ these four omnibus tests each have similar high powers for each dimension.

Taken overall, the simulation results indicate that the ISE test statistic would generally be a good choice for testing multivariate normality. The 5% critical values for the ISE statistic for dimensions 1 to 6 and for sample sizes up to 500 are listed in table 5.7. These results also show that Fisher's method would generally make a better choice than just using the Cramer von-Mises or Rayleigh tests on their own. This is because its performance was close to the better of the other two, and which of the two is better depends upon the underlying distribution.

Table 5.1. Observed % points for the Cramer-von Mises statistic simulated from 5000 samples in each case.

Sample size		% point								
	n	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0	99.5
p = 2	25	0.034	0.048	0.072	0.113	0.166	0.211	0.251	0.321	0.368
	50	0.033	0.047	0.071	0.113	0.167	0.217	0.263	0.326	0.384
	100	0.034	0.048	0.074	0.115	0.175	0.222	0.265	0.327	0.361
p = 3	25	0.033	0.046	0.069	0.107	0.158	0.196	0.229	0.279	0.339
	50	0.033	0.046	0.070	0.107	0.160	0.202	0.247	0.315	0.348
	100	0.033	0.047	0.072	0.109	0.159	0.200	0.242	0.295	0.340
p = 4	25	0.034	0.047	0.069	0.107	0.152	0.190	0.229	0.279	0.315
	50	0.034	0.046	0.069	0.105	0.158	0.194	0.226	0.278	0.316
	100	0.033	0.046	0.069	0.106	0.154	0.195	0.236	0.299	0.350
p = 5	25	0.035	0.048	0.071	0.109	0.162	0.196	0.235	0.283	0.314
	50	0.034	0.046	0.068	0.103	0.153	0.186	0.222	0.281	0.317
	100	0.032	0.045	0.067	0.103	0.152	0.188	0.224	0.273	0.317
p = 6	25	0.035	0.049	0.072	0.111	0.164	0.203	0.244	0.296	0.353
	50	0.033	0.046	0.069	0.105	0.153	0.192	0.230	0.282	0.325
	100	0.032	0.045	0.067	0.103	0.150	0.190	0.224	0.277	0.296

Table 5.2. Observed % points for the Rayleigh statistic simulated from 5000 samples in each case.

Sample size		% point								
	n	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0	99.5
p = 2	25	0.186	0.540	1.364	2.698	4.457	5.637	6.838	8.198	9.439
	50	0.208	0.556	1.397	2.773	4.694	6.062	7.395	9.142	10.659
	100	0.217	0.576	1.426	2.858	4.713	6.066	7.653	9.105	10.739
p = 3	25	0.519	1.072	2.153	3.802	5.745	7.184	8.624	10.515	11.778
	50	0.552	1.191	2.323	4.008	6.159	7.713	9.696	11.582	13.580
	100	0.576	1.170	2.351	4.038	6.053	7.629	9.299	11.411	13.310
p = 4	25	0.697	1.738	3.099	4.948	7.087	8.688	9.988	11.957	13.497
	50	1.018	1.852	3.233	5.229	7.470	9.070	10.902	12.886	14.580
	100	1.013	1.843	3.260	5.362	7.771	9.467	10.892	12.960	14.490
p = 5	25	1.338	2.343	3.792	5.742	8.044	9.600	11.153	13.355	14.327
	50	1.534	2.493	4.133	6.306	8.837	10.647	12.195	14.288	16.120
	100	1.568	2.591	4.291	6.390	9.052	10.799	12.360	14.686	16.225
p = 6	25	1.779	2.762	4.415	6.472	8.904	10.725	12.320	14.241	16.018
	50	1.988	3.089	4.831	7.198	9.908	11.557	13.237	15.368	16.906
	100	2.215	3.418	5.155	7.369	9.774	11.724	13.659	16.511	16.867

Table 5.3

Powers of the density based, Cramer-von Mises, Rayleigh and Fisher'scombined tests, estimated from 1000 simulated samples in each case, when the underlying distribution is Cauchy (0,1) in each margin.

	Dimension					
	1	2	3	4	5	6
Sample size 25						
ISE fixed	0.948	0.997	0.995	1.000	0.999	1.000
Entropy (fixed)	0.950	0.997	1.000	1.000	0.999	1.000
Entropy ($\alpha = 1/p$)	0.926	0.986	1.000	1.000	0.999	1.000
CVM	–	0.986	0.995	1.000	0.999	0.991
RAY	–	0.897	0.965	0.981	0.995	0.997
FISHER	–	0.990	0.997	1.000	1.000	0.999
Sample size 50						
ISE fixed	0.998	1.000	1.000	1.000	1.000	1.000
Entropy (fixed)	0.998	1.000	1.000	1.000	1.000	1.000
Entropy ($\alpha = 1/p$)	0.998	1.000	1.000	1.000	1.000	1.000
CVM	–	1.000	1.000	1.000	1.000	1.000
RAY	–	0.969	0.992	0.996	0.999	1.000
FISHER	–	1.000	1.000	1.000	1.000	1.000
Sample size 100						
ISE fixed	1.000	1.000	1.000	1.000	1.000	1.000
Entropy (fixed)	1.000	1.000	1.000	1.000	1.000	1.000
Entropy ($\alpha = 1/p$)	1.000	1.000	1.000	1.000	1.000	1.000
CVM	–	1.000	1.000	1.000	1.000	1.000
RAY	–	0.983	1.000	1.000	1.000	1.000
FISHER	–	1.000	1.000	1.000	1.000	1.000

Table 5.4

Powers of the density based, Cramer-von Mises, Rayleigh and Fisher's combined tests, estimated from 1000 simulated samples in each case, when the underlying distribution is Cauchy (0,1) in each margin.

	Dimension					
	1	2	3	4	5	6
Sample size 25						
ISE fixed	0.966	0.996	0.999	1.000	0.999	0.998
Entropy (fixed)	0.959	0.992	0.995	0.999	0.996	0.997
Entropy ($\alpha = 1/p$)	0.964	0.990	0.995	0.991	0.993	0.998
CVM	–	0.816	0.875	0.910	0.901	0.879
RAY	–	0.955	0.975	0.980	0.990	0.992
FISHER	–	0.960	0.984	0.985	0.989	0.986
Sample size 50						
ISE fixed	1.000	1.000	1.000	1.000	1.000	1.000
Entropy (fixed)	1.000	1.000	1.000	1.000	1.000	1.000
Entropy ($\alpha = 1/p$)	1.000	1.000	1.000	1.000	1.000	1.000
CVM	–	0.991	0.997	0.999	0.998	0.999
RAY	–	1.000	1.000	1.000	1.000	1.000
FISHER	–	1.000	1.000	1.000	1.000	1.000
Sample size 100						
ISE fixed	1.000	1.000	1.000	1.000	1.000	1.000
Entropy (fixed)	1.000	1.000	1.000	1.000	1.000	1.000
Entropy ($\alpha = 1/p$)	1.000	1.000	1.000	1.000	1.000	1.000
CVM	–	1.000	1.000	1.000	1.000	1.000
RAY	–	1.000	1.000	1.000	1.000	1.000
FISHER	–	1.000	1.000	1.000	1.000	1.000

Table 5.5

Powers of the density based, Cramer–von Mises, Rayleigh and Fisher's combined tests, estimated from 1000 simulated samples in each case when the underlying distribution is a bimodal normal mixture

	Dimension					
	1	2	3	4	5	6
Sample size 25						
ISE fixed	0.621	0.671	0.440	0.336	0.176	0.151
Entropy (fixed)	0.488	0.420	0.192	0.185	0.095	0.072
Entropy ($\alpha = 1/p$)	0.715	0.661	0.349	0.181	0.098	0.092
CVM	–	0.452	0.221	0.128	0.107	0.074
RAY	–	0.117	0.077	0.064	0.033	0.049
FISHER	–	0.370	0.126	0.065	0.049	0.046
Sample size 50						
ISE fixed	0.962	0.998	0.996	0.847	0.619	0.488
Entropy (fixed)	0.918	0.934	0.708	0.392	0.213	0.158
Entropy ($\alpha = 1/p$)	0.966	0.988	0.863	0.480	0.295	0.164
CVM	–	0.845	0.489	0.243	0.173	0.123
RAY	–	0.117	0.090	0.063	0.043	0.037
FISHER	–	0.787	0.343	0.170	0.093	0.063
Sample size 100						
ISE fixed	0.999	1.000	1.000	1.000	0.999	0.980
Entropy (fixed)	0.999	1.000	1.000	0.953	0.692	0.423
Entropy ($\alpha = 1/p$)	0.999	1.000	1.000	0.967	0.697	0.387
CVM	–	0.997	0.880	0.527	0.298	0.187
RAY	–	0.125	0.086	0.061	0.037	0.040
FISHER	–	0.997	0.812	0.368	0.183	0.118

Table 5.6

Powers of the density based, Cramer-von Mises, Rayleigh and Fisher's combined tests, estimated from 1000 simulated samples in each case when the underlying distribution is Gamma (2,1) in each margin

	Dimension					
	1	2	3	4	5	6
Sample size 25						
ISE fixed	0.588	0.688	0.680	0.697	0.603	0.599
Entropy (fixed)	0.551	0.629	0.633	0.650	0.526	0.492
Entropy ($\alpha = 1/p$)	0.586	0.574	0.572	0.590	0.483	0.485
CVM	–	0.199	0.247	0.234	0.164	0.128
RAY	–	0.474	0.589	0.536	0.553	0.524
FISHER	–	0.449	0.545	0.507	0.473	0.432
Sample size 50						
ISE fixed	0.880	0.951	0.964	0.961	0.960	0.958
Entropy (fixed)	0.908	0.939	0.937	0.933	0.901	0.900
Entropy ($\alpha = 1/p$)	0.900	0.930	0.908	0.910	0.910	0.904
CVM	–	0.424	0.546	0.595	0.612	0.563
RAY	–	0.812	0.904	0.928	0.929	0.943
FISHER	–	0.810	0.889	0.916	0.921	0.934
Sample size 100						
ISE fixed	0.995	1.000	1.000	1.000	1.000	1.000
Entropy (fixed)	0.998	0.999	1.000	1.000	1.000	0.999
Entropy ($\alpha = 1/p$)	0.998	1.000	0.997	1.000	0.997	0.994
CVM	–	0.727	0.875	0.917	0.933	0.926
RAY	–	0.990	0.996	0.999	1.000	1.000
FISHER	–	0.886	0.998	0.999	1.000	0.999

Table 5.7

5% points of the integrated squared error test statistic, simulated from 1000 samples in each case

The critical region lies above the tabulated value

	Dimension					
Sample size	1	2	3	4	5	6
25	0.0109	0.00735	0.00378	0.00165	0.000685	0.000263
50	0.00766	0.00551	0.00282	0.00128	0.000533	0.000206
100	0.00567	0.00371	0.00203	0.000931	0.000395	0.000160
150	0.00453	0.00304	0.00163	0.00076	0.000340	0.000138
200	0.00380	0.00261	0.00142	0.000678	0.000300	0.000124
250	0.00332	0.00229	0.00126	0.000614	0.000275	0.000113
300	0.00301	0.00200	0.00114	0.000563	0.000253	0.000106
350	0.00272	0.00188	0.00105	0.000529	0.000236	0.0000995
400	0.00232	0.00176	0.000991	0.000498	0.000223	0.0000945
500	0.00205	0.00151	0.000877	0.000440	0.000201	0.0000864

5.5 Examples.

Example 1. Some haematology data.

These data, given in full Royston (1983), consist of six measurements made on each of 103 black (West Indian or African) paint sprayers in a car assembly plant. The six variables considered are:

1. Haemoglobin concentration.
2. Packed cell volume.
3. White blood cell count.
4. Lymphocyte count.
5. Neutrophil count.
6. Serum lead concentration.

Variables 3, 4, 5 and 6 have skewed empirical distributions and were therefore logarithmically transformed before analysis.

This dataset was tested for consistency with multivariate normality by Royston (1983) who extended Shapiro and Wilks' (1965) univariate W test to a multivariate setting (the 'H test') and also to a normal probability plot of the (square-root transformed) squared radii, $\hat{R}_i$ (the "Ω test"). He reports that the normal probability plots for each of the six variables are reasonably linear with the H test of combined W ranks having a p-value of 0.08. However, the Ω test indicates a strong departure from 6-normality with a p-value of 0.0004. To further investigate this the Ω test was carried out on all pairs and triples of the variables. The only significant result is for 3, 4 and 5 with $p = 6 \times 10^{-6}$. Three outliers are identified in the space of these three variables (cases 21, 47 and 52) and on their removal the Ω test on variables 3-5 has a p-value of 0.52.

The values of the test statistics discussed in detail in this

chapter are presented in the following table.

Table 5.6. Results of tests of normality for the haematology data.

Test	Data Set			
	Vars 1-6	Vars 1-6 ex. cases 21,47,52	Vars 3,4,5	Vars 3,4,5 ex. cases 21,47,52
CVM	0.253 ($p \simeq 0.017$)	0.118 ($p \simeq 0.20$)	0.406 ($p < 0.005$)	0.0894 ($p \simeq 0.39$)
RAY	33.323 ($p < 0.005$)	22.450 ($p < 0.005$)	35.329 ($p < 0.005$)	23.265 ($p < 0.005$)
FISHER	20.13 ($p < 0.005$)	15.20 ($p < 0.005$)	23.97 ($p < 0.005$)	13.92 ($p \simeq 0.008$)
ISE	0.000168 ($p < 0.05$)	0.000164 ($p < 0.05$)	0.00396 $p << 0.05$	0.00317 $p << 0.05$

For the full dataset each test results in a significant result. The χ^2 probability plot (figure 5.3) shows most of the points near the straight line but identifies cases 10, 21, 47, 52 and 80 as possible outliers - in particular 21 and 52 are very extreme. Removing 21, 47 and 52 in accordance with Royston and then repeating the tests results in a fairly marked decrease in the value of each test statistic except for ISE. The CVM test is now non-significant but for the others the results are qualitatively the same. This is in contrast to Royston's results which provide no evidence of departure from 6-normality on removal of these three individuals. If, in addition, cases 10 and 80 are also removed the values of the four test statistics (not included) are fairly similar to when only 21, 47 and 52 are removed thus leading to the same conclusions.

The results for variables 3, 4 and 5 are highly significant for each test thus confirming Royston's conclusions. The χ^2 probability

plot (figure 5.4) again shows 21, 47 and 52 to be extreme. On their removal the values of each statistic decrease markedly but again only CVM is non-significant - the others are still all highly significant thereby indicating strong departures from 3-normality.

Example 2. Plasma lipid data

A group of 371 males were selected from patients with chest pain, referred to a hospital cardiology unit, to be part of a study into the relationship between plasma cholestrol and plasma triglyceride concentrations (mg/100ml) and coronary artery disease. These patients were then divided into two groups according to the criteria "diseased" or "normal". 320 patients were included in the "diseased" group with the remaining 51 being classed as "normal". Details of the experimental methods are described by Scott et al (1978).

To analyse the risk of coronary artery disease associated with higher levels of plasma triglyceride Scott et al use the "odds form" of Baye's rule which depends on the likelihood ratio of the joint bivariate density for diseased patients to that for normal patients. By using Kolmogorov-Smirnov two-tailed tests on the marginal distributions they reject the hypothesis that the underlying density functions are normal. The bivariate densities are therefore estimated nonparametrically using a kernel approach which is described in the paper.

For the diseased patients a scatter plot and contour plot based on a kernel density estimate are illustrated in Silverman (1986, p.81-82). These show the distribution to be both bimodal and highly skewed. Indeed, omnibus tests of normality using Fisher's combined test and the ISE statistic are both highly significant (see table

5.7). The data were therefore transformed using the natural log function. A scatter plot with the contours of a kernel density estimate superimposed is illustrated in figure 5.5 The density estimate is based on standard normal fixed kernels using a smoothing parameter, h, given by (5.3.6) scaled by a robust estimate of scale (Hogg (1979)) in each co-ordinate direction. The distribution is now unimodal with much of the skewness removed but both omnibus tests are still highly significant (see table 5.7).

Table 5.8 Results of tests of Normality for the plasma lipid data.

Data	Tests			
	CVM	RAY	FISHER	ISE
Diseased patients				
Original data (n = 320)	1.937 ($p << 0.005$)	81.320 ($p << 0.005$)	23.97 ($p << 0.005$)	0.0158 ($p << 0.05$)
Log_e transformed (n = 320)	0.154 ($p \simeq 0.14$)	9.545 ($p < 0.005$)	15.92 ($p < 0.005$)	0.00292 ($p < 0.05$)
Normal patients				
Original data (n = 51)	0.450 ($p < 0.005$)	6.727 ($p \simeq 0.0375$)	18.550 ($p < 0.005$)	0.00853 ($p < 0.05$)
Original data (n = 50)	0.206 ($p \simeq 0.0617$)	3.822 ($p \simeq 0.168$)	9.139 ($p \simeq 0.0602$)	0.00385 ($p > 0.05$)

For the 51 normal patients a scatter plot with contours superimposed is shown in figure 5.6. The contours are again based on using fixed standard normal kernels with h given by (5.3.6) scaled by robust estimates of scale. The distribution is unimodal and does not have the same spread, especially for the plasma triglyceride concentration, as for the data from the diseased patients. The tests of normality are all significant though (see table 5.7). The

scatter plot enables a clear outlier to be identified who has a high plasma cholestrol concentration and a very high plasma triglyceride concentration relative to the other normal patients. When this patient is omitted none of the tests are significant at the 5% level (see table 5.7) but the low p-value for the Cramer-von Mises and Fisher's combined test indicates some caution is required in accepting that the underlying distribution is bivariate normal.

The results of the more extensive testing of bivariate normality detailed above therefore vindicates the use of kernel density estimates in the likelihood ratio in the analysis of Scott et al.

Figure 5.1. Powers of the goodness-of-fit tests based on 1000 samples of size 50 from the $0.5Np((-1.5,\ldots,-1.5)^T, Ip) + 0.5Np((1.5,\ldots,1.5)^T, Ip)$ distribution.

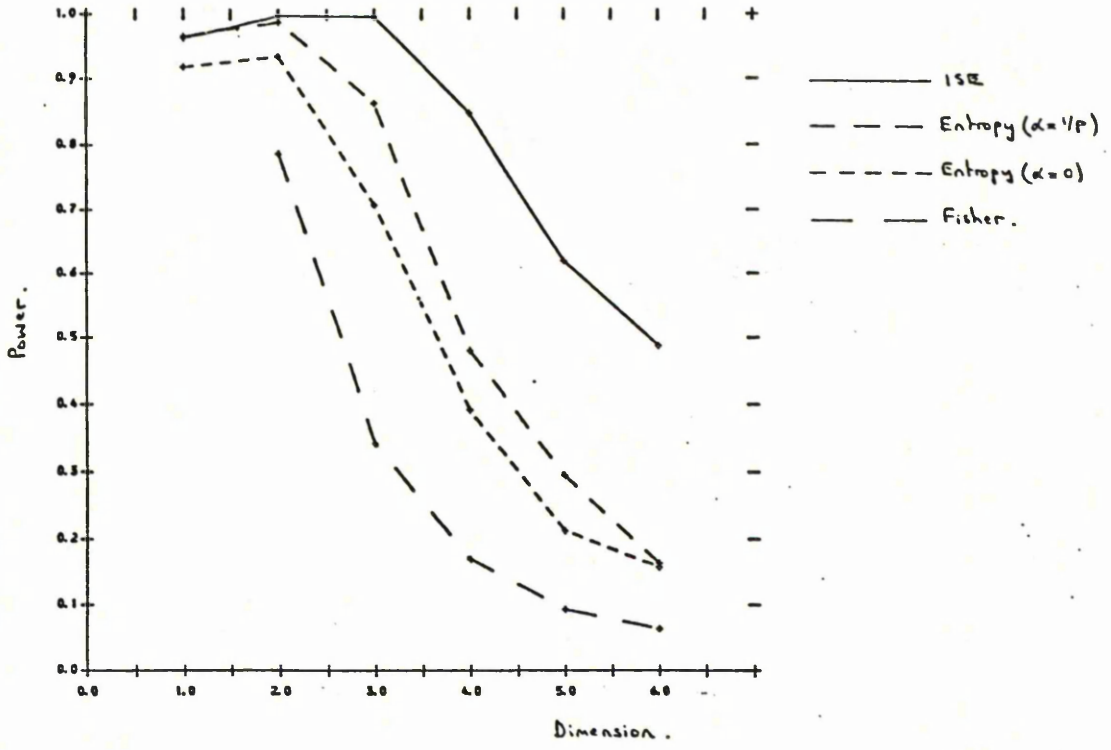


Figure 5.2. Powers of the goodness-of-fit tests based on samples of size 50 from a distribution which is Gamma(2,1) in each margin.

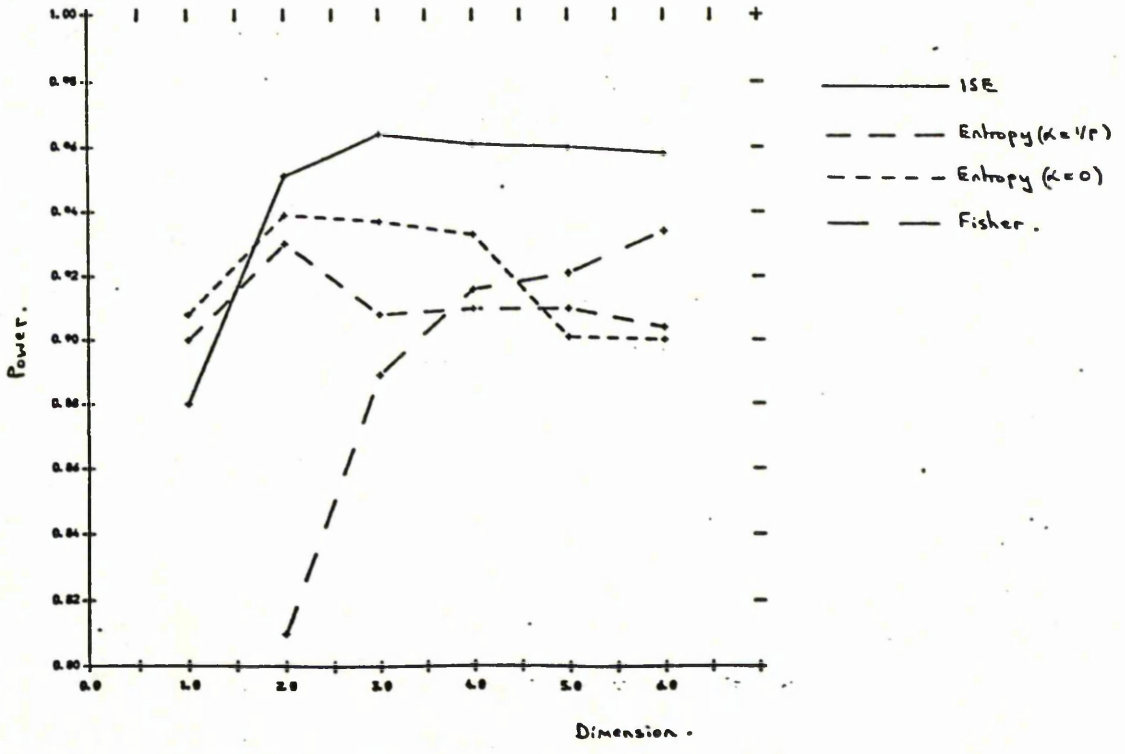


Figure 5.3. χ^2 probability plot of the Mahalanobis squared distances for the haematology data (variables 1-6).

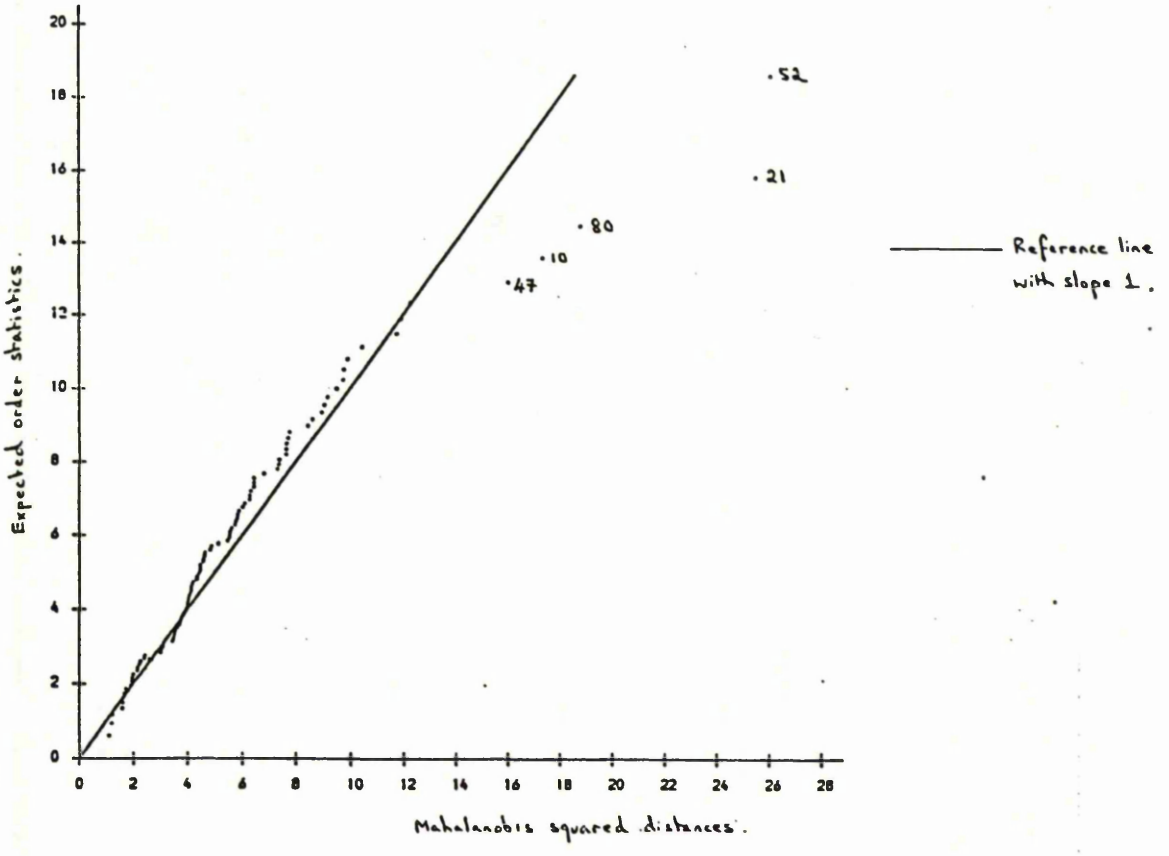


Figure 5.4. χ^2 probability plot of the Mahalanobis squared distances for the haematology data (variables 3,4,5)

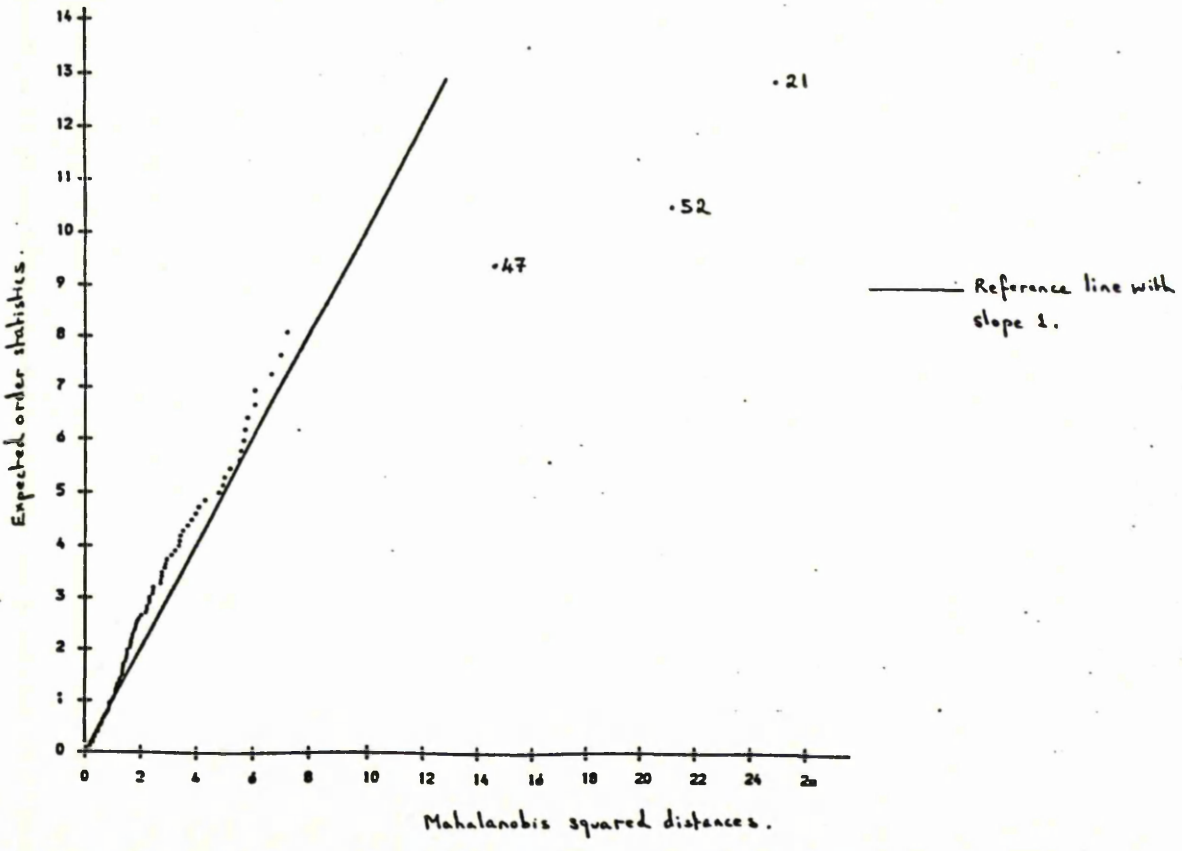


Figure 5.5 Scatter plot and kernel estimate for the log-transformed plasma lipid data. (Diseased patients only).

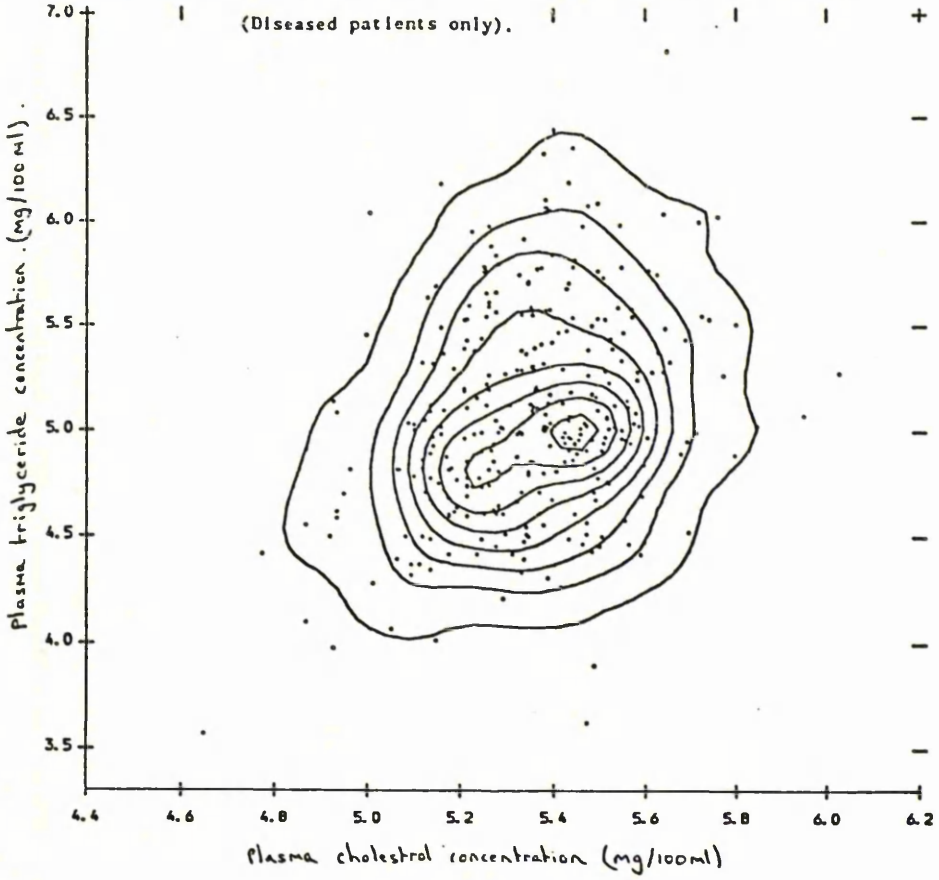
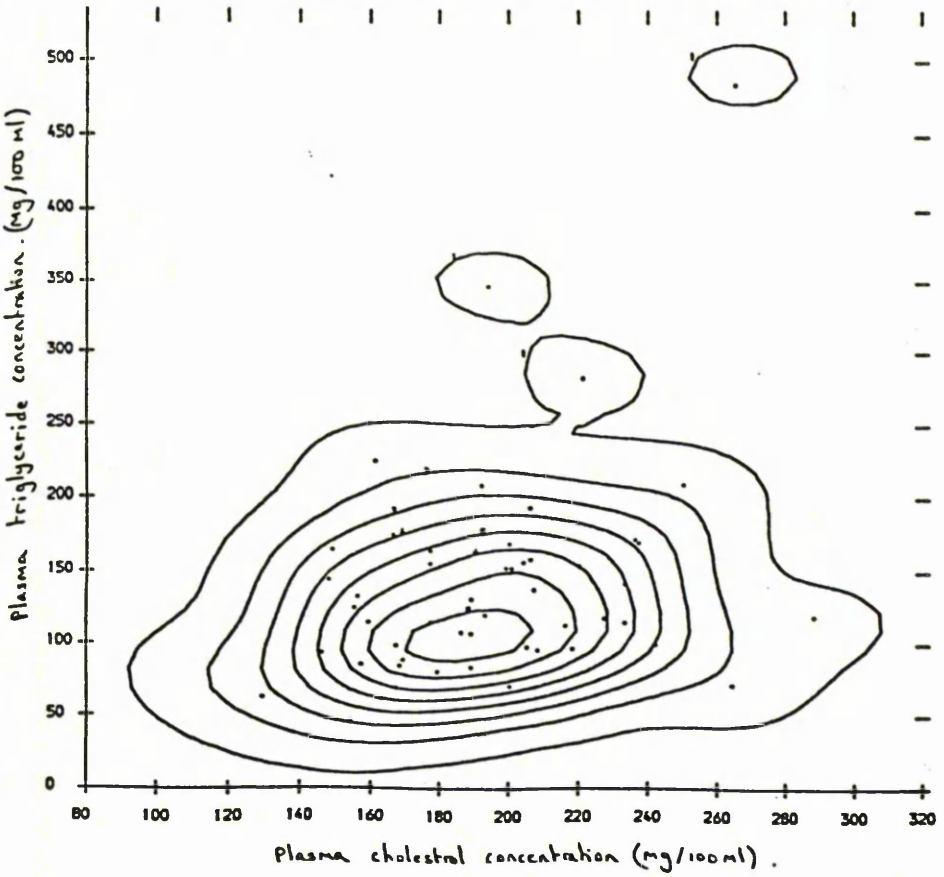


Figure 5.6. Scatter plot and kernel estimate for the plasma lipid data. (Normal patients only)



Chapter 6. Finding Directions of High Multivariate Density

6.1. Introduction.

Given a multivariate data set $\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n\}$, the initial analysis is often exploratory with the hope that any important features such as clusters, skewness or long "arms" will be revealed. This is generally approached by looking at a variety of graphical displays which therefore necessitates a presentation of the data in one or two dimensions. The most common technique used to achieve this is the application of linear projections. While these are straightforward to implement and are certainly useful their interpretation may not be so easy as discussed by Gower in the discussion of Jones and Sibson (1987). Also, projections obscure either partly or totally actual structure of the full dimensional data. Two methods for finding "interesting" projections are principal components analysis and projection pursuit.

Principal components analysis (PCA) is a rigid rotation of the original axes $X_1,X_2,\ldots,X_p$ to new positions $Y_1,Y_2,\ldots,Y_p$ such that the orthogonal projections of the data onto them have decreasing spread. The first q ($< p$) components define the best fitting q-dimensional subspace to the data in terms of minimising the sum of squared orthogonal distances from the sample points to this subspace. In particular Y_1 and Y_2 define the best fitting plane. To be successful, PCA requires that large variation corresponds to interesting structure which may easily fail to be true in practice.

Projection pursuit (PP) methods seek projections of the data which maximise some index of "interestingness". PCA is therefore a PP

procedure where the index corresponds to the proportion of the total variation explained by the projected data. Friedman and Tukey (1974) constructed a PP index based on the product of a measure of spread and the local density of the projected data which has to be optomised numerically. Recent authors such as Huber (1985), Jones and Sibson (1987) and Friedman (1987) have approached the problem by considering the converse idea of uninteresting projections and present heuristic arguments that normally distributed projected data is least interesting. Their projection indexes are then based on indices of normality such as entropy with the numerical optimisation procedure then seeking maximum divergence from this criteria. The two main problems with PP methods are firstly the difficulty in successfully implementing the numerical optimisation procedure which may take considerable computing effort and also be trapped by local maxima and secondly that the structure apparently revealed is just the result of random variation.

In this Chapter a different exploratory approach based on finding directions of high multivariate density will be described. There are three main aims in doing this. Firstly, to explore the shape of a multivariate density function when it cannot be readily plotted for $p > 2$. Secondly, to find non-linear features such as clustering in the data and thirdly, to use pairs of directions for the construction of 2-dimensional representations. The technique will also be useful in determining the reasons for the rejection of the hypothesis of normality following one of the density based tests described in Chapter 5.

As is normally the case before carrying out a PCA or PP analysis the data will be recentred using the sample mean vector and the scale

effects removed by standardising each variable to have unit variance. In the following discussion it will be assumed that the data have already been transformed in this way.

6.2. Finding the directions.

Features such as clusters, skewness and long "arms" in the data will clearly be indicated by modes and long tails in the underlying density function. A cross-section of the density in a direction through such a feature will therefore have large cross-sectional area. Hence a criteria for identifying these effects is to find directions $\underline{v}$ from the centre which maximise

$$\int_0^\infty f(c.\underline{v})dc \qquad (6.2.1)$$

where f is the underlying density function, $\underline{v} = (v_1, v_2, \ldots, v_p)^T$ is a unit vector and $c \geqslant 0$ is a scalar.

For bivariate data a line a little longer than the distance from the centre to the furthest data point can be swept round in a circle in small angular steps. For any such line the quantity

$$\sum_i \hat{f}(\underline{z}_i) \qquad (6.2.2)$$

is calculated where the $\underline{z}_i$'s form an equally spaced grid of points along this line and $\hat{f}(\cdot)$ is a nonparametric estimate of the true density. The values of (6.2.2) are then plotted against $\theta \in [0, 2\pi]$ with the modes in the plot corresponding to interesting directions in the data. In Section 6.4 it is proved that if the underlying distribution is bivariate normal then the two directions of highest density are equivalent to the positive and negative directions of the first PC.

This method could also be applied to multivariate data in general but for $p \geq 3$ is computationally rather infeasible and a different approach is therefore needed.

As we are interested only in directions, an alternative strategy is first to scale all vectors $\{\underline{x}_i\}$ to have unit length, i.e.

$$\underline{v}_i = (\|\underline{x}_i\|_2)^{-1} . \underline{x}_i \qquad (6.2.3)$$

where $$\|\underline{x}_i\|_2 = \sqrt{\sum_{j=1}^{p} x_{ij}^2} \quad .$$

The elements of $\underline{v}_i$ correspond to the direction cosines of the vector $\underline{x}_i$. Interesting directions of high density now correspond to modes of the distribution which has been induced on the p-dimensional unit hypersphere because the mass g at a point $\underline{v}$ on this hypersphere equals $\int_0^\infty f(c.\underline{v})dc$.

Given a random sample of data on the real line a number of authors have considered methods of estimating the mode θ of the underlying density f based on this data. The direct method of Chernoff (1964) is to choose the centre of the interval of length 2a, for some constant a, that contains the largest number of observations. Venter (1967) also based his estimates on that point around which the greatest "clustering" of observations occurs but uses instead a function of the order statistics. Grenander's (1965) estimate is also based on the spacings between the data but uses a different function of the order statistics. Alternatively, an indirect estimate of θ can be based on a kernel density estimate $\hat{f}$ of f. The mode, assumed unique, is then defined to be $\max_x \hat{f}(x)$. The properties of such an estimate have been studied by Parzen (1962) and Eddy (1980). Both the direct

and indirect approaches are reviewed by Rao (1983).

Direct and indirect methods for estimating the mode in the multivariate case are also described by Rao (1983). In the discussion of a direct estimate it is assumed that the density f is right quadrant continuous, i.e.

$$\lim_{\underline{y}\to\underline{x},\ \underline{y} > \underline{x}} f(\underline{y}) = f(\underline{x}) \qquad (6.2.4)$$

and there is a $\underline{\theta}$ such that $f(\underline{\theta}) > f(\underline{x})$ for all $\underline{x} \neq \underline{\theta}$. Let $a_n \longrightarrow 0$ as $n \longrightarrow \infty$ and let $[\underline{l}_n, \underline{l}_n + a_n.\underline{1}]$ be an interval of "length" a_n containing the largest number of data points $\{\underline{x}_i\}$ among all intervals of "length" a_n. Here, $\underline{l}_n$ denotes a p-vector and $\underline{1}$ a p-vector of 1's. The direct estimate of θ is then given by $\underline{l}_n + (a_n/2).\underline{1}$. The consistency of this estimator is shown but in practice it would be very difficult to use because of the considerable computational problem of searching for a "best" interval. Sager (1979) discusses a related estimate based on finding a sequence of nested convex sets containing given numbers of observations but points to the practical difficulties of finding such sets.

The properties of an indirect estimate of the unique mode of a unimodal density f based, as in the univariate case, on finding $\max_{\underline{x}} \hat{f}(\underline{x})$, where $\hat{f}$ is a kernel estimate of f, are also discussed.

In the context of direction finding we will generally need to be able to locate several modes of the distribution on the p-dimensional hypersphere. The multivariate methods discussed above are not really practical or applicable so we develop the following algorithm which partitions the data into groups around directions of high density and then calculates the mean direction of each group as an estimate of the local mode.

i) Estimate $g(\underline{v}_i)$, $i = 1,\ldots,n$ where $g(\cdot)$ is the underlying induced density on the hypersphere.

ii) Rank the observations from largest to smallest according to the size of the estimates $\hat{g}(\cdot)$, i.e.

$$\hat{g}_{(n)},\cdots,\hat{g}_{(1)}$$

iii) Characterise a first group G_1 by the direction of the observation corresponding to $\hat{g}_{(n)}$. Call this direction $\underline{d}_1$

iv) Assign the observation corresponding to $\hat{g}_{(n-1)}$ to G1 if the angle between this observation and $\underline{d}_1$ is less than α radians (e.g. $\alpha = \pi/2$). Otherwise, form a new group, G_2, whose characteristic direction, $\underline{d}_2$, is defined by the direction cosines of the observation corresponding to $\hat{g}_{(n-1)}$.

v) In descending order of density height, successively assign observations to the closest existing group if the angle of separation is less than α. If the angle between the current observation and all existing groups is greater than or equal to α form a new group.

The end result is k groups $G_1,G_2,\ldots,G_k$ containing $n_1,n_2,\ldots,n_k$ observations such that $\sum_i n_i = n$.

iv) Find the mean direction, using the $\hat{g}(\cdot)$ as weights, for each of the groups, i.e. for G_ℓ the mean direction is

$$\Big[\sum_{i=1}^{n_\ell} \hat{g}_i^{(\ell)}\Big]^{-1} \cdot \sum_{i=1}^{n_\ell} \hat{g}_i^{(\ell)} \cdot \underline{v}_i^{(\ell)}$$

where $\underline{v}_i^{(\ell)}$ denotes the ith vector of direction cosines with weight $\hat{g}_i^{(\ell)}$ in the ℓ^{th} group.

The lengths of the mean vectors provide a measure of concentration of the directions in a group about the mean direction. A value close to one will indicate tight clustering about the mean. (Mardia (1972)).

vii) Scale the mean directions to have unit length.

It has been found in practice that by choosing $\alpha = \pi/2$ the procedure is often able to find the main features of the data. However, results could be compared for different values of α especially if it is believed that a number of distinct features are fairly close together. As α is reduced the number of groups tends to increase, particularly for high dimensional data, and hence the numbers of observations forming the groups decreases. Conversely, as α is increased the number of groups tends to decrease.

If required, only a specified proportion (e.g. 0.75) of the observations with highest density could be used to find the modes thus using the ideas of sharpening by Tukey and Tukey (1981).

Before trying to find the modes of a density on the hypersphere, a preliminary test that this density is uniform, which corresponds to a radially symmetric scatter of the data in Euclidean space, could be carried out to try and avoid finding spurious directions.

At each step, when an observation is assigned to a group, the characteristic direction of that group could then be updated by calculating the mean direction of all the current members of the group. The final number of groups and their characteristic directions will not necessarily be the same as when no updating takes place. In practice though, no advantages of doing this as opposed to averaging

at the end have generally been found.

Kittler (1976) estimates the density of each data point and then uses this information in the construction of a path through the data which passes through as many of the data points as possible about one mode before passing onto those near another. The aim is not to specifically identify where the modes are but to classify the data into clusters based on these modes.

The algorithm described above depends on being able to estimate a hyperspherical density so in the next Section the construction of an appropriate kernel estimate will be described.

6.3. Estimation of the density on a hypersphere.

Given a random sample of unit vectors $(\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n)$ where

$$\underline{v}_i = (v_{i1}, v_{i2}, \ldots, v_{ip})^T \qquad (6.3.1)$$

it is required to estimate the density as the point $\underline{v}$.

It is proposed to use as a kernel function with an appropriate density defined on a hypersphere, to avoid any problems with choice of origin, i.e.

$$\hat{g}(\underline{v}) = \frac{1}{n} \sum_{i=1}^{n} K(\underline{v} \; ; \; \underline{v}_i, \; c) \qquad (6.3.2)$$

where $K(.,v_i,c)$ is the kernel function centered at $\underline{v}_i$ and c is a smoothing parameter. As a choice of kernel consider the rotationally symmetric unimodal function

$$K(\underline{l};\underline{m},k) = \frac{k^{(p/2-1)} \; e^{k.\underline{l}^T\underline{m}}}{(2\pi)^{p/2} \; I_{(p/2-1)}(k)} \qquad (6.3.3)$$

where $\underline{l}$ is a vector of direction cosines, $\underline{m}$ is the vector of direction cosines corresponding to the mean direction, k is a scale parameter such that the larger its value the more concentrated $K(.;\underline{m},k)$ is about the mean/modal direction and $I_p(k)$ is the modified Bessel function of order p evaluated at k. The quantity $d_0(k) = k^{(p/2-1)}/\{(2\pi)^{p/2} I_{(p/2-1)}(k)\}$ can be regarded as the normalising constant for the function $e^{k\underline{l}^T\underline{M}}$. If a change of variable to polar co-ordinates is made then the p.d.f. for $p = 2$ is the Von-Mises distribution and for $p = 3$ is the Fisher distribution.

When using (6.3.3) as a kernel in (6.3.2) large values of c will mean each density in the summation is concentrated and around its mode at $\underline{v}_i$ and hence only a small amount of smoothing is carried out. The opposite is true for small values of c until when $c = 0$ it becomes the uniform distribution on the hypersphere.

For the purposes of the mode seeking algorithm it is not necessary to include the normalising constant in the evaluation of $\hat{g}$ because when using fixed kernels it will be the same for each $\hat{g}(\underline{v}_i)$ and it is only the relative size of the density weights which are important. This avoids having to calculate the function $I_p(c)$.

To calculate the estimate it is necessary to choose a value for the smoothing parameter c. One approach would be to carry out an analysis separately for several subjectively chosen values of c and compare the results. It has been found that results can be fairly similar for a broad range of smoothing parameters. However, it is probably more useful to have a more accurate assessment of the degree of smoothing required based on a suitable loss function.

One approach described in the discussion of optimal smoothing in Bowman (1988) for circular data and Diggle and Fisher (1984) for

spherical data approximates the Von Mises and Fisher distributions by wrapped normal distributions. The results for optimal smoothing, in terms of minimising the MISE when the data is from a normal distribution, are then adapted to the angular case. While this analogy is reported to work well for unimodal, approximately symmetric distributions, in the present context we are generally expecting the underlying density to be multimodal and so using this approach may result in considerable oversmoothing.

The integrated squared error (ISE) of the estimator $\hat{g}$ is

$$\int (\hat{g}-g)^2 = \int \hat{g}^2 - 2\int \hat{g}\, g + \int g^2 . \qquad (6.3.4)$$

The last term does not depend on $\hat{g}$ so choosing c to minimise ISE corresponds to that choice which minimises the function $R(\hat{g})$ defined by

$$R(\hat{g}) = \int \hat{g}^2 - 2 \int \hat{g}\, g . \qquad (6.3.5)$$

The idea of least squares cross-validation is to construct an estimate of $R(\hat{g})$ using the data and then minimise the value of this estimate over c to give the choice of smoothing parameter. Using arguments analogous to those of Rudemo (1982) and Bowman (1984) for data on the real line the quantity to be minimised for angular data is

$$M(c) = \frac{1}{n^2} \sum_{ij} K2\ (\underline{v}_i\ ;\ \underline{v}_j,\ c) - \frac{2}{n^2} \sum_{i \neq j} \sum K(\underline{v}_i\ ;\ \underline{v}_j, c) \qquad (6.3.6)$$

where $K2(\cdot)$ is the convolution of K with itself. When using (6.3.3) as a kernel function this convolution does not result in another function of the same type as is the case with normal densities. When $p = 2$ Mardia (1972) used wrapped normal approximations to the Von

Mises distributions to obtaining an approximation to K2. The result is another Von Mises density but with concentration parameter k_3 which is the solution to

$$A(k_3) = A(k_1).A(k_2) \qquad (6.3.7)$$

where $A(k) = I_1(k)/I_0(k)$. This is a complex relationship to have to invert and the possibility of having to do this many times in the minimisation of $M(c)$, in addition to calculating the normalising constant which depends on the concentration parameter, makes least squares cross-validation computationally impractical for $p = 2$. Similarly, it will also be impractical for $p > 2$.

The most practical approach is to use log-likelihood cross-validation which chooses c to maximise

$$L(c) = \frac{1}{n} \sum_{i=1}^{n} \ln \left\{ \hat{g}_{-i}(\underline{v}_i) \right\} \qquad (6.3.8)$$

where $\hat{g}_{-i}(\underline{v}_i)$ is the kernel estimator (6.3.2) evaluated at $\underline{v}_i$ using all the data points except $\underline{v}_i$. The normalising constant d_0 also depends on c and so needs to be calculated but, because fixed kernels are being used, it is the same for each term in the summation and it is therefore only necessary to evaluate it once. Using the results in Mardia (1972, Sect. 8.8) it can be expressed as:

$$d_0(c)^{-1} = 2.\int_0^{\pi} e^{c.\cos\theta_1} \, d\theta_1, \quad p = 2.$$

$$d_0(c)^{-1} = \left[\int_0^{\pi} \sin^{p-2}\theta_1 \, e^{c.\cos\theta_1} d\theta_1\right].\left[\int_0^{\pi} \sin^{p-3}\theta_2 \, d\theta_2\right].....$$

$$\left[\int_0^{\pi} \sin\theta_{p-2} \, d\theta_{p-2}\right] \left[\int_0^{\pi} 1.d\theta_{p-1}\right], \; p = 3,4... \qquad (6.3.9)$$

The integral

$$\int_0^{\pi} \sin^{p-2}\theta_1, \, e^{c\,\cos\theta_1} \, d\theta_1 = J_p, \quad \text{say} \qquad (6.3.10)$$

needs to be evaluated using numerical integration but the others in the product can be evaluated analytically. The resulting constants for dimensions 2 to 13 are given in table 6.1.

Table 6.1. Normalising constants of integration for the function $\exp(c.\underline{1}^T.\underline{m})$

$\underline{p}$	$d_0(c)^{-1}$
2	$2.J_2$
3	$2\pi.J_3$
4	$4\pi.J_4$
5	$2\pi^2.J_5$
6	$8\pi^2/3.J_6$
7	$\pi^3.J_7$
8	$16\pi^3/15.J_8$
9	$\pi^4/3.J_9$
10	$32\pi^4/105.J_{10}$
11	$\pi^5/12.J_{11}$
12	$74\pi^5/945.J_{12}$
13	$\pi^6/60.J_{13}$

On the real line the performance of likelihood cross-validation has been noted to be sensitive to outliers. (Scott and Factor (1981)). This will carry over to angular data because an outlier will contribute a large negative value to $L(c)$ unless c is small. Hence, data sets which contain at least one outlier will tend to be oversmoothed.

Schuster and Gregory (1981) show that if the density f of data on the real line has a tail which is monotonic and dies off either exponentially or slower then using likelihood cross-validation will not result in a consistent estimate of f. This is because as n tends to infinity extreme observations will be recorded in the tails and the sensitivity of the method to outliers ensures that the value of the smoothing parameter does not tend to zero. However, for

angular data the domain of the observations is bounded and so the gaps between observations should shrink towards zero as n goes to infinity and hence the argument of Schuster and Gregory will not apply.

6.4. Results for normal data.

Consider the bivariate normal distribution with mean vector $\underline{0}$ and correlation matrix $\sum = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ so that the p.d.f. for $\underline{x} = (x_1, x_2)^T$ is

$$f(x) = \frac{1}{2\pi} |\sum|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \underline{x}^T \sum^{-1} \underline{x}\right)$$

$$= \frac{1}{2\pi} |\sum|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{1}{1-\rho^2}\right](x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\} \quad (6.4.1)$$

If a transformation to polar co-ordinates is made then $x_1 = r.\cos\theta$ and $x_2 = r.\sin\theta$ and

$$f(r,\theta) = \frac{1}{2\pi} |\sum|^{-\frac{1}{2}} \exp\left[\frac{-r^2}{2(1-\rho^2)}(1-\rho\sin 2\theta)\right] \quad (6.4.2)$$

Therefore,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\underline{x})\, d\underline{x} = \int_0^{2\pi}\int_0^{\infty} r.f(r,\theta).dr.d\theta. \quad (6.4.3)$$

(The Jacobian of the transformation is r).

Because of the correlation ρ between the variables, f is not radially symmetric, and so the integral is not constant over θ. The values of θ for which the cross-sectional integral along the corresponding radius from the origin is maximised is when the function

$$\exp\left\{-\frac{r^2}{2(1-\rho^2)}(1-\rho\sin 2\theta)\right\} \quad (6.4.4)$$

is a maximum. This occurs at the values of θ which minimise $(1-\rho\sin 2\theta)$ which are easily found to be $\theta = \pi/4$ and $\theta = 5\pi/4$.

Therefore, the unit vectors defining the directions along which the cross-sectional integral of the density is a maximum are

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} . \qquad (6.4.5)$$

Chatfield and Collins (1980) show that this vector is the first principal component for the above correlation matrix.

Hence, for bivariate normal data the method for finding directions of highest density is equivalent to finding the first principal component which in turn corresponds to the principal axis any elliptical contour of the density.

6.5. Implementation of the method and presentation of the results.

With multivariate data the linear structure between variables can be removed by performing an eigenvalue-eigenvector decomposition of the sample covariance matrix S and then defining new variables Z by

$$\underline{Z} = S^{-\frac{1}{2}}.\underline{X} = \underline{U}D^{-\frac{1}{2}}\underline{U}^T\underline{X}. \qquad (6.6.1)$$

The diagonal matrix D contains the nonnegative eigenvalues of S arranged in descending order of magnitude and the columns of U are the corresponding normalised eigenvectors.

We now have $E[\underline{Z}] = \underline{0}$, ($X$ was preliminary recentred to have zero mean) and $Cov(\underline{Z}) = I_p$. Data transformed in this way are called "sphered" data. For data from a distribution which is completely specified by the linear relationships between the variables, such as the Normal, the underlying density of $\underline{Z}$ will be spherically symmetric and hence integrals of cross-sections in all directions will be the same.

The main aim is to find non-linear effects so in practice, as a first step, the linear structure should be removed from a set of data using (6.6.1). Directions of high multivariate density for the $\underline{Z}$ data will then correspond to directions containing non-linear effects in the original $\underline{X}$ data. The transformation (6.6.1) is 1-1 so these solutions in $\underline{Z}$ co-ordinates can be transformed back to reference the X-co-ordinates by using

$$\underline{X} = UD^{\frac{1}{2}}U^T\underline{Z} \qquad (6.6.2)$$

Sphering is a commonly used technique in PP analysis but the motivation is usually computational efficiency during the numerical optimisation.

The method finds k directions (in X space) defined by k unit vectors. They are generally non-orthogonal so it is useful to calculate the angles between all possible pairs where the angle between pair d_i and d_j is defined to be:

$$\theta_{ij} = \cos^{-1}(\underline{d}_i^T.\underline{d}_j) \ . \qquad (6.6.3)$$

It is also useful to see how the directions compare with the principal components based on the X data by again calculating the angular separation between each of the directions and principal components.

Cross-sectional profiles of the density can be estimated using nonparametric density estimation with features such as modes indicating clusters and a fairly high density slowly tailing-off indicating long arms in the data. The density heights along a radius can then be used to evaluate the cross-sectional area by using numerical integration. If the data are, or can be, divided into groups then a cross-sectional plot calculated separately for each group indicates which data are contributing to a feature in a

particular direction.

It is also useful to have cross-sectional density estimates in the planes defined by pairs of directions. This can take the form of a contour plot. The directions are generally non-orthogonal so it is necessary to find two orthogonal axes with which to define the plane. Consider then two unit vectors $\underline{d}_i$ and $\underline{d}_j$ defining two directions. It is required to find a third unit vector, $\underline{a}$, which is orthogonal to $\underline{d}_i$ and in the same plane as both $\underline{d}_i$ and $\underline{d}_j$. The vector $\underline{a}$ must therefore satisfy the following conditions:

$$\begin{aligned} \underline{a}^T.\underline{d}_i &= 0 \\ \underline{a} &= \alpha.\underline{d}_1 + \beta.\underline{d}_2 \quad \text{where } \alpha \text{ and } \beta \text{ are consants} \\ \|\underline{a}\|_2 &= 1 \ . \end{aligned} \tag{6.6.4}$$

Using these conditions it is found that $\alpha = \frac{-1}{\tan\gamma}$ and $\beta = \frac{1}{\sin\gamma}$ where γ is the angle between $\underline{d}_i$ and $\underline{d}_j$ i.e.

$$\underline{a} = \frac{-1}{\tan\gamma} \cdot \underline{d}_1 + \frac{1}{\sin\gamma} \cdot \underline{d}_2 \ . \tag{6.6.5}$$

Therefore, by using both linear and planar cross-sections a picture of the main features of the data and its density can be built up. Inspection of the variable loadings for a particular solution will indicate the relative strength of each of the corresponding variables to the observed effect.

Further, in order to reduce the dimensionality of the data to two it can be projected onto the plane defined by two of the directions. This should indicate the relationship between the actual data points and directions and may also result in "interesting" views as a consequence of viewing the data orthogonal to "interesting" high density regions - we will be looking at a configuration of the data giving rise to these regions. Also, using directions of high

density should mean that the resulting plane provides a reasonably good fit to the data because for directions to have high density a large number of data points must have a small Euclidean distance from the defining line. The fit of a projection plane can be compared with that defined by the first two principal components by calculating the sum of squared orthogonal distances from the data points to the plane and comparing it with the minimising value of the principal component solution.

In the next Section this approach to exploring a multivariate dataset will be illustrated through three examples. The first involves simulated bivariate data and will illustrate the relationship between non-linear features in the data and the resulting modes of the induced density on the circle. The second examines a set of data on flea-beetles (Lubischew (1962)) for which principal component and projection pursuit analyses have been carried out and published in the literature. Finally, a dataset concerned with several measurements of deprivation in each of the 56 Scottish local government districts is analysed.

6.6. Examples.

Example 1. Simulated bivariate data.

In order to illustrate the relationship between structure in a dataset and the induced density, which in this bivariate case will be on the unit circle, 200 observations were simulated according to the following scheme:

Number of simulated observations	Distribution from which observations are simulated
60	$N_2\left[(0,2)^T, \begin{bmatrix} 0.40 & 0.15 \\ 0.15 & 0.10 \end{bmatrix}\right]$

$$60 \qquad N_2\left[(-2,-1)^T, \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.30 \end{bmatrix}\right]$$

$$60 \qquad N_2\left[(3,-1)^T, \begin{bmatrix} 0.40 & -0.30 \\ -0.30 & 0.30 \end{bmatrix}\right]$$

$$20 \qquad N_2\left[(0,0)^T, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right]$$

The observations were firstly centred and standardised using the overall mean vector and standard deviations and then sphered as described in Section 6.5. A scatterplot of the transformed data is shown in figure 6.1 which clearly indicates the three main groupings. The observations were then projected onto the unit circle and weights proportional to an estimate of the induced density evaluated using the smoothing parameter $c = 65$ found by log-likelihood cross-validation. Figure 6.2 shows the weights plotted against the corresponding observation's polar co-ordinate. The three large modes indicate three directions which correspond to non-linear features in the standardised data. Using the algorithm described in Section 6.2 three directions $\underline{d}_1$, $\underline{d}_2$ and $\underline{d}_3$ are in fact found with polar co-ordinates 331.8°, 93.8° and 212.7° respectively. These correspond to the modes of figure 6.2. The dashed lines on figure 6.1 indicate that these three directions do indeed pass through each of the main features.

Figures 6.3, 6.4 and 6.5 show density cross-sections along these directions based on a fixed-kernel estimate with a subjectively chosen smoothing parameter of $0.3\,\hat{\sigma}_i$ $(i = 1,2)$ where the $\hat{\sigma}_i$'s are estimates of scale in the two co-ordinate directions. The robust estimate

$$\sigma_i = \text{median}[\,|(x_{ij}-\text{median}(x_{ij})|\,]/0.6745$$

was used (Hogg (1979)).

In each of these cross-sections the directions found correspond to the positive part of the x-axis. The modes indicate clustering of observations while the broader base of the peak for $\underline{d}_3$ points to a spreading out of data in this direction. The smaller modes near the origin for $\underline{d}_1$ and $\underline{d}_2$ indicate the presence of a smaller less well defined cluster.

The standard bivariate normal cross-sections in figures 6.3-6.5 clearly show that these 200 observations are not from a single bivariate normal population. The cross-sectional integrals from the origin for $\underline{d}_1$, $\underline{d}_2$ and $\underline{d}_3$ calculated by numerical integration are 0.384, 0.281 and 0.321 respectively whereas for a standard bivariate normal the equivalent area is 0.216.

Example 2. Flea-beetles.

The genus of flea-beetles, Chaetocnema, includes species which are very difficult to distiguish by visual examination. An entimologist has collected and taken certain physical measurements from male specimens of the three species, concinna Marsh, heikertinger Lubisch and heptapotamica Lubisch, coded 1, 2 and 3 respectively. Tables 4, 5 and 6 of Lubischew (1962) give six particular measurements for 21 beetles from species 1, 31 from species 2 and 22 from species 3. (i.e. 74 cases in total). The six variables are:

X_1 : width in microns on the first joint of the first tarsus.

X_2 : the same for the second joint.

X_3 : the maximal width in microns of the aedeagus in the forepart.

X_4 : the front angle of the aedeagus (1 unit = 7.5 degrees).

X_5 : the maximal width in 0.01mm of the head between the external

edges of the eyes.

X_6 : the aedeagus width from the side in microns.

The data were first mean-corrected and standardised before sphering. They were then projected onto the unit hypersphere and individual weights determined using a log-likelihood cross-validatory smoothing parameter of 1. The mode finding algorithm split the data into five groups containing 25, 18, 20, 7 and 4 observations. The co-ordinates of the modal directions of these five groups were then transformed back to reference the centered, standardised data and are as follows :

Direction :	1	2	3	4	5
	-0.506	0.132	-0.644	0.306	-0.064
	-0.003	0.311	0.023	-0.649	0.218
	-0.308	0.179	0.154	-0.177	0.792
	-0.484	0.750	-0.149	-0.475	0.018
	0.455	0.346	-0.621	0.111	-0.391
	-0.456	0.413	0.392	-0.464	-0.409
Cross-sectional integral	0.018	0.019	0.017	0.010	0.008

(Note that the cross-sectional integrals are calculated in the positive direction from the origin only).

The angles in degrees between the directions are:

Direction :	1	2	3	4
2	112.6			
3	155.0	89.5		
4	45.3	134.7	124.1	
5	106.1	95.1	75.4	99.4

The closest are $\underline{d}_1$ and $-\underline{d}_3$ with an angle of only 25.0 between them. $\underline{d}_1$, $\underline{d}_2$ and $\underline{d}_3$ are also not widely separated from $\underline{d}_4$ but all other pairs are within about 20.0 of being orthogonal to each other.

Linear cross-sections of the multivariate density for $\underline{d}_1$, $\underline{d}_2$ and $\underline{d}_3$, the three with the highest cross-sectional integrals, are illustrated in figures 6.6-6.8. A fixed kernel density estimator was employed with a subjectively chosen smoothing parameter of $0.5.\hat{\sigma}_i$ $(i = 1,\ldots,6)$ where the $\hat{\sigma}_i$'s are as described in Example 1. The distinct mode in each of the positive halves of these plots clearly indicates that these three directions pass through three clusters in the data. The smaller modes in the negative halves of the plots for $\underline{d}_1$ and $\underline{d}_3$ are due to those data contributing to the larger modes in $\underline{d}_3$ and $\underline{d}_1$ respectively evidenced by the small angle of 25.0 between $\underline{d}_1$ and $-\underline{d}_3$. The density cross-sections for $\underline{d}_4$ and $\underline{d}_5$ also contain modes but their heights are much lower than for the first three.

Planar cross-sections of the six-dimensional density were also constructed using pairs of directions to define a plane. That based on $\underline{d}_2$ and $\underline{d}_3$, using the same smoothing parameters as above is illustrated in figure 6.9 and clearly shows the six-dimensional density to have three modes in this plane. This trimodal feature was also evident in other planar cross-sections such as when using $\underline{d}_1$ and $\underline{d}_2$.

To see which species are contributing to the modes in figure 6.9 the data were projected onto this plane and the points labelled 1, 2 and 3 according to species - see figure 6.10. The data divides up into three distinct groups according to species with $\underline{d}_2$ and $\underline{d}_3$ corresponding to species 1 and 3 respectively. Projection onto the

plane containing $\underline{d}_1$ and $\underline{d}_2$ (not illustrated) indicates that $\underline{d}_1$ corresponds to species 2. The sum of squared perpendicular distances from the data to the plane containing $(\underline{d}_2, \underline{d}_3)$ is 123.0 which compares with the optimal value for the plane defined by the first two principal components of 87.8 and indicates a reasonable fit.

The data projected onto the plane defined by the first two principal components and labelled by species is shown in figure 6.11. Again the division into three groupings according to species is evident but the distinction between species is perhaps not as strong as in figure 6.10.

The following table gives the angles between the first three directions and first two principal components:

		PC 1	PC 2
	1	155.3	110.9
Direction:	2	54.9	125.8
	3	39.8	51.1

The strongest similarity is between $\underline{d}_1$ and -PC1. On the other hand $\underline{d}_1$ and $\underline{d}_2$ are well separated from PC2.

This data has also been analysed using projection pursuit by Jones and Sibson (1987). They also provide a planar solution which divides the data into the three species but at the expense of considerably more computing effort than that involved in finding the above directions.

6.6.3. Scottish deprivation data.

There are 56 local government districts in Scotland. For each of

these districts the values of the following seven variables have been recorded:

X_1 : Standardised Mortality Ratio. (SMR)

X_2 : Persons in private households with economically active (i.e. in work, seeking work or temporarily sick) head in social class 1 (higher managerial and professional) as a proportion of all persons in private households with economically active heads. (S1)

X_3 : Same as in social class I but for social class V (unskilled manual). (S5)

X_4 : Proportion of persons in private households living in overcrowded accommodation (i.e. > 1.5 persons per room). (OV)

X_5 : Proportion of persons in private households with no car. (NC)

X_6 : Proportion of children (under 16) in private households with only one adult. (PC)

X_7 : Proportion of economically active males seeking work. (UN)

The SMR for a district is defined as the ratio of observed to expected deaths in the district. The expected number of deaths is obtained by applying the national age-sex-specific mortality rates to the population structure of the district.

The data used are given in Amfoh (1988). He obtained the SMR data from the 1981 annual report of the Registrar-General of Scotland while that for the other six variables was extracted from the 1981 Scottish census small area statistics. All the variables are indicators of deprivation in a district but, in contrast to the others, small values of S1 denote deprivation and large values affluence.

Amfoh (1988) provides boxplots and histograms for each of the variables. Those for SMR are fairly symmetric but for the other six they show varying degrees of skewness to the right.

The data were first mean-corrected before sphering. Application of the algorithm of Section 6.2 to the sphered data using $c = 7$, found by log-likelihood cross-validation, determined five directions based on 12, 12, 17, 8 and 7 observations. The density profiles for these directions (not illustrated) indicate clustering about the origin as well as two smaller but marked clusters away from the origin in directions 2 and 5. These profiles also clearly indicate the non-normality of the data.

These directions were then transformed to back reference the mean-corrected standardised data. The unit vectors defining each of them are as follows :

Direction :	1	2	3	4	5
	0.625	0.405	-0.350	-0.474	-0.361
	-0.220	-0.051	-0.401	0.062	0.529
	-0.197	0.377	-0.289	-0.074	-0.461
	-0.347	0.592	-0.420	-0.318	-0.135
	0.303	0.460	-0.425	0.381	-0.295
	0.087	0.133	0.293	-0.455	-0.475
	0.550	0.334	-0.441	-0.559	-0.217
Cross-sectional integral.	0.038	0.029	0.060	0.024	0.039

The following table gives all the pairwise angles in degrees between the directions:

Direction :	1	2	3	4
2	71.4			
3	106.0	141.3		
4	130.3	146.5	55.2	
5	117.0	134.2	79.4	43.1

They are all reasonably well separated. The closest pairs are $(\underline{d}_2, -\underline{d}_3)$ and $(\underline{d}_2, -\underline{d}_4)$ with angles of 38.7 and 33.5 between the pairs respectively.

Each of the one-dimensional profiles is quite similar in shape to the corresponding one for the sphered data. Those for $\underline{d}_2$, $\underline{d}_3$ and $\underline{d}_5$ are illustrated in figures 6.12-6.14. Each of these shows a large mode near the origin indicating clustering there while $\underline{d}_2$ and $\underline{d}_5$ both have smaller modes away from the origin as well. The profiles for $\underline{d}_1$ and $\underline{d}_4$, not included, are both unimodal with modal points to the right of the origin.

Examination of the variable loadings enables menaingful interpretations to be given to some of the directions. For example, $\underline{d}_2$ can be interpreted as an index of SMR, S5, OV, NC and UN (positive coefficients) while $\underline{d}5$ is a contrast between S1 (positive) and SMR, S5, NC and PC (negative). Also $\underline{d}_4$ is an index of a group of the deprivation variables with negative coefficients while $\underline{d}_1$ is a contrast between SMR and UN (positive) and OV (negative). In arriving at these interpretations coefficients with absolute values less than half the largest absolute coefficient have been ignored.

The planar cross-section containing $(\underline{d}_2, \underline{d}_5)$, figure 6.15, shows the full seven-dimensional density to be trimodal in this plane with the largest mode near the origin and the smallest in direction 5.

That containing $(\underline{d}_4, \underline{d}_5)$, figure 6.16, again shows modes in $\underline{d}_5$ and around the centre but also indicate two "arms" in the density. Positive $\underline{d}_4$ and $\underline{d}_5$ are both directions of increasing affluence. For the plane containing $(\underline{d}_1, \underline{d}_2)$, figure 6.17, there is a separate mode in $\underline{d}_2$ and a region of high density near the centre containing one large peak and two smaller ones.

The data were then projected onto each of these three planes. That for the plane containing $(\underline{d}_2, \underline{d}_5)$, figure 6.18, shows most of the points to be around the origin but with some spread along positive $\underline{d}_2$ and $\underline{d}_5$. The point with the largest value along $\underline{d}_2$ and smallest along $\underline{d}_5$ corresponds to Glasgow whereas at the other extreme Bearsden and Eastwood have the largest values on $\underline{d}_5$ and smallest on $\underline{d}_2$. In fact, examination of the original data reveals that Glasgow has the highest values for SMR, S5, OV, NC, PC and UN out of all the districts together with a small, well below average, value for S1. In contrast Bearsden and Eastwood have between them the lowest values for SMR, S5, OV and PC, very small values for NC and UN and by far the largest values for S1. These outliers are again clearly shown in the projection onto the plane containing $(\underline{d}_4, \underline{d}_5)$, figure 6.19, and also that containing $(\underline{d}_1, \underline{d}_2)$, figure 6.20.

All these plots indicate that the data can be divided up into three groupings. The largest is around the origin which also contains some finer additional structure. The other two in directions 2 and 5 are due to more deprived and affluent districts respectively.

A principal component analysis was also carried out on this dataset. The first two principal components together explain 75.1% of the total variation and are as follows :

	PC1	PC2
	-0.370	0.220
	0.261	0.697
	-0.383	-0.084
	-0.415	0.364
	-0.445	0.126
	-0.295	-0.526
	-0.437	0.184
% of total variation explained.	61.70	13.41

PC1 can be interpreted as a contrast between positive S1 (i.e. affluence) and all the other variables (i.e. deprivation). It is difficult to give PC2 a meaningful interpretation which is also the case for the other principal components.

The angles in degrees between the five directions and first two principal components are as follows:

		PC1	PC2
	1	118.0	91.8
Direction:	2	160.6	73.3
	3	52.9	140.7
	4	25.8	94.7
	5	29.5	63.1

$\underline{d}_2$ has an angle of only 19.4° with -PC1 while $\underline{d}_4$ and $\underline{d}_5$ are both less than 30.0° from +PC1. Otherwise, each of the directions is well separated from PC1 and PC2. This is also the case for the other principal components with many of the pairwise comparisons indicating near orthogonality.

A plot of the 7-dimensional density in the plane of (PC1,PC2), figure 6.21, again shows the contours to be concentrated about the origin with a much lower separate contour in the -PC1 direction due to the more deprived areas.

Projecting the data onto the plane defined by PC1 and PC2, figure 6.22, again shows Bearsden and Eastwood to be extreme values for +PC1 and Glasgow to be extreme for -PC1. However, the relative scores amongst the data in terms of PC2 are not helpful. Taking other pairs of principal components to define a plane reveals only a single mode near the origin in the planar cross-sections and little of the relationships between the districts in the projections.

The sum of squared perpendicular distances from the data to the (PC1,PC2) plane is 95.8. For the planes containing $(\underline{d}_1, \underline{d}_2)$, $(\underline{d}_2, \underline{d}_5)$ and $(\underline{d}_4, \underline{d}_5)$ the sum of squares are 138.1, 104.1 and 126.0 respectively. The differences in the fits as compared with the optimal principal component solution are not great but more of the structure in the data is revealed.

Figure 6.1. Scatterplot of the sphered simulated data with the three directions of high bivariate density ($\underline{d}_i$, i = 1,2,3) indicated.

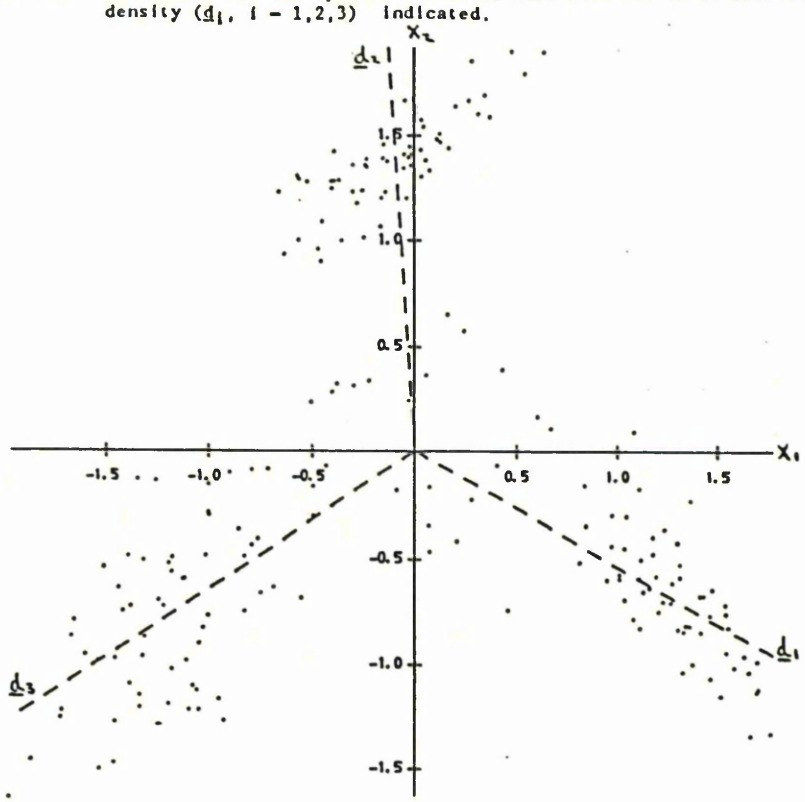


Figure 6.2. Weights proportional to an estimate of the induced circular density (c = 65) for the sphered simulated data.

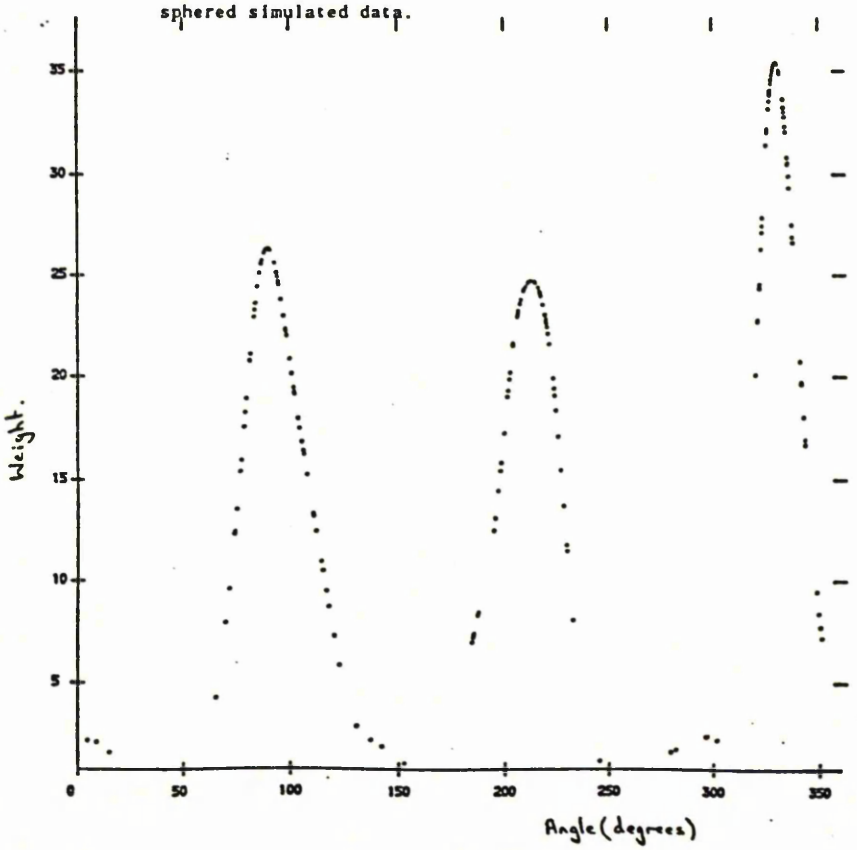


Figure 6.3. Fixed kernel estimate, $h_i = 0.3\,\hat{\sigma}_i$ $(i = 1,2)$, for a cross-section along $\underline{d}_1$ and an $N_2(\underline{0}, I_2)$ density profile for the sphered simulated data.

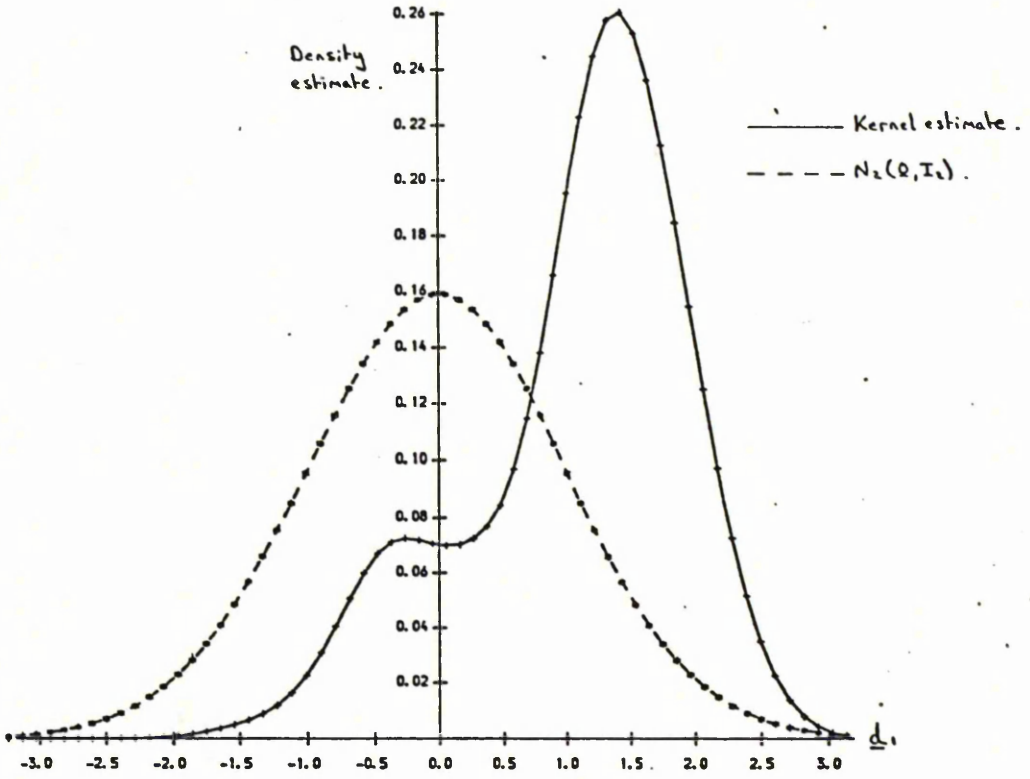


Figure 6.4. Fixed kernel estimate, $h_i = 0.3\,\hat{\sigma}_i$ $(i = 1,2)$, for a cross-section along $\underline{d}_2$ and an $N_2(\underline{0}, I_2)$ density profile for the sphered simulated data.

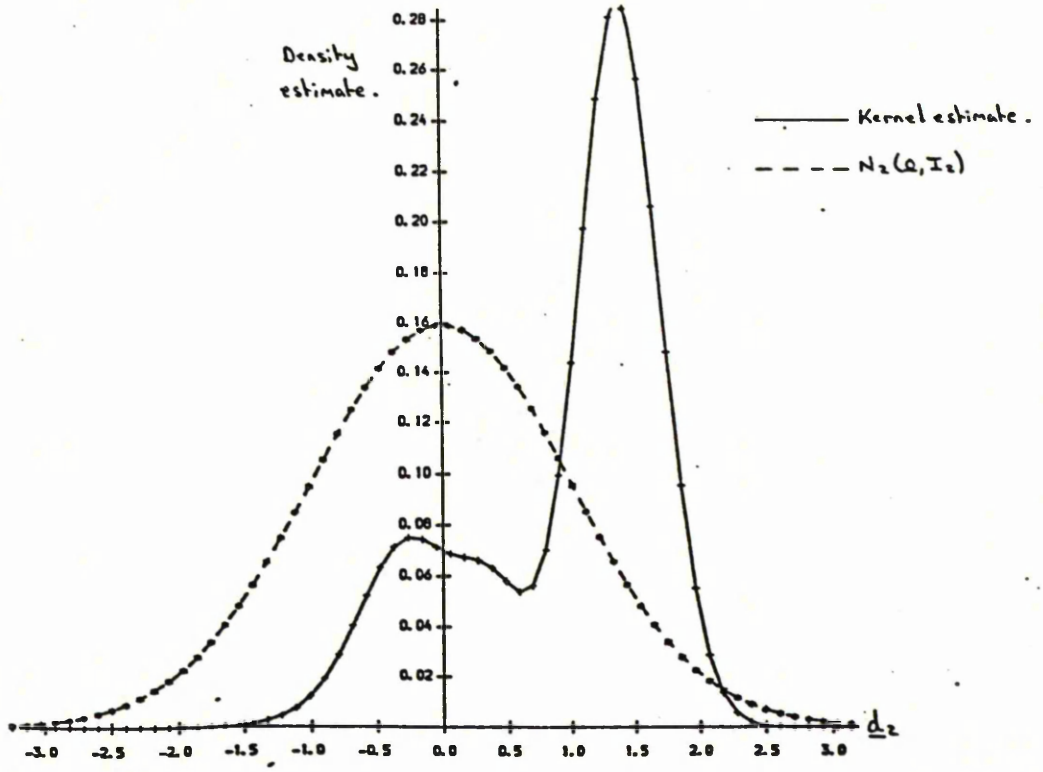


Figure 6.5. Fixed kernel estimate, $h_l = 0.3\,\hat{\sigma}_l$ $(l = 1,2)$, for a cross-section along $\underline{d}_3$ and an $N_2(\underline{0}, I_2)$ density profile for the sphered simulated data.

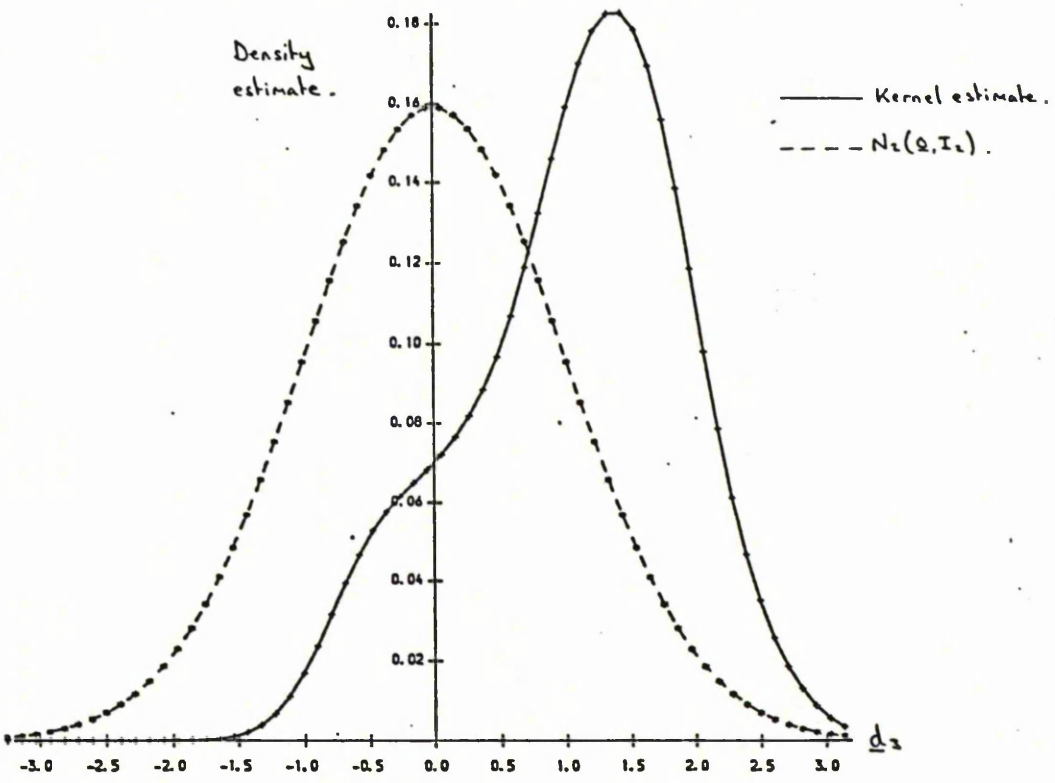


Figure 6.6. Fixed kernel estimate, $h_l = 0.5\,\hat{\sigma}_l$ $(l = 1,\ldots,6)$, for a cross-section along $\underline{d}_1$ for the standardised flea-beetle data.

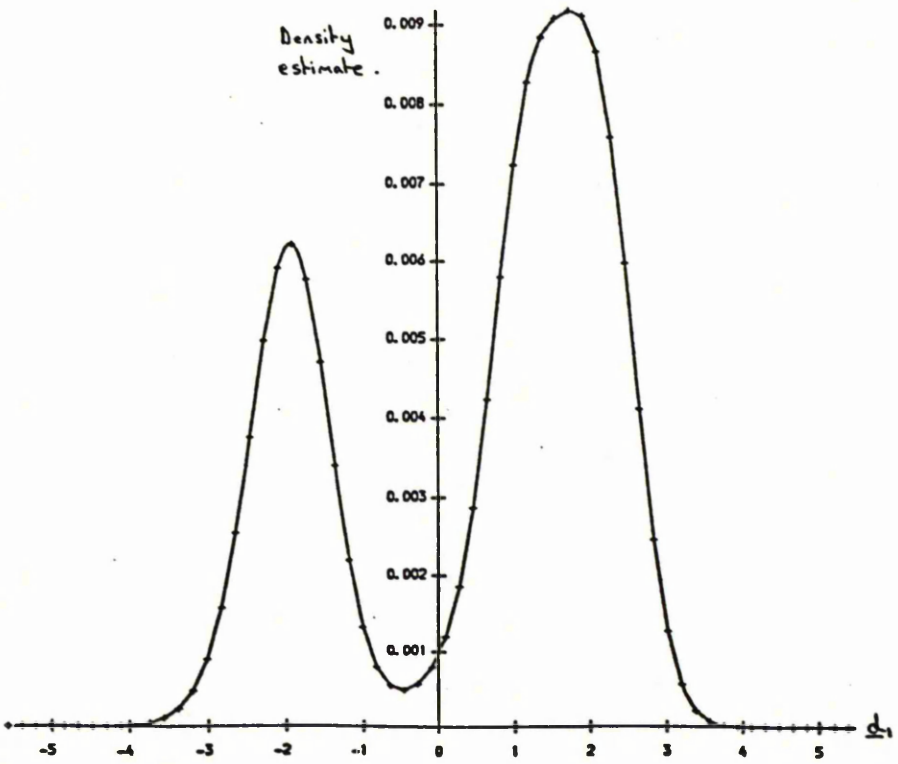


Figure 6.7. Fixed kernel estimate, $h_i = 0.5\ \hat{\sigma}_i$ $(i = 1,\ldots,6)$, for a cross-section along $\underline{d}_2$ for the standardised flea-beetle data.

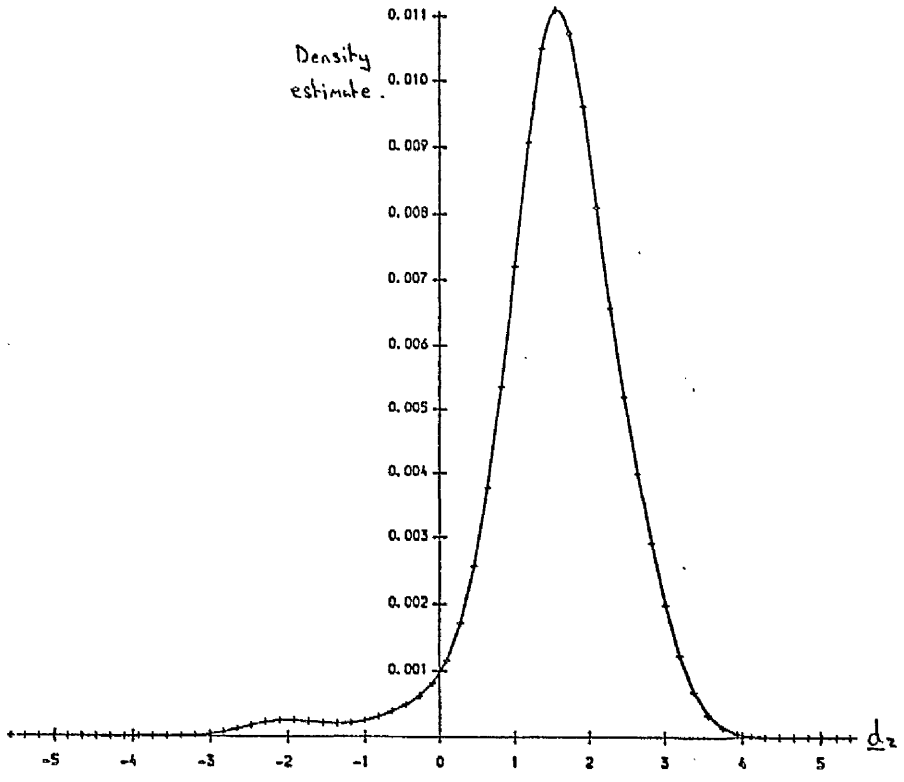


Figure 6.8. Fixed kernel estimate, $h_i = 0.5\ \hat{\sigma}_i$ $(i = 1,\ldots,6)$, for a cross-section along $\underline{d}_3$ for the standardised flea-beetle data.

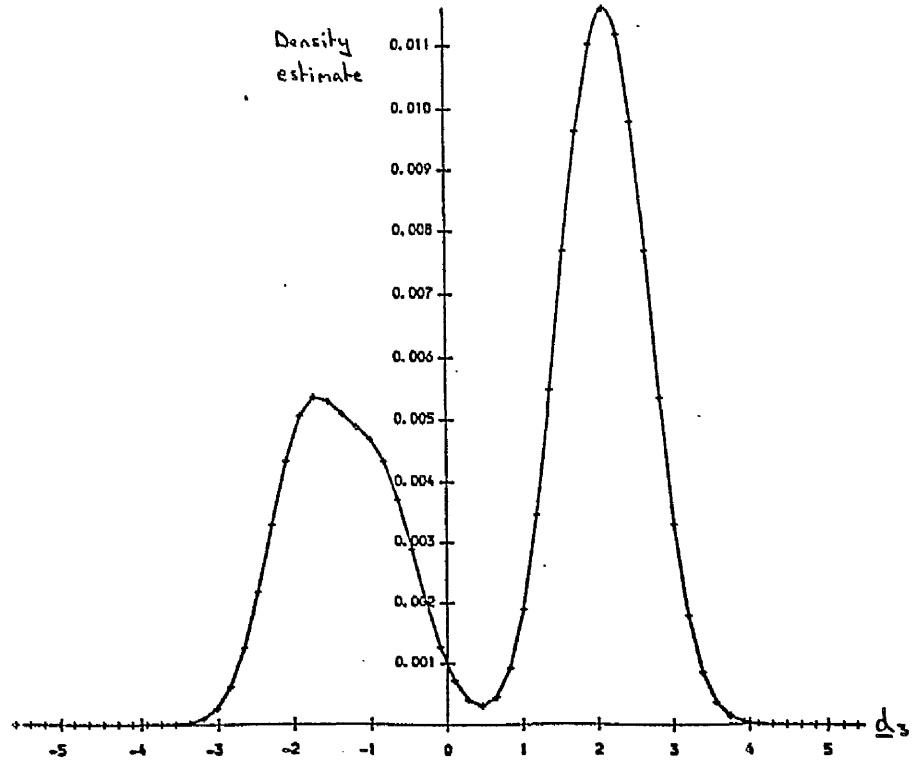


Figure 6.9. Fixed kernel estimate, $h_l = 0.5\,\hat{\sigma}_l$ $(l = 1,\ldots,6)$, for a cross-section in the plane containing $\underline{d}_2$ and $\underline{d}_3$ for the standardised flea-beetle data.

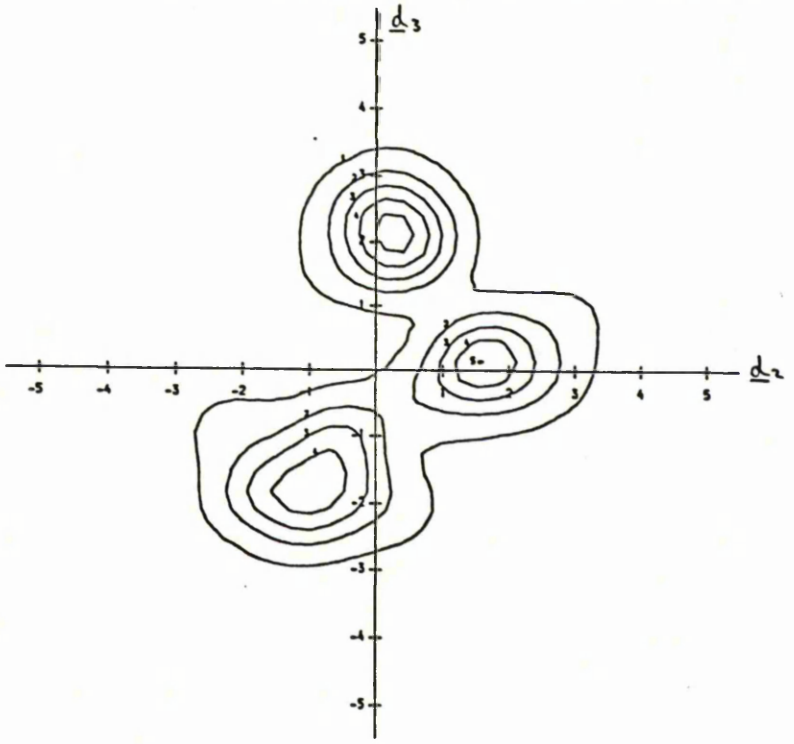


Figure 6.10. The standardised flea-beetle data, with data points labelled by species, projected onto the plane containing $\underline{d}_2$ and $\underline{d}_3$.

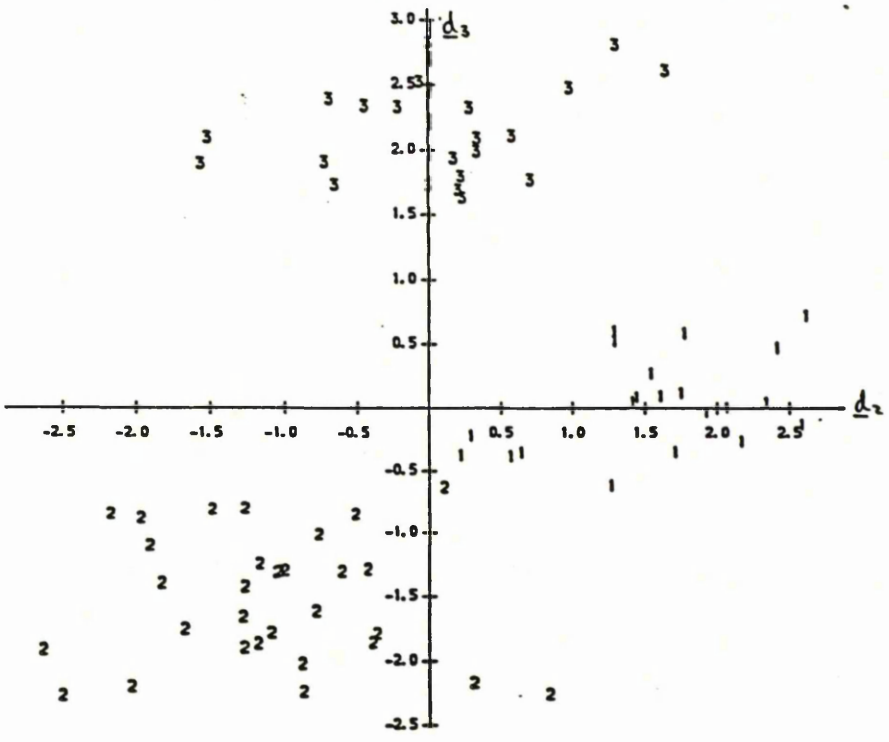


Figure 6.11. The standardised flea-beetle data, with data points labelled by species, projected onto the plane defined by the first two principal components.

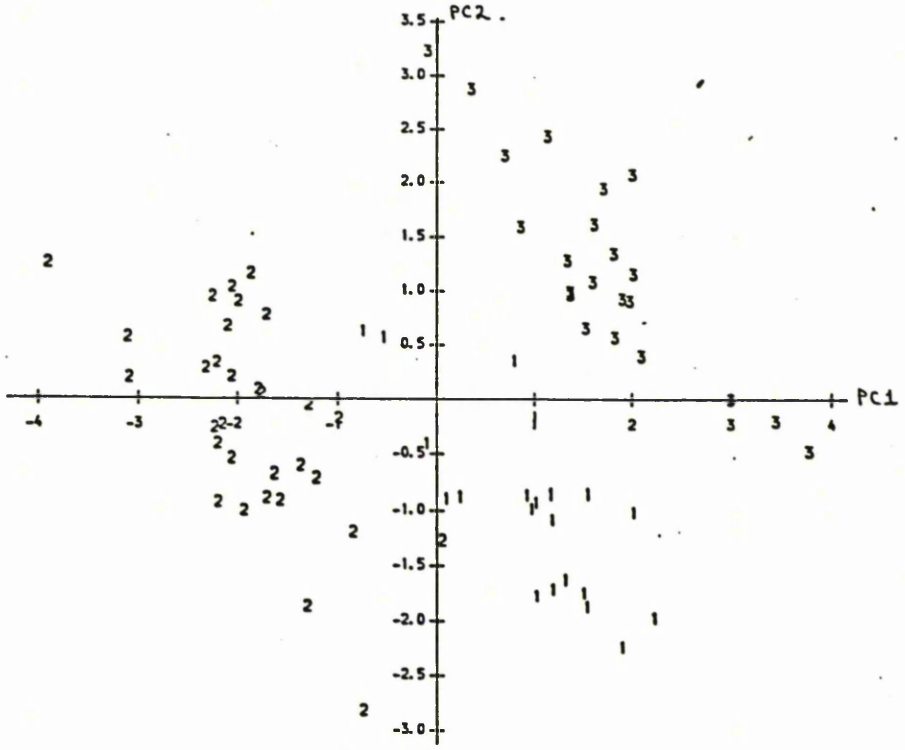


Figure 6.12. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section along $\underline{d}_2$ for the standardised Scottish deprivation data.

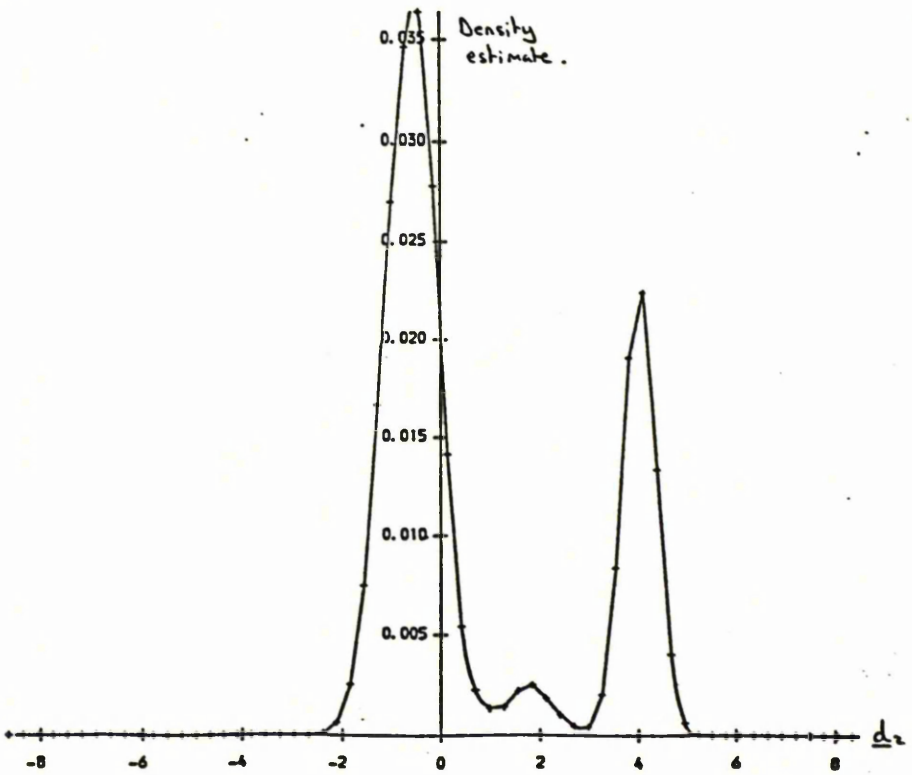


Figure 6.13. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section along $\underline{d}_3$ for the standardised Scottish deprivation data.

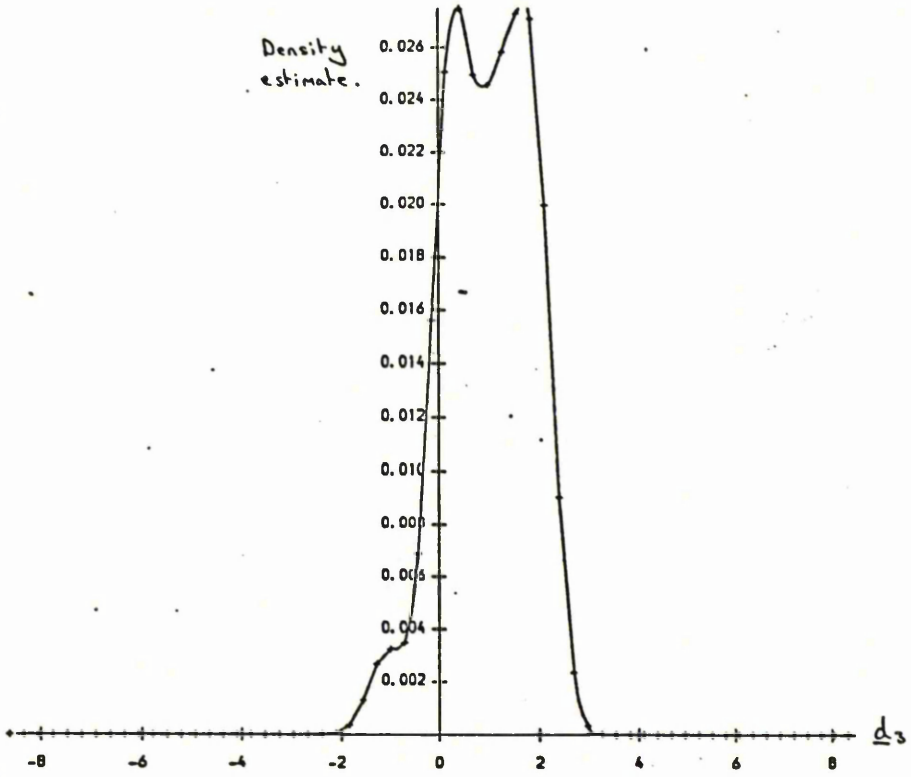


Figure 6.14. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section along $\underline{d}_5$ for the standardised Scottish deprivation data.

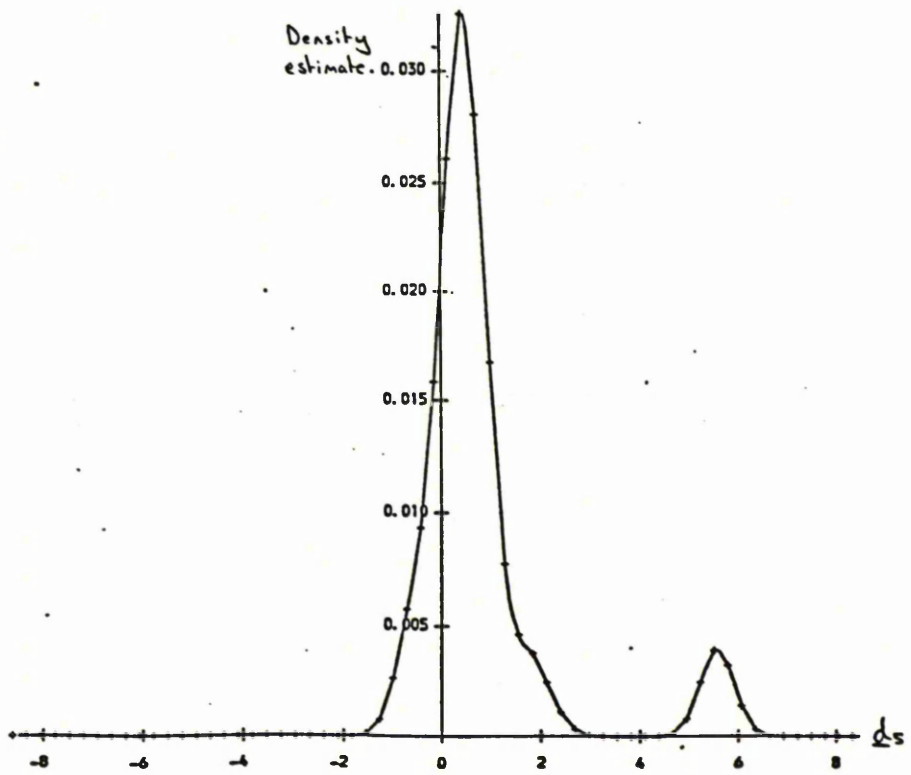


Figure 6.15. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section in the plane containing $\underline{d}_2$ and $\underline{d}_5$ for the standardised Scottish deprivation data.

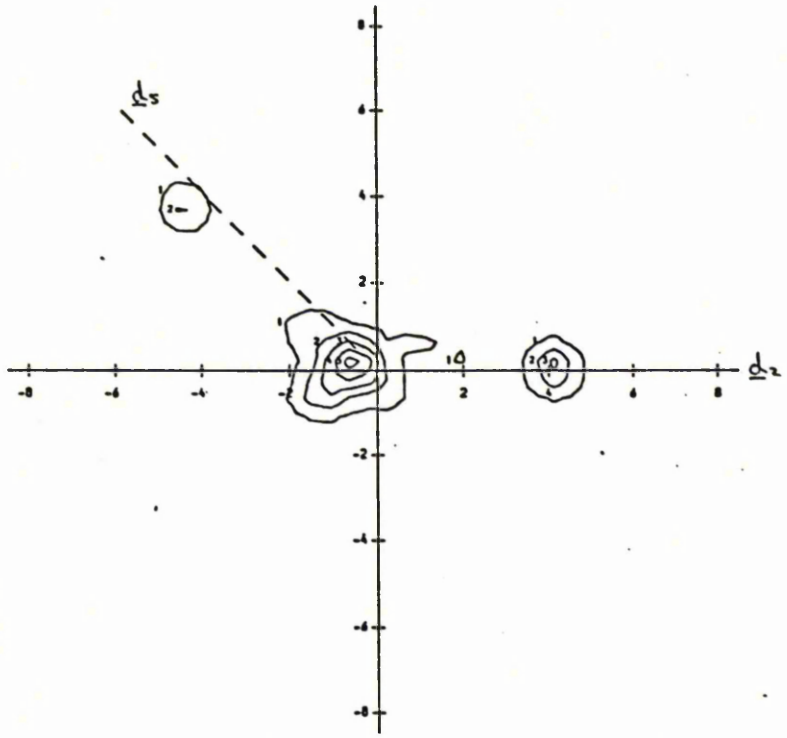


Figure 6.16. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section in the plane containing $\underline{d}_4$ and $\underline{d}_5$ for the standardised Scottish deprivation data.

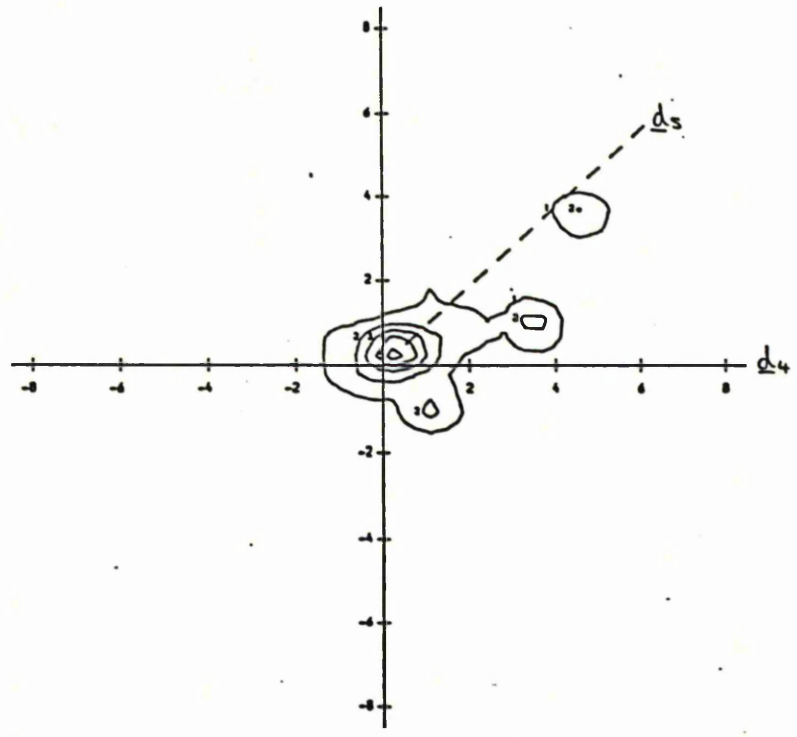


Figure 6.17. Fixed kernel estimate, $h_i = 0.5\,\hat{\sigma}_i$ $(i = 1,\ldots,7)$, for a cross-section in the plane containing $\underline{d}_1$ and $\underline{d}_2$ for the standardised Scottish deprivation data.

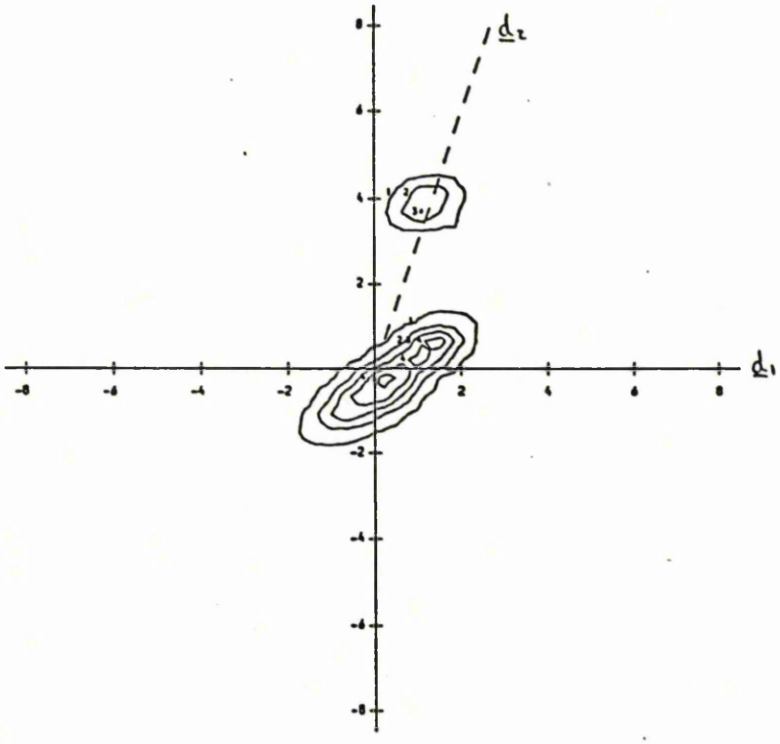


Figure 6.18. The standardised Scottish deprivation data projected onto the plane containing $\underline{d}_2$ and $\underline{d}_5$.

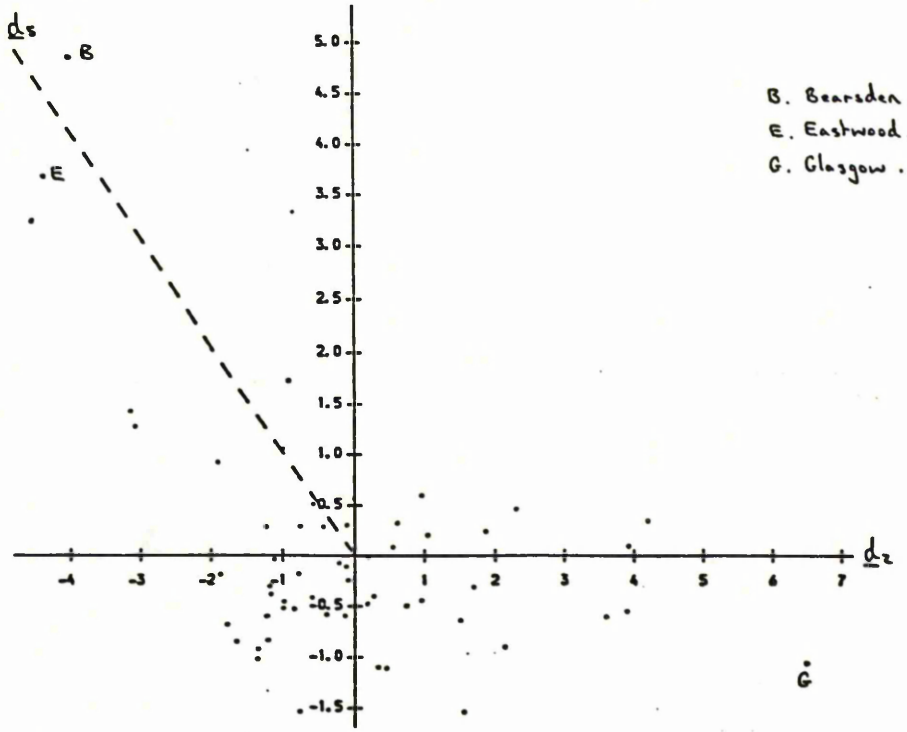


Figure 6.19. The standardised Scottish deprivation data projected onto the plane containing $\underline{d}_4$ and $\underline{d}_5$.

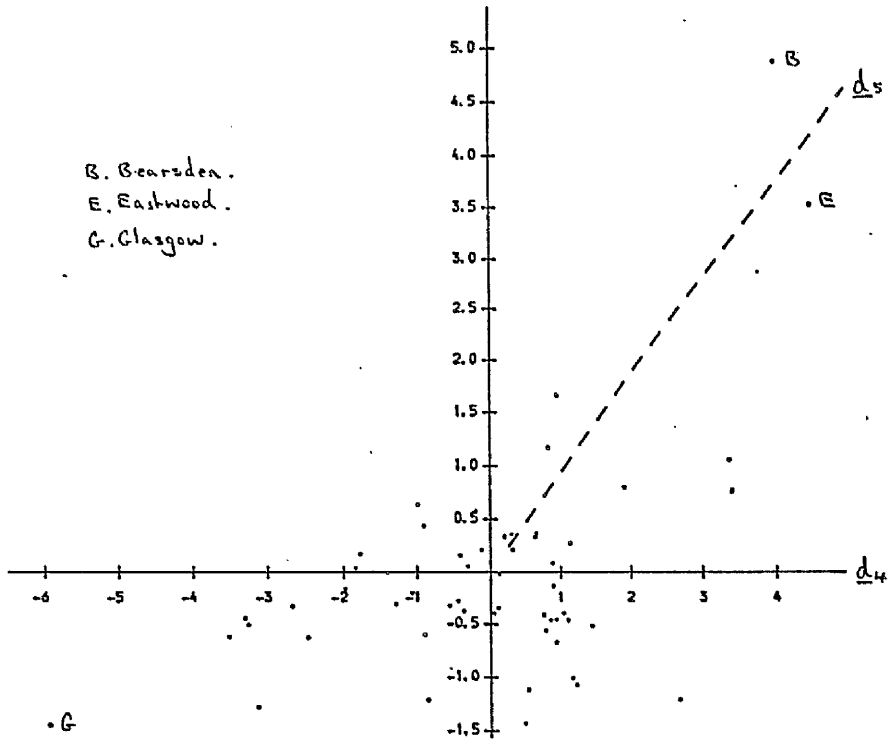


Figure 6.20. The standardised Scottish deprivation data projected onto the plane containing $\underline{d}_1$ and $\underline{d}_2$.

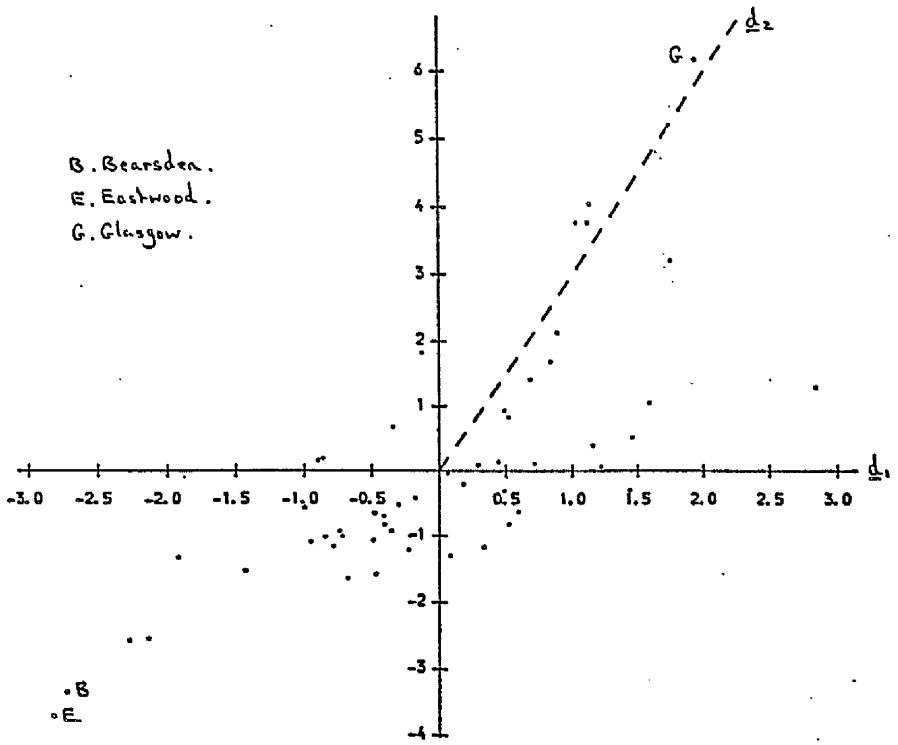


Figure 6.21. Fixed kernel estimate, $h_l = 0.5\,\hat{\sigma}_l$ $(l = 1,\ldots,7)$, for a cross-section in the plane defined by PC1 and PC2 for the standardised Scottish deprivation data.

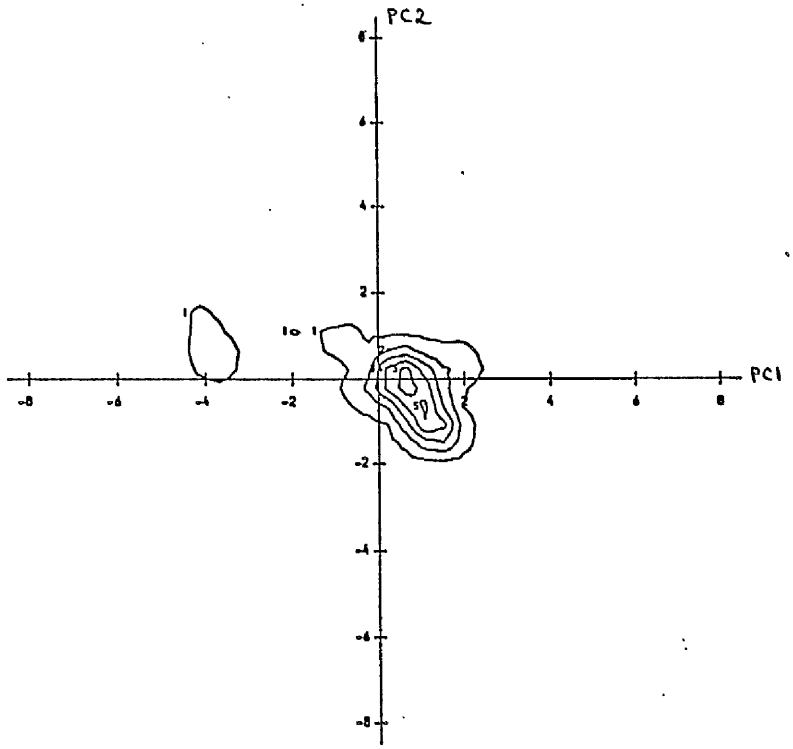
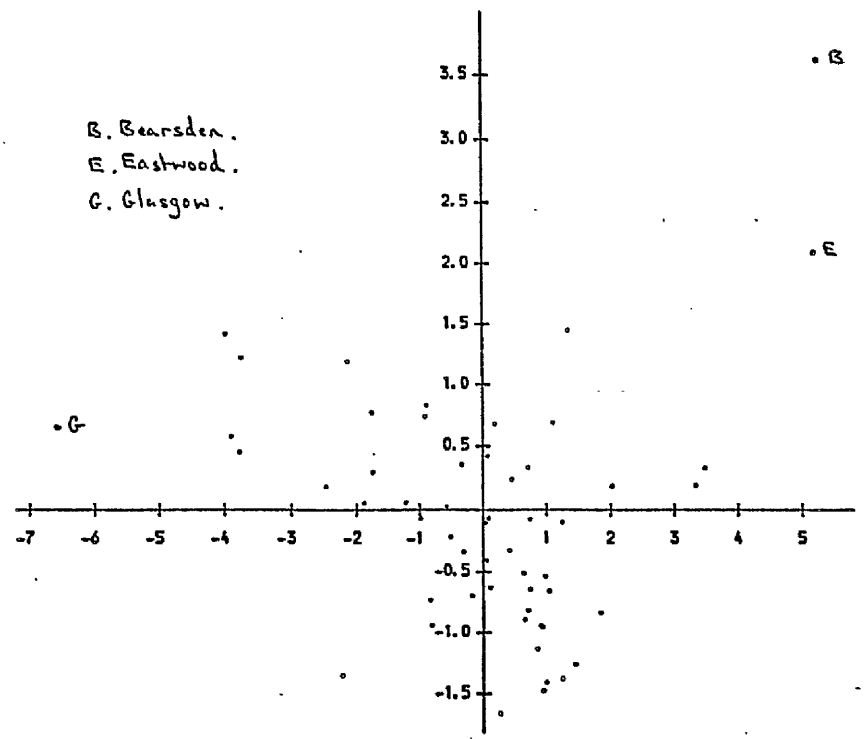


Figure 6.22. The standardised Scottish deprivation data projected onto the plane defined by the first two principal components.



Chapter 7. Assessing Logistic Regression Models.

7.1. Introduction.

Suppose that we have N sets of observations in the form $(Y_1,n_1,x_1),\ldots,(Y_N,n_N,x_N)$. For each individual or object in the study we have observed the outcome of a binary response variable which is coded 1 for "success" and 0 otherwise. Y_i denotes the number of individuals having positive response (i.e. equal to one) in the i^{th} set which comprises n_i individuals each with a common covariate value x_i and probability $p(x_i)$ (also to be denoted by p_i) of a positive response. If the outcomes for each of the individuals in set i are independent then $Y_i \sim B_i(n_i,p_i)$ with $E[Y_i] = n_i.p_i = \mu_i$. Ungrouped data comprises a special case and have $n_i = 1$ for $i = 1,\ldots,N$.

The aim is to investigate the dependence of a positive reponse on the measured covariate which may be either categorical or on a continuous scale. Such analyses have many areas of application such as in medical, economic and educational studies. For example, we may be interested in the probability that a patient with a particular disease will survive for at least five years following surgery when he or she is x years old at the time of the operation.

A common method of modelling such dependencies is to fit linear functions of parameters to some transformation of the probabilities. One of the most widely used techniques in this respect is logistic regression. The logistic model postulates that

$$\log(p_i/(1-p_i)) = \alpha + \beta x_i \qquad (7.1.1)$$

or $$p_i = p(x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \qquad (7.1.2)$$

where α and β are unknown parameters to be estimated from the data and $p(x_i)$ denotes the dependence of p_i on x_i. In generalised linear model terminology the right hand side of (7.1.1) is called the "linear predictor" and the left hand side the "link function" as it links the expected response for an individual in the ith set to the linear predictor for that individual.

Under the above conditions the likelihood function for the observed data is

$$L = \prod_{i=1}^{N} \binom{n_i}{x_i} p_i^{y_i} (1-p_i)^{n_i-y_i} \qquad (7.1.3)$$

so that the log likelihood is

$$\log(L) = \ell = \text{const.} + \sum_{i=1}^{N} [y_i \,.\, \log(p_i) + (n_i-y_i).\log(1-p_i)] \qquad (7.1.4)$$

$$= \text{const.} + \sum_{i=1}^{N} y_i \,.\, (\alpha+\beta x_i) - \sum_{i=1}^{N} n_i \log(1+e^{\alpha+\beta x_i}) \qquad (7.1.5)$$

by using (7.1.2).

To obtain estimates of α and β the method of maximum likelihood is used so that ℓ is partially differentiated with respect to α and β and the results set equal to zero. This gives a system of non-linear equations in α and β which need to be solved iteratively. Second derivatives can easily be computed so the Newton-Raphson method can be employed. McCullagh and Nelder (1983) show that this can be expressed as iteratively reweighted least squares i.e.

$$\underline{\theta}^{(t)} = (X^T W X)^{-1} . (X^T W \underline{Z}) \qquad (7.1.6)$$

where,

$$\underline{\theta} = (\alpha, \beta)^T,$$

$$X^T = \begin{bmatrix} 1, \ldots, 1 \\ x_1, \ldots, x_N \end{bmatrix},$$

$$W = \text{diag}\ (n_i p_i (1-p_i)),$$

$$\underline{Z} = X\underline{\theta} + W^{-1}\underline{s}$$

and $\underline{s}$ has elements $s_i = y_i - n_i . p_i$, $i = 1, \ldots, N$.

The right hand side of (7.1.6) is evaluated at $\underline{\theta}^{(t-1)}$ (i.e. using the estimates of α and β calculated at the $(t-1)^{th}$ iteration). At convergence the elements of $\underline{\theta}$ correspond to the maximum likelihood estimates.

The theory readily extends to the values of p covariates being available for each individual so that the model becomes

$$\log(p_i/(1-p_i)) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p \qquad (7.1.7)$$

or

$$p(\underline{x}) = \exp(\underline{\theta}^T \underline{x})/(1+\exp(\underline{\theta}^T \underline{x})) \qquad (7.1.8)$$

where $\underline{\theta}$ is now $(\alpha, \beta_1, \ldots, \beta_p)^T$. The x_i's can be a mixture of both categorical and continuous variables.

The usual asymptotic results associated with maximum likelihood estimators apply so that

$$\underline{\theta} \sim N_{p+1}\ (\underline{\theta},\ (X^T W X)^{-1}). \qquad (7.1.9)$$

In this context $(X^T W X)^{-1}$ is equivalent to the inverse of the Fisher information matrix evaluated at $\hat{\theta}$.

Once a logistic model has been fitted it is then useful to then

assess how well it actually does fit the data. Such checks can be difficult to do visually, either by examining the fitted values or by using certain graphical techniques, and therefore some measure of goodness-of-fit is desirable. Goodness-of-fit tests may be effective in detecting problems but they will not necessarily indicate their nature. They do, however, help to prevent inappropriate conclusions being made as a consequence of a poorly fitting model. In Section 2 some commonly used goodness-of-fit measures are reviewed. A number of authors have suggested estimating some or all of the covariate effects non-parametrically and incorporating these into the logistic model. Some of these approaches are reviewed in Section 3 while in Section 4 the psuedo-likelihood ratio test, which aims to compare the logistic model with a nonparametric alternative, is described in detail. Lack of fit may result from omitting important covariates or by incorrectly specifying the functional form of a covariate effect. In Section 5, the use of partial residuals for establishing the correct functional form is investigated.

7.2. Some measures of goodness-of-fit.

In generalised linear models a common measure of discrepancy between the data $\underline{y}$ and fitted values $\hat{\underline{\mu}}$ is the scaled deviance which is defined to be

$$\frac{D(\underline{y}, \hat{\underline{\mu}})}{\varphi} = -2.(\ell - \ell_0) \qquad (7.2.1)$$

where ℓ is the log-likelihood evaluated at the MLE's $\hat{\underline{\theta}}$, ℓ_0 is the maximum achievable log-likelihood for a saturated model with the number of parameters equal to the number of observations and the scaling factor φ is the dispersion parameter which equals 1 for

binomial models.

For binomial data (7.2.1) takes the form

$$D(\underline{y},\hat{\underline{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \left[\frac{y_i}{\hat{\mu}_i}\right] + (n_i - y_i) \log \left[\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right]\right\} . \tag{7.2.2}$$

If the n_i's are large then the distribution of (7.2.2) is approximately $\chi^2(n-p)$. The asymptotics require that the model remains fixed as the number of observations increases so that the $\mu_i \longrightarrow \infty$. While this is the case for binomial data, for binary observations the μ_i will remain small and the chi-squared approximation will be invalid. Williams (1983) shows that for logistic models of binary data the deviance can be written as

$$D(y,\hat{\underline{\mu}}) = -2 \sum_{i=1}^{N} \left\{ \hat{\mu}_i \log (\hat{\mu}_i) + (1-\hat{\mu}_i)\log(1-\hat{\mu}_i)\right\} \tag{7.2.3}$$

which is a function of the fitted values only and is therefore uniformative about the goodness-of-fit. A model for binary data can still be tested against a non-saturated alternative by comparing the differences in deviances with a χ^2 in the usual way but the actual accuracy of this χ^2 approximation is unknown.

Another widely used measure is Pearsons χ^2 statistic defined by

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} . \tag{7.2.4}$$

For large n_i's this also has an $\chi^2(n-p)$ distribution.

McCullagh (1986) argues that for the deviance and Pearson's χ^2 the appropriate reference distribution is conditional on the sufficient statistic, S, for the unknown $\underline{\theta}$ rather than considering

the marginal distributions. For binomial and binary data $S = X^T\underline{Y}$ of which $\underline{\theta}$ is a 1-1 function. For binary data this conditional distribution of the deviance is a function of S and so contains no information regarding lack of fit whereas for binomial data it is asymptotically normal with the moments given in the paper. The unconditional mean and variance of χ^2 are derived and it is also shown that χ^2 and S are independent to first order in n.

Hosmer and Lemeshow (1980, 1982) describe and discuss a number of statistics for assessing goodness-of-fit. To calculate their recommended statistic the estimated probabilities $\{\hat{p}(\underline{x}_i),\ i = 1,\ldots,N\}$ are firstly ranked and then grouped into deciles so that the first decile contains the smallest N/10 values of $\hat{p}(x_i)$, etc. If N is not a multiple of 10 then an extra observation is assigned to an appropriate number, (N-[n/10]), of the deciles containing the largest $\hat{p}(x_i)$'s. Either fewer or more than 10 groups can be used provided the number is greater than p+1 but most applications do use 10. Their test statistic compares the observed and estimated frequencies in each decile and is given by

$$C = \sum_{k=0}^{1} \sum_{\ell=1}^{10} \frac{(o_{k\ell}-e_{k\ell})^2}{e_{k\ell}} \qquad (7.2.5)$$

where,

$$o_{0\ell} = \sum_{i \in D_\ell} (n_i - y_i),$$

$$o_{1\ell} = \sum_{i \in D_\ell} y_i,$$

$$e_{0\ell} = \sum_{i \in D_\ell} n_i(1-\hat{p}(x_i)),$$

$$e_{1\ell} = \sum_{i \in D_\ell} n_i\hat{p}(x_i),$$

and D_ℓ denotes the set of individuals in the ℓth decile. They show in their 1980 paper via computer simulations that provided $g > p+1$ then under the null hypothesis C is approximately distributed as a $\chi^2(g-2)$ where g is the number of groups used.

7.3. Incorporating smooth functions of the covariates into the model.

Hastie and Tibshirani (1987) generalise the linear model (7.1.1) to one which models logit(p) as the sum of smooth functions of the covariates which they call a generalised additive model i.e.

$$\text{logit}(p(x)) = \alpha + \sum_{j=1}^{p} f_j(x_j) \qquad (7.3.1)$$

where the $f_j(\cdot)$'s are unspecified smooth one-dimensional functions. The log-odds are thus modelled in an additive but nonparametric manner. The model (7.3.1) is fitted using a 'local scoring algorithm' which is a generalisation of iteratively reweighted least squares and the $f_j(\cdot)$'s are estimated using a scatterplot smoother in another iterative procedure. Full details are contained in the paper. Here all the non-linearities are estimated simultaneously and may suggest suitable parametric transformations of the covariates. Pairwise interactions can be included by allowing bivariate functions into the model. They also describe a number of inferential tools to assist in assessing the relevance and significance of the estimated functions. These include asymptotic confidence intervals, degrees of freedom and hypothesis tests but there is an absence of appropriate distribution theory to accompany these.

While these models are certainly useful there are also some problems and unanswered questions. Firstly, the effect of the dependence between two or more covariates on the fitting algorithm, standard deviations and degrees of freedom is not clear. Secondly, no account is taken of the inherent bias in the function estimates and the effect this has when constructing confidence intervals and comparing fits. Thirdly, categorical variables have not been modelled in this framework and finally the model only exists on the computer so that estimating $p(\underline{x})$ for a new individual with covariate vector $\underline{x}$ may involve a large amount of computation.

O'Sullivan et al (1986) consider modelling the covariates non-parametrically by spline functions using a penalised likelihood with a Laplacian penalty function. They do not use an additive model but estimate $\text{logit}(p(\underline{x}))$ directly using multi-dimensional splines. These are however difficult to display and interpret for more than two covariates. They do not consider any inference based on their estimates.

Green and Yandell (1985) consider the case when all but one of the covariates are linear. They also use spline functions in a penalised likelihood approach and solve the normal equations explicitly. Goodness-of-fit is assessed by the deviance with the asymptotic expectation providing an estimate of it's degrees of freedom. They use further approximate asymptotics to derive an estimated covariance matrix for $\hat{\underline{\theta}}$ from which standard errors can be calculated. However, they recommend that these results should only be used informally due to a lack of appropriate distribution theory.

7.4. The Pseudo-Likelihood Ratio Test.

7.4.1. Introduction.

The methods described in the following sections are most appropriate for binary data or data for which all the n_i's are small but are also applicable to binomial data with large group sizes.

The aim is to construct a nonparametric estimate of the logistic regression function and then quantify and assess the significance of its global discrepancy from the logistic model estimate given by (7.1.8) evaluated at $\hat{\underline{\theta}}$ and now to be denoted either by $p(\underline{x},\hat{\underline{\theta}})$ or $\hat{p}^L$. This can be expressed by the following hypotheses:

$$H_0 : p(x) = p(\underline{x},\underline{\theta}) \quad \text{for some } \underline{\theta}$$

$$H_1 : p(x) \quad \text{is a smooth function.}$$

Azzalini et al (1989) tested these hypotheses using a likelihood ratio approach for the case of a single continuous covariate x. The likelihood under H_0 is evaluated using the logistic model maximum likelihood estimate of θ while the likelihood under the alternative is evaluated at $\hat{p}(\cdot)$, a nonparameteric kernel regression estimate of the logistic curve. The pseudo-likelihood ratio test statistic is then given by:

$$T = \sum_{i=1}^{N} \left[y_i \log \left[\frac{\hat{p}(x_i)}{p(x_i,\hat{\theta})} \right] + (n_i - y_i) \log \left[\frac{1 - \hat{p}(x_i)}{1 - p(x_i,\hat{\theta})} \right] \right] \qquad (7.4.1)$$

H_0 and H_1 are nested hypotheses but because the model under H_1 is not fitted by maximum likelihood T could take a negative value. The statistic estimates the Kulback-Leibler distance between the two models and so the test will be consistent because as the number of

groups increases the normalised test statistic will converge to zero when H0 is true and to some non-zero value when H_0 is false. In an as yet unpublished paper W. Hardle and E. Mammen show that the asymptotic distribution of T is normal but for finite sample sizes the significance of an observed value needs to be determined by simulation which will be described in more detail later in this section.

When a p-dimensional ($p > 1$) vector $\underline{x}$ of covariates are available for each individual then the pseudo-likelihood ratio test statistics (7.4.1) can still be used but the nonparametric estimates of $p(\underline{x})$ need to be constructed using a multivariate kernel. As the number of covariates increases the quality of such an estimate is likely to deteriorate and the following approach based on an analogy with the Hosmer-Lemeshow test maybe used instead.

In the Hosmer-Lemeshow test the $\hat{p}^L$'s are ranked and grouped into deciles. Under H_0 there is a 1-1 relationship between $\hat{p}^L$ and $\hat{Z} = \hat{\underline{\theta}}^T\underline{X}$. Therefore, the Hosmer-Lemeshow test is equivalent to ranking the $\hat{z}$'s and grouping into deciles. The relationship between the $\hat{p}^L$'s and $\hat{z}$'s is given by (7.1.8) and is essentially represented by a step function in the Hosmer-Lemeshow test statistic. An estimate of p under H_1 can therefore be constructed by regarding z as a single covariate and smoothing the data in the form (y_i, n_i, z_i) to give a nonparametric smooth estimate of the relationship in (7.1.8). This estimate, unlike a multivariate one, can always be plotted and can also be used in the test statistic to assess the goodness-of-fit of multivariate data. Because the data are no longer being grouped this test may have greater power than the Hosmer-Lemeshow test.

7.4.2. Smooth nonparametric regression for estimating $p(x)$.

In the following discussion x is a single covariate which can be replaced by z for the construction of the test statistic. If each of the n_i's are large then a useful estimate of $p(x_i)$ is given by the maximum likelihood estimate y_i/n_i. However, if the n_i's are small or equal to one then meaningful proportions are not available and it is necessary to average over neighbouring values of x. This approach was first suggested by Copas (1983) with his proposed estimate being essentially a nonparametric regression function. Under the assumption of smoothness we have the following kernel estimator of $p(x)$:

$$\hat{p}(x) = \frac{\sum_{i=1}^{N} Y_i \, W\left[\frac{x-X_i}{h}\right]}{\sum_{i=1}^{N} n_i \, W\left[\frac{x-X_i}{h}\right]} \qquad (7.4.2)$$

where $w(\cdot)$ is a symmetric non-negative kernel function with a mode at zero and h is a smoothing parameter controlling the degree of local averaging. If h is very small then (7.4.2) will just interpolate the data while when it is very large the resulting estimate approaches the sample mean $N^{-1} \sum_{i=1}^{N} y_i/n_i$. In the examples to follow a standard normal kernel has been used.

Copas suggests plotting $\hat{p}$ for a range of differently chosen h values. While this is satisfactory for a simple graphical inspection of the data a more precise choice, satisfying some optimality criterion, is needed if $\hat{p}$ is to be used for inferential purposes.

Azzalini et al (1989) suggest choosing h to maximise the likelihood function

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \hat{p}_{-i}(x_i)^{y_i} \, (1 - \hat{p}_{-i}(x_i))^{n_i - y_i} \qquad (7.4.3)$$

which is equivalent to minimising

$$\sum_{i=1}^{N} \left\{ - y_i \log \hat{p}_{-i}(x_i) - (n_i - y_i) \log (1 - \hat{p}_{-i}(x_i)) \right\} \qquad (7.4.4)$$

Here $\hat{p}_{-i}(\cdot)$ denotes the nonparametric estimate constructed from all the data points except (y_i, n_i, x_i). In general nonparametric regression with continuous response data asymptotic optimality properties have been derived for a least squares cross-validatory choice of h which, under the assumption of normally distributed errors, can be regarded as a likelihood criterion. (Hardle and Marron (1985)). The properties of a likelihood based choice of h in the binomial context have not been investigated but it is expected that some of the justification in the continuous case will carry over. This selection rule may also be regarded as choosing h to make each $n_i.\hat{p}_i$ an effective predictor of y_i and so the resulting curve will be quite sensitive to variation in the data.

For binary data the density function of y_i is:

$$g(y_i) = p_i^{y_i} \, (1-p_i)^{1-y_i} \, , \quad y_i = 0,1 \qquad (7.4.5)$$

and its expected value is:

$$E[g(y_i)] = p_i^2 + (1-p_i)^2 \qquad (7.4.6)$$

Kapperman (1987) proposes choosing h as the solution to the non-linear equation obtained by equating an estimate of $\sum_{i=1}^{N} g(y_i)$ to its expectation again using a cross-validatory approach to ensure the existence of a solution. This approach generalises to grouped data.

An iterative procedure is required to find a solution and it is reported that more than one may exist in which case it is argued that the smaller one should be chosen. Again it seems likely that this method will be quite data sensitive. He does remark that it works well but does not illustrate or describe in what sense this is so.

A further possibility is to choose an h which is suited more to the model rather than to the observed data. This is particularly appropriate if the estimate is to be used in the pseudo-likelihood ratio test statistic and also in view of the simulations required to assess significance. The model optimal criterion is:

$$\underset{h}{\text{Min}} \sum_{i=1}^{N} \frac{E[(Y_i - n_i \hat{p}_i^s)^2]}{V(Y_i)} \qquad (7.4.7)$$

where $\hat{p}_i^s \equiv \hat{p}(x_i)$.

This is equivalent to

$$\underset{h}{\text{Min}} \sum_{i=1}^{N} \frac{E[[(Y_i - n_i p_i) - (n_i \hat{p}_i^s - n_i p_i]^2]}{V(Y_i)} . \qquad (7.4.8)$$

It can be seen that the value of (7.4.8) will approach zero as h approaches zero because $n_i . \hat{p}_i^s$ will become closer to y_i. It is therefore necessary to use a cross-validatory choice to obtain a feasible solution. Expression (7.4.8) is therefore redefined to be:

$$\underset{h}{\text{Min}} \sum_{i=1}^{N} \frac{E[[(Y_i - n_i p_i) - (n_i \hat{p}_{-i}^s - n_i p_i]^2]}{V(Y_i)} \qquad (7.4.9)$$

where again $\hat{p}_{-i}^s$ is the nonparametric estimate evaluated at x_i, but calculated without the ith data point. Consider the numerator:

$$E[[(Y_i-n_ip_i)-(n_i\hat{p}^s_{-i}-n_ip_i)]^2]$$

$$= E[(Y_i-n_ip_i)^2] - 2E[(Y_i-n_ip_i)(n_i\hat{p}^s_{-i}-n_ip_i)]$$

$$+ E[(n_i\hat{p}^s_{-i}-n_ip_i)^2].$$

Now, $E[(Y_i-n_ip_i)^2] = V(Y_i)$ and does not depend on h.

Also, $E[(Y_i-n_ip_i)(n_i\hat{p}^s_{-i}-n_ip_i)]$

$$= n_i\, E[Y_i\hat{p}^s_{-i}] - n_ip_i\, E[Y_i] - n_i^2p_i\, E[\hat{p}^s_{-i}] + n_i^2\, p_i^2$$

$$= n_i\, E[Y_i\hat{p}^s_{-i}] - n_i^2p_i\, E[\hat{p}^s_{-i}]$$

$$= n_i \frac{\sum_{j\neq i} E[Y_i\, Y_j]W_{ij}}{\sum_{j\neq i} n_jW_{ij}} - n_i^2\, p_i \frac{\sum_{j\neq i} E[Y_j]W_{ij}}{\sum_{j\neq i} n_jW_{ij}}$$

where $W_{ij} = W\left[\frac{X_i-X_j}{h}\right]$

$$= n_i^2\, p_i \frac{\sum_{j\neq i} n_jp_jW_{ij}}{\sum_{j\neq i} n_jW_{ij}} - n_i^2\, p_i \frac{\sum_{j\neq i} n_jp_jW_{ij}}{\sum_{j\neq i} n_jW_{ij}}$$

$$= 0.$$

Finally, $E[(n_i\hat{p}^s_{-i}-n_ip_i)^2]$

$$= n_i^2\, [E[(\hat{p}^s_{-i})^2] - 2p_i\, E[\hat{p}^s_{-i}] + p_i^2]$$

$$= n_i^2\, [[V(\hat{p}^s_{-i}) + E(\hat{p}^s_{-i})^2] - 2p_i\, E[\hat{p}^s_{-i}] + p_i^2]$$

$$\neq 0$$

where

$$V(\hat{p}^s_{-i}) = \frac{\sum_{j\neq i} n_jp_j(1-p_j)W_{ij}^2}{\left[\sum_{j\neq i} W_{ij}\right]^2}.$$

Therefore, the cross-validatory model optimal method is to choose h such that

$$\underset{h}{\text{Min}} \sum_{i=1}^{N} \frac{n_i E[(\hat{p}^s_{-i} - p_i)^2]}{p_i(1-p_i)} \quad . \tag{7.4.10}$$

In practice the unknown logistic model probabilities are replaced by their maximum likelihood estimates. It has generally been found that using (7.4.10) results in a larger value of h and hence a smoother curve than that arising from the likelihood based choice. However, the positive benefits of using a model optimal choice in the goodness-of-fit statistic will be discussed later in this section.

If we now have a p-vector $(p > 1)$ of covariates and are using the linear predictor, $\hat{Z}$, as a single covariate (Section 7.4.1) then the model optimal criterion (7.4.10) will not be strictly correct. This is because now $W_{ij} = W((Z_i - Z_j)/h)$ which is a function of the Y_i's through Z_i and Z_j so that the expected value of the numerator will not be as given above. However, if we assume that the fitted model is correct (i.e. $p_i = \hat{p}^L_i$, $i = 1,\ldots,N$) then the W_{ij}'s are non-random and (7.4.10) is again obtained as the criterion. This assumption about the fitted method will also be made in the simulation procedure to determine the significance of the pseudo-likelihood ratio test statistic for a set of data.

As discussed in chapter 3, nonparametric curve estimates are biased. In the present context the estimates are to be used to test the goodness-of-fit of a parametric logistic model. Therefore, the bias can be evaluated under H_0 and subtracted from the original estimate to give an unbiased estimate if H_0 is true. The practical version of the bias corrected estimate is then:

$$\hat{p}^s_i - \frac{\sum_{j=1}^{N} n_j \hat{p}^L_j W_{ij}}{\sum_j n_j W_{ij}} + \hat{p}^L_i \tag{7.4.11}$$

where W_{ij} is based on the x_i's or $\hat{z}_i$'s as appropriate. The value of (7.4.11) may not necessarily lie between 0 and 1 in which case it should be reset to zero or one as appropriate.

An alternative bias corrected estimate is given by

$$\hat{p}_i^s \cdot \frac{\hat{p}_i^L \left(\sum_j W_{ij}\right)}{\sum_j n_j \hat{p}_j^L W_{ij}} \qquad (7.4.12)$$

but again this does not necessarily lie between 0 and 1.

A bias corrected estimate of the form (7.4.11) will be used in all subsequent analyses.

To illustrate the regression estimates 40 random binary observations were simulated from the model

$$\log(p_i/(1-p_i)) = -1 + X_1 + X_2 + 3X_2^2 \qquad (7.4.13)$$

where X_1 and X_2 are independently uniformly distributed on the interval (-1,1). The simulation procedure used is as follows:

i) Sample x_1 and x_2 independently from the U(-1,1) distribution.

ii) Calculate the value of the linear predictor z using these values of x_1 and x_2.

iii) Compute $p = \exp(z)/(1+\exp(z))$.

iv) Sample a U(0,1) random number, u. If u is less than or equal to p set the response y to 1, otherwise set to zero.

v) Repeat N times to obtain a random sample of (y_i,n_i,x_i)'s where $\underline{x}_i = (x_{1i},x_{2i})^T$ and n_i is 1 for each i.

If binomial data had been required then step iv) would have been repeated n_i times.

A model just involving linear terms was then fitted to these data using the maximum likelihood procedure described in Section 7.1. The estimated logistic model is:

$$\log(\hat{p}^L/(1-\hat{p}^L)) = \hat{z} = 0.489 + 0.335.x_1 + 1.118.x_2. \qquad (7.4.14)$$

The data in the form $(y_i, n_i, \hat{z}_i)$ were then smoothed using both likelihood and model optimal choices of h which are 0.11 and 0.54 respectively. The resulting bias corrected curves are shown, together with the logistic regression estimate in figures 7.1 and 7.2. Both nonparametric estimates indicate a lack of fit throughout the whole range of z but this is far more marked when using the more data sensitive likelihood based h.

To take into account the heterogeneity in the variances of the y_i's a weighted nonparametric regression estimate was also considered. The variance of Y_j under H_0 is $n_j p_j^L(1-p_j^L)$ so an appropriate set of weights summing to one are:

$$\frac{\hat{\sigma}_j^{-1}}{\sum_j \hat{\sigma}_j^{-1}} \quad , \quad j = 1,\ldots,N \qquad (7.4.15)$$

where $\hat{\sigma}_j = \{n_j\, \hat{p}_j^L(1-\hat{p}_j^L)\}^{\frac{1}{2}}$.

These are then used to scale the kernel in the regression estimate resulting in the following weighted estimate:

$$\hat{p}^{ws} = \frac{\sum_j y_j W_j \hat{\sigma}_j^{-1}}{\sum n_j W_j \hat{\sigma}_j^{-1}} \qquad (7.4.16)$$

where W_j represents the kernel function evaluated at $(x-x_j)/h$. Greater weight is therefore applied to data with low variance reflecting that more emphasis should be given to stable data and thereby hopefully constructing a better estimate. The scaling of the W_j's is by an estimate of

$$\{n_j p_j(1-p_j)\}^{-\frac{1}{2}} = \left[\frac{(1+e^{z_j})^2}{n_j\, e^{z_j}}\right]^{\frac{1}{2}} = \frac{1+e^{z_j}}{\sqrt{n_j}\; e^{z_j/2}}$$

$$= \frac{2}{\sqrt{n_j}} \left[\frac{e^{-z_j/2} + e^{z_j/2}}{2}\right]$$

$$= \frac{2}{\sqrt{n_j}} .\cosh(z_j/2) .$$

The minimum weight of $2/\sqrt{n_j}$ is therefore when $z_j = 0$ with increasing weight applied in a symmetric manner as z_j increases and decreases away from zero.

The effect of weighting will however be downplayed by the kernel weights W_j, especially when h is small as then the local averaging is carried out only over a short interval with data in that interval tending to have similar variance and hence weights. Similarly, for an unweighted estimate the smoothing parameter, especially when small, should minimise the effect of the unequal variances. This is illustrated when using the data simulated from (7.4.13) where, for a weighted estimate, both the optimal likelihood and model based choice of h were found to be almost unchanged. The resulting weighted estimates differ very little from the weighted ones shown in figures 7.1 and 7.2. Such similarities were also found for a number of other data sets. As a result weighted estimates will not be used in any subsequent analyses.

7.4.3. Assessing the significance of the pseudo-likelihood ratio test statistic.

The significance of an observed test statistic needs to be determined by simulation because of the unknown finite sample distribution. To achieve this a simulated sample $(y_1^*,\ldots,y_N^*]$ is firstly obtained from the fitted model such that each y_i^* has a $Bi(n_i,\hat{p}_i^L)$ distribution. For the ith group this is achieved by sampling n_i independent $U(0,1)$ random numbers and counting the number which are less than or equal to $\hat{p}_i^L$ to give y_i^*. A new set of nonparametric estimates $\hat{p}_i^{s*}$ are then calculated and hence a a simulated value of the test statistic can be obtained. This process can then be repeated a large number of times to give B, say, simulated test statistic values. These B values can then be ordered and the significance of the original value determined by its position among these ordered values.

There are two important considerations in this simulation process which concern the fit of a simulated data set to the models under the null and alternative hypotheses. The first is whether to choose a new smoothing parameter for each simulated sample and the second is whether to refit the logistic model to each new sample.

The value of h clearly does affect how the nonparametric estimate fits the data but very large samples are needed before any optimal properties the choice of h may have come strongly into play. This imprecision means that in most practical situations choosing a new h for each sample will add another source to, and hence increase, the total variability in the simulated distribution of the test statistic. The calculation of a new smoothing parameter by one of the cross-validatory methods for each new sample also involves a great deal of

extra computational effort, especially for large data sets. Therefore, it is appropriate to use the original h value for smoothing each set of simulated data and hence using a model optimal choice should provide a set of more stable estimates than the more data sensitive likelihood choice.

The test statistic T (7.4.1) for the original data is the difference between the log likelihoods calculated under the alternative non-parametric smooth and null parametric logistic hypotheses (i.e. $\ell_1-\ell_0$). The parameter estimates, $\hat{\underline{\theta}}$, for the parametric model are based on maximum likelihood. However, in the subsequent simulated samples $\hat{\underline{\theta}}$ will clearly not maximise the likelihood of the new data and hence an extra source of variability is introduced into the simulations if $\hat{\underline{\theta}}$ is always kept the same. To remove this, $\underline{\theta}$ should be re-estimated for each new sample. This results in the ℓ_0 component of the test statistic being maximised and hence the value of $\ell_1-\ell_0$ is minimised. The result is that the observed test statistic becomes more extreme with respect to the simulated values. The calculation of $\hat{\underline{\theta}}$ is iterative as described in Section 7.1. However, a good estimate is available after the first iteration, especially if $\hat{\underline{\theta}}$ from the original data is used as a starting value. Therefore, to reduce the computation level only one iteration need be carried out.

To illustrate this discussion the pseudo likelihood ratio test was carried out on the data simulated from model (7.4.13). 500 simulations were performed to assess significance. Each nonparametric estimate was based on the original h value and bias corrected as in (7.4.11) while the parametric model was refitted using only one iteration. The results are given in the following table:

Table 7.1

Smoothing parameter	Test statistic	Significance level
Likelihood, h = 0.11.	T = 21.39	0.020
Model optimal, h = 0.54	T = 6.58	0.012

The value of T is much smaller for h = 0.54 because the resulting estimate is much closer to the logistic curve as can be seen in figures 7.1 and 7.2. The significance levels are comparable for the two tests because the same seed for the random number generator was used for both sets of simulations. A more significant result is obtained when using the model optimal h, probably because the likelihood choice produced much more variable estimates from the simulated data sets for reasons discussed earlier and thus made the observed value less extreme.

The Hosmer-Lemeshow test statistic for these data has the value 20.75 on 8 degrees of freedom and is also highly significant when compared with $\chi^2(8,0.95) = 15.51$.

7.5. Assessing the functional form of covariates.

7.5.1. Introduction.

If a logistic model is found to have a poot fit, for example after using the pseudo-likelihood ratio test, it is natural to then investigate the possible reasons for this. As mentioned in Section 7.1 it may be the result of incorrectly specifying the functional form of one or more of the covariates. For example a quadratic contribution may also be required from a variable which is currently only included linearly.

If, through knowledge of the data, the non-linear functional form of a variable is known or can be hypothesised then a term with the appropriately transformed variable should be included in the parametric model. However, if this approach cannot be adequately justified but a non-linear form is still suspected then the semi-parametric approach of Green and Yandell (1985) or a generalised additive model with all but the variable of interest included linearly (Hastie and Tibshirani (1987)) could be used.

It is often not clear a priori that non-linear contributions are required and a useful first step is to fit a model containing only linear effects. It is therefore useful to have available methodology for checking whether the effects are linear and to identify forms which may improve the fit.

For assessing departures from linearity associated with a variable X_m the approach of Minkin (1989) is to choose $k-1$ values $v_1, v_2, \ldots, v_{(k-1)}$ and define k new variables as follows:

$$Z_{i1} = \begin{cases} x_{im} & , \ x_{im} < v_1, \\ v_1 & , \ \text{otherwise.} \end{cases}$$

$$Z_{ij} = \begin{cases} 0 & , \ x_{im} < v_{j-1}, \\ x_{im} - v_{j-1} & , \ v_{ij} \leqslant x_{im} < v_j \\ v_j - v_{j-1} & , \ x_{im} > v_j. \end{cases}$$

$$Z_{ik} = \begin{cases} x_{im} - v_{k-1}, & x_{im} > v_{k-1}, \\ 0 & , \ \text{otherwise.} \end{cases} \qquad (7.5.1)$$

The term $\beta_m x_{im}$ in the linear predictor is then replaced by $\sum_{j=1}^{b} \gamma_j Z_{ij}$ and the model fitted using maximum likelihood in the usual way. The effect of x_m is therefore approximated by segments joined at the points $v_1, v_2, \ldots, v_{(k-1)}$ and a graphical representation

is provided by plotting the $\sum_{j=1}^{k} \hat{\gamma}_j Z_{ij}$ against the x_{im}. If the v_i's are selected independently from the response then the likelihood ratio test statistic for testing if the segments all have the same slope has an asymptotic $\chi^2(k-1)$ distribution. An automatic procedure for selecting the join points based on goodness-of-fit is also proposed and a simulation study shows the validity of the $\chi^2(k-1)$ distribution as a reference for the test when the sample size is large. However, for moderately sized samples it is not reliable.

Another, older procedure, which will be concentrated on for the rest of this section, is the use of partial residual plots. For multiple linear regression the plot was first proposed by Ezekiel (1924) and then much later advocated by Larsen and McCleary (1972). In this context when we have data from the model

$$E(\underline{Y}) = X\underline{\beta} + f(\underline{Z}) \qquad \text{cov}(\underline{Y}) = \sigma^2 I_n \qquad (7.5.2)$$

where f is a smooth, but unknown, function of the covariate Z and the vector $f(\underline{Z}) = f(Z_1),\ldots,f(Z_n))^T$ the partial residual vector $\underline{r}^p$ for Z is defined by:

$$\begin{aligned} \underline{r}^p &= (\underline{y} - X\hat{\underline{\beta}} - \underline{z}\hat{\gamma}) + \underline{z}\hat{\gamma} \\ &= \underline{r} + \underline{z}\hat{\gamma} \\ &= \underline{y} - X\hat{\underline{\beta}} \end{aligned} \qquad (7.5.3)$$

where the vector $\underline{r}$ contains the ordinary residuals. It can be regarded as the dependent variable vector corrected for all the independent variables except Z. The coefficient estimates $\hat{\underline{\beta}}$ and $\hat{\gamma}$ are obtained by fitting the complete model. The plot is then obtained by plotting $\underline{y} - X\hat{\underline{\beta}}$ against $\underline{z}$.

There are two orthogonal components which make up the partial residual. The first, $\underline{r}$, represents scatter while $\underline{z}\hat{\gamma}$ is the systematic part. In a least squares regression of $\underline{r}^P$ on $\underline{z}$ the intercept is 0 and the slope is $\hat{\gamma}$. If $f(\underline{z})$ is linear then the expected configuration of the plotted points will be linear with slope γ but when it is non-linear the shape of the plotted points should suggest the form of f. In the partial residual plot the variance of $\hat{\gamma}$ may underestimate the estimated variance of $\hat{\gamma}$ from the full regression because any effect due to fitting the other variables is ignored. Thus the plot may give a spurious impression of the accuracy of the slope of the fitted line and will be most marked when the correlation between Z and the other regressor variables is high. (Cook and Weisberg (1982)).

The expected error for the partial residual plot is defined as

$$\begin{aligned} \underline{e} &= f(\underline{Z}) - E[\underline{r}^P] \\ &= f(\underline{Z}) - [X\underline{\beta} + f(\underline{Z}) - X(\underline{\beta} + \text{bias}(\hat{\underline{\beta}}))] \\ &= X.(\text{bias}(\hat{\underline{\beta}})) \end{aligned} \qquad (7.5.4)$$

where the $\text{bias}(\hat{\underline{\beta}})$ is given in Mansfield and Conerly (1987) as

$$\text{bias}(\hat{\underline{\beta}}) = (X^TX)^{-1}X^T[I - \underline{Z}(\underline{u}^T\underline{u})^{-1}\underline{u}^T]\underline{f} \qquad (7.5.5)$$

where $\underline{u} = [I - X(X^TX)^{-1}X^T]\underline{Z}$ and the value of $(\underline{u}^T\underline{u})^{-1}\underline{u}^T\underline{f}$ is the slope that occurs if $\underline{f}$ is regressed on $\underline{Z}$ adjusted for X (i.e. on $\underline{u}$). When f is linear then $(\underline{u}^T\underline{u})^{-1}\underline{u}^T\underline{f} = \gamma$ and hence the $\text{bias}(\hat{\underline{\beta}})$ and error are zero. If there is no linear association between $\underline{u}$ and $\underline{f}$ then $(\underline{u}^T\underline{u})^{-1}\underline{u}^T\underline{f} = 0$. This occurs when $\underline{u}^T\underline{f} = 0$ or more specifically when $\underline{z}^T\underline{f} = 0$ and $X^T\underline{f} = \underline{0}$ and implies that neither $\underline{Z}$ nor X contain information about f. The $\text{bias}(\hat{\underline{\beta}})$ and hence the error are

again zero. When the bias$(\hat{\beta})$ is non-zero its size depends on the correlation between Z and the variables in X and on how well $\underline{Z}$ and X represent f across the range of relevant data.

The use of partial residual plots in logistic regression was first suggested by Landwehr et al (1984). To construct the partial residuals they exploited the relationship between logistic and weighted linear regression - see equation (7.1.6). They regard

$$\underline{y}^* = x\hat{\underline{\theta}} + z\hat{\gamma} + W^{-1}(\underline{y}-\underline{n}\,\hat{\underline{p}}^L) \qquad (7.5.6)$$

where $W = \text{diag}(n_i.\hat{p}_i^L.(1-\hat{p}_i^L))$ as "logistic observations" and a straightforward application of the normal linear model results gives the logistic partial residuals as

$$\underline{r}^{pL} = W^{-1}(\underline{y} - \underline{n}\,\hat{\underline{p}}^L) + \underline{z}\hat{\gamma} \qquad (7.5.7)$$

and the plot is obtained by plotting $\underline{r}^{pL}$ against $\underline{z}$. Here, $\underline{y}$ and $\underline{n}\hat{\underline{p}}^L$ are vectors of length N containing the y_i's and $n_i\hat{p}_i^L$'s respectively. For binary data the points will fall into two separate clouds depending on whether $y_i = 1$ or 0 thus obscuring the functional form of f. It is therefore useful to smooth the plot using a kernel regression estimate of the form

$$\hat{r}^{pL}(z) = \frac{\sum_{i=1}^{N} K\left[\frac{z-z_i}{b}\right] \cdot r_i^{pL}}{\sum_{i=1}^{N} K\left[\frac{z-z_i}{b}\right]} \qquad (7.5.8)$$

where $K(\cdot)$ is a symmetric p.d.f. such as the standard normal density and b is the bandwidth or smoothing parameter (Watson, (1964)). A choice of b can be made by using least squares cross-validation, i.e.

$$\underset{b}{\text{Min}} \sum_{i=1}^{N} (r_i^{pL} - \hat{r}_{-i}^{pL})^2 \qquad (7.5.9)$$

An alternative model based approach to selecting b can also be derived and will be discussed further later in this Section in the context of testing for linearity.

To remove the structure resulting from the binary nature of the observations Fowlkes (1987) uses partial residuals with y_i replaced by $\hat{p}_i^s$ in the definition where $\hat{p}_i^s$ is calculated using a multivariate kernel function.

7.5.2. Testing the linearity of the partial residuals.

If the variable z in the logistic regression model has been specified correctly as being linear then the partial residuals should lie about a straight line with slope approximately equal to $\hat{\gamma}$. Smoothing the plot allows a subjective impression to be made of the degree of linearity but it is desirable to be able to quantify the strength of linearity, i.e. to test the null hypothesis

$$H_0 : E[r_i^{pL}] = \gamma z_i, \qquad i = 1,\ldots,N$$

against the alternative

$$H_1 : E[r_i^{pL}] = f(z_i), \quad i = 1,\ldots,N \text{ for some smooth function } f.$$

To do this consider the residual sum of squares of the partial residuals about the line $\hat{\gamma}z$ and curve $\hat{r}^{pL}(z)$ which are estimates of the line γz and the curve $f(z)$. The residual sum of squares provides a measure of the discrepancy between the data and fitted line or curve. In general $RSS_L > RSS_C$ where the subscripts L and C refer to the line and curve respectively, as the curve has more freedom

to explain the data. A test statistic can then be defined by considering the difference in residual sums of squares relative to RSS_C, i.e.

$$S = (RSS_L - RSS_C)/RSS_C \qquad (7.5.10)$$

and rejecting H_0 for large values of S.

This is analogous to testing nested hypotheses in a normal linear model framework where the numerator and denominator are divided by appropriate degrees of freedom so that the test statistic, under H_0, has an F distribution. It is not necessary to divide by estimates of degrees of freedom here, but as the distribution of S is unknown and difficult to evaluate it is therefore necessary to use simulation to assess the signficance of an observed value. The simulation procedure is the same as that described in Section 7.4 but with the additional step of calculating the partial residuals either from the new y_i's or $\hat{p}_i^S$'s based on using the linear predictors in (7.4.2) rather than Foulkes's multivariate kernel approach. The test statistics calculated from each simulated data set are then ordered and significance assessed by evaluating the proportion of simulated values greater than that observed.

In order to be able to correct for bias and add bands to the plot at ±2 standard deviations of the smooth estimate, the mean and variance of the smooth curve can be calculated under H_0. As in Section 7.4.2 it will be assumed that the fitted model is correct. The following results are exact if the data did indeed arise from such a model but if not they should still provide a good approximation provided the $\hat{p}_i^L$'s are good estimates of the p_i's. When the discrete response is used we have:

$$r_j^{pL} = \frac{Y_j - n_j \hat{p}_j^L}{n_j \hat{p}_j^L (1-\hat{p}_j^L)} + \hat{\gamma}\, z_j \qquad (7.5.11)$$

so that

$$E[r_j^{pL}] = \hat{\gamma}\, z_j \qquad (7.5.12)$$

and

$$\text{Var}\ (r_j^{pL}) = [n_j \hat{p}_j^L (1-\hat{p}_j^L)]^{-1}. \qquad (7.5.13)$$

Using (7.5.8) to obtain smooth estimates $\{\hat{r}_i^{pL};\ i = 1,\ldots,N\}$ gives

$$E[\hat{r}_i^{pL}] = \frac{\sum_j \hat{\gamma}\, z_j\, K_{ij}}{\sum K_{ij}} \qquad (7.5.14)$$

where K_{ij} denotes the value of the weight function K at $(z_i - z_j)/b$ and

$$\text{Var}\ (\hat{r}_i^{pL}) = \frac{\sum_j K_{ij}^2 \cdot [n_j \hat{p}_j^L (1-\hat{p}_j^L)]^{-1}}{\left[\sum_j K_{ij}\right]^2}. \qquad (7.5.15)$$

When the smoothed response is used then:

$$r_j^{ps} = \frac{\hat{p}_j^s - \hat{p}_j^L}{\hat{p}_j^L (1-\hat{p}_j^L)} + \hat{\gamma}\, z_j \qquad (7.5.16)$$

so that,

$$E[r_j^{ps}] = \frac{E[\hat{p}_j^s] - \hat{p}_j^L}{(\hat{p}_j^L (1-\hat{p}_j^L)} + \hat{\gamma}\, z_j \qquad (7.5.17)$$

where

$$E[\hat{p}_j^s] = \begin{cases} \dfrac{\sum_k n_k\, \hat{p}_k^L\, W_{kj}}{\sum_k n_k\, W_{kj}} & \text{, without bias correction,} \\ \hat{p}_j^L & \text{, with bias correction} \end{cases} \qquad (7.5.18)$$

where $W_{kj} = W(z_k' - z_j')/h)$ where z_k' denotes the linear predictor for the k^{th} observation. Also,

$$V(r_j^{ps}) = \frac{V(\hat{p}_j^s)}{[\hat{p}_j^L(1-\hat{p}_j^L)]^2} \qquad (7.5.19)$$

where

$$V(\hat{P}_j^s) = \frac{\sum_k n_k \hat{p}_k^L (1-\hat{p}_k^L) W_{kj}^2}{(\sum_k n_k W_{kj})^2} \qquad (7.5.20)$$

$Var(\hat{p}_j^s)$ is the same whether $\hat{p}_j^s$ has been bias corrected or not. Using these results gives

$$E[\hat{r}_i^{ps}] = \frac{\sum_j E(r_j^{ps}).K_{ij}}{\sum_j K_{ij}} = \frac{\sum_j \hat{\gamma} z_j K_{ij}}{\sum_j K_{ij}} \qquad (7.5.21)$$

if $\hat{p}_j^s$ is bias corrected whereas if it has not been then expression (7.5.17) should be substituted for $E[r_j^{ps}]$. Also,

$$V(\hat{r}_i^{ps}) = \frac{\sum_k K_{ij}^2 V(r_j^{ps})}{(\sum_k K_{ij})^2} \qquad (7.5.22)$$

where expression (7.5.19) should be substituted for $Var(r_j^{ps})$.

A bias corrected smooth estimate is then:

$$\hat{r}_i^{ps} - E(\hat{r}_i^{ps}) + \hat{\gamma} z_i \qquad (7.5.23)$$

and approximate confidence bands can be added at $\pm 2\sqrt{V(\hat{r}_i^{ps})}$. In subsequent examples bias corrected estimates will be used.

As mentioned earlier the smoothing parameter may be chosen by least squares cross-validation or by using a model based choice similar in nature to that used for constructing $\hat{p}^s$. In particular, when using the discrete response, we can choose b to

$$\underset{b}{\text{Min}} \sum_{i=1}^{N} \frac{E[(r_i^{pL} - \hat{r}_i^{pL})^2]}{V(r_i^{pL})} . \qquad (7.5.24)$$

As before it is necessary to use a cross-validatory approach because otherwise the expression is minimised as $b \longrightarrow 0$ so that the $\hat{r}_i^{pL}$ interpolate the data. It will again be assumed that the fitted model is correct so that expression 7.5.24 can be rewritten as:

$$\underset{b}{\text{Min}} \sum_{i=1}^{N} \frac{E[(r_i^{pL}-\hat{\gamma}Z_i)^2]-2E[(r_i^{pL}-\hat{\gamma}Z_i)(\hat{r}_{-i}^{pL}-\hat{\gamma}Z_i)(\hat{r}_{-i}^{pL}-\hat{\gamma}Z_i)]+E[(\hat{r}_{-i}^{pL}-\hat{\gamma}Z_i)^2]}{V(r_i^{pL})} \qquad (7.5.25)$$

where $\hat{r}_{-i}^{pL}$ denotes the estimate calculated using all but the ith observation.

The first term does not depend on b and the second term is zero which can be shown as follows:

$$E\left[(r_i^{pL} - \hat{\gamma}Z_i)(\hat{r}_{-i}^{pL} - \hat{\gamma}Z_i)\right]$$

$$= E\left[r_i^{pL}\, \hat{r}_{-i}^{pL}\right] - \hat{\gamma}Z_i\, E\left[\hat{r}_{-i}^{pL}\right] - \hat{\gamma}Z_i\, E\left[r_i^{pL}\right] + \hat{\gamma}^2\, Z_i^2 .$$

Now,

$$E\left[r_i^{pL}.\ \hat{r}_{-i}^{pL}\right] = E\left[\frac{\sum_{j\neq i} K_{ij}\, r_j^{pL}\, r_i^{pL}}{\sum_{j\neq i} K_{ij}}\right]$$

$$\simeq E\,(r_i^{pL}) \cdot \frac{\sum_{j\neq i} K_{ij}\, E(r_j^{pL})}{\sum_{j\neq i} K_{ij}}$$

$$= \hat{\gamma}\, Z_i \frac{\sum_{j\neq i} K_{ij}\, \hat{\gamma}\, Z_j}{\sum_{j\neq i} K_{ij}} .$$

where the $\simeq$ sign is due to the fact that the i^{th} observation

contributes to the fitting of the logistic model and hence to r_j^{pL}. Therefore, r_i^{pL} and r_j^{pL} are not strictly independent but this dependence should be small in practice.

Also,

$$E\left[\hat{r}_{-i}^{pL}\right] = \frac{\sum_{j\neq i} K_{ij}\,\hat{\gamma}\,Z_j}{\sum_{j\neq i} K_{ij}}$$

and

$$E\left[r_i^{pL}\right] = \hat{\gamma}\,Z_i.$$

Therefore,

$$E\left[(r_i^{pL} - \hat{\gamma}\,Z_i)(\hat{r}_{-i}^{pL} - \hat{\gamma}\,Z_i)\right]$$

$$= \hat{\gamma}\,Z_i\,\frac{\sum_{j\neq i} K_{ij}\,\hat{\gamma}\,Z_j}{\sum_{j\neq i} K_{ij}} - \frac{\hat{\gamma}\,Z_i \sum_{j\neq i}\hat{\gamma}\,Z_j}{\sum_{j\neq i} K_{ij}} - \hat{\gamma}^2\,Z_i^2 + \hat{\gamma}^2 Z_i^2 = 0.$$

The criterion therefore simplifies to:

$$\underset{b}{\text{Min}} \sum_{i=1}^{N} \frac{E[(\hat{r}_{-i}^{pL} - \hat{\gamma}z_i)^2]}{V(r_i^{pL})}\,. \qquad (7.5.26)$$

or,

$$\underset{b}{\text{Min}} \sum_{i=1}^{N} \frac{V(\hat{r}_{-i}^{pL}) + E[(\hat{r}_{-i}^{pL})^2] - 2\hat{\gamma}z_i\,E[\hat{r}_i^{pL}] + (\hat{\gamma}z_i)^2}{\{n_i\hat{p}_i^L(1-\hat{p}_i^L)\}^{-1}} \qquad (7.5.27)$$

Choosing b by using (7.5.27) tends to result in a larger value of b and hence a smoother curve estimate than when using least squares cross-validation.

To illustrate the use of partial residual plots data simulated from the model

$$\text{logit}(p_i) = -1 + X_1 + X_2 + 3X_2^2$$

will be used again. As before, a model involving linear terms only was fitted to this data and partial residual plots then constructed for X_1 and X_2.

Figures 7.3 and 7.4 show the plots based on using the discrete response for variables X_1 and X_2. Both plots show how the points fall into the two separate clouds corresponding to values of $y = 1$ and 0. However, smoothing, with b chosen by least squares cross-validation, clearly shows that X_1 has been specified correctly whereas the quadratic shaped curve for X_2 indicates the need for an additional quadratic term. The reference line $1.118X_2$ lies outside the pointwise confidence bands near the origin. When the partial residuals are smoothed using a model optimal b-value the results are qualitatively similar to those in figures 7.3 and 7.4.

When the smoothed response ($h = 0.11$) is used the points no longer divide into two groups. The plot for X_1, figure 7.5, again shows a clear linear trend while that for X_2, figure 7.6, shows marked curvature made clearer by the smooth regression curve superimposed. This time the linear reference line lies above the upper confidence band in the centre and below the lower confidence band on both sides of the plot. The need for a quadratic contribution from X_2 is clear. The confidence bands when using the smooth response are much narrower than when using the discrete response.

The plots obtained when using the model optimal $h = 0.54$ to smooth the response are not included but are very similar to those illustrated using $h = 0.11$ except the curvature is perhaps not as marked for X_2.

The partial residuals do not all have the same variance and so a weighting scheme similar to that used for smoothing y was considered with weights equal to the inverse of the variance of a partial residual. However, the resulting plots show little difference to those constructed without weighting and so subsequent examples will not use weighting. Again, the smoothing parameter, especially if small, ensures that local averaging is carried out using observations with similar variance. Also the range of weights is small. For example when using the discrete response for binary data the minimum and maximum possible weights are 0 and 0.25 respectively. This small range also reduces the effects weighting may have.

As remarked earlier the significance of an observed value of the statistic S (7.5.10) for testing linearity needs to be determined by simulation. The value of the bandwidth b for smoothing the partial residuals for the original data will be used to smooth each set of simulated partial results with the argument for this following that for the pseudo-likelihood ratio test (Section 7.4). When using discrete responses the model optimal choice makes a better one due to the greater stability of the resulting estimates over those resulting from the more data sensitive least squares method.

For each simulated data set the RSS needs to be calculated. If the logistic model is not refitted for a simulated data set then the RSS will be larger than when comparing the simulated partial residuals with the correct linear reference line. The effect of not re-fitting will therefore be to increase the value of each simulated S value with the overall result of making the observed S less extreme. Therefore, the logistic model should be re-fitted to each simulated data set but to reduce computation one iteration of the fitting

algorithm will give good estimates if the original parameter estimates are used as starting values.

The linearity of variables X_1 and X_2 of the previously described and analysed data from model (7.4.13) were tested using both discrete and smooth response partial residuals. All smooth estimates were bias corrected and the test was based on 500 simulations with the model re-fitted to each simulated data set using one iteration only. The results obtained were as follows:

Table 7.2

Response	X_1	X_2
Discrete	b = 1.17 (model opt.) S = 0.00371 sig. level = 0.25	b = 0.51 (model opt.) S = 0.211 sig. level = 0.000
Discrete	b = 7.18 (least sq.'s) S = 0.00361 sig. level = 0.166	b = 0.25 (least sq.'s) S = 0.542 sig. level = 0.000
Smooth (h = 0.11)	b = 6.97 (least sq.'s) S = 0.00867 sig. level = 0.39	b = 0.20 (least sq.'s) S = 2.016 sig. level = 0.046
Smooth (h = 0.54)	b = 0.33 (least sq.'s) S = 0.0899 sig. level = 0.748	b = 0.11 (least sq.'s) S = 5.523 sig. level = 0.41

When using the discrete response the test is highly significant for X_2 but not for X_1 when using both the model optimal and least squares smoothing parameters to smooth the partial residuals.

When using a smooth response the results for X_1 are not significant for both h = 0.11 and 0.54 with that for h = 0.54

being by far the least significant. However, for X_2, when using the likelihood based $h = 0.11$ the test is significant at the 5% level but a clearly non-significant result is obtained for the model optimal $h = 0.54$. The partial residual plot in this case does show much less curvature that when using $h = 0.11$, the difference being most marked on the left of the plot.

Using smooth responses in the partial residuals does make clearer the functional form of a covariate than when using the discrete response which results in a discretised plot. However, the addition of a non-parametric smooth regression curve makes the form much clearer and indeed the smooth curves for X_2 in the above example when using the discrete responses and smooth responses ($h = 0.11$) are very similar. Partial residuals based on the discrete responses are also much simpler and quicker to compute than those based on the smooth responses.

7.6. Examples.

In the following two examples the significance of the observed values of the pseudo-likelihood ratio test statistic and the statistic for testing linearity are assessed through 500 simulations. Each new nonparametric estimate is constructed using the original model optimal smoothing parameter and appropriately bias corrected while the parametric logistic model is refitted using only one iteration.

Example 1: Finney's data.

These data were originally analysed by Finney (1947) and subsequently by several other authors. They consist of 39 observations on the effect of two covariates rate and volume of air inspired, on the occurrence ($Y = 1$) or non-occurrence ($Y = 0$) of a transient

vasoconstriction response in the skin of the fingers. These observations were collected from only three individuals but the experiment was designed to try and ensure that observations on the same subject were independent. Finney (1947) log transformed each of the covariates and this has also been done here. An additive logistic model was fitted by maximum likelihood to give:

$$\text{logit } (p) = \hat{z} = -2.924 + 5.330 \log(\text{vol}) + 4.631 \log(\text{rate}) \qquad (7.6.1)$$

The deviance for this model is 29.264 on 36 d.f. but due to the binary nature of the response this is uninformative about goodness-of-fit as discussed in Section 7.2.

The data in the form $(y_i, n_i, \hat{z}_i)$ were then smoothed using (7.4.2) with a model optimal h-value of 1.02. The resulting bias corrected estimate (figure 7.7) has a peak near $\hat{z} = -2.5$ due in the main to observations 4 and 18. These both have $Y = 1$ at points where the fitted probability of "success" is small i.e. 0.073 for case 4 and 0.103 for case 18. The plot also indicates that the logistic model may not be approaching its upper asymptote rapidly enough to ensure a good fit.

The observed value of the pseudo-likelihood ratio test statistic T (7.4.1) is 9.029. The associated p-value of 0.004 confirms the lack of fit of the logistic model indicated in figure 7.7.

Partial residual plots using the binary responses were obtained for the two covariates and are displayed in figures 7.8 and 7.9. In both cases the nonparametric estimates of functional form based on model optimal b-values closely follow the linear reference lines and are well within the approximate confidence bounds. Also, the

fact that cases 4 and 18 are outliers is clearly indicated. Application of the test of linearity (7.5.10) gives $S = 0.031$ ($p = 0.696$) for log (volume) and $S = 0.046$ ($p = 0.380$) for log (rate) thus confirming that they have both been specified correctly. The poor fit of the logistic model is therefore not due to the misspecification of the functional form of the covariates.

Example 2: Cardiff bronchitis data.

The data consist of 212 observations on two covariates and a binary response obtained in a study of male chronic bronchitis in Cardiff conducted by Jones (1975). The variables are CIG, the number of cigarettes smoked per day and POLL, the smoke level near the respondent's home obtained by interpolating the levels at 13 air pollution monitoring stations in the city. The response Y takes the value 1 if the respondent suffered from chronic bronchitis and 0 if he did not.

The fitted logistic model is:

$$\text{logit}(p) = \hat{z} = -10.085 + 0.212\ \text{CIG} + 0.132\ \text{POLL} \qquad (7.6.2)$$

which has deviance of 174.214 on 209 d.f. The bias corrected non-parametric estimate based on the model optimal h of 0.54 is illustrated in figure 7.10. Considerable lack of fit is indicated by the large trough for values of the linear predictor between about 2 and 4. In fact only nine individuals have values of $\hat{z}$ greater than 1.0 and the size of the trough can be attributed in particular to case 147 who has $\hat{z} = 2.828$, $Y = \emptyset$ but CIG = 24.9 and POLL = 58.0. The logistic model also does not appear to fit particularly well for values of $\hat{z}$ less than 0.5. It tends to overestimate the probability of chronic bronchitis for $\hat{z} < -1.8$ and underestimate it for

$-1.8 \leqslant \hat{z} < 0.5$. Indeed, examination of the data show that there are no individuals who smoke less than 3 cigarettes per day and with any air pollution level who have chronic bronchitis. On the other hand though there are sufferers among non-smokers. This apparent lack of fit is confirmed by the pseudo-likelihood ratio test: $T = 15.881$ with an associated p-value of 0.000.

The partial residual plots for CIG and POLL are illustrated in figures 7.11 and 7.12 respectively. That for CIG indicates that a quadratic term with a negative coefficient may improve the fit while the marked trough on the right is due again to case 147. Case 122 is also indicated as an outlier - he is a non-smoker living in an area with fairly low air pollution but is suffering from chronic bronchitis. The plot for POLL suggests that this term has been specified correctly with cases 122 and 147 again being clear outliers. The test of linearity gives $S = 0.298$ ($p = 0.010$) for CIG and $S = 0.134$ ($p = 0.084$) for POLL thus confirming the subjective impression although the small p-value for POLL suggests it may also have some curvature.

Aitken et al (1989) consider logistic models for these data. They extend the simple additive model to a third degree response surface and eliminate unnecessary terms by comparing differences in deviance with the appropriate χ^2 values. The above approach certainly complements this. However, they conclude that the complexity indicates a systematic failure in representing the data and they go on to obtain a more satisfactory analysis by grouping both CIG and POLL into classes and modelling the number of bronchitis sufferers in each cell by a binomial distribution.

Figure 7.1. Logistic regression curve and bias corrected nonparametric estimate (h = 0.11) for 40 observations simulated from the model $\text{logit}(p) = -1 + X_1 + X_2 + + 3X_2^2$. The $\hat{Z}$'s are based on linear terms only.

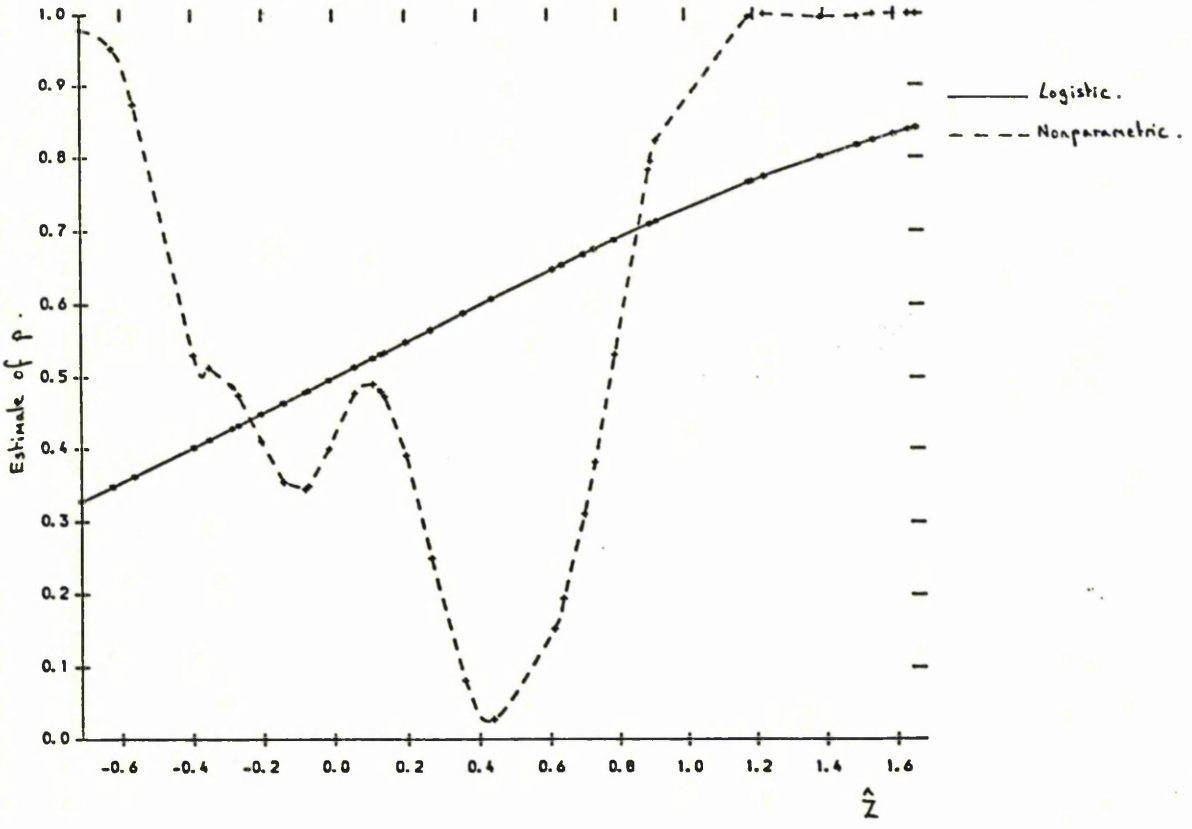


Figure 7.2. Logistic regression curve and bias corrected nonparametric estimate (h = 0.54) for 40 observations simulated from the model $\text{logit}(p) = -1 + X_1 + X_2 + + 3X_2^2$. The $\hat{Z}$'s are based on linear terms only.

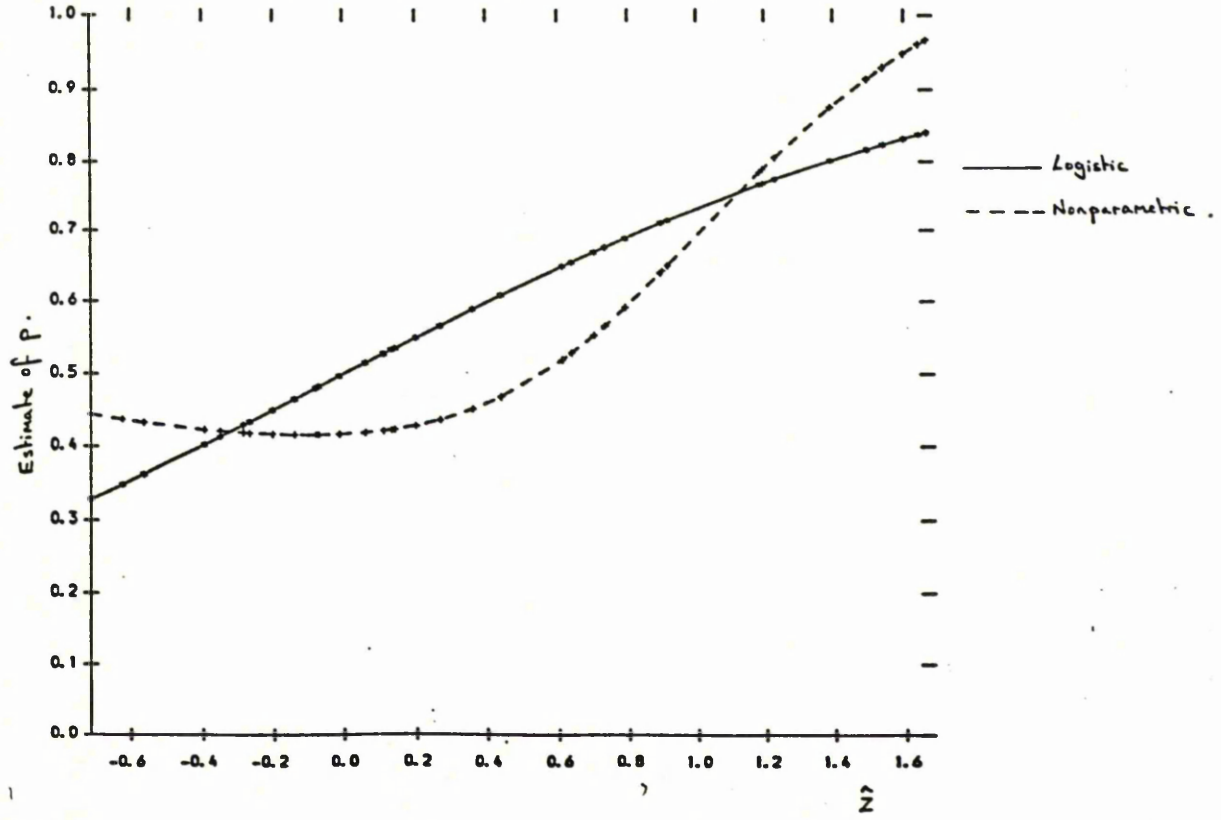


Figure 7.3. Partial residual plot based on the discrete responses for X_1 of the simulated data.

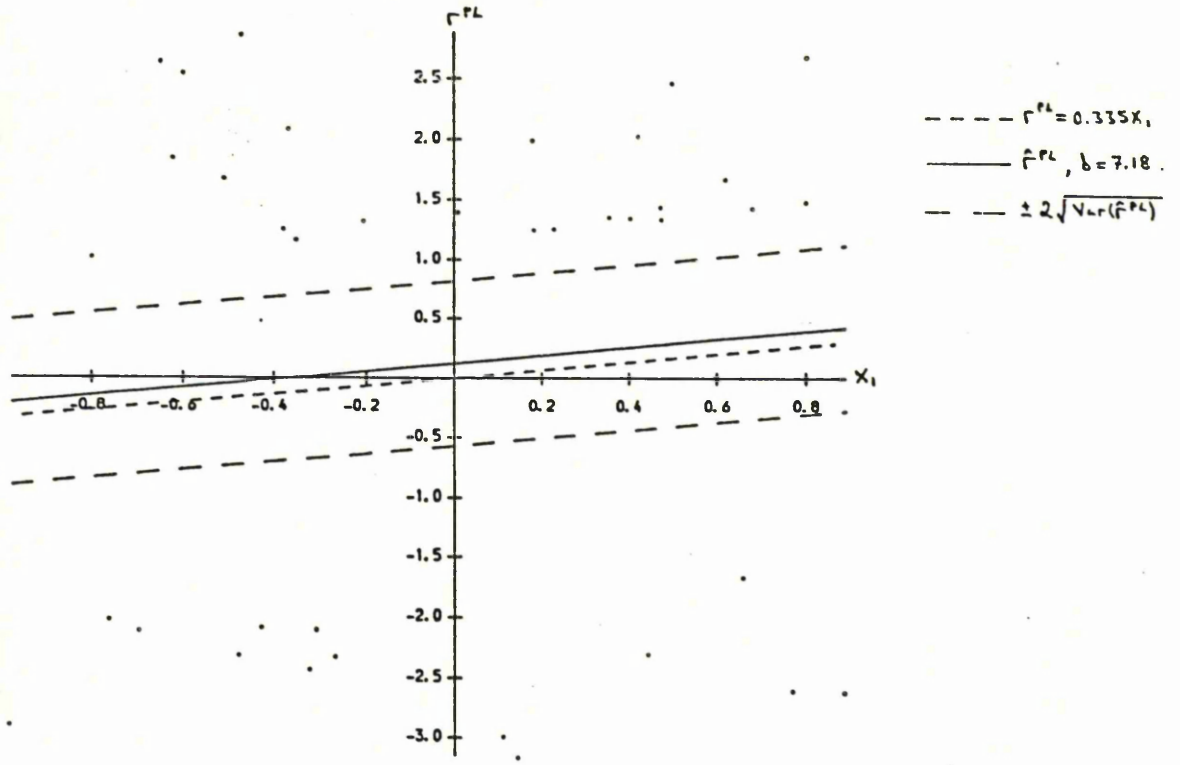


Figure 7.4. Partial residual plot based on the discrete responses for X_2 of the simulated data.

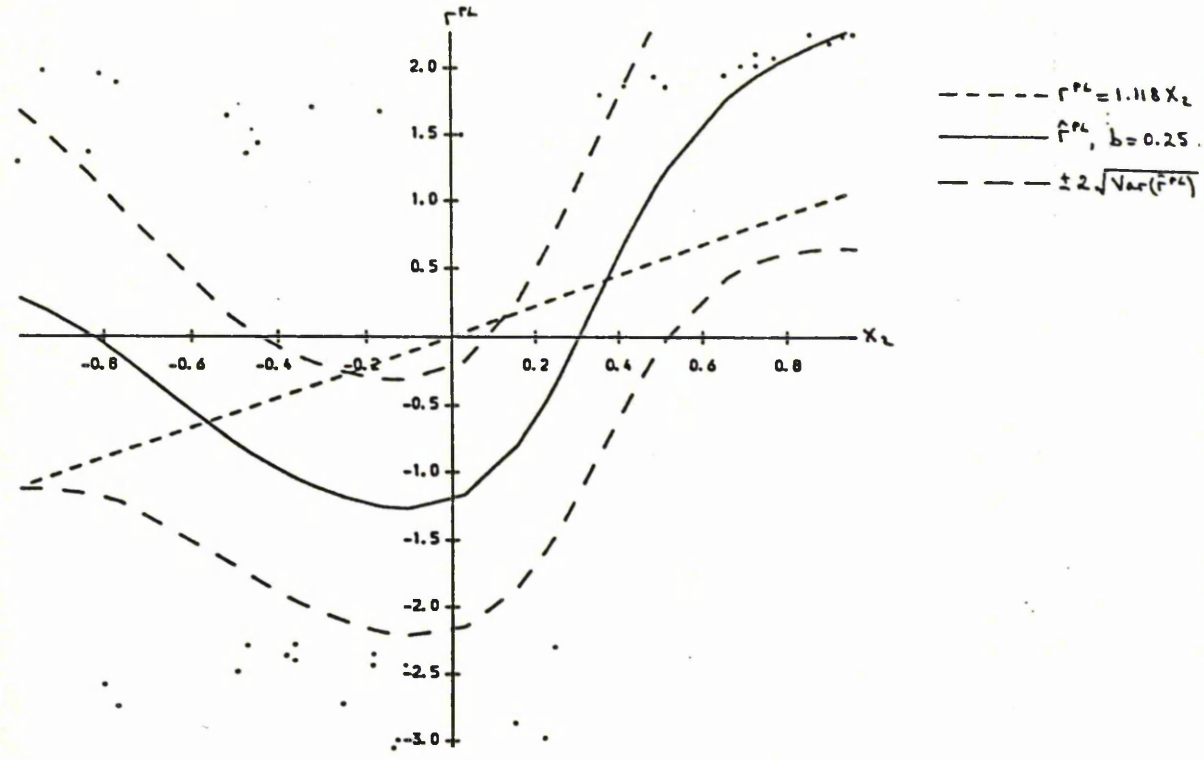


Figure 7.5. Partial residual plot based on the smoothed responses (h = 0.11) for X_1 of the simulated data.

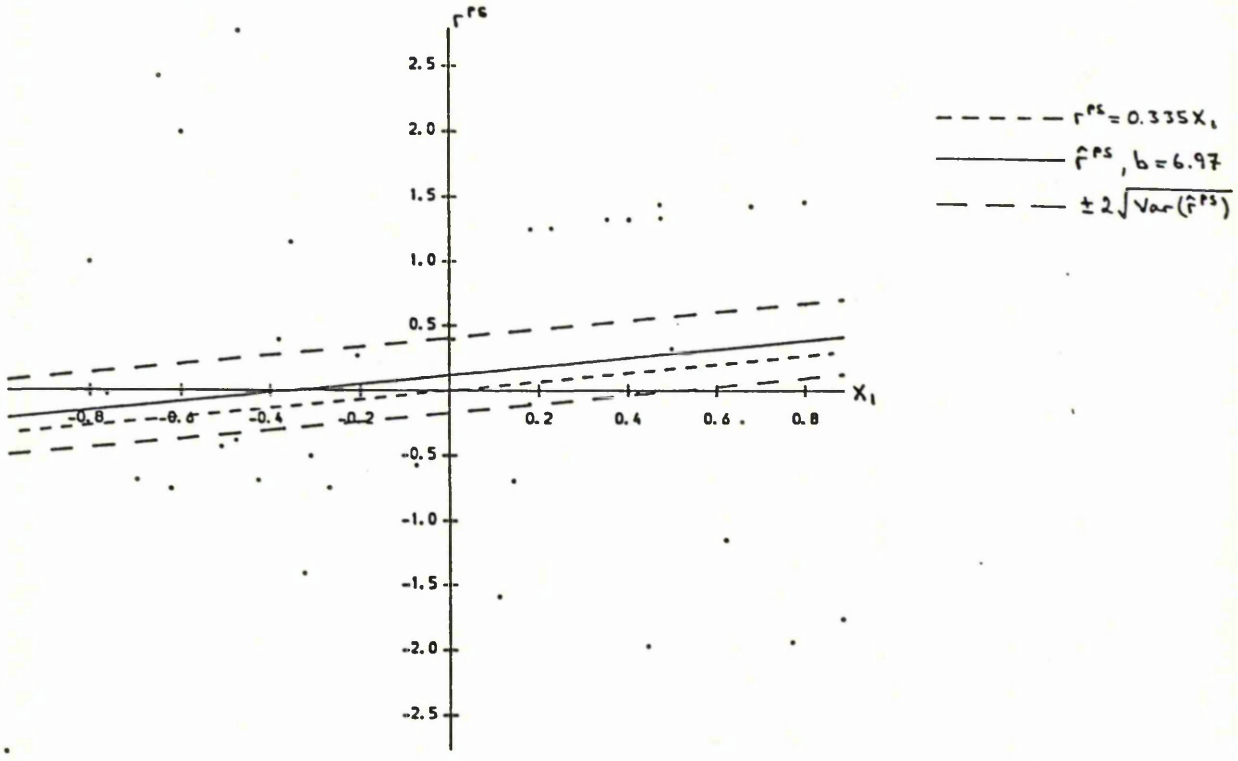


Figure 7.6. Partial residual plot based on the smoothed responses (h = 0.11) for X_2 of the simulated data.

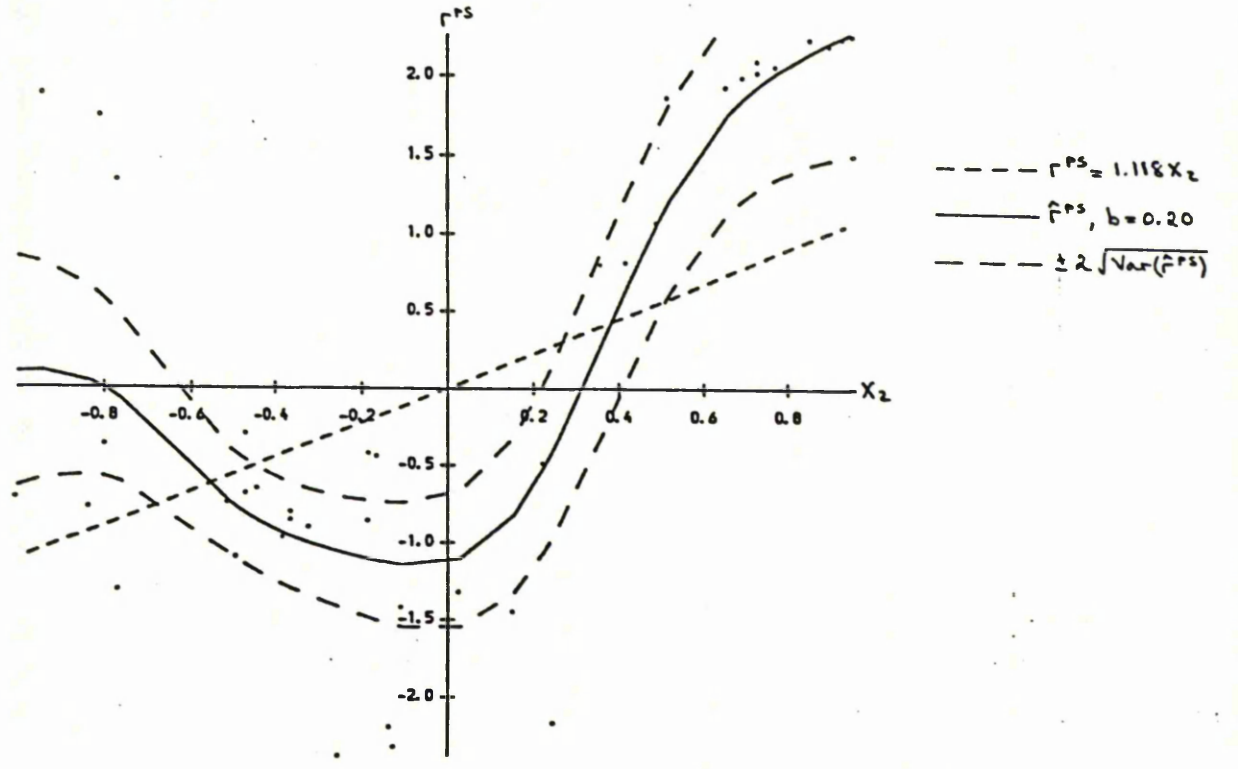


Figure 7.7. Logistic regression curve and bias corrected nonparametric estimate (h = 1.02) for Finney's data.

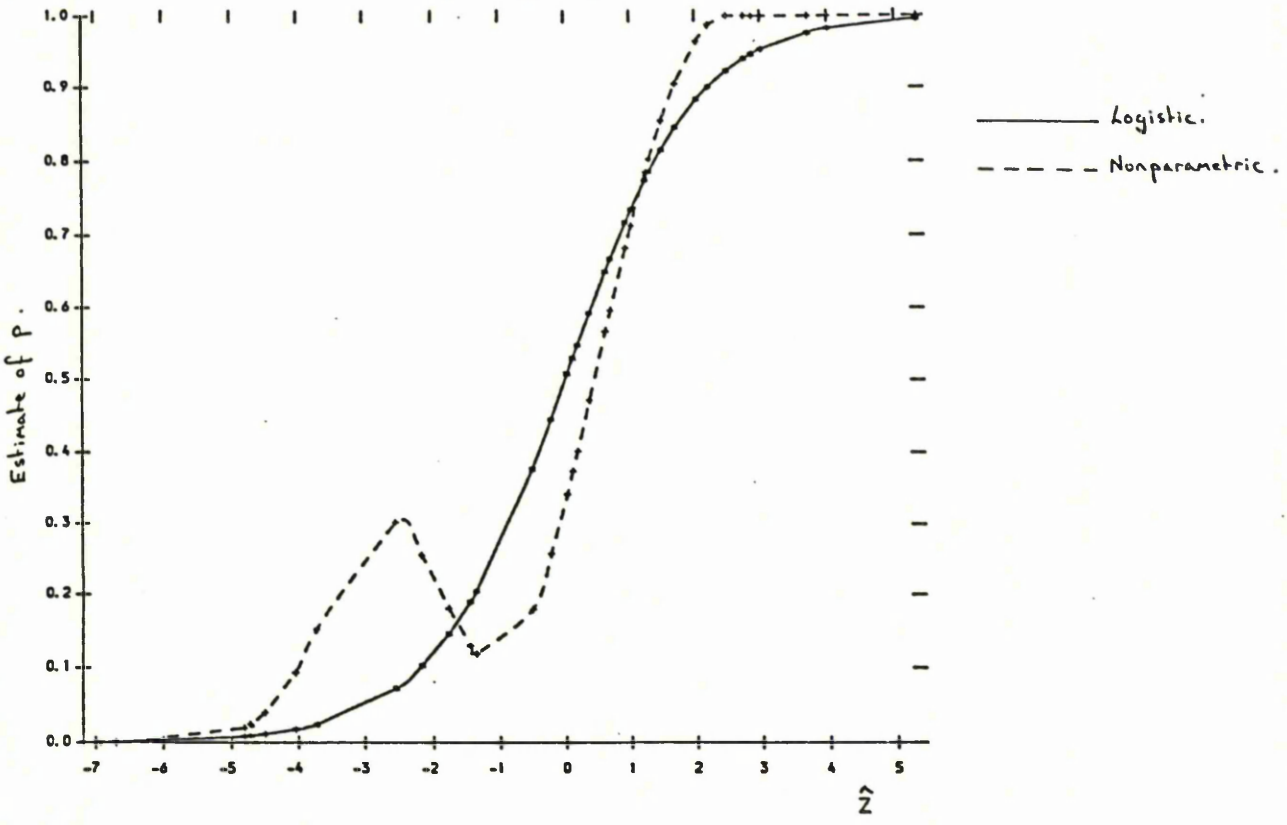


Figure 7.8. Partial residual plot based on the discrete responses for log(volume) of Finney's data.

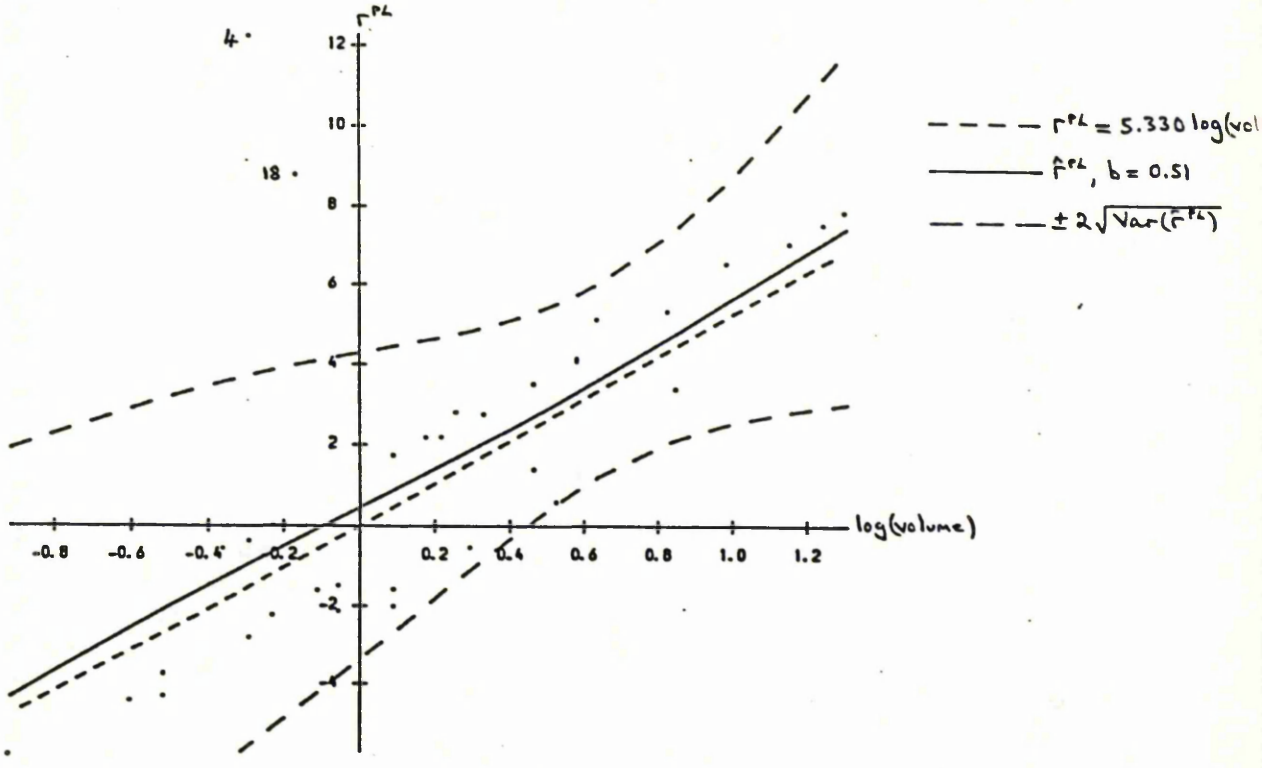


Figure 7.9. Partial residual plot based on the discrete responses for log(rate) of Finney's data.

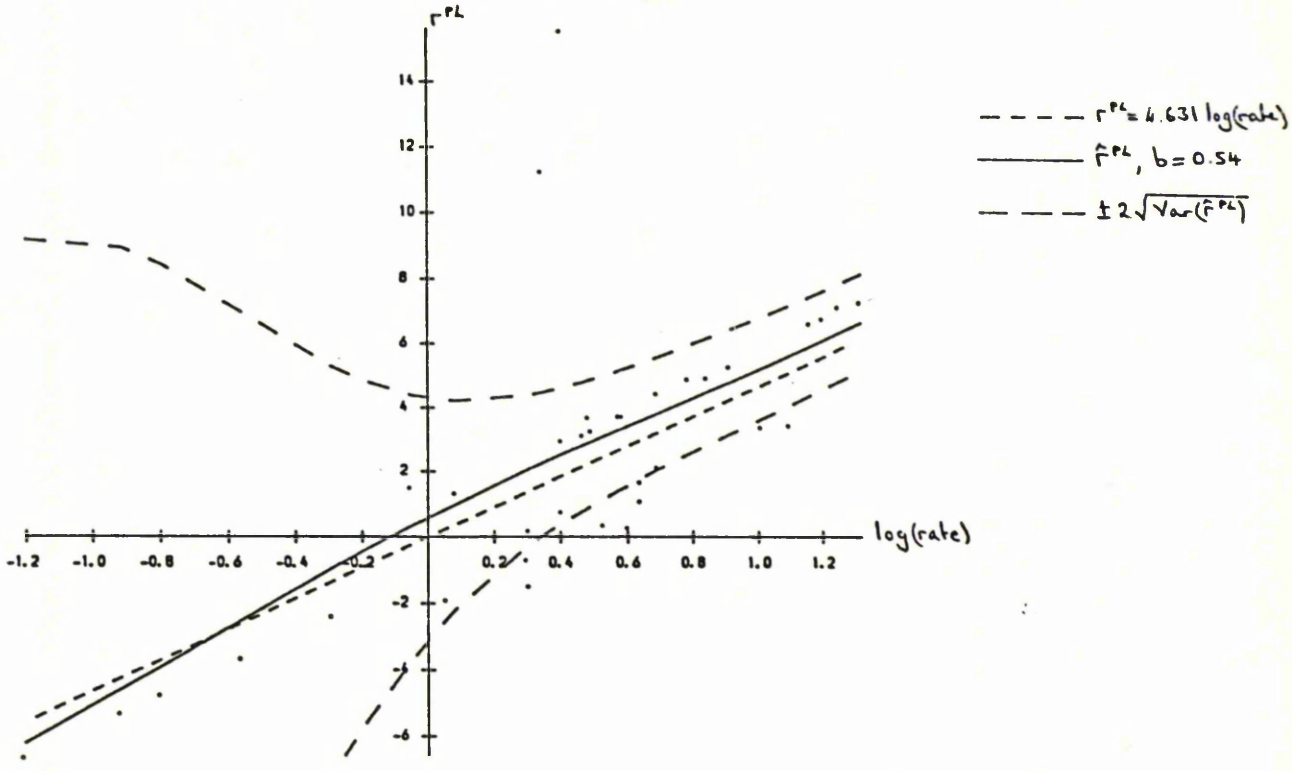


Figure 7.10. Logistic regression curve and bias corrected nonparametric estimate (h = 0.54) for the Cardiff bronchitis data.

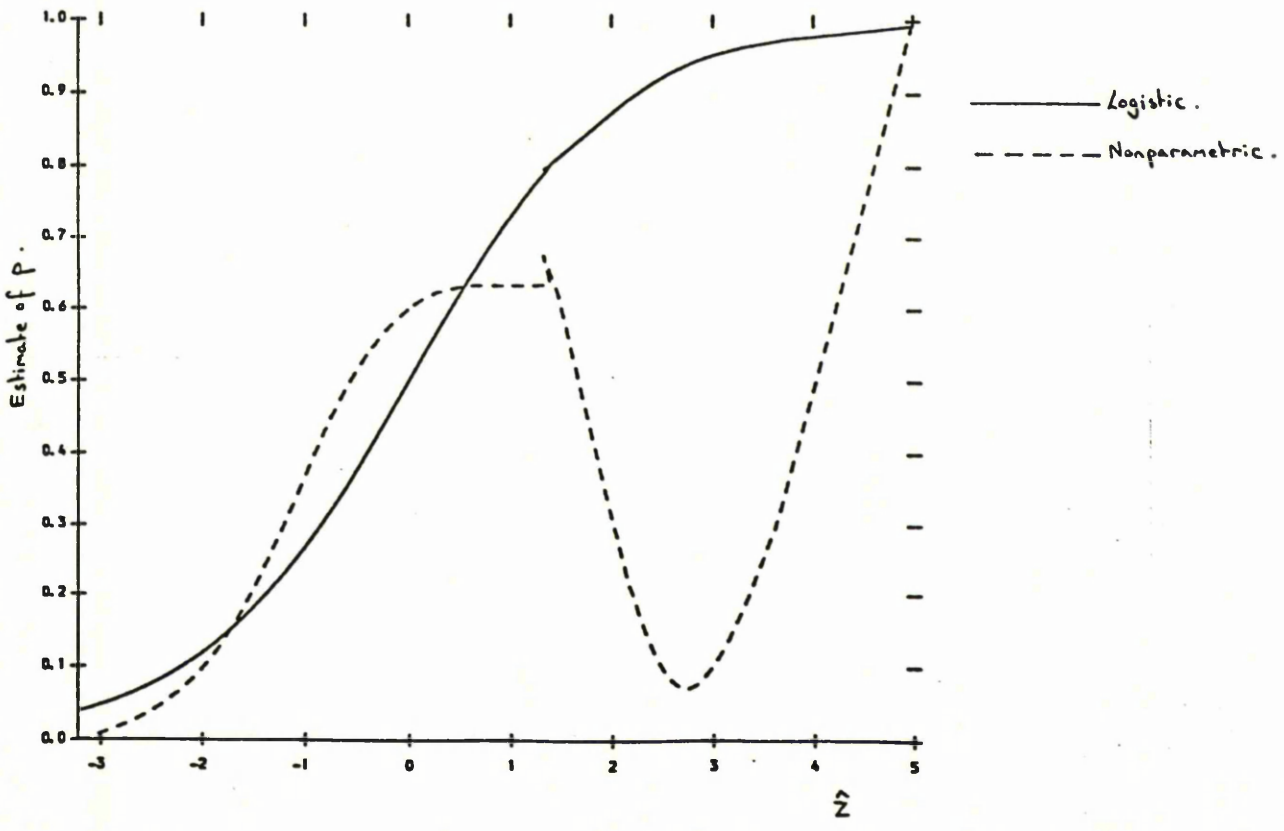


Figure 7.11. Partial residual plot based on the discrete responses for the variable CIG of the Cardiff bronchitis data.

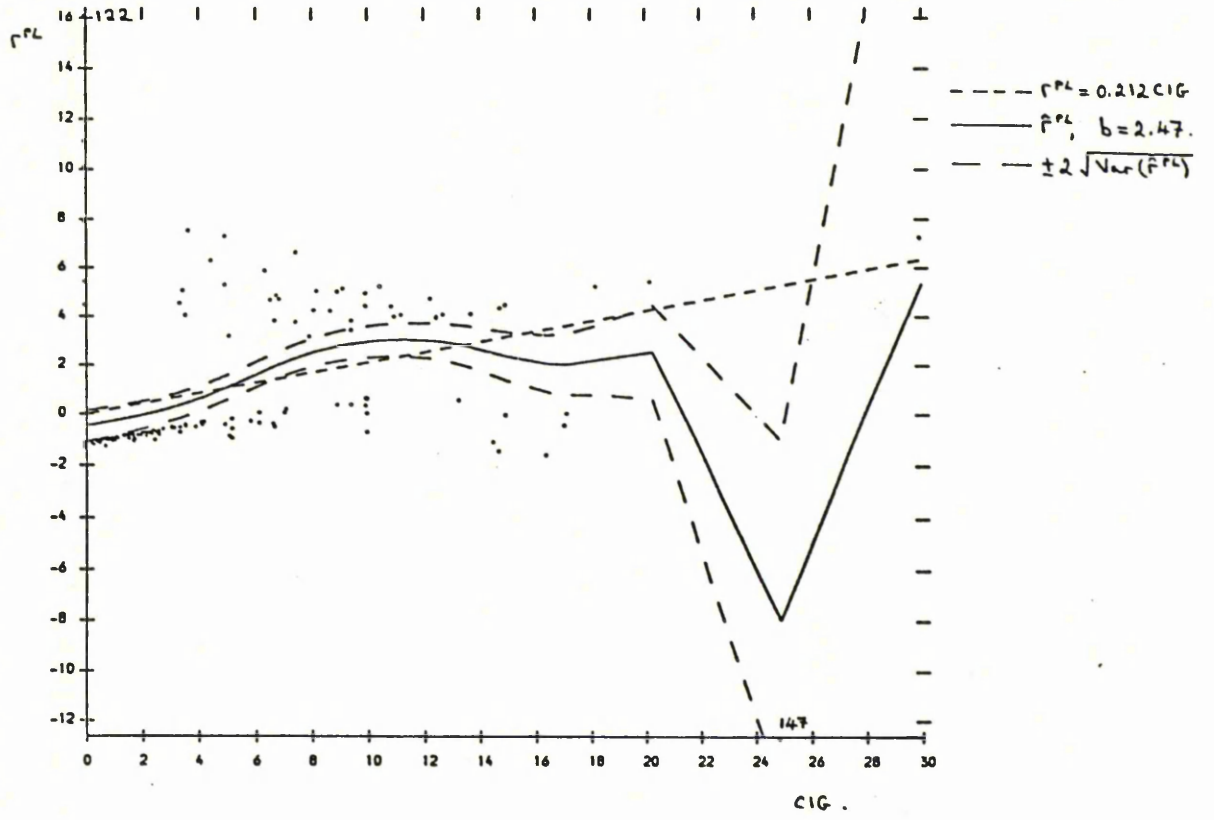
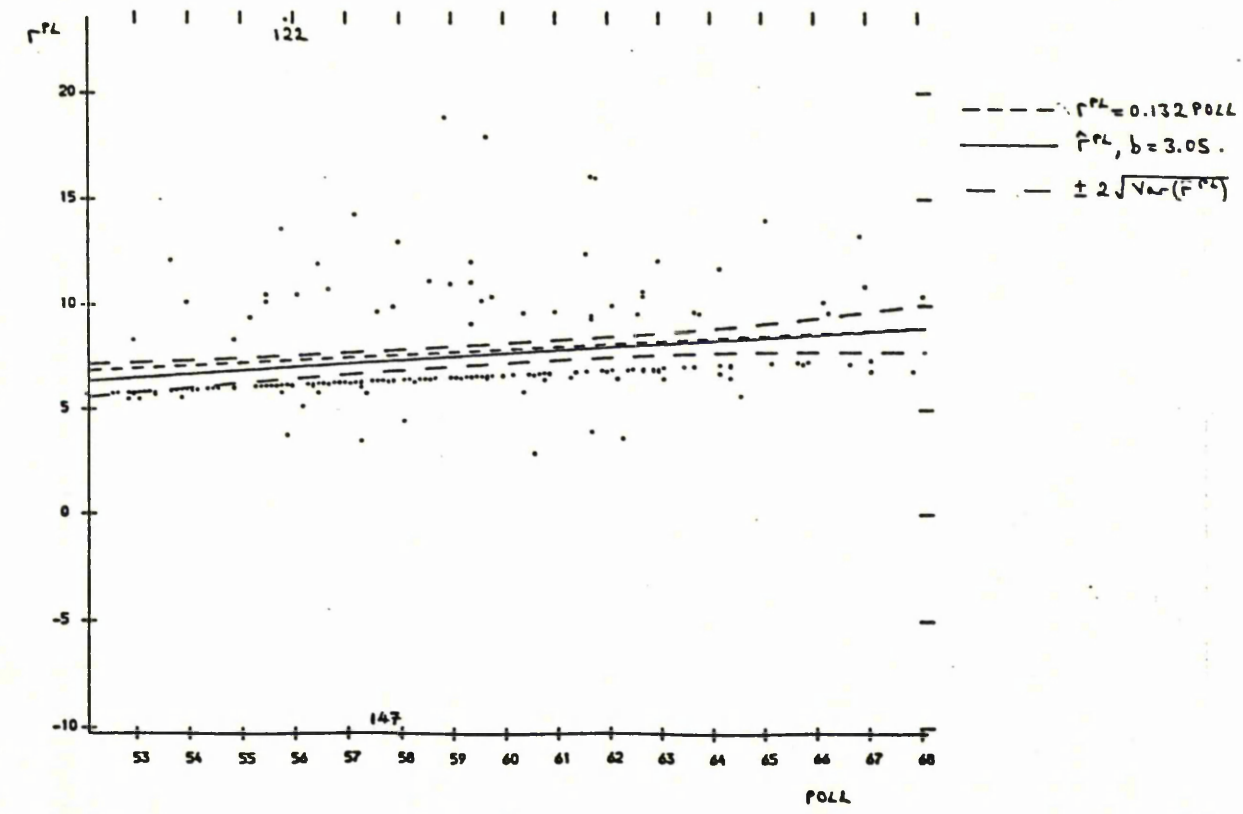


Figure 7.12. Partial residual plot based on the discrete responses for the variable POLL of the Cardiff bronchitis data.



REFERENCES

Abramson, I.S. (1982). On bandwidth variation in kernel estimates - a square root law. Ann. Statist., 10, 1217-1223.

Aitken, M., Anderson, D., Francis, B. and Hinde, J. (1989). Statistical modelling in GLIM. Oxford: Clarendon Press.

Amfoh, K.K. (1988). Statistical investigation of spatial models of mortality variation in Scotland and its relation to social deprivation. M. Phil. Thesis. University of Edinburgh.

Azzalini, A., Bowman, A.W. and Hardle, W. (1989). On the use of nonparametric regression for model checking. Biometrika, 76, 1-11.

Bartlett, M.S. (1963). Statistical estimation of density functions. Sankhya Ser. A, 25, 245-254.

Bean, S.J. and Tsokos, C.P. (1982). Bandwidth selection procedures for kernel density estimates. Comm. Stat. (T and M), 11, 1045-1069.

Benedetti, J.K. (1977). On the nonparametric estimation of regression functions. J. Roy. Statist. Soc. B, 39, 248-253.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. Biometrika, 74, 457-468.

Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. Ann. Statist., 1, 1071-1095.

Bowman, A.W. (1985). A comparative study of some kernel-based non-parametric density estimators. J. Statist. Comput. Simul., 21, 313-327.

Bowman, A.W. (1988). Density based tests of distributional shape. Unpublished manuscript.

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

Burden, R.L., Faires, J.D. and Reynolds, A.C. (1981). Numerical Analysis. Boston: Prindle, Weber and Schmidt.

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

Chatfield, C. and Collins, A.J. (1980). Introduction to Multivariate Analysis. London: Chapman and Hall.

Chernoff, H. (1964). Estimation of the mode. Ann. Inst. Statist. Math. 16, 31-41.

Clark, R.M. (1980). Calibration, cross-validation and Carbon-14 II. J. Roy. Statist. Soc. A, 143, 177-194.

Cook, R.D. and Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.

Copas, J.B. (1983). Plotting p against x. Appl. Statist., 32, 25-31.

Devroye, L. (1985). A note on the L_1 consistency of variable kernel estimates. Ann. Statist., 13, 1041-1049.

Diggle, P.J. and Fisher, N.I. (1984). Sphere: a contouring program for spherical data. Computers and Geosciences.

Eddy, W.F. (1980). Optimum kernel estimators of the mode. Ann. Statist. 8, 870-882.

Epanechnikov, V.A. (1969). Nonparamertric estimation of a multivariate probability density. Theor. Prob. Appl. 14, 153-158.

Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. J. Amer. Statist. Assoc., 19, 431-453.

Finney, D.J. (1947). The estimation from individual records of the relationship between dose and quantal response. Biometrika, 34, 320-334.

Fisher, R.A. (1958). Statisical Methods for Research Workers. London: Oliver and Boyd.

Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. Biometrika, 74, 503-515.

Friedman, J.H. (1987). Exploratory projection pursuit. J. Amer. Statist, Assoc. 82, 249-266.

Friedman, J.H. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput., Ser. C. 23, 881-889.

Fryer, M.J. (1976). Some errors associated with the nonparametric estimation of density functions. J. Inst. Maths. Applics., 18, 371-380.

Gajek, L. (1986). On improving density estimators which are not bona fide functions. Ann. Statist, 14, 1612-1618.

Gasser, T. and Muller, H-G. (1979). Kernel estimation of regression functions. In: Smoothing techniques for curve estimation. (eds. Gasser, T. and Rosenblatt, M). Lecture Notes in Mathematics, 757.

Gasser, T., Muller, H-G., Kohler, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves. Ann. Statist. 12, 210-229.

Gasser, T., Muller, H-G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. J. Roy. Statist. Soc. B, 47, 238-252.

Green, P.J. and Yandell, B. (1985). Semi-parametric generalised linear models. In Proceedings of the GLIM 1985 Conference, Springer-Verlag Lectures Notes in Statistics, 32.

Grenander, U. (1965). Some direct estimates of the mode. Ann. Math. Statist. 36, 131-138.

Habbema, J.D.F., Hermans, J. and Remme, J. (1978). Variable kernel density estimation in discriminant analysis. Compstat. 1978, Proceedings in Computational Statistics. Vienna: Physica Verlag.

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

Hardle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. Ann. Statist., 13, 1465-1481.

Hastie, T. and Tibshirani, R. (1987). Generalised additive models: some applications. J. Amer. Statist. Assoc., 82, 371-386.

Healy, M.J.R. (1968). Multivariate normal plotting. Appl. Statist. 17, 157-161.

Hogg, R.V. (1979). Statistical Robustness: One view of its use in applications today. American Statistician, 33, 108-116.

Hosmer, D.W. and Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. Commun. Statist., A, 9, 1043-1069.

Hosmer, D.W. and Lemeshow, S. (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. Amer. J. Epidem., 115, 92-106.

Huber, P.J. (1981). Projection pursuit. Ann. Statist., 13, 435-475.

Jones, K. (1975). A geographical contribution to the aetiology of chronic bronchitis. B.Sc. dissertation, University of Southampton.

Jones, M.C. and Sibson, R. (1987). What is projection pursuit? J. Roy. Statist. Soc. A, 150, 1-36.

Kappenman, R.F. (1987). Nonparametric estimation of dose-response curves with application to ED50 estimation. J. Statist. Comput. Simul., 28, 1-13.

Kittler, J. (1976). A locally sensitive method for cluster analysis. Pattern Recognition, 8, 23-33.

Koutrouvelis, I.A. and Kellermeier (1981). A goodness-of-fit test based on the empirical characteristic function when parameters must be estimated. J. Roy. Statist. Soc. B, 43, 173-176.

Koziol, J.A. (1982). A class of invariant procedures for assessing multivariate normality. Biometrika, 69, 423-427.

Koziol, J.A. (1983). On assessing multivariate normality. J. Roy. Statist. Soc. B, 45, 358-361.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models (with discussion). J. Amer. Statist. Assoc., 79, 61-83.

Larsen, W.A. and McCleary, S.J. (1972). The use of partial residual plots in regression analysis. Technometrics, 14, 781-789.

Littell, R.C. and Folks, J.L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. J. Amer. Statist. Assoc., 66, 802-806.

Loftsgaarden, D.O. and Quensenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36, 1049-1051.

Lubischew, A.A. (1962). On the use of discriminant functions in taxonomy. Biometrics, 18, 455-477.

Mack, Y.P. and Rosenblatt, M. (1979). Multivariate K-nearest neighbour density estimates. J. Multivariate Anal., 9, 1-15.

Mansfield, E.R. and Conerly, M.D. (1987). Diagnostic value of residual and partial residual plots. Amer. Statistician, 41, 107-116.

Mardia, M.V. (1972). Statistics of Directional Data. London: Academic Press.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. J. Amer. Statist. Assoc., 81, 104-107.

McCullagh, P. and Nelder, J.A. (1983). Generalised linear models. London: Chapman and Hall.

McLachlan, G.J., Lawoko, C.R.O. and Ganesaligam, S. (1982). On the likelihood ratio test for compound distributions for homogeneity of mixing proportions. Technometrics, 24, 331-334.

Minkin, S. (1989). Fit assessment and identification of functional form in logistic regression. Appl. Statist., 38, 343-350.

Moore, D.S. and Yackel, J.W. (1977). Consistency properties of nearest neighbour density estimators. Ann. Statist., 5, 143-154.

Muller, H-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. Ann. Statist. 12, 766-774.

Muller, H-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. Scandanavia J. Statist., 12, 221-232.

Muller, H-G. and Gasser, T. (1979). Optimal convergence properties of kernel estimates of derivatives of a density function. In: Smoothing techniques for curve estimation (eds. Gasser, T. and Rosenblatt, M). Lectures Notes in Mathematics, 757.

Muller, H-G. and Stadtmuller, U. (1987). Variable bandwidth kernel estimators of regression functions. Ann. Statist. 15, 182-201.

Muller, H-G., Stadtmuller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika, 74, 743-9.

NAG (1988). The Numerical Algorithms Group Fortran Library Manual - Mark 13. Oxford: NAG Ltd.

O'Sullivan, F., Yandel, B. and Raynor, W. (1986). Automatic smoothing of regression functions in generalised linear models. J. Amer. Statist. Assoc., 81, 96-103.

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

Parzen, E. (1979). Nonparametric statistical data modelling. J. Amer. Statist. Assoc., 74, 105-131.

Paulson, A.S., Roohan, P. and Sullo, P. (1987). Some empirical distribution function tests for multivariate normality. J. Statist. Comput. Simul., 28, 15-30.

Prakasa Rao, B.L.S. (1983). Nonparametric Functional Estimation. New York: Academic Press.

Priestly, M.B. and Chao, M.T. (1972). Nonparametric function fitting. J. Roy. Statist. Soc. B, 34, 385-392.

Rice, J.A. (1984). Boundary modification for kernel regression. Comm. Statist. - Theor. Meth., 13, 893-900.

Rice, J.A. (1986). Bandwidth choice for differentiation. J. Mult. Analysis, 19, 251-264.

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

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann. Statist., 3, 1-14.

Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Appl. Statist., 32, 121-133.

Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. J. Amer. Statist. Assoc, 74, 329-339.

Schvcany, W.R., Gray, H.L. and Owen, D.B. (1971). On bias reduction in estimation. J. Amer. Statist. Assoc. 66, 524-533.

Schvcany, W.R. and Sommers, J.P. (1977). Improvement of kernel type density estimators. J. Amer. Statist. Assoc., 72, 420-423.

Schuster, E.F. and Gregory, C.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In Eeddy, W.F. (ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. New York: Springer-Verlag, pp.295-298.

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

Scott, D.W., Gotto, A.M., Cole, J.S. and Gorry, G.A. (1978). Plasma lipids as collateral risk factors in coronary artery disease - a study of 371 males with chest pain. J. Chron. Dis. 31, 337-345.

Scott, D.W. and Thompson, J.R. (1983). Probability density estimation in higher dimensions. In Gentle, J.E. (ed.), Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface. Amsterdam: North Holland, pp.173-179.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 65, 591-611.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

Stephens, M.A. (1974). EDF statistics for goodness-of-fit and some comparisons. J. Amer. Statist. Assoc. 69, 730-737.

Stvetzle, W. and Mittal, Y. (1979). Some comments an the asymptotic behaviour of robust smoothers. In: Smoothing techniques for curve estimation (eds. Gasser, T. and Rosenblatt, M). Lecture Notes in Mathematics, 757.

Terrell, G.R. and Scott, D.W. (1980). On improving convergence rates for nonnegative kernel density estimators. Ann. Statist. 8, 1160-1163.

Tukey, J.W. (1977). Exploratory Data Analysis. Reading, Mass.: Addison-Wesley.

Tukey, P.A. and Tukey, J.W. (1981). Graphical display of data sets in 3 or more dimensions. In Barnet, V. (ed.), Interpreting multivariate data. Chichester: Wiley, 189-275.

Vasicek, O. (1976). A test for normality based on sample entropy. J. Roy. Statist. Soc. B, 38, 54-59.

Venter, J.H. (1967). On estimation of the mode. Ann. Math. Statist. 38, 1446-1455.

Watson, G.S. (1964). Smooth regression analysis. Sankhya A, 26, 359-372.

Williams, D.A. (1983). The use of the deviance to test the goodness-of-fit of a logistic-linear model to binary data. Glim Newsletter, 6.

Yamoto, H. (1972). Some statistical properties of estimators of density and distribution functions. Bull. Math. Statist., 19, 113-131.